1. Objeto de la ciencia económica

Antes de estudiar cómo conoce la economía, conviene precisar qué estudia: qué se entiende por objeto de una disciplina y cuál es el de la economía.

1.1. Concepto de objeto de una disciplina

Definición: el objeto de una ciencia son los fenómenos que estudia. Siguiendo la tradición escolástica, se distingue el **objeto material**, la parcela de la realidad que se estudia, del **objeto formal**, la perspectiva desde la que se estudia.

Características: para delimitar el objeto conviene tener en cuenta tres rasgos.

- *Acotación*: el objeto es la materia de estudio y no debe confundirse con el objetivo, la finalidad de la disciplina.
- *Coincidencia con otras disciplinas*: el objeto material suele solaparse con el de otras ciencias, por lo que la delimitación descansa sobre todo en el formal.
- *Carácter histórico*: el objeto se precisa a medida que la disciplina se desarrolla y puede variar de una escuela a otra. Cabe entender esa delimitación como una definición nominal y no real, que fija una convención y se juzga por su utilidad para ordenar la investigación.

1.2. Objeto de la economía

1.2.1. Concepto

Definición: el objeto material de la economía es la actividad humana relacionada con la producción, la distribución, el intercambio y el consumo de bienes y servicios. Lo comparte en parte con la sociología, la ciencia política, la psicología o la historia. Lo que la distingue es su objeto formal, que ha cambiado con las escuelas y que hoy es la elección en condiciones de escasez.

1.2.2. Evolución histórica

Escuela clásica ortodoxa: Smith y Ricardo se ocupan del crecimiento, los precios relativos y la distribución. Su objeto es la renta y la riqueza de los países. Smith (1776) concibe la economía política como una rama de la ciencia del legislador. Say (1803) y Mill (1836) la convierten en una ciencia de las leyes de la riqueza. Se pasa así de un arte de gobierno a una ciencia positiva. Al aislar el motivo de la riqueza, Mill define ya la disciplina por un aspecto de la conducta y no solo por una clase de objetos, como hará Robbins.

Escuela clásica heterodoxa: Malthus, Sismondi, List o Marx se centran en las relaciones sociales, históricas, demográficas y de poder. Para Marx, el objeto es descubrir la ley económica del movimiento de la sociedad moderna (1867), y la economía política estudia relaciones sociales de producción propias de un periodo histórico, no leyes naturales y eternas.

Escuela neoclásica: Walras y Marshall incorporan la utilidad marginal a la teoría del valor, y el objeto pasa de las clases y las naciones a los agentes individuales y la asignación óptima. Marshall (1890) define la economía como el estudio de la humanidad en los asuntos ordinarios de la vida, definición *material* que comparte Cannan y que Robbins criticará. La otra rama marginalista prepara la definición actual. Menger (1871) llama actividad económica a la *economización* de los bienes cuya cantidad es inferior a las necesidades, y Wicksteed (1910) y Mises (1922) extienden la lógica de la elección a toda administración de recursos.

Escuelas macroeconómicas: la keynesiana, la nueva macroeconomía clásica, la nueva economía keynesiana y la nueva síntesis neoclásica (Goodfriend y King, 1997) no redefinen la disciplina, sino su problema central. Keynes desplaza la atención de la asignación de recursos plenamente empleados al nivel de producción y empleo, y las escuelas posteriores discuten si las fluctuaciones son desequilibrios o respuestas óptimas a perturbaciones. Para Keynes, la economía consiste en pensar en términos de modelos, unida al arte de elegir los relevantes para el mundo contemporáneo (1938).

1.2.3. Concepto actual

La cuestión es si existe un objeto universal y atemporal, independiente de los autores.

Definición de Robbins (1932): "la economía es aquella ciencia que estudia la conducta humana como una relación entre fines y medios escasos que tienen usos alternativos".

Esta definición es **analítica y no clasificatoria** y define la economía por su objeto formal, es decir, no selecciona un tipo de conducta, como el bienestar material de Marshall, sino un aspecto de toda conducta, la forma que le impone la escasez. Los manuales la extienden a la sociedad (Samuelson y Nordhaus, 2010), y la **escasez relativa** pasa a ser el tema central del pensamiento moderno. La definición es neutral entre fines, lo que la separa de la ética, y no exige suponer deseos ilimitados, porque basta con que los medios no alcancen para todos los fines a la vez. Tardó en imponerse, y los manuales no la adoptan de forma general hasta los años sesenta.

Características: tiene *generalidad*, porque se abstrae del contexto histórico de cada autor y ha resultado coherente con la expansión posterior de la disciplina, y *universalidad*, porque no depende de un tipo de bienes ni de un sistema económico. Vale así para cualquier elección con medios escasos, en una economía de mercado, planificada o doméstica.

Críticas: su generalidad ha recibido cuatro objeciones.

- *Amplitud excesiva*: si toda elección entre fines y medios es economía, nada la distingue por su objeto material, sino solo por su objeto formal. De ahí el *imperialismo económico*, que extiende el análisis a la familia, el delito o la política (Becker, 1976).
- *Recursos ociosos*: con desempleo involuntario, el problema no es asignar medios escasos sino movilizar los ociosos, y la escasez deja de ser la restricción relevante.
- *Intercambio e instituciones*: Buchanan (1964) propone centrar la economía en el intercambio y las instituciones del mercado, la *catalláctica*, y no en la asignación, que sería un problema técnico de un decisor único.
- *Fines y medios dados*: deja fuera la formación de las preferencias y el descubrimiento de medios nuevos (Kirzner, 1973; Nelson y Winter, 1982).

De ahí surgen alternativas como "lo que hacen los economistas", atribuida a Viner, que refleja la distancia entre las definiciones y la práctica (Backhouse y Medema, 2009).

Microeconomía y macroeconomía: la microeconomía estudia los agentes individuales y su interacción en mercados concretos, en equilibrio parcial o general, mientras que la macroeconomía estudia la economía en su conjunto y los fenómenos agregados, como la inflación o el crecimiento. La distinción, que suele atribuirse a Frisch (1933), es cada vez más difusa por la microfundamentación de la macroeconomía.

Positiva y normativa: la **economía positiva** estudia lo que es (J. N. Keynes, 1891), y Friedman (1953) la juzga por sus predicciones. Su fundamento es la *ley de Hume* (1740): de lo que es no se deduce lo que debe ser. La **economía normativa** estudia lo que debe ser, con juicios de valor como la eficiencia o la equidad. La normativa no es independiente de la positiva, porque toda recomendación combina un objetivo con una predicción sobre los efectos de cada medio. J. N. Keynes añadía un tercer ámbito, el *arte* de la economía, el sistema de reglas para alcanzar un fin dado, antecedente de la actual política económica.

La frontera es menos nítida de lo que sugiere Hume, porque para Myrdal (1930), los juicios de valor intervienen ya en la elección de problemas y conceptos, y para Putnam (2002), conceptos como bienestar o pobreza son a la vez descriptivos y valorativos. Por eso se exige hacerlos explícitos, no eliminarlos. Con todo, para Putnam se desmorona la dicotomía entre hechos y valores, no la distinción, y según Sen (1970a) muchos juicios de valor son *no básicos*, porque dependen de circunstancias de hecho y la evidencia puede obligar a revisarlos.

En suma, la definición de Robbins se acepta por su generalidad, pero esa misma generalidad deja abiertas las fronteras con otras ciencias sociales, encaja mal con los recursos ociosos y relega las instituciones del intercambio.

2. Método de la ciencia económica

Definido el objeto, se estudian el método científico general, su evolución y su concepción actual, y, en el mismo orden, el método de la economía.

2.1. Método científico

2.1.1. Concepto

Definición: el **método científico** (del griego *méthodos*, "camino para llegar a un fin") es el procedimiento ordenado con el que se obtiene y se justifica el conocimiento científico. El objeto dice qué se analiza, y el método, cómo.

Conviene distinguir las *técnicas*, como la encuesta o la regresión; el *método*, la lógica que ordena su uso, y la *metodología*, el estudio crítico de los métodos (Blaug, 1980). Dentro de la metodología, Reichenbach (1938) separa el *contexto de descubrimiento*, cómo se llega a concebir una hipótesis, del *contexto de justificación*, cómo se decide si es aceptable, que es el que ocupa sobre todo a las metodologías.

Inferencias: el razonamiento científico combina tres formas.

- *Deducción*: las conclusiones se siguen necesariamente de las premisas. Preserva la verdad, pero no amplía el conocimiento.
- *Inducción*: generaliza a partir de casos observados. Amplía el conocimiento, pero sin certeza.
- *Abducción*: infiere la hipótesis que explicaría un hecho sorprendente. Peirce la sitúa en la fase en que se generan las conjeturas, que después la deducción desarrolla y la inducción somete a contraste.

Método hipotético-deductivo (Hempel, 1966): es la reconstrucción habitual de cómo se combinan.

1. *Problema*: la observación plantea un hecho que las teorías no explican.
2. *Hipótesis*: se propone una conjetura, sin regla para inventarla.
3. *Deducción de consecuencias*: con hipótesis auxiliares y condiciones iniciales, se deducen predicciones observables.
4. *Contrastación*: si fallan, la hipótesis o alguna auxiliar queda refutada. Si se cumplen, solo queda confirmada provisionalmente.

Según el modelo nomológico-deductivo (Hempel y Oppenheim, 1948), explicar es deducir de leyes y condiciones iniciales, de modo que explicación y predicción comparten estructura (la *tesis de la simetría*, afín al peso que Friedman da a la predicción).

Nociones: la *contrastación* confronta las consecuencias con la evidencia; la *verificación*, prueba concluyente, solo es posible para enunciados singulares; la *confirmación* es gradual y nunca definitiva (Carnap, 1936-1937); la *falsación*, por *modus tollens*, permite que una sola observación contraria refute un enunciado universal, y la *demarcación* separa la ciencia de lo que no lo es.

2.1.2. Evolución histórica

Aristóteles concibe la ciencia como demostración silogística desde primeros principios, a los que se llega por inducción. El método moderno nace con la revolución científica del s. XVII, de Galileo a Newton (1687), y su reflexión filosófica arranca con Bacon (1620) y Descartes (1637).

Racionalismo (Descartes): el método deductivo aplica reglas lógicas a axiomas evidentes para obtener teoremas, y la ciencia avanza por la razón. No amplía el contenido de sus premisas, y su valor depende de axiomas que en ciencias sociales suelen ser discutibles. Descartes parte de la duda metódica, dudar de todo, para liberar la razón de nociones preconcebidas.

Empirismo (Bacon, Hume): el método inductivo parte de la observación, descubre regularidades y las generaliza en leyes. Pero ningún número finito de observaciones justifica un enunciado universal, y apelar al éxito pasado de la inducción es inducción, dificultad conocida como **problema de la inducción** (Hume, 1739).

Criticismo (Kant, 1781): todo conocimiento comienza con la experiencia, pero no todo procede de ella. Las formas *a priori* de la mente permiten *juicios sintéticos a priori*, idea que anticipa la *carga teórica de la observación* (Hanson, 1958; Kuhn). El positivismo lógico rechazó después esos juicios, porque la geometría no euclídea y la relatividad general mostraron que la euclídea, que Kant tenía por sintética *a priori*, no es una verdad necesaria sobre el espacio físico.

Positivismo y positivismo lógico (Comte; Círculo de Viena, 1924-1936): según el *criterio verificacionista del significado*, un enunciado no analítico solo tiene sentido si es verificable, y la metafísica plantea pseudoproblemas. Esta *concepción heredada* dominó hasta los sesenta. Tropieza con tres dificultades: las leyes universales no son verificables, el criterio no supera su propia prueba y Quine (1951) negó la separación entre analítico y sintético y sostuvo que los enunciados se contrastan en bloque. La primera dificultad es el punto de partida de Popper.

Falsacionismo (Popper, 1934): parte de la asimetría entre verificación y falsación. La **falsabilidad** es el criterio de demarcación, y no de significado, que deja fuera la metafísica, el psicoanálisis y el marxismo. La ciencia avanza por **conjeturas y refutaciones** (1963), con hipótesis audaces sometidas a pruebas severas, y las que resisten quedan *corroboradas*, no verificadas. Tiene tres problemas.

- *Tesis de Duhem-Quine*: una hipótesis se contrasta junto con auxiliares, y no se sabe cuál falla. Popper prohíbe los supuestos *ad hoc*, las *estratagemas convencionalistas*, que Albert llamó después *inmunizadoras*.
- *Falsacionismo ingenuo*: Lakatos (1970) distingue tres versiones del falsacionismo. El dogmático tiene por concluyente toda refutación, y el metodológico ingenuo acepta por convención los enunciados de observación, pero descarta toda teoría refutada, lo que haría errática la evolución de la ciencia. El sofisticado solo abandona una teoría cuando otra explica lo mismo y predice hechos nuevos corroborados.
- *Hipótesis probabilísticas*: no son falsables en sentido estricto, porque ninguna muestra finita es lógicamente incompatible con ellas. Popper propone tratarlas como refutadas si los datos son muy improbables, una decisión metodológica análoga a la de los contrastes estadísticos.

Giro historicista (Kuhn, Lakatos): en esta corriente, llamada también pospositivismo, la filosofía de la ciencia atiende a cómo procede de hecho la ciencia, y la unidad de análisis deja de ser la teoría aislada.

- *Kuhn* (1962): las teorías se integran en **paradigmas**, realizaciones reconocidas que proporcionan modelos de problemas y soluciones a una comunidad. Tras un periodo preparadigmático, la *ciencia normal* resuelve enigmas hasta que las anomalías se acumulan y llevan a la *crisis*. Si un paradigma nuevo se impone, hay una *revolución científica*, decidida también por la persuasión. Por la amplitud del concepto, Kuhn lo sustituyó después (1969) por el de *matriz disciplinaria*, y reservó el de paradigma para los *ejemplares*, soluciones concretas que sirven de modelo.
- *Lakatos* (1970): la unidad es el **programa de investigación**, con un *núcleo duro* irrefutable por decisión metodológica y un *cinturón protector* de hipótesis auxiliares que se reajustan. El cambio es gradual: los programas son teórica o empíricamente progresivos si predicen hechos nuevos, corroborados en el segundo caso, y degenerativos si acumulan supuestos *ad hoc*. La *heurística negativa* prohíbe dirigir la refutación contra el núcleo, y la *positiva* indica cómo desarrollar el cinturón. Solo son científicos, para Lakatos, los programas progresivos.

Ambos explican por qué no se abandona una teoría ante la primera refutación. A Kuhn se le reprochan la vaguedad del paradigma, con una veintena de acepciones (Masterman, 1970), la *inconmensurabilidad*, que para Lakatos convierte el cambio en "psicología de masas", y no explicar cómo se pasa de las anomalías a la ruptura. A Lakatos, Feyerabend le reprocha no decir cuándo abandonar un programa degenerativo: un "anarquismo disfrazado".

Anarquismo epistemológico (Feyerabend, 1975): ninguna regla metodológica resiste la historia de la ciencia, y el único principio que no inhibe el progreso es "todo vale". Evita el dogmatismo, pero menosprecia patrones de descubrimiento eficaces.

2.1.3. Concepto actual

Definición: hoy, el método es un procedimiento sistemático, público y autocorrector, sin una versión única. Exige hipótesis falsables, revisión por pares, que institucionaliza el *escepticismo organizado* de Merton (1942), y replicabilidad, reforzada tras la crisis de replicación de la década de 2010, con pluralismo de enfoques. En la práctica, la falsación estricta ha dado paso a un enfoque próximo al de Lakatos.

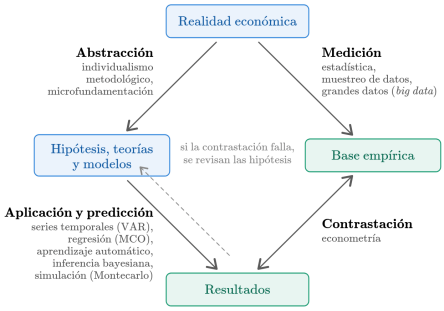
Características (Gibson, 1960): abstracción, generalidad, neutralidad entre juicios de hecho y de valor, objetividad y evidencia empírica. Gibson las defiende para las ciencias sociales frente a quienes niegan su carácter científico.

2.2. Método económico

2.2.1. Concepto

Definición: el método de la economía relaciona cuatro elementos, la realidad económica, los modelos, la base empírica y los resultados, mediante cuatro operaciones que forman un ciclo.

- *Abstracción*: de la realidad se extraen hipótesis, teorías y modelos. Descansa en el individualismo metodológico y la microfundamentación.
- *Medición*: la realidad se mide con técnicas estadísticas y forma la base empírica. Incluye el muestreo de microdatos y el tratamiento de grandes datos.
- *Aplicación y predicción*: los modelos se aplican para obtener resultados. Se usan series temporales (VAR), regresión (MCO), aprendizaje automático, inferencia bayesiana y simulación.
- *Contrastación*: los resultados se contrastan con la base empírica mediante la econometría. Si fallan, se revisan las hipótesis.



El método de la economía como ciclo: abstracción, medición, aplicación y predicción, y contrastación.

2.2.2. Evolución histórica

La reflexión explícita empieza con Senior (1836) y Mill (1836), sigue con Cairnes (1857) y se sistematiza con J. N. Keynes (1891). Entre las cuatro etapas cambian el modo de razonar, el lenguaje y la forma de contrastar las teorías.

Escuela clásica (s. XVIII-1870): predomina el deductivismo lógico-verbal, con elementos inductivos.

- *Smith* combina el razonamiento deductivo con la descripción de las instituciones.
- *Ricardo*, teórico puro, enuncia leyes generales sin interés por la contrastación y es pionero del modelo económico sistemático, con el ricardiano de distribución.
- *Mill* concibe la economía como ciencia abstracta e hipotética, de método *a priori*: deduce desde supuestos sobre la conducta, y sus conclusiones son *tendencias* que otras causas pueden contrarrestar. Describe lo que luego se llamó *homo economicus*. Su verificación *a posteriori* solo comprueba si en el caso concreto se ha omitido alguna causa relevante, sin contrastar las premisas, y por eso Blaug (1992) llama *verificacionistas* a Mill, Cairnes y J. N. Keynes.
- *Malthus y Marx*, heterodoxos, dan más peso al análisis histórico y empírico. Marx lo hace con el método dialéctico y el materialismo histórico.

Marginalismo e historicismo (1870-1910): la Revolución Marginalista generaliza las matemáticas y refuerza el carácter deductivo. El historicismo, de List a Schmoller, es inductivo y se apoya en la historia y la estadística. La **disputa del método** (*Methodenstreit*) enfrenta a Menger con Schmoller a raíz de una obra de Menger de 1883. De esta etapa proceden el equilibrio general de Walras, el parcial *ceteris paribus* de Marshall y el uso de la estadística de Jevons. Entre los marginalistas, Menger se apoya en la lógica deductiva sin matemáticas, y Walras confía en las matemáticas para entender la interdependencia.

Revolución formalista (1910-1970): nace la econometría, con los primeros modelos de Tinbergen (1936, 1939) y los de ecuaciones simultáneas de la Cowles Commission, base de los grandes modelos keynesianos. Se generalizan las matemáticas, con la optimización en micro y el IS-LM (Hicks, 1937) en macro, y con Arrow y Debreu (1954) el equilibrio general walrasiano desplaza al marshalliano.

Keynes razona con relaciones entre agregados que no se derivan de la decisión de cada agente, concibe la economía como "una ciencia moral" y desconfía de la econometría en su reseña de Tinbergen (1939). Frente al verificacionismo, cuatro autores trasladan la prueba a las predicciones, que han de poder ser refutadas.

- *Hutchison* (1938): plantea la exigencia. Ataca el apriorismo de Robbins, para quien los postulados son hechos indiscutibles de la experiencia, y sostiene que en la economía dominan las proposiciones tautológicas, que no superan el criterio de Popper.
- *Samuelson* (1947): muestra cómo cumplirla. Busca "teoremas operacionalmente significativos", refutables al menos en condiciones ideales, y por el *principio de correspondencia* obtiene el signo de la estática comparativa de las condiciones de segundo orden, si hay optimización, y de la estabilidad dinámica, si no la hay.
- *Friedman* (1953): hace de la predicción el único criterio. Coincide con Popper en que los hechos solo pueden no refutar una hipótesis, pero se aparta de él con la **tesis de irrelevancia de los supuestos**, el *F-Twist* de Samuelson, que es *instrumentalista*. Musgrave (1981) distingue supuestos de irrelevancia, de dominio y heurísticos, y el dictum de Friedman no vale para ninguno.
- *Lipsey* (1963; 2.ª ed., 1966): la lleva a los manuales y propone una falsación estadística y comparada, porque una teoría solo se abandona cuando predice peor que una hipótesis alternativa.

Microfundamentación y nuevas técnicas (desde 1970): los métodos se renuevan en tres campos.

- *Microeconomía*: al modelo estático de optimización se suma la teoría de juegos, lenguaje común de la organización industrial desde los ochenta (Tirole, 1988), y el laboratorio permite contrastar sus predicciones.
- *Macroeconomía*: los modelos pasan a ser dinámicos y estocásticos. El control óptimo y la programación dinámica entran en los sesenta con el crecimiento óptimo, y la crítica de Lucas (1976) los generaliza.
- *Econometría*: de la crítica de Lucas procede la estimación de modelos estructurales; de la de Sims (1980), los modelos VAR, y de la de Leamer (1983), los diseños cuasiexperimentales.

2.2.3. Concepto actual

Características: sus leyes son *inexactas*, tendencias que solo se cumplen *ceteris paribus*, y su contrastación es difícil porque los datos proceden casi siempre de la observación y no del experimento. Además, explica su ámbito con pocos factores causales deducidos de la conducta individual. Hausman (1992) la describe por ello como una ciencia **inexacta y separada**, heredera de Mill.

Buenos modelos: según Gabaix y Laibson (2008), reúnen siete propiedades agrupables en tres requisitos: *formalización* (profundidad, generalización y parsimonia), *aplicabilidad* (falsables y tratables) y *utilidad* (consistencia empírica y precisión predictiva). No siempre son compatibles, porque un modelo más tratable suele ajustarse peor a los datos. Elegir un modelo exige ponderarlas según la pregunta que se quiere responder.

Debates: entre los que persisten, destacan cuatro.

- *Monismo o pluralismo*: si el distinto objeto de las ciencias exige métodos distintos.
- *Falsacionismo inocuo*: los economistas lo suscriben, pero rara vez lo practican (Blaug, 1992). Hausman (1992) critica el dogmatismo con que la disciplina, fiel a Mill, protege sus leyes básicas frente a la evidencia, y Caldwell (1982) propone un pluralismo crítico.
- *Individualismo y holismo*: la economía dominante explica lo social desde los individuos, aunque el individualismo estricto no justifica el agente representativo (Kirman, 1992). El *holismo* de Durkheim o Marx, en cambio, atribuye a las clases o a la sociedad regularidades propias.
- *Formalismo y retórica*: los economistas persuaden con recursos retóricos (McCloskey, 1985), y P. Romer (2015) llama *mathiness* a las matemáticas que dejan holgada su conexión con los conceptos y los datos.

En suma, el método económico combina abstracción y contrastación en un ciclo que va de la realidad a los modelos y de estos a los datos.

3. Cuestiones y debates actuales

Tres líneas discuten dos supuestos tradicionales, que la economía no puede experimentar y que sus agentes son plenamente racionales y están en equilibrio. Las tres amplían el núcleo neoclásico sin sustituirlo. A ellas se suma el debate sobre el método de la macroeconomía, de la crítica de Lucas a la crisis de 2008.

3.1. Economía experimental (1960-actualidad)

La **economía experimental** crea entornos artificiales en el laboratorio o estudia entornos reales con asignaciones aleatorias, diseñadas o fruto de las circunstancias.

3.1.1. Experimentos de laboratorio (Smith, 1962)

Idea: Vernon Smith (Nobel 2002) asignaba a cada participante un valor de reserva, de modo que las curvas de oferta y demanda inducidas eran conocidas. Después los dejaba negociar en una *subasta doble*, en la que compradores y vendedores anuncian públicamente pujas y ofertas.

Evidencia: con pocos participantes y sin que nadie conociera las curvas, precios y cantidades convergían rápido al equilibrio competitivo, sin necesidad de información perfecta ni de muchos agentes. Smith fijó además la **teoría del valor inducido** (1976): si la recompensa depende de las decisiones, el investigador controla las preferencias. Roth (Nobel 2012) combinó la teoría de emparejamientos, el laboratorio y el estudio de campo en el diseño de mercados.

Ofrece *validez interna*, pero su *validez externa* es discutida: estudiantes, cantidades pequeñas y entorno artificial.

3.1.2. Ensayos controlados aleatorios

Idea: Banerjee, Duflo y Kremer (Nobel 2019) descompusieron los grandes interrogantes de la pobreza en preguntas acotadas y comprobables. Los participantes se asignan al azar a un grupo de tratamiento o de control, y el azar garantiza que sean iguales en promedio salvo en el tratamiento. La diferencia de medias estima sin sesgo el efecto causal medio:

$$\widehat{ATE} = \bar{Y}_{T=1} - \bar{Y}_{T=0}, \qquad ATE = E[Y_i(1) - Y_i(0)]$$

Extensiones: antes de su auge en el desarrollo, se usaron en EE. UU. desde 1968 para evaluar programas sociales, como el impuesto negativo sobre la renta o el seguro sanitario de RAND, y hoy evalúan también políticas activas de empleo.

Elimina el sesgo de selección con supuestos mínimos y ha orientado la evaluación de políticas hacia la evidencia medible. Pero su validez externa es discutida, son costosos, plantean dilemas éticos y solo responden a preguntas que admiten aleatorización.

3.1.3. Experimentos naturales

Origen: Card compartió el Nobel de 2021 con Angrist e Imbens.

Idea: un factor ajeno, natural o institucional, expone a unos individuos a un tratamiento y a otros no. El investigador debe comprobar la *comparabilidad* de los grupos y la *plausibilidad* de que la exposición sea *como si* aleatoria. Angrist e Imbens muestran que, cuando los individuos eligen en parte el tratamiento, se estima el **efecto medio local** (LATE, 1994) sobre los *cumplidores*. Resultados de Card:

- *Salario mínimo* (Card y Krueger, 1994): su subida en Nueva Jersey, frente a Pensilvania, no redujo el empleo en la comida rápida.
- *Inmigración* (Card, 1990): el éxodo del Mariel no tuvo efectos apreciables sobre los salarios locales, resultado discutido por Borjas (2017).

Con variables instrumentales, diferencias en diferencias y regresión discontinua han dado la **revolución de la credibilidad** (Angrist y Pischke, 2010), respuesta a la crítica de Leamer (1983). Para Deaton (2010) y Heckman (2010), identifican efectos locales y, sin teoría, no explican por qué se producen. La identificación a partir de una variación creíblemente exógena ha pasado, con todo, a ser el estándar de la disciplina.

3.2. Economía conductual (1970-actualidad)

La **economía conductual** incorpora los hallazgos de la psicología sobre cómo deciden realmente las personas y documenta desviaciones sistemáticas y predecibles de la racionalidad, sesgos con dirección definida que no se compensan en el agregado. Su principal antecedente es la **racionalidad limitada** de Simon (1955; Nobel 1978). Con capacidad limitada de procesar información, los agentes se conforman con alternativas satisfactorias. Con todo, la economía conductual actual, desde Kahneman y Tversky, conserva la optimización como referencia y mide las desviaciones respecto de ella (Sent, 2004).

3.2.1. Descuento hiperbólico

Origen: según la *ley de igualación* de Herrnstein (1961), la frecuencia relativa de una conducta se iguala a la frecuencia relativa de la recompensa que obtiene. De ella se deriva que el valor de una recompensa diferida cae de forma hiperbólica, idea que Ainslie (1975) aplicó a la elección humana.

Idea: con el descuento exponencial de Samuelson (1937), las preferencias son coherentes en el tiempo, mientras que con el hiperbólico, de tasa alta a corto plazo y baja a largo, son **dinámicamente inconsis-**

tentes. Un individuo prefiere hoy 110 euros en 31 días a 100 en 30, pero el día 30 prefiere los 100 inmediatos. Laibson (1997) lo formaliza con el descuento cuasihiperbólico de Phelps y Pollak (1968), que explica el ahorro insuficiente para la jubilación y la demanda de mecanismos de compromiso, como los planes de pensiones con penalización por rescate anticipado.

Un solo sesgo hacia el presente explica así la inconsistencia, con un único parámetro más que el exponencial. Pero sus predicciones dependen de que el individuo sea *sofisticado* o *ingenuo* (O'Donoghue y Rabin, 1999), algo difícil de observar, y en los experimentos con dinero la preferencia por el pago inmediato puede reflejar desconfianza en el pago futuro.

3.2.2. Teoría de la perspectiva y heurísticas (Tversky y Kahneman, 1974; Kahneman y Tversky, 1979)

Teoría de la perspectiva (1979): sustituye la utilidad esperada por una función de valor sobre ganancias y pérdidas y una ponderación de probabilidades.

- *Punto de referencia*: se valoran cambios, no niveles de riqueza.
- *Aversión a la pérdida*: una pérdida pesa más que una ganancia igual.
- *Sensibilidad decreciente*: el valor es cóncavo en ganancias y convexo en pérdidas.
- *Ponderación*: se sobreponderan las probabilidades bajas y se infraponderan las moderadas y altas.

Heurísticas (1974): atajos útiles que llevan a errores sistemáticos: la *representatividad*, que juzga la probabilidad de un suceso por su parecido con un estereotipo e ignora la probabilidad de partida; la *disponibilidad*, que juzga la frecuencia por lo que viene a la mente, y el *anclaje y ajuste*, con un ajuste insuficiente desde el valor inicial.

Efectos: el *efecto marco* (1981), por el que "salvar 200 vidas de 600" y "dejar morir a 400" cambian la elección contra el principio de invarianza, y el *efecto justicia* (Kahneman, Knetsch y Thaler, 1986). Por este último, las decisiones dependen de lo que se considera justo, como muestra el rechazo a subir el precio de las palas quitanieves tras una gran nevada.

La teoría de la perspectiva explica, con supuestos contrastados en el laboratorio, elecciones que la utilidad esperada no recoge, como la aversión al riesgo en ganancias y su búsqueda en pérdidas. Pero no dice cómo fijar el punto de referencia, y su versión de 1979 puede violar la dominancia estocástica de primer orden, defecto que corrige la acumulativa (1992).

3.2.3. Contabilidad mental, efecto dotación y *nudge* (Thaler, 1980; 1985; Thaler y Sunstein, 2008)

Aportaciones: Thaler (Nobel 2017) incorpora supuestos psicológicamente realistas.

- *Contabilidad mental* (1985): el dinero se clasifica en cuentas no fungibles. Como ya mostraron Tversky y Kahneman (1981), quien pierde una entrada de teatro es menos propenso a comprar otra que quien pierde su equivalente en efectivo (46 % frente a 88 %).
- *Nudge* (Thaler y Sunstein, 2008): la presentación de las opciones influye sin restringirlas. Quienes diseñan ese contexto son *arquitectos de la elección*, y la inscripción automática en planes de pensiones es *paternalismo libertario*.
- *Efecto dotación* (1980): se valora más lo que se posee, con una brecha entre disposición a pagar y a aceptar. Thaler (1985) añade la *utilidad de transacción*, el valor que se atribuye a la calidad del trato.
- *Autocontrol* (Thaler y Shefrin, 1981): un *planificador* de largo plazo y un *ejecutor* miope explican las reglas y compromisos.

Su obra explica de forma sencilla y contrastable conductas de consumo y ahorro que el modelo estándar trata como anomalías. Pero la brecha entre aceptar y pagar desaparece al controlar los errores de comprensión (Plott y Zeiler, 2005), y el *nudge* se critica por influir sin que el individuo lo advierta.

3.2.4. Conclusiones

Aspectos positivos: ha dado lugar a políticas basadas en el comportamiento, desde el Behavioural Insights Team británico (2010), y se ha incorporado al modelo estándar como extensiones que se contrastan caso a caso.

Limitaciones: su éxito abre cuatro debates.

- *Teoría unificada*: acumula sesgos sin predecir cuál domina. Para Gigerenzer, muchas heurísticas son *racionalidad ecológica*.
- *Replicación*: de 18 experimentos publicados entre 2011 y 2014, se replicaron 11, el 61 % (Camerer et al., 2016).
- *Escala*: en dos grandes unidades de comportamiento de EE. UU., un *nudge* elevó la respuesta 1,4 puntos porcentuales frente a 8,7 en revistas, sobre todo por el sesgo de publicación (DellaVigna y Linos, 2022).
- *Bienestar*: si las elecciones no reflejan el bienestar, hace falta otro criterio, y el regulador también tiene sesgos.

3.3. Economía de la complejidad (desde 1987)

La **economía de la complejidad** estudia la economía como un **sistema complejo adaptativo**, con agentes heterogéneos que interactúan y aprenden, y cuyo comportamiento agregado surge sin que nadie lo diseñe. Frente al análisis de equilibrio, se interesa por los procesos, por cómo se forman las pautas agregadas y cómo cambian.

3.3.1. Mercado bursátil artificial (Arthur et al., 1997)

Origen: en 1987 el Instituto Santa Fe reunió, por iniciativa de Arrow y Anderson, a economistas y a científicos de la física, la biología y la informática. En 1988 nació su programa de economía, dirigido por W. Brian Arthur.

Idea: agentes heterogéneos forman expectativas con reglas que prueban y sustituyen según su éxito. Con aprendizaje lento, los precios convergen al equilibrio de expectativas racionales, mientras que con aprendizaje rápido aparecen burbujas, crisis y volatilidad agrupada. El equilibrio es solo uno de los regímenes posibles, y de ahí proceden los *modelos basados en agentes*, hoy usados, entre otros ámbitos, en el análisis de la estabilidad financiera.

Ofrece una alternativa al agente representativo, pero sus resultados dependen de reglas y parámetros elegidos por el investigador, difíciles de contrastar y sin resultados analíticos generales.

3.3.2. Propiedades (Arthur, Durlauf y Lane, 1997)

Propiedades: el sistema complejo adaptativo se reconoce por seis rasgos: interacción dispersa, ausencia de un controlador global, organización jerárquica transversal, adaptación continua, novedad perpetua y dinámica fuera del equilibrio. Dan un marco común a modelos muy distintos, pero describen más que predicen.

3.3.3. Complejidad económica (Hidalgo y Hausmann, 2009)

Idea: la riqueza depende de las capacidades productivas, que no se observan, pero se reflejan en las exportaciones. El **índice de complejidad económica** combina la *diversidad*, cuántos productos exporta un país con ventaja comparativa revelada, y la *ubicuidad*, cuántos países los exportan, de modo que un país es complejo si exporta muchos productos que pocos países son capaces de producir. Según el *Atlas* (2011), predice el crecimiento mejor que los indicadores convencionales, y los países cuya complejidad es alta en relación con su renta tienden a crecer más deprisa.

El índice es un indicador operativo para la diversificación, pero omite servicios y producción interior, puede reflejar solo ensamblaje en las cadenas de valor y no prueba causalidad.

3.4. El método de la macroeconomía (1976-actualidad)

3.4.1. La crítica de Lucas (1976)

Idea: Lucas (Nobel 1995) cuestiona que las políticas se evalúen simulándolas en los grandes modelos macroeconométricos. Sus ecuaciones recogen reglas de decisión de los agentes, que dependen de la política que esperan, de modo que, cuando la política cambia, cambian los coeficientes estimados, como ocurrió con la curva de Phillips.

Razonamiento: solo sirven para evaluar políticas los modelos con *parámetros profundos*, las preferencias y la tecnología, derivados de la optimización de agentes con expectativas racionales. De ahí nacen los modelos de ciclo real, que se *calibran*, y los DSGE. La **crítica de Lucas** suele describirse como un cambio de paradigma.

3.4.2. La crítica de Sims y los modelos VAR (1980)

Idea: Sims (Nobel 2011) critica la *identificación* de esos modelos, que excluían variables y declaraban exógenas otras con restricciones "increíbles". Propone los **modelos VAR**, en los que todas las variables son endógenas y cada una se explica por sus retardos y por los de las demás, y que se interpretan con las *funciones de impulso-respuesta*. Se les reprocha ser *ateóricos* y no escapar a la crítica de Lucas.

3.4.3. La macroeconomía tras la crisis de 2008

Crítica: la Gran Recesión reabrió el debate, porque los DSGE no la anticiparon, en su mayoría carecían de fricciones financieras, descansaban en un agente representativo y trataban mal el límite inferior de los tipos. Romer (2016) sostuvo que la disciplina había antepuesto la coherencia formal a la evidencia.

Respuesta: ampliar el marco antes que abandonarlo, con fricciones financieras y agentes heterogéneos (modelos HANK, Kaplan et al., 2018), y aceptar un pluralismo de herramientas. Blanchard (2018) distingue modelos fundacionales, DSGE, de política, de juguete y de predicción. A ello se suman los datos masivos y el aprendizaje automático, útil para predecir, pero que no sustituye la identificación causal (Mullainathan y Spiess, 2017).

1. Formación y madurez del pensamiento clásico

Entre *La riqueza de las naciones* (1776) y la revolución marginalista de la década de 1870 transcurre el periodo clásico. Sus autores comparten la idea de que, gracias a la competencia, el sistema económico se autorregula con mínima intervención estatal, aunque Ricardo subraya el conflicto distributivo entre clases.

1.1. Adam Smith (1776)

Importancia: catedrático de filosofía moral en Glasgow y autor de la *Teoría de los sentimientos morales* (1759), Smith integra en un sistema el conocimiento económico de su generación, y con *La riqueza de las naciones* la economía adquiere el estatus de disciplina académica. Muchas de sus ideas, como su "sistema de libertad natural", coinciden con las de los fisiócratas, como Quesnay, pero deben más a Hutcheson y a Hume. Su método combina la deducción con el análisis inductivo de la historia.

1.1.1. Valor: agua y diamantes y valor-coste

Valor de uso y de cambio: el valor de cambio es la cantidad de otros bienes que una mercancía permite adquirir; el valor de uso, su aptitud para satisfacer deseos, su utilidad total. La **paradoja del valor** consiste en que nada es más útil que el agua y apenas se obtiene nada a cambio de ella, mientras que el diamante, casi inútil, compra muchos bienes. En *La riqueza de las naciones* Smith no trata de resolverla y se centra en el valor de cambio. La paradoja nace de entender el valor de uso como utilidad total y no marginal, lo que deja fuera la escasez relativa.

Precios: a corto plazo, el precio de mercado depende de la cantidad llevada al mercado y de la *demanda efectiva*. A largo plazo, con libre competencia, gravita hacia el **precio natural**, que en la sociedad primitiva, sin capital ni apropiación de la tierra, depende del trabajo incorporado y en la sociedad avanzada, de los costes de producción. Como medida del valor, Smith propone el *trabajo comandado*, la cantidad de trabajo ajeno que una mercancía permite comprar. El valor-coste es circular, porque los precios dependen del salario real, y este de los precios.

1.1.2. Distribución: el fondo de salarios

Smith no tiene una teoría de la distribución, sino un análisis de cada remuneración.

- *Salarios*: según la **doctrina del fondo de salarios**, que sistematizarán McCulloch y Mill, el ahorro de los capitalistas forma un fondo que, dividido por la población trabajadora, da el salario a corto plazo, $w = F_w/L$. El salario depende de la negociación, favorable a los patronos, y del ritmo de aumento de la riqueza, no de su nivel.
- *Beneficios e interés*: el beneficio incluye el interés, y la competencia hace caer secularmente la tasa de beneficios. Hay regiones en *estado progresivo*, como las colonias de Norteamérica, con salarios y beneficios altos, y otras en *estado estacionario*, como la China de su tiempo. Smith legitima el beneficio por una función útil, pero también lo describe como deducción del producto del trabajo, ambigüedad de la que parte Marx.
- *Renta*: es un precio de monopolio de la tierra, y por ello Smith reprocha a los terratenientes que "gustan de cosechar donde nunca sembraron".

1.1.3. Crecimiento: la división del trabajo

Conceptos: la *riqueza* es un flujo anual de bienes, precursor del PIB, frente al metal de los mercantilistas y el excedente agrario de los fisiócratas. El *capital* es circulante, con el fondo de salarios, o fijo. El *trabajo* es productivo si se fija en una mercancía vendible y acumulable, e improductivo en otro caso, como el de criados o soldados.

Determinantes: el producto anual depende de dos factores.

1. *Productividad del trabajo*: depende de la **división del trabajo**, ilustrada con la fábrica de alfileres, que aumenta la destreza, ahorra tiempo y favorece la maquinaria. Está limitada por la *extensión del mercado* y por la *acumulación de capital*.
2. *Proporción de trabajo productivo*: cuanto mayor es el capital respecto de la renta, mayor es esa proporción.

Ambos remiten a la acumulación, que procede del ahorro de los capitalistas: "la parsimonia, y no la laboriosidad, es la causa inmediata del aumento del capital". La acumulación exige un marco institucional favorable al mercado y la propiedad privada. La idea del ahorro como fuente del crecimiento pasa a los modelos clásicos y neoclásicos, aunque en el de Solow un mayor ahorro solo eleva el crecimiento durante la transición.

1.1.4. Comercio: la ventaja absoluta

Contra los mercantilistas, para quienes el comercio es de suma cero, Smith lo considera de suma positiva. Un país tiene **ventaja absoluta** en un bien si su coste unitario es menor o su Pme_L mayor. Le conviene especializarse totalmente en él si la relación real de intercambio está entre los precios de autarquía. La apertura amplía el mercado, eleva la división del trabajo y da salida a los excedentes. No explica por qué comercia un país con ventaja absoluta en todo ni cómo se fija la relación de intercambio, y solo admite especialización total.

1.1.5. Dinero: billetes reales y flujo-especie

Smith acepta la teoría cuantitativa en economía cerrada, pero en una economía pequeña y abierta con moneda convertible la cantidad de dinero es endógena. El papel moneda libera capital, y la **doctrina de los billetes reales** evita la sobreemisión: los bancos emiten solo contra letras comerciales reales. El **mecanismo flujo-especie** de Hume (1752), que Smith recoge en sus *Lectures*, corrige automáticamente los desequilibrios exteriores. El país con déficit pierde oro, bajan sus precios y ganan competitividad sus exportaciones.

1.1.6. Política: mano invisible y *laissez faire*

Laissez faire: el Gobierno tiene tres funciones: defensa, justicia y obras e instituciones públicas que la iniciativa privada no provee. Se apoya en la **mano invisible**: cada individuo es el mejor juez de su interés y, al perseguirlo, promueve sin proponérselo el bienestar colectivo. Smith no usa la expresión *laissez faire* y admite excepciones, como las leyes de navegación.

Hacienda: los impuestos siguen cuatro cánones, igualdad, certeza, comodidad y economía, y el tributo preferido grava la renta de la tierra. Rechaza la deuda, que convierte trabajo productivo en improductivo.

1.1.7. Conclusiones

Sobre la mano invisible se asienta el liberalismo económico. Pero es una intuición, no una demostración, porque el primer teorema del bienestar exige competencia perfecta y sin fallos de mercado, y garantiza la eficiencia, no la equidad. Quedan además sin resolver la paradoja del valor y la circularidad del valor-coste.

1.2. Thomas Robert Malthus (1798)

Importancia: clérigo y primer catedrático de economía política de Inglaterra, su *Ensayo sobre el principio de la población* influyó en Ricardo, con quien polemizó sobre la renta, el valor y el exceso general de oferta. Su método es más inductivo que el de Ricardo.

1.2.1. Crecimiento: la tesis sobre la población

Contexto: la población presionaba sobre los alimentos, la pobreza encarecía las leyes de pobres, y Malthus quiere refutar a Godwin y Condorcet, que atribuían la miseria a las instituciones.

Tesis: sin frenos, la población se duplica cada veinticinco años, en progresión geométrica, mientras los alimentos crecen como mucho en progresión aritmética. El ajuste llega por dos frenos: los *positivos*, que elevan la mortalidad (hambres, guerras, epidemias), y los *preventivos*, que reducen la natalidad, como la *restricción moral*, el retraso del matrimonio. Rechaza los anticonceptivos, que reivindicarán los neomalthusianos.

1.2.2. Ciclos: el subconsumismo malthusiano

En sus *Principios de economía política* (1820), Malthus se pregunta qué ocurre si la demanda agregada es menor que la oferta, contra la **ley de Say** (1803), que niega el exceso general de oferta. Lo motiva la crisis de sobreproducción tras las Guerras Napoleónicas.

Razonamiento: los bienes básicos generan su propia demanda, pero en el resto, cuya demanda depende de los hábitos de gasto de los grupos improductivos, puede haber sobreproducción. Se razona en tres pasos.

1. Los trabajadores no perciben todo el producto y su demanda no basta.
2. Los capitalistas pueden comprar, pero ahorran, y si acumulan amplían la capacidad sin demanda que la absorba.
3. La diferencia debe cubrirla el *consumo improductivo* de terratenientes, funcionarios o sirvientes, porque en caso contrario no se alcanza el pleno empleo.

Conclusiones: Keynes (1936) lo reivindicó como precursor de la demanda efectiva. Pero Malthus no explicó con claridad por qué esta había de ser insuficiente. Como admitía que todo lo ahorrado se invierte, su argumento es de exceso de capacidad y no de atesoramiento, lo que lo separa de Keynes, y además justificaba las rentas de los terratenientes.

1.2.3. Política: la ley de hierro de los salarios

Como toda subida del salario real hace crecer la población, mejorar la situación de los pobres está condenado al fracaso (la *ley de hierro de los salarios*). Malthus propuso abolir gradualmente las leyes de pobres y, a diferencia de Ricardo, defendió las leyes de los cereales (1815).

1.2.4. Conclusiones

Ricardo incorporó su tesis a la economía clásica y Keynes lo reivindicó. Sus presagios no se cumplieron, porque el progreso técnico agrario superó a la población y la transición demográfica redujo la natalidad, pero asociaron la economía a la "ciencia lúgubre".

1.3. David Ricardo (1817)

Importancia: agente de bolsa y diputado, en los *Principios de economía política y tributación* construye el primer sistema analítico de la economía política y es precursor de los modelos económicos. Escribe para criticar las **leyes de los cereales**, que prolongaron la protección agrícola tras las Guerras Napoleónicas. Su método es estrictamente deductivo, con rendimientos decrecientes en la agricultura, población malthusiana y, en la versión simplificada, un único bien salario.

1.3.1. Valor: el valor-trabajo al 93 %

Ricardo acepta la distinción de Smith: la utilidad no mide el valor de cambio, aunque es imprescindible. Busca una *medida invariable del valor* y se centra en los bienes reproducibles. Los no reproducibles, como los cuadros raros, valen por su escasez.

Novedad: la renta de la tierra no forma parte del coste de producción. El precio natural depende del trabajo y del capital, que es trabajo acumulado, de modo que el **trabajo incorporado** es su determinante principal. Admite otros, como el capital fijo o los salarios, pero de efecto pequeño. Por eso Stigler (1958a) la llamó "teoría del valor-trabajo al 93 %".

1.3.2. Distribución: la renta diferencial

Salario: a corto plazo lo fija el fondo de salarios. A largo, la población malthusiana tiende a devolverlo a la subsistencia.

Beneficio: es el residuo tras pagar renta y salarios. Su caída se debe a los rendimientos decrecientes agrícolas.

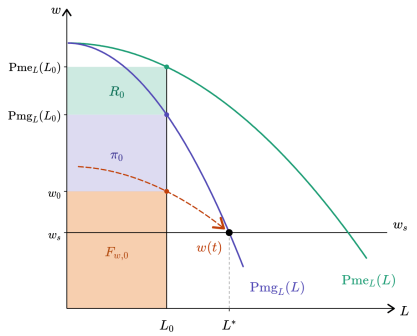
Renta diferencial: es "la parte del producto de la tierra que se paga al terrateniente por el uso de las facultades originarias e indestructibles del suelo". Iguala la tasa de beneficio entre tierras de distinta fertilidad. Con iguales dosis de capital y trabajo, es la diferencia entre el producto de la fértil, A, y el de la marginal, B, que no paga renta, $R_A = Y_A - Y_B$, y retribuye la simple propiedad, no un servicio útil.

Razonamiento: las leyes de los cereales fomentan el cultivo local, agotan el margen intensivo y llevan el cultivo a tierras peores, el margen extensivo. Los rendimientos decrecientes elevan el precio del grano, y con él la renta de las tierras fértiles y el salario monetario de subsistencia, que comprime el beneficio. La mera propiedad de un recurso limitado y desigual genera así una renta, y las leyes de los cereales elevan las rentas y aceleran la llegada del estado estacionario.

Gráficamente: con empleo L_0 y salario de mercado w_0, el producto se reparte en renta, beneficio y salarios:

$$\begin{aligned} R_0 &= [\mathrm{Pme}_L(L_0) - \mathrm{Pmg}_L(L_0)]\, L_0 \\ \pi_0 &= [\mathrm{Pmg}_L(L_0) - w_0]\, L_0 \\ F_{w,0} &= w_0\, L_0 \end{aligned}$$

Al crecer la población, el salario de mercado, $w(t)$, converge al de subsistencia, w_s, y en el estado estacionario, L^*, donde $\mathrm{Pmg}_L(L^*) = w_s$, el beneficio cae a su mínimo, que el modelo simplificado iguala a cero.



Teoría ricardiana de la distribución: en L_0, renta, R_0, beneficio, π_0, y salarios, $F_{w,0}$; el salario de mercado, $w(t)$, converge a w_s en el estado estacionario, L^*.

1.3.3. Crecimiento: el estado estacionario "lúgubre"

La acumulación, alimentada por los beneficios, impulsa el crecimiento. Cuando se agotan, se llega al **estado estacionario**: beneficio mínimo, salario de subsistencia y todo lo demás para los terratenientes. En 1821 admite, en "Sobre la maquinaria", que las máquinas pueden perjudicar a los trabajadores.

1.3.4. Comercio: la ventaja comparativa

Con el paño y el vino de Inglaterra y Portugal, Ricardo introduce la **ventaja comparativa**. Un país la tiene en un bien si su coste de oportunidad es menor. Aunque Portugal tenga ventaja absoluta en ambos, hay ganancias mutuas si la relación de intercambio está entre los precios de autarquía. El patrón lo explican las diferencias tecnológicas relativas. Resuelve el problema de la desventaja absoluta, pero no determina la relación de intercambio, mantiene la especialización total y descansa en un solo factor.

1.3.5. Dinero: el debate bullonista

Durante la restricción del Banco de Inglaterra (1797-1821) subió el oro y se depreció la libra. Ricardo (1810) lo atribuyó al exceso de emisión, frente a los antibullonistas y su doctrina de los billetes reales, e inspiró, con Thornton, el *Bullion Report*. Propuso ligar la oferta monetaria a lingotes de oro, el *ingot plan*, precursor del patrón lingote-oro, y su plan de banco nacional (1824), que separaba la emisión de la banca, inspiró a la escuela monetaria.

1.3.6. Política: el *laissez faire*

Para retrasar el estado estacionario, propone el libre comercio de cereales y confía en el progreso técnico agrario. En Hacienda, prefiere los impuestos a la deuda y, aunque en teoría son equivalentes, duda de que lo sean en la práctica. Barro (1974) revivió la idea, conocida hoy como **equivalencia ricardiana**. Analiza también la incidencia, porque un impuesto sobre la renta de la tierra recae en el terrateniente, y uno sobre los salarios, en el beneficio.

1.3.7. Conclusiones

Aspectos positivos: es el primer modelo coherente de cómo varía el reparto entre trabajadores, capitalistas y terratenientes, y muestra que sus intereses se contraponen, idea que influirá en Marx.

Limitaciones: su método abstracto, el *vicio ricardiano* de Schumpeter (1954), extrae conclusiones de política de supuestos muy simples, y el estado estacionario no llegó, porque el progreso técnico compensó los rendimientos decrecientes.

En suma, Smith sitúa el crecimiento en la división del trabajo, Malthus le opone la población y la demanda insuficiente, y Ricardo integra la población malthusiana y los rendimientos decrecientes en un sistema en el que la renta comprime el beneficio hasta detener la acumulación.

2. Transición al pensamiento neoclásico

Tras Ricardo, los socialistas ricardianos usan el valor-trabajo para reclamar todo el producto para el trabajo. Senior y Mill revisan el sistema y preparan la transición neoclásica.

2.1. Nassau Senior (1836)

Primer catedrático Drummond de Oxford (1825) y coautor del informe de la ley de pobres de 1834, Senior defiende en su *Esbozo de la ciencia de la economía política* una economía positiva, separada del consejo político, distinción que J. N. Keynes (1891) sistematizará entre economía positiva, normativa y arte. Deduce la ciencia de cuatro proposiciones: obtener la mayor riqueza con el menor sacrificio, el principio de la población, la productividad del capital y los rendimientos decrecientes agrícolas.

El valor depende de la utilidad y de la limitación de la oferta, a través de la cual actúa el coste de producción, que incluye el trabajo y la **abstinencia**. El interés remunera la renuncia al consumo inmediato, lo que distingue el beneficio del interés y anticipa la espera de Marshall y la preferencia temporal.

2.2. John Stuart Mill (1848)

Importancia: pensador ecléctico, es el puente entre clásicos y neoclásicos. Sus *Principios de economía política* fueron el texto de referencia hasta finales del s. XIX. Su método es deductivo, *a priori*, desde un agente que solo desea la riqueza, el futuro *homo economicus*.

2.2.1. Valor: el valor-coste de producción

Mill, continuador de Ricardo, hace depender el valor del coste de producción, con el trabajo como elemento principal, y reconoce el papel de la demanda según tres tipos de bienes:

- *Oferta inelástica*, como las obras de arte: el precio lo fija la demanda.
- *Oferta totalmente elástica*, como las manufacturas: lo fijan los costes, constantes.
- *Oferta con pendiente positiva*, como la agricultura: oferta y demanda, con costes crecientes.

Fundamenta así la crítica marginalista y la concepción de la oferta y la demanda como curvas.

2.2.2. Distribución: rechazo del fondo de salarios

Mill acepta la doctrina del fondo de salarios, pero se retracta en 1869, en su reseña de Thornton. El fondo no está predeterminado, y unos sindicatos con poder pueden elevar el salario a costa del beneficio. Mantiene la caída de los beneficios y la ley de Say: puede haber excesos de oferta parciales, pero no permanentes y generales.

2.2.3. Crecimiento: el estado estacionario "feliz"

Mill separa las leyes de la **producción**, inmutables como las de la física, de las de la **distribución**, que dependen de las instituciones y pueden reformarse. Educando a los trabajadores para limitar la natalidad, el salario puede superar la subsistencia, y el estado estacionario puede ser amable: "una condición estacionaria del capital y de la población no implica un estado estacionario del progreso humano".

2.2.4. Comercio: las demandas recíprocas

Mill define con precisión la relación real de intercambio y la explica con la **teoría de las demandas recíprocas**: depende de la fuerza relativa de la demanda de importaciones de cada país. Identifica el mecanismo dentro de los límites ricardianos, pero él mismo advirtió que varias relaciones de intercambio pueden satisfacer la ecuación de la demanda internacional, indeterminación que aclararon las curvas de oferta de Marshall y Edgeworth. Admite, como única excepción económica al libre comercio, la protección temporal de una *industria naciente* (tema B07).

2.2.5. Teoría monetaria y bancaria

Tras las crisis de 1825 y 1836-1839, la **escuela monetaria** de Overstone pide limitar la emisión y respaldarla con oro, y la **escuela bancaria** de Tooke, a la que se adhiere Mill, confía en la doctrina de los billetes reales. Mill matiza su adhesión, porque los billetes apenas influyen al comienzo de la especulación, aunque pueden alimentarla después, y concluye que la Ley de 1844, al impedir prestar en el pánico, agrava las crisis. Además, explica las crisis por la especulación y el crédito y muestra que, con crédito, puede haber un exceso general de oferta, aunque breve.

2.2.6. Política: bienestar e imposición

Mill admite excepciones al *laissez faire*, muchas de ellas hoy fallos de mercado. Distingue funciones *necesarias* del Estado, como la justicia o la defensa, y *facultativas*, como regular el trabajo infantil, primer esbozo del Estado del bienestar. En impuestos, completa los cánones de Smith con la capacidad de pago, como igualdad de sacrificio, y la exención de un mínimo. La progresividad por escala le parece un impuesto sobre la laboriosidad y el ahorro, salvo en las herencias.

2.2.7. Conclusiones

Mill es el clásico más atento a la vertiente social y pionero en el análisis de la intervención pública. Pero su eclecticismo carece de principio unificador y su valor-coste quedó superado por la revolución marginalista.

En suma, Senior y Mill mantienen el valor-coste y corrigen el sistema clásico: separan la ciencia positiva, explican el interés por la abstinencia, abandonan el fondo de salarios y hacen del estado estacionario una situación compatible con el progreso.

3. Crítica al pensamiento clásico: Marx

Doctor en filosofía, periodista y revolucionario exiliado en Londres, Marx parte del valor-trabajo, por lo que suele situarse al final de la escuela clásica. Su obra principal es *El capital* (1867), cuyos libros II y III editó Engels. Su método es dialéctico e histórico.

3.1. Rasgos generales y materialismo histórico

Diferencias con los clásicos: Marx reduce las clases a dos, capitalistas, dueños de los medios de producción, y proletarios, que solo venden su fuerza de trabajo. Aunque en el libro III conserva a los terratenientes, el conflicto central enfrenta a esas dos. Y frente a unas leyes "naturales", sostiene que las leyes económicas son históricas y encierran relaciones de poder y lucha de clases.

Modos de producción: su teoría económica aplica su teoría de la historia, el **materialismo histórico**, expuesta en la *Contribución a la crítica de la economía política* (1859) y en el *Manifiesto comunista* (1848). Cada época, esclavista, feudal o capitalista, se corresponde con un modo de producción, con dos clases antagónicas. Se compone de *fuerzas productivas*, la tecnología, que son dinámicas, y *relaciones de producción*, la propiedad y el reparto, que son estáticas. Sobre la estructura económica se levanta la *superestructura* jurídica y política, que mantiene el *statu quo*. El *fetichismo de la mercancía* presenta las relaciones sociales como relaciones entre cosas, y la *alienación* separa al trabajador de lo que produce y de su propia actividad.

Dialéctica materialista: Marx invierte la dialéctica de Hegel, y son las condiciones materiales las que mueven la historia. La tríada que sigue es una lectura didáctica, no de Marx. A la *armonía* entre fuerzas y relaciones sigue la *contradicción*, en forma de lucha de clases, y la *síntesis*, la revolución social. El capitalismo lleva al socialismo y este, al comunismo, sin clases ni Estado.

3.2. Teoría del valor

Marx usa el valor sobre todo para explicar el origen del beneficio, y trata los precios en el libro III. Se apoya en Ricardo y en la distinción entre productos, con solo valor de uso, y mercancías, con valor de uso y de cambio. El **tiempo de trabajo socialmente necesario** determina el valor. El *trabajo concreto*, útil y distinto en cada actividad, crea el valor de uso, mientras que el *trabajo abstracto*, gasto homogéneo de energía en el que el trabajo cualificado es un múltiplo del simple, crea el valor.

Composición del valor: $W = c + v + s$.

- *Capital constante*, c: el valor de los medios de producción consumidos, que se transfiere sin crear valor nuevo.
- *Capital variable*, v: los salarios. Se llama variable porque la fuerza de trabajo crea más valor del que cuesta.
- **Plusvalía**, s: el valor creado por encima de v, del que se apropia el capitalista, $s = W - (c + v)$.

Toda economía capitalista produce, así, más valor del necesario para cubrir los costes sociales reales de producción.

3.3. Distribución: plusvalía y explotación

El capital recorre el circuito D-M-D , y su incremento nace de distinguir el *trabajo* de la **fuerza de trabajo**, que se vende por su valor, el de los bienes de subsistencia, con un "elemento histórico y moral". En el *trabajo necesario* el trabajador reproduce su salario y en el *plustrabajo* crea la plusvalía, sin ningún intercambio desigual. La plusvalía *absoluta* alarga o intensifica la jornada, mientras que la *relativa* abarata los bienes de subsistencia. Como los medios de producción están en manos de los capitalistas, el trabajador no percibe la plusvalía que genera, y de ahí proceden, según Marx, la alienación y la explotación. En el libro III, la plusvalía se reparte entre beneficio industrial, interés y renta de la tierra.

3.4. Leyes del movimiento capitalista

Tasa de ganancia decreciente: con la tasa de plusvalía, $s_v = s/v$, y la composición orgánica, $\theta = c/(c+v)$, la tasa de ganancia es:

$$r = \frac{s}{c+v} = s_v\,(1-\theta)$$
$$\uparrow c \;\Rightarrow \uparrow \theta \;\Rightarrow \downarrow r$$

Quien mecaniza primero obtiene una plusvalía extraordinaria, pero cuando todos lo hacen sube θ y cae r, salvo que s_v suba lo bastante. La ley supone constante s_v. Las subidas salariales también reducen r, pero pertenecen al ciclo de la distribución. Como Smith y Ricardo, Marx predice la caída del beneficio, pero por la composición orgánica, no por la competencia ni los rendimientos decrecientes.

Concentración: la gran empresa mecanizada desplaza a los pequeños talleres. La mayor capacidad puede desembocar en sobreproducción y caída de precios que solo resisten los más eficientes, y el capital se reúne en menos manos.

Ejército industrial de reserva: frente a Malthus, la mecanización y las crisis crean una sobrepoblación relativa que contiene los salarios dentro de límites compatibles con la acumulación. Hay desempleo involuntario, lo que equivale a rechazar la ley de Say en la lectura de Keynes (1936).

Miseria creciente: cuanto mayor es el ejército de reserva, peor es la situación del proletariado, y los capitalistas endurecen las condiciones de trabajo para frenar la caída de sus beneficios.

Crisis: sin teoría única, Marx señala la caída de la ganancia, la desproporción entre sectores (la reproducción simple exige $v_1 + s_1 = c_2$) y el subconsumo, con la acumulación como motor común. Como en una economía monetaria la venta y la compra se separan, rechaza la ley de Say. El ejército de reserva genera además un ciclo de la distribución. Con salarios altos cae la plusvalía, y con ella la ganancia y la acumulación. La implicación es el fin del capitalismo, porque la acumulación agrupa y organiza a una clase obrera creciente que acaba expropiando a los expropiadores, sin que Marx deduzca un derrumbe automático.

Conclusiones: Marx trata el ciclo y el cambio técnico como endógenos, algo que le reconoció Schumpeter (1942). Pero cada ley tiene objeciones.

- *Ganancia*: el capital puede elevar la productividad y la tasa de plusvalía, por lo que no es seguro que una mayor dotación de capital rebaje la tasa de ganancia.
- *Concentración*: se ha cumplido en parte, matizada por la competencia internacional.
- *Ejército de reserva*: en Marx, el ajuste del salario es el propio ciclo y el cambio técnico recrea el ejército de reserva, pero la mecanización no ha generado un paro creciente a largo plazo.
- *Miseria*: en renta real no se ha cumplido. En Gran Bretaña, los salarios reales crecieron al ritmo de la productividad desde la década de 1840.
- *Crisis*: no hay una teoría rigurosa del ciclo.

3.5. Política: el *Manifiesto comunista* (1848)

Marx y Engels plantean un programa de transición, con el que el proletariado, en el poder, arrebataría poco a poco el capital a la burguesía y lo centralizaría en el Estado. Para los países más avanzados enumeran diez medidas: abolir la propiedad de la tierra y la herencia, confiscar los bienes de emigrados y rebeldes, centralizar el crédito en un banco nacional y los transportes en el Estado, extender las fábricas estatales, un impuesto muy progresivo, trabajo obligatorio, fusión de agricultura e industria y educación pública gratuita.

La propiedad estatal y la planificación definieron las economías socialistas. Otras medidas, como el impuesto progresivo, la escuela pública gratuita o un banco central con el monopolio de emisión, acabaron en las economías de mercado.

3.6. Fracaso del comunismo y del valor-trabajo

Marx pone en primer plano la distribución y el conflicto de clases, y sus esquemas de reproducción y su ciclo de la distribución anticipan los modelos de Feldman (1928) y Goodwin (1967). Pero sus leyes se han cumplido solo en parte, el carácter normativo de la explotación es discutido y su teoría ha recibido dos críticas decisivas.

3.6.1. El problema de la planificación central

Cálculo económico: sin propiedad privada no hay precios de los bienes de capital y no puede calcularse la asignación eficiente (Mises, 1920). Lange (1936) propuso un socialismo de mercado con precios por ensayo y error, y Hayek (1945) replicó que la información está dispersa y solo los precios la agregan. Ya en *Camino de servidumbre* (1944) había sostenido que la planificación central acaba con la libertad política.

Incentivos: con una *restricción presupuestaria blanda*, el Estado cubre las pérdidas y las empresas no controlan costes ni innovan (Kornai,

1980).

3.6.2. El problema de la transformación (Böhm-Bawerk)

Anunciado por Engels (1885), planteado por Böhm-Bawerk (1896) y resuelto formalmente por Bortkiewicz (1907), el **problema de la transformación** es que el valor-trabajo no explica cómo se convierten los valores en precios de producción cuando la composición orgánica difiere entre industrias y la tasa de beneficio es uniforme. Corregido el error de Marx, que valoraba los insumos a sus valores, la suma de precios y de valores y la de beneficios y plusvalías no se igualan a la vez en general. Para la mayoría de los economistas, el problema debilitó seriamente el conjunto de la teoría. La ortodoxia vio en él la confirmación del análisis marginal, y el marxismo quedó en la heterodoxia, centrado en los fenómenos sociales. Sraffa (1960) mostró que precios y ganancia dependen de la técnica y del salario real, y los valores-trabajo resultan redundantes (Steedman, 1977).

1. Esfera austro-alemana

En el mundo germanohablante domina la escuela historicista, hostil a la teoría abstracta. La tradición que Menger inaugura en Viena en 1871 se caracteriza por el subjetivismo del valor, el razonamiento deductivo verbal, sin matemáticas, y la atención al tiempo, la incertidumbre y el empresario. Su identidad la fijan dos controversias, la *Methodenstreit* con el historicismo y el debate sobre el cálculo socialista.

1.1. Antecedentes: proto-marginalistas

1.1.1. Von Thünen (1826; 1850)

Von Thünen aplicó el cálculo al análisis marginal: el empresario contrata trabajadores hasta que el producto del último iguala su salario, lo que anticipa en medio siglo la teoría marginal de la distribución. En *El Estado aislado*, de influencia explícita sobre Marshall, formuló el primer modelo espacial, con cultivos en anillos concéntricos según el coste de transporte, que explica la renta por la localización y no solo por la fertilidad.

1.1.2. Gossen (1854)

Ignorado en su época y rescatado por Jevons (1879), Gossen formula dos leyes.

1. *Utilidad marginal decreciente*: el placer de cada nuevo acto de consumo es menor que el del anterior, idea ya planteada por Bernoulli o Bentham.
2. *Equimarginalidad*: el consumidor maximiza su utilidad cuando la última unidad monetaria gastada en cada bien le reporta la misma utilidad. Formalmente:

$$\frac{u_i'(x_i)}{p_i} = \frac{u_j'(x_j)}{p_j}, \quad \forall i, j$$

1.2. Carl Menger (1871)

Importancia: en sus *Principios de economía*, Menger convierte la utilidad marginal en el principio del que se deriva todo el sistema y funda una escuela que le da continuidad. Su método deductivo, sin matemáticas, lo enfrentó a Schmoller en la *Methodenstreit* (1883).

1.2.1. Teoría del valor

Menger resuelve con claridad la **paradoja del agua y los diamantes**, que nace de confundir la utilidad total, ligada al valor de uso, con la marginal. El valor de cambio lo fija la utilidad marginal, y como decrece con la cantidad, la última unidad de agua, abundante, aporta muy poco, mientras que la del diamante, escaso, aporta mucho. La paradoja la había planteado Smith, al observar que el agua tiene más valor de uso que los diamantes y, sin embargo, menos valor de cambio.

1.2.2. Producción y distribución

Órdenes de bienes: los de *primer orden* son bienes finales, como el pan; los de *orden superior* son factores que contribuyen a producirlos (la harina es de segundo orden, el trigo y los molinos de tercero). Son *complementarios*, porque solo son productivos conjuntamente, y su valor se obtiene por **imputación**: derivan su valor del bien final que ayudan a producir. Así, la utilidad marginal del bien final, y no el trabajo o el coste, determina el valor, idea central de la escuela austriaca. Si se deja de demandar pan, se dejará de demandar el trigo para panificar, y así sucesivamente.

Tiempo e incertidumbre: la transformación lleva tiempo y el futuro es incierto, de modo que los bienes presentes valen más que los futuros, lo que prefigura la **preferencia temporal**. Con incertidumbre, los mercados no tienen por qué estar en equilibrio, de ahí el papel del empresario, sobre el que los austriacos posteriores construirán la competencia como proceso dinámico, frente a la visión estática de Jevons y Walras. El empresario es el agente que busca información y detecta oportunidades de beneficio.

1.2.3. El origen del dinero

El dinero surge de forma espontánea, sin Estado. En el trueque, cada agente acepta los bienes más fáciles de revender, y la práctica imitada convierte en medio de cambio a la mercancía más vendible, normalmente los metales preciosos. Mises partió de esta explicación para fundar el valor del dinero en su teorema de la regresión.

1.2.4. Conclusiones

Su valor subjetivo y la imputación fundan la escuela austriaca. Pero aplica tímidamente el análisis marginal a la oferta y, centrado en la demanda, no determina completamente los precios.

1.3. Sistematización: Wieser y Böhm-Bawerk

Menger no explica cómo repartir el valor entre los factores ni de dónde procede el rendimiento del capital. Sus discípulos completan el sistema, Wieser con la imputación y el coste de oportunidad, y Böhm-Bawerk con la teoría del capital y el interés.

1.3.1. Coste de oportunidad e imputación (Wieser)

Wieser (1884; 1889) imputa el valor del bien final a los factores y es el primero en introducir explícitamente el **coste de oportunidad**, el valor de la mejor alternativa a la que se renuncia al destinar un factor a un uso. Apuntada antes por von Thünen y Mill, su desarrollo sistemático hace que se conozca como ley de Wieser. Prolonga así un paso más la cadena causal de Menger, que va de la utilidad marginal al precio del bien final.

1.3.2. El problema del interés (Böhm-Bawerk, 1889)

Origen: para Böhm-Bawerk, los únicos factores originarios son la tierra y el trabajo. El capital es trabajo y tierra acumulados que permiten **métodos de producción indirectos** o *roundabout*, más productivos cuanto más largos. Si el valor de los medios deriva del valor del producto, ¿por qué este los supera y deja un interés?

Idea: en la *Teoría positiva del capital*, el interés es positivo porque los bienes presentes valen más que los futuros, por tres motivos.

- *Circunstancias*: quien espera estar mejor provisto en el futuro valora más el presente, como en una economía en crecimiento, en la que los bienes presentes son relativamente más escasos.
- *Miopía*: se infravaloran las necesidades futuras.
- *Superioridad técnica*: los bienes presentes permiten iniciar métodos indirectos más productivos, como pescar con red frente a pescar con la mano. Sin una remuneración del capital, ningún empresario elegiría esos métodos.

El interés es el *agio* que iguala bienes presentes y futuros, y Böhm-Bawerk da primacía al tercer motivo. Usó su teoría para criticar el valor-trabajo y la explotación de Marx, a lo que replicó Hilferding.

Combina preferencia temporal y productividad, base de la teoría de Fisher y del ciclo austriaco. Pero Fisher mostró que el tercer motivo no es independiente de los otros dos.

1.4. Evolución posterior: Mises y Hayek

Con la tercera generación, el análisis se desplaza a la moneda, el ciclo, la crítica de la planificación y la teoría del conocimiento.

1.4.1. Ciclo, praxeología y cálculo socialista (Mises)

Dinero y ciclo: para Mises (1912), el *teorema de la regresión* resuelve la circularidad del valor del dinero retrocediendo hasta cuando era una mercancía ordinaria. Sienta las bases de la teoría austriaca del ciclo. La expansión del crédito sitúa el tipo de mercado por debajo del natural y orienta la inversión hacia procesos demasiado largos (*malinvestment*), que la escasez de ahorro obliga a liquidar. El teorema extiende al dinero la utilidad marginal, y en ese ciclo se rompe la neutralidad monetaria clásica.

Cálculo socialista (1920): sin propiedad privada ni precios de mercado, el socialismo no puede asignar racionalmente los recursos. Le replicaron Lange y Lerner, que proponían simular el mercado por ensayo y error, sobre la base teórica previa de Barone (1908).

Praxeología: para Mises (1949), la economía estudia la acción humana a partir de axiomas *a priori*. La sistematiza a partir de sus *Grundprobleme der Nationalökonomie* (1933), como enfoque metodológico de la economía.

1.4.2. Ciclo y conocimiento disperso (Hayek)

Ciclo: en *Precios y producción* (1931) formaliza la intuición de Mises con el triángulo hayekiano de la estructura intertemporal de la producción y el tipo natural de Wicksell. De ahí su oposición a los estímulos de demanda, que solo posponen la liquidación, y su polémica con Keynes (1931-1932). Desde la London School of Economics, a la que se incorpora en 1931, convierte el aparato austriaco en la principal alternativa teórica al keynesianismo emergente.

Conocimiento: en *The Use of Knowledge in Society* (1945), frente a Lange, el cálculo socialista es un problema de conocimiento, porque la información está dispersa, es en parte tácita y solo se revela en el mercado a través de los precios. Ningún planificador puede, por tanto, reunirla de antemano.

En la Viena de entreguerras, Haberler, Machlup, Rosenstein-Rodan y Morgenstern, coautor con von Neumann de la teoría de juegos, se integran en la corriente principal.

En suma, los austriacos desplazan el valor-coste con la utilidad marginal, la imputación, el coste de oportunidad y la competencia como proceso. Pero su rechazo de las matemáticas y de la contrastación, y su negativa a determinar los precios simultáneamente, los separan de Marshall y Walras justo cuando convergen, y pese a su recuperación en los setenta, con Kirzner, Rothbard y el Nobel de Hayek (1974), quedan en la heterodoxia. La imputación de Wieser solo funciona con coeficientes fijos, y el periodo medio de producción de Böhm-Bawerk está mal definido y exige conocer el interés que pretende determinar.

2. Esfera anglosajona

La esfera anglosajona, británica en su origen, fija el canon de la microeconomía. El marginalismo surge aquí frente a la escuela clásica más sólida, la de Ricardo y Mill. Jevons es paralelo a Menger, con un método opuesto, y Marshall reintroduce la oferta y funda la escuela de Cambridge, a la que pertenecen Pigou, Robertson, Robinson y el propio Keynes. En Estados Unidos, una generación posterior, con Clark (formado en Alemania), Fisher y Knight, extiende el principio marginal a la distribución y al capital.

2.1. Antecedentes británicos

Los antecedentes son numerosos pero inertes. Bentham fundamenta el utilitarismo en el cálculo del placer y el dolor. Lloyd (1833), Longfield (1834) y Senior (1836) enuncian la utilidad decreciente y su relación con el precio. Bernoulli, precursor común a todas las esferas, recurre ya en 1738 a una utilidad marginal decreciente para resolver la paradoja de San Petersburgo, con lo que introduce la utilidad esperada. En un medio dominado por el valor-trabajo y el coste de producción, ninguno de estos autores alteró el paradigma vigente.

2.2. William Stanley Jevons (1871)

Importancia: en *The Theory of Political Economy*, Jevons reconstruye la teoría del intercambio y la distribución sobre la utilidad marginal para desplazar a Ricardo y Mill. Como Menger, se centra en la demanda, pero su método es matemático y estadístico: si ha de ser ciencia, la economía debe ser matemática. Aplica la estadística al estudio de los precios, con números índice, y del ciclo.

2.2.1. Teoría del valor

El valor depende del *grado final de utilidad*. La "**cadena de Jevons**" es causal y unidireccional: el coste determina la oferta, esta el grado final de utilidad y este el valor. En último término, la demanda fija el precio. Marshall, Edgeworth y Lausana la criticaron por ignorar la determinación simultánea.

Intercambio: dos "cuerpos comerciales" precio-aceptantes alcanzan el equilibrio por la *ley de la indiferencia*, por la que una mercancía no puede tener dos precios a la vez, y comercian hasta que sus RMS se igualan entre sí y con la relación de intercambio:

$$\mathrm{RMS}^A_{YX} = u'^A_X/u'^A_Y = p_X/p_Y = u'^B_X/u'^B_Y = \mathrm{RMS}^B_{YX}$$

2.2.2. Trabajo, capital e interés

El trabajador prolonga la jornada hasta que la utilidad del producto de la última hora iguala su penosidad. Jevons concibe además el capital como tiempo: el interés es el incremento relativo del producto al prolongar la inversión, $F'(t)/F(t)$, lo que anticipa a Böhm-Bawerk. Entiende el capital como el adelanto de subsistencias a los trabajadores, que supone una renuncia temporal y cuyo fruto, el interés, solo aparece más tarde.

2.2.3. El carbón y las manchas solares

En 1865 advierte que mejorar la eficiencia en el uso del carbón aumenta su consumo, la **paradoja de Jevons**, antecedente del efecto rebote. Con la **teoría de las manchas solares** atribuye las fluctuaciones a la actividad solar, uno de los primeros estudios empíricos del ciclo.

2.2.4. Conclusiones

Asentó las matemáticas y la estadística en la economía. Pero Edgeworth mostró que su intercambio entre dos partes, un monopolio bilateral, deja el precio indeterminado, y la teoría de las manchas solares no se sostuvo con los datos. La indeterminación del intercambio se reduce al crecer el número de agentes.

2.3. Alfred Marshall (1890)

Importancia: en *Principios de economía* desarrolla el marco de la microeconomía moderna, generaliza el análisis gráfico y lidera la escuela de Cambridge. Llega al análisis marginal por su cuenta, desde Cournot y von Thünen, y lo reconcilia con la tradición clásica: con él, el marginalismo se convierte en escuela neoclásica. Prioriza el razonamiento verbal y gráfico y relega las matemáticas a notas y apéndices.

2.3.1. Teoría del valor

Idea: oferta y demanda determinan simultáneamente el precio, como las dos hojas de unas tijeras: discutir si el valor lo fija la utilidad o el coste es como discutir qué hoja corta el papel. Concilia así la postura clásica, de oferta, y la marginalista, de demanda. El análisis es de *equilibrio parcial* y estudia un mercado aislado, a diferencia de Walras, que considera la interdependencia de todos.

Planteamiento: un programa para cada lado del mercado.

- *Demanda*: con utilidad cardinal, valor en el margen, utilidad separable y aditiva, cuasilineal en la renta no gastada, m, y utilidad marginal de la renta constante, el consumidor resuelve:

$$\max_{\mathbf{x},\,m} \sum_{i=1}^{n} u_i(x_i) + m \quad \text{s.a.:} \sum_{i=1}^{n} p_i x_i + m = M_0$$

- *Oferta*: la empresa, precio-aceptante, elige la cantidad que maximiza su beneficio dados sus costes:

$$\max_{x_i} \pi(x_i) = p_i x_i - C(x_i)$$

CPO: el consumidor iguala la utilidad marginal de cada bien a su precio por la utilidad marginal de la renta, $u'_i(x_i) = \lambda p_i$ con $\lambda = 1$, y la empresa amplía la producción hasta que el precio iguala el coste marginal. Con $u''_i < 0$ y $\mathrm{Cmg}' > 0$, bastan para un máximo:

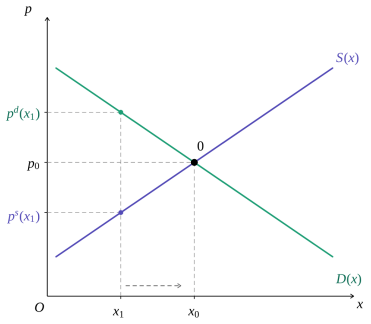
$$u'_i(x_i) = p_i \qquad p_i = \mathrm{Cmg}(x_i)$$

Solución:

- *Demanda*: la demanda inversa es la utilidad marginal y, como decrece, la demanda individual y la de mercado, su suma horizontal, tienen pendiente negativa (la **ley general de la demanda**). Sobre ella construye la elasticidad-precio y la distinción entre movimientos y desplazamientos de la curva. Marshall da por buena la utilidad marginal de la renta constante si el bien absorbe una fracción reducida del presupuesto.
- *Oferta*: la de la empresa es el tramo creciente del coste marginal por encima del mínimo del coste variable medio, y la de la industria a corto plazo, la suma horizontal de las ofertas de las empresas. Distingue cuatro periodos analíticos: el *de mercado*, con oferta fija, en el que la demanda fija el precio; el *corto plazo*, con algún factor fijo y oferta según el coste marginal; el *largo plazo*, con todos los factores variables y oferta horizontal, creciente o decreciente según los costes, y el *secular*, con tecnología variable. Las **economías internas y externas** moldean los costes. El coste pesa más en el precio cuanto más largo es el periodo. Como las economías internas son incompatibles con la competencia perfecta, Marshall atribuye los rendimientos crecientes a las externas, incoherencia de la que parte Sraffa (1926). Añade la organización como cuarto factor, el *excedente del productor* y la *empresa representativa*, la que se gestiona con capacidad normal y accede normalmente a las economías internas y externas de su industria.

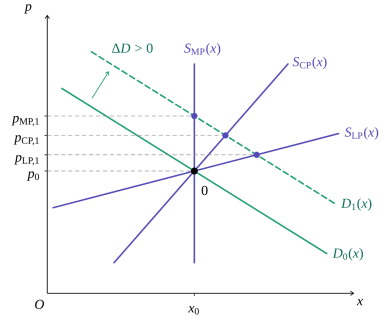
Equilibrio: la cantidad es la variable independiente. Para cada cantidad, la demanda da el máximo que pagarían los compradores y la oferta el mínimo que aceptarían los vendedores. Se cortan en la **cruz marshalliana**, y el ajuste es por cantidades: si el precio de demanda supera al de oferta, la producción aumenta. Con demanda decreciente y oferta creciente el cruce, si existe, es único, pero Marshall admitía que, con rendimientos crecientes, la oferta a largo plazo puede ser decreciente y cortar varias veces la demanda, con equilibrios estables e inestables. En el equilibrio, el precio de demanda, igual a la utilidad marginal, y el de oferta, igual al coste marginal, coinciden:

$$p^d(x_0) = p^s(x_0) = p_0$$



La cruz marshalliana: demanda (utilidad marginal) y oferta (coste marginal); en x_1, $p^d > p^s$ y la cantidad sube hasta x_0.

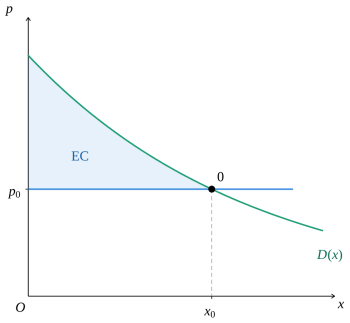
Estática comparativa: ante un aumento de la demanda, en el periodo de mercado solo sube el precio; a corto plazo sube menos, porque aumenta la cantidad, y a largo, aún menos, por la entrada de empresas. Con costes constantes, todo el ajuste recae en la cantidad.



Periodos de Marshall: respuesta del precio a un aumento de la demanda en el periodo de mercado, el corto y el largo plazo.

Bienestar: el **excedente del consumidor** es el área entre la demanda y el precio: las unidades intramarginales se pagan por menos de lo que se estaría dispuesto a pagar. Como suma utilidades, interpreta el excedente de mercado como bienestar colectivo. Con el precio p_0 y la cantidad de equilibrio x_0:

$$\mathrm{EC} = \int_0^{x_0} D(x)\,dx - p_0\,x_0$$



Excedente del consumidor: área entre la demanda y el precio de mercado hasta x_0.

Conclusiones: reconcilia coste y utilidad, pero la teoría fue cuestionada desde 1900 por tres motivos.

- *Medición*: la utilidad cardinal es exigente e innecesaria, porque basta con ordenar las cestas, como mostrará la teoría ordinal.
- *Utilidad marginal de la renta constante*: elimina el efecto renta, de modo que el excedente solo es exacto con preferencias cuasilineales, y no acomoda el caso Giffen, que Marshall recogió en 1895.
- *Separabilidad*: ignora sustitución y complementariedad, que la utilidad general de Edgeworth, base de sus curvas de indiferencia, sí recoge.

2.3.2. Producción y distribución

Su *producto marginal neto* da curvas de demanda de factores y hace de la distribución ricardiana un caso particular de la productividad marginal. Con la **cuasirrenta**, el exceso sobre el coste que obtiene un factor mientras su oferta no se ajusta, muestra que a corto plazo la remuneración depende del precio del bien y a largo plazo lo determina. El producto marginal neto es el incremento del producto ante un aumento del factor, con las cantidades óptimas de los demás, y el marco temporal importa porque influye en la elasticidad de la oferta de los factores. La cuasirrenta se aplica por igual a la renta de la tierra, los salarios, los beneficios y los intereses.

2.3.3. Crecimiento y ciclos

La naturaleza tiende a rendimientos decrecientes y el hombre, con la organización, a crecientes. Acepta con Mill que no cabe una superproducción general permanente y atribuye el ciclo a la confianza empresarial, postura que Hobson criticó desde el subconsumismo.

2.3.4. Teoría del comercio

En 1879 traduce la demanda recíproca de Mill a **curvas de oferta neta**, que determinan gráficamente la relación real de intercambio. Cada curva muestra cuántas exportaciones ofrecería un país a cambio de importaciones según la relación real de intercambio.

2.3.5. Dinero: la versión de Cambridge

Con un análisis de saldos de caja, publicado por Pigou (1917) y Marshall (1923), da fundamentos microeconómicos a la teoría cuantitativa: $M = kPY$, donde k es la fracción de la renta monetaria que se desea mantener en dinero. Con k estable y pleno empleo, más dinero eleva proporcionalmente los precios. La renta real está en su nivel de pleno empleo, independiente de M, por la flexibilidad del salario real y la ley de Say.

2.3.6. Conclusiones

Buena parte del utillaje micro es marshalliano: la cruz, la elasticidad, los excedentes, el corto y el largo plazo, la cuasirrenta y las economías externas. Pero sus costes en U con competencia perfecta los supera la organización industrial, los periodos son un artificio sin correlato real, y la versión de Cambridge descansa en una k estable y la ley de Say, que la Gran Depresión desmintió. La *Teoría general* sustituiría ambos supuestos por una demanda de dinero sensible al tipo de interés y por el principio de la demanda efectiva.

2.4. Evolución hasta la revolución ordinal

2.4.1. Intercambio puro (Edgeworth, 1881)

En *Mathematical Psychics*, con una utilidad general no separable, introduce las curvas de indiferencia y la **curva de contratos**, con múltiples asignaciones eficientes. Con la *recontratación* anticipa el **núcleo**, el conjunto de asignaciones alcanzables no bloqueadas por ninguna coalición, que se reduce a las asignaciones competitivas al crecer el número de agentes, conjetura demostrada por Debreu y Scarf (1963).

2.4.2. Economía del bienestar (Pigou, 1920)

Pigou, sucesor de Marshall, es el padre de la economía del bienestar de equilibrio parcial: el bienestar, aunque subjetivo, es medible en su parte económica, la que admite medida monetaria. En *The Economics of Welfare* generaliza las economías externas como divergencia entre el producto neto marginal privado y el social, origen de la **externalidad**, y propone corregirla con impuestos y subvenciones pigouvianos. La obra es el punto de partida de la teoría de los fallos de mercado.

2.4.3. Competencia imperfecta (1933)

Por caminos independientes, Robinson, desde la crítica de Sraffa en Cambridge, y Chamberlin, desde Harvard, llegan a resultados en parte coincidentes. Robinson difunde el *monopsonio*, analiza las magnitudes medias y marginales del ingreso y del coste y los casos de discriminación de precios. Chamberlin acuña la *competencia monopolística*, con diferenciación de producto y costes de venta, y el teorema del exceso de capacidad. La organización industrial los superó, pero la competencia monopolística resurgió con Dixit y Stiglitz (1977), cuyo modelo es la base de la nueva teoría del comercio.

2.4.4. Teoría ordinal (Hicks y Allen, 1934)

Las objeciones a Marshall se resuelven primero en Yale, Lausana y Kiev y definitivamente en la London School of Economics. Fisher, Pareto, Slutsky y Hicks y Allen concluyen que la utilidad cardinal es innecesaria: basta ordenar cestas, con curvas de indiferencia cuya pendiente es la RMS. Hicks define además la variación compensatoria y la equivalente (1939-1942), base de los principios de compensación. Queda así cerrado el aparato de la teoría del consumidor tal y como hoy se enseña.

2.5. Evolución en Estados Unidos: Clark, Fisher y Knight

2.5.1. John Bates Clark (1899)

Clark separa el **capital social**, un fondo abstracto y permanente, de los bienes de capital concretos en que se encarna sucesivamente (la *transmigración del capital*). Los bienes concretos obtienen rentas absolutas, y el fondo, un interés en porcentaje, justificado por la propia transmigración, que permite reponer los bienes concretos. En estado estacionario, el interés iguala la productividad marginal del capital.

Si cada factor cobra su productividad marginal, no está garantizado que las remuneraciones agoten el producto (el problema del **agotamiento del producto**).

1. *Clark* (1891; 1899): primer intento, verbal y sin demostración formal, para defender el capitalismo frente a George y el marxismo. La teoría marginal de la distribución la formula en *The Distribution of Wealth* (1899).
2. *Wicksteed y Flux* (1894): con competencia perfecta y rendimientos constantes a escala, el teorema de Euler garantiza el agotamiento. Con $f(K, L)$ homogénea de grado uno, el producto se reparte exactamente entre los factores:
$$f(K, L) = \frac{\partial f}{\partial K} K + \frac{\partial f}{\partial L} L$$
3. *Wicksell* (1900; 1901): los rendimientos constantes son suficientes pero no necesarios. La condición necesaria y suficiente es la competencia perfecta y producir en el mínimo de los costes medios a largo plazo, es decir, rendimientos constantes en la dimensión óptima.

2.5.2. Irving Fisher (1930)

El autor: Fisher integra en un sistema ideas dispersas y en parte incompatibles: la preferencia temporal de Böhm-Bawerk, la productividad de Clark y la teoría cuantitativa. Ya en su tesis (1892) representó el equilibrio general con un modelo hidráulico y advirtió que no exige una utilidad medible. Combina formalización y medición: pionero de los números índice, fue el primer presidente de la Econometric Society (1930).

Teoría del capital y el interés: formula la teoría aceptada hoy: el valor del capital es el valor descontado de sus rentas futuras, y el interés deja de ser un residuo para ser el precio que relaciona flujos en el tiempo. Integra la productividad y la **impaciencia**, pero el interés solo es positivo si se prefieren los bienes presentes. De la impaciencia sale su modelo de consumo en dos periodos, del que parten las teorías de la renta normal, y de la productividad, la *tasa de retorno sobre el coste*, que Keynes llamará eficiencia marginal del capital.

Teoría monetaria y de la inflación: añade la *hipótesis de expectativas puras* (1896) y la **ecuación de Fisher**, de $(1 + i) = (1 + r)(1 + \pi^e)$:
$$r \approx i - \pi^e$$

En 1911 da sentido causal a la ecuación de cambio, $MV = PT$, con V y T independientes de M, de modo que el dinero es neutral. En 1933 rompe esa neutralidad con la **deflación por deudas**: la liquidación de activos contrae precios y crédito, y como las deudas son nominales, desapalancarse eleva su carga real. V y T solo dependen de M en los periodos de transición.

Conclusiones: su teoría del interés microfundamenta los fondos prestables y los modelos intertemporales, y de ella procede el *teorema de separación*: con mercados de capitales perfectos, la empresa maximiza su valor presente con independencia de las preferencias temporales de sus propietarios. Pero los datos documentan primas por plazo no nulas y variables, que rechazan las expectativas puras (el tipo a largo como media de los tipos a corto esperados), y la Depresión desmintió la estabilidad de V y T, que el propio Fisher matizó en 1933.

2.5.3. Frank Knight (1921)

El beneficio existe por la **incertidumbre**, no asegurable, y remunera a quien la asume, frente al **riesgo**, asegurable y con probabilidades medibles, que el seguro convierte en coste.

En suma, la esfera anglosajona fija el canon instrumental de la microeconomía y, con su rama norteamericana, cierra el sistema con la productividad marginal y la teoría del capital y el interés. Pero la Gran Depresión derribó la ley de Say y el ajuste automático, el equilibrio parcial fue superado por el general, y la productividad marginal arrastra la objeción de la medición, porque el producto es conjunto. Su uso como justificación de la distribución, explícito en Clark, excede lo que el aparato permite sostener.

3. Esfera franco-italiana

Nace en la tradición francesa de matemáticos e ingenieros. Walras, desde Lausana, sustituye la determinación causal del precio por un sistema de ecuaciones simultáneas (el equilibrio general).

3.1. Antecedentes: proto-marginalismo francés

3.1.1. Cournot (1838)

En sus *Recherches*, Cournot expresa por primera vez la demanda como función decreciente del precio, la *loi du débit*, y resuelve el monopolio, que iguala ingreso y coste marginal, y el duopolio que lleva su nombre. Sin teoría de la utilidad marginal, la condición del monopolista es:

$$\text{Img}(x) = \text{Cmg}(x)$$

3.1.2. Dupuit (1844)

Dupuit identifica la curva de demanda con la de utilidad marginal en dinero. De ahí mide el bienestar con el excedente, su *utilité relative*, que retomará Marshall, y estudia la tarificación discriminatoria de puentes y ferrocarriles. La curva de demanda tiene pendiente negativa porque la utilidad marginal decrece.

3.2. Léon Walras (1874)

Importancia: en *Elementos de economía política pura* integra la demanda y la oferta en la teoría del **equilibrio general** y lidera la escuela de Lausana. Cournot y Dupuit resolvían problemas de un solo mercado, mientras que Walras convierte en objeto la interdependencia de todos, resuelta con un sistema de ecuaciones.

3.2.1. Teoría del valor

Idea: llama *rareté* a la utilidad marginal y, a diferencia de Jevons y Menger, no la liga causalmente al valor, sino que la inserta en un sistema de interrelaciones entre todos los mercados.

Planteamiento: el modelo parte del intercambio puro y se amplía a la producción, con un programa para cada lado del mercado.

- *Demanda*: cada uno de los H consumidores maximiza una utilidad cardinal y aditiva, cuya derivada es la *rareté*, sujeta a su renta, el valor de su dotación inicial:

$$\max_{\mathbf{x}^h} \sum_{i=1}^{n} u_i^h(x_i^h) \quad \text{s.a.:} \quad \sum_{i=1}^{n} p_i x_i^h = M_0^h$$

- *Oferta*: con coeficientes de producción fijos, a_{ji}, la empresa maximiza un beneficio lineal en la producción:

$$\max_{x_i} \pi_i = \left(p_i - \sum_{j=1}^{m} a_{ji}\, p_j^z\right) x_i$$

CPO: cada consumidor iguala su valoración subjetiva al precio relativo y, en la empresa, solo cabe una producción positiva y finita si el precio iguala el coste unitario:

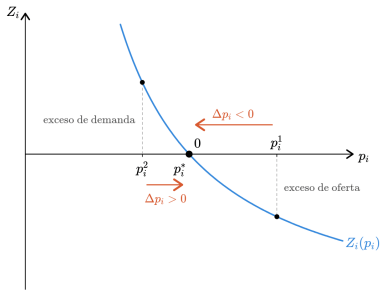
$$\text{RMS}_{ji}^h = \frac{p_i}{p_j} \qquad\qquad p_i = \sum_k a_{ki}\, p_k^z$$

Con *rareté* decreciente, la tangencia es un máximo. El empresario no obtiene ni beneficio ni pérdida, y la demanda fija la escala.

Solución: las demandas, $x_i^h(\mathbf{p}, M_0^h)$, dependen de todos los precios, de modo que la interdependencia está ya en el consumidor. La demanda de cada servicio productivo, $\sum_i a_{ji}\, x_i$, se deriva de la de los productos.

Equilibrio: un vector de precios que vacía todos los mercados, $Z_i(\mathbf{p}^*) = 0$ para todo i. Con producción, el sistema reúne las demandas de productos, precio igual a coste unitario, las ofertas de servicios y su vaciado. Permite enunciar cuatro resultados, de desigual solidez:

- *Ley de Walras*: el valor de los excesos de demanda suma cero, $\sum_i p_i Z_i(\mathbf{p}) = 0$.
- *Existencia*: con un numerario y quitando la ecuación redundante, coinciden ecuaciones e incógnitas, lo que sugiere la existencia, pero no la demuestra.
- *Unicidad*: no está garantizada sin condiciones como la sustituibilidad bruta.
- *Estabilidad*: Walras la sostiene con un argumento heurístico, que el efecto directo de cada revisión de precio sobre su propio mercado domina a los indirectos. Solo se demostró bajo condiciones como la sustituibilidad bruta, y Scarf (1960) mostró que no está garantizada en general.



Exceso de demanda del bien i: se anula en p_i^*, y el tanteo baja el precio con exceso de oferta y lo sube con exceso de demanda.

Gráficamente, $Z_i(p_i)$ es decreciente y se anula en p_i^*: por encima, el exceso de oferta baja el precio, y por debajo, el de demanda lo sube.

Dinámica: en el *tâtonnement* se anuncian precios. Si hay excesos, no se comercia y los precios suben donde hay exceso de demanda y bajan donde hay exceso de oferta, hasta vaciar todos los mercados, $\Delta p_i = \lambda_i Z_i(\mathbf{p})$. El precio es la variable de ajuste, frente a la cantidad en Marshall, en un tanteo con información completa y sin costes de transacción que la tradición personifica en el *subastador walrasiano*.

3.2.2. Producción y distribución

El empresario no es un factor aparte, sino un papel. En equilibrio no obtiene *ni bénéfice ni perte*, porque los precios de los productos igualan sus costes de producción. Sus coeficientes de producción, fijos al principio, pasan a ser variables desde la tercera edición (1896), con la productividad marginal. Los precios de los servicios (renta, salario, interés) salen del mismo sistema, sin una teoría de la distribución aparte.

3.2.3. Política económica

No es un defensor convencido del *laissez faire*: quiere que el Estado cree por ley mercados competitivos y propone nacionalizar la tierra para financiar al Estado sin impuestos. Veía su economía pura como instrumento de política económica.

3.2.4. Conclusiones

El equilibrio general supera al parcial, funda la economía del bienestar y sigue vigente en los modelos de equilibrio general aplicado y DSGE. Pero el *tâtonnement* exige centralización, información completa y ausencia de intercambios fuera del equilibrio, y el recuento de ecuaciones no garantiza precios no negativos.

3.3. El sistema paretiano (Pareto, 1906)

Pareto, sucesor de Walras en Lausana, depura el sistema de la utilidad cardinal y le añade un criterio normativo, con lo que el equilibrio general funda la economía del bienestar.

1. *Utilidad ordinal*: el bienestar no es medible ni comparable. Acuña la *ofelimidad* (1896), la aptitud de un bien para satisfacer un deseo, legítimo o no.
2. *Criterio de Pareto* (*Manual*, 1906): una asignación es eficiente si no se puede mejorar a un agente sin empeorar a otro, sin comparaciones interpersonales. Busca así una teoría del bienestar con un mínimo de juicios de valor.
3. *Ley de Pareto* (1896): el número de perceptores por encima de una renta disminuye de forma potencial con ella, con exponente parecido entre países y épocas. Sirve hoy de base para estimar la concentración de las rentas más altas con datos fiscales.

3.4. Difusión y formalización

Barone (1908) aplica el sistema a una economía planificada, punto de partida formal del debate sobre el cálculo socialista. Además, la escuela resuelve dos problemas. Cassel (1918b) supera la inaccesibilidad de la obra simplificando a Walras (modelo Walras-Cassel). La demostración de existencia llega con Schlesinger y Wald (1933-1936), en el coloquio de Viena, y en general con Arrow y Debreu (1954) y McKenzie (1954).

La formalización alcanza también otros campos. Frisch axiomatiza la demanda (1926; 1932), Bergson (1938) y Samuelson (1947) crean la función de bienestar social, que responde a la indeterminación de los criterios paretianos, y se desarrolla la dualidad, con Hotelling (1932), Roy (1942) y Shephard (1953).

En suma, la esfera pasa del equilibrio parcial al general, con un ajuste por precios, y Pareto lo convierte en criterio para juzgar asignaciones. Por su lenguaje y la falta de una prueba de existencia, tardó en imponerse, y la microeconomía de entreguerras se construyó sobre el aparato marshalliano.

1. Macromodelo neoclásico: una reconstrucción

Los neoclásicos se centraron en microfundamentar el pensamiento clásico, pero su macroeconomía, con teorías del ciclo, del crecimiento y monetarias, no estaba unificada antes de los años veinte. El modelo, que reconstruyeron después Keynes (1936), Modigliani (1944) y Patinkin (1956), se asocia a Cassel (1918b), Marshall (1923), Fisher (1930), Hicks (1932) y Pigou (1933). Hicks, Fisher y Pigou ponen en contacto las teorías del ciclo con el resto de la teoría.

1.1. Principales rasgos

Supuestos: la microfundamentación permite aplicar el equilibrio general al conjunto de la economía. El entorno es "utópico": se cumple la **ley de Say**, por la que toda oferta genera su propia demanda; no hay fallos de mercado, porque la información es perfecta y no hay poder de mercado, bienes públicos ni rigideces, de modo que precios y salarios son flexibles y todos son precio-aceptantes, y rige la **dicotomía clásica**, que separa la economía real de la monetaria. Los agentes son plenamente racionales, como el *homo economicus* clásico.

Ajuste y política: de esas instituciones y de la racionalidad se deriva un ajuste walrasiano, vía precios plenamente flexibles, el salario real y el tipo de interés real, y de ahí procede la defensa de la no intervención. Las políticas de demanda no alteran la renta real ni el empleo.

1.2. Mercado de trabajo

El modelo neoclásico es un **modelo de oferta**, porque la causalidad nace en la oferta. Con el análisis marginal de Hicks (1932) y Pigou (1933), la oferta se obtiene del mercado de trabajo, que determina el empleo, y de la función de producción agregada, que lo convierte en producto ofrecido.

Oferta de trabajo: el trabajador reparte su tiempo entre trabajo y ocio e iguala su RMS entre consumo y ocio al salario real. Si el efecto sustitución domina, la oferta agregada de trabajo crece con el salario real, $L^s(w/P)$. Con o el ocio, x el consumo, l_0 la dotación de tiempo y M_0 la renta exógena, el programa del consumidor-trabajador y su condición de óptimo son:

$$\max_{o,x}\ u(o,x) \quad \text{s.a.: } Px \le w(l_0-o)+M_0 \ \rightarrow\ \text{RMS}_{xo} \equiv \frac{u_o'}{u_x'} = \frac{w}{P}$$

Demanda de trabajo: es derivada de la demanda de bienes. Con capital fijo, la empresa contrata hasta que $f_L' = w/P$, y la demanda agregada de trabajo decrece con el salario real, $L^d(w/P)$. La causalidad va del salario al empleo, uno de los puntos de ruptura de Keynes. La condición resulta del programa de la empresa, con r_K el precio del capital:

$$\max_{L}\ \Pi = Pf(\bar{K},L) - wL - r_K\bar{K} \ \rightarrow\ f_L' = \frac{w}{P}$$

Equilibrio: es único, estable y Pareto-eficiente, con **pleno empleo**. Solo quedan el paro friccional y los no participantes. El desempleo involuntario exige una rigidez que mantenga el salario real demasiado alto. El salario real es un indicador perfecto de la escasez de trabajo. El cruce de ambas curvas fija el empleo y el salario real de pleno empleo:

$$L^s\left(\frac{w}{P}\right) = L^d\left(\frac{w}{P}\right) \ \rightarrow\ L^*,\ \left(\frac{w}{P}\right)^*$$

Estática comparativa y dinámica: las políticas de demanda mueven precios y salarios nominales en la misma proporción y las empresas mantienen sus planes. Un aumento de la productividad, por progreso técnico o más capital, desplaza la demanda de trabajo y eleva el salario real y el empleo. Cualquier perturbación se corrige sola: un exceso de oferta de trabajo baja el salario real hasta el equilibrio. La flexibilidad del salario real es la **primera línea de defensa** de la no intervención.

1.3. Mercado de bienes y servicios

Consumo: solo cubre las necesidades del momento y es casi un residuo del ahorro. Para Fisher (1930), crece con la renta y, si domina el efecto sustitución, decrece con el tipo de interés, $C(Y^+, r^-)$, de modo que el ahorro crece con él, $S_F(r^+)$. Con esta formulación se cumplen la dicotomía y la neutralidad del dinero. Con el efecto Pigou (1943) depende también de la riqueza, $C(Y^+, r^-, W^+)$. La forma de Fisher procede del modelo de consumo intertemporal en dos periodos, y en ella el dinero no forma parte de la riqueza.

Inversión: nadie atesora, y todo el ahorro se invierte. En la "segunda aproximación" de Fisher, la tasa de retorno marginal sobre el coste se iguala al tipo de interés real. Con productividad marginal del capital decreciente, la inversión decrece con él, $I(r)$.

Gasto agregado: la demanda real de bienes es la suma del consumo, la inversión y el gasto público, G_0, exógeno, como los ingresos públicos netos de transferencias, T_0:

$$Y^d = C(Y,r) + I(r) + G_0$$

Ley de Say: con la oferta de pleno empleo, el mercado de bienes se vacía cuando la inversión se financia con el ahorro privado y el público, $I = S_F + (T_0 - G_0)$, es decir, $S = I$. El mercado de trabajo garantiza el pleno empleo y la ley de Say el vaciado de bienes, de modo que la renta de equilibrio es la de pleno empleo a corto y a largo plazo. Formalmente, con la identidad que reparte la renta entre consumo, ahorro e impuestos:

$$Y = C(Y,r) + S_F(r) + T_0, \qquad Y^d = Y \ \rightarrow\ I = S_F + (T_0 - G_0) = S_F + S_G$$

Segunda línea de defensa: la **flexibilidad del tipo de interés real** en el mercado de fondos prestables, una teoría no monetaria del interés. La oferta de fondos es el ahorro, $S_F(r^+)$, y el interés remunera la espera. La demanda es la inversión, $I(r^-)$, más el déficit público. El equilibrio:

$$S_F(\bar{r}) = I(\bar{r}) + (G_0 - T_0)$$

es único y estable. $\bar{r}$ es el **tipo de interés natural** de Wicksell (1898), que fijan fuerzas reales, el ahorro y la productividad del capital, y hacia el que gravita el tipo monetario. Solo varía con perturbaciones reales: más déficit lo eleva y expulsa inversión y consumo, más propensión al ahorro lo reduce y más productividad del capital lo eleva. Pervive como referencia de la política monetaria, pero Keynes objetó que el ahorro depende sobre todo de la renta, de modo que su igualdad con la inversión se alcanza por variaciones de la renta y no del tipo de interés, y que el interés es un fenómeno monetario.

1.4. Mercado de dinero: la teoría cuantitativa (Fisher y Cambridge)

Teoría cuantitativa: por la dicotomía, la economía monetaria se explica aparte. Anunciada por Azpilcueta y la escuela de Salamanca (1556), Bodin (1568) y Hume (1752), sostiene que a largo plazo los precios son proporcionales a la cantidad de dinero, exógena. Fisher (1911) da sentido causal a la ecuación de cambio, con un enfoque macro e institucional, y Cambridge (1917) parte de la fracción k de la renta que se desea mantener en saldos:

$$MV = PY \qquad \Longleftrightarrow \qquad M^d = kPY, \quad k = \frac{1}{V}$$

La demanda de dinero es mecanicista y estable, porque el dinero solo es medio de cambio, y V depende de factores institucionales. Como la renta se fija en el mercado de trabajo, el dinero es un velo y solo mueve precios y salarios nominales. Explica la relación de largo plazo entre dinero y precios, confirmada sobre todo con inflación alta, y es base del monetarismo, pero la velocidad ha resultado inestable y depende del tipo de interés.

1.5. Macromodelo neoclásico OA-DA

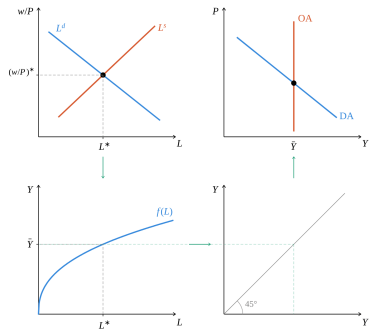
Oferta agregada: llevando L^* a la función de producción con capital fijo:

$$Y = f(\bar{K},L) \ \rightarrow\ Y^s = f(\bar{K},L^*)$$

Es rígida respecto al tipo de interés y a los precios (vertical).

Oferta y demanda agregadas: en el plano renta-precios, la OA es vertical en la renta de pleno empleo, $\bar{Y}$, y la DA no es el gasto real (con la dicotomía, el tipo de interés lo iguala con $\bar{Y}$ a cualquier nivel de precios), sino las combinaciones de renta y precios compatibles con la ecuación de cambio, $PY = MV$, y solo fija el nivel de precios. Es decreciente porque, con MV dado, un nivel de precios más alto compra menos producción:

$$\text{DA}: \quad Y = \frac{MV}{P}$$



Macromodelo neoclásico: el mercado de trabajo fija el pleno empleo, la OA es vertical y la DA solo fija los precios.

Política económica:

- *Monetaria*: puede bajar transitoriamente el tipo de mercado bajo el natural, el proceso acumulativo de Wicksell, pero precios y salarios suben en proporción y el dinero es **neutral**, sin alterar el tipo natural. Debe ceñirse a dar liquidez y contener la inflación.
- *Fiscal*: eleva el tipo de interés y expulsa consumo e inversión por el mismo importe, un *crowding out* total. Solo cambia el tipo y la composición del gasto, no la renta, el empleo ni, con V fija, los precios. De ahí procede la regla del presupuesto equilibrado.

– *Oferta*: la intervención solo se justifica para eliminar obstáculos institucionales, como las rigideces salariales. Los tributos sobre el trabajo, las subvenciones o la formación sí afectan a la renta y al empleo.

En suma, el modelo neoclásico es un primer intento de modelar lo agregado, pero la ley de Say solo es plenamente válida en una economía de trueque, la dicotomía estricta es poco plausible y el *laissez faire* puede no ser aconsejable. En una economía monetaria, la ley de Say exige suponer que el dinero sirve solo como medio de pago y que nadie lo atesora.

2. Pensamiento de Keynes: macromodelo keynesiano

La Gran Depresión, con su desempleo involuntario, desmiente las predicciones neoclásicas. La *Teoría general de la ocupación, el interés y el dinero* (1936) es una sustitución de paradigmas en términos de Kuhn. Frente al *Tratado sobre el dinero* (1930), en el que el ajuste era vía precios y beneficios, ahora la variable de ajuste es la producción. Como Keynes no formalizó su teoría, se expone con el modelo de la síntesis: el IS-LM de Hicks (1937), difundido por Hansen (1953), el salario nominal rígido de Modigliani (1944) y el aspa de Samuelson (1948a). Se supone corto plazo, economía cerrada y cuatro mercados, trabajo, bienes, dinero y bonos, este implícito por la ley de Walras. Con precios fijos, la oferta se adapta a la demanda y $r = i$. El modelo perseguía un doble objetivo, formalizar las ideas de Keynes y aunarlas con las neoclásicas.

2.1. La sustitución de paradigmas

Metodología: combina inducción y deducción en una teoría "realítica", realista y analítica.

Supuestos:

– *Entorno*:
  · *Ley de Say invertida*: la **demanda efectiva**, los ingresos que los empresarios esperan de sus ventas, determina la producción y el empleo. Si es insuficiente, hay desempleo involuntario.
  · *Incertidumbre radical*: el futuro no se reduce a riesgo cuantificable, y las expectativas son centrales. Keynes no ofrece, sin embargo, un modelo de su formación.
  · *Rigidez del salario nominal a la baja*, que Keynes atribuye al salario relativo, con precios flexibles y competencia perfecta en bienes. Aunque los salarios fueran flexibles, su bajada no garantizaría el pleno empleo.
  · *No dicotomía* entre economía real y monetaria, que Keynes no distingue de forma explícita.
– *Agentes*: buscan su interés, pero deciden bajo incertidumbre con convenciones y *animal spirits*, y la lectura habitual añade la ilusión monetaria.

Ajuste: inspirado en Malthus y Marshall, es **marshalliano**, vía cantidades: las existencias, la producción y la renta. Keynes aboga por intervenir, sobre todo por el lado de la demanda.

2.2. Mercado de alquiler de trabajo

Para Keynes, los salarios son también renta, de modo que la demanda agregada es esencial en el mercado de trabajo, y cabe desempleo involuntario incluso con competencia perfecta en bienes.

Oferta de trabajo: la *Teoría general* prescinde de ella. Suele leerse con **ilusión monetaria**: los trabajadores fijan el salario nominal. Si bajan los precios, rechazan rebajas nominales, y el salario es rígido a la baja con un mínimo, w_m, por debajo del cual no se ofrece trabajo. No es un salario de reserva, sino una norma institucional:

$$L^s = 0 \ \ \forall w < w_m, \qquad L^s > 0 \ \ \forall w \geq w_m$$

– La oferta tiene un *tramo horizontal* en w_m, con mucho paro involuntario; uno *creciente*, al reactivarse la economía y ganar fuerza los sindicatos, y uno *vertical* en el pleno empleo.

Si suben los precios, en cambio, aceptan un salario real menor. La asimetría se explica por el *salario relativo*, pues una subida general de precios no altera la posición de nadie en la escala. La rigidez no es condición necesaria del desempleo involuntario, que puede deberse a otras rigideces, ni suficiente, porque con demanda suficiente el desempleo es voluntario. Así, aunque precios y tipo de interés sean flexibles y haya competencia perfecta, el pleno empleo no está garantizado.

Hace depender el empleo de la demanda sin atribuir irracionalidad a los trabajadores, pero justifica la rigidez de forma informal, y la ilusión monetaria no resiste las expectativas de inflación, como mostrarán Friedman y Phelps.

Demanda de trabajo: sigue a los neoclásicos, $f'_L = w/P$, que en el plano empleo-salario nominal da una curva decreciente, pero la causalidad va del empleo al salario, y una demanda de bienes insuficiente origina desempleo.

Equilibrio: si la demanda agregada es insuficiente, el equilibrio está en el tramo horizontal, con paro involuntario igual a la diferencia entre el empleo al salario mínimo y el nivel en que la oferta empieza a crecer. Si es suficiente, está en el creciente, con paro voluntario, aunque aún con recursos ociosos. El equilibrio se da en el plano empleo-salario nominal, y no en el del salario real, como en el modelo neoclásico.

Política económica: las políticas de demanda expansivas son fundamentales mientras haya recursos ociosos. Keynes no apoya rebajar el salario nominal, que solo elevaría el empleo indirectamente, por el efecto Keynes, como una expansión monetaria.

2.3. Bienes y dinero: el modelo IS-LM

Como el desempleo depende de la demanda y no hay dicotomía, bienes y dinero se tratan juntos. La oferta de bienes se adapta a la demanda, con precios fijos, y la de dinero es exógena.

Consumo: frente a su papel residual en los clásicos, es el fin de la actividad económica y la base del multiplicador. Por una "ley psicológica fundamental", al aumentar la renta corriente el consumo sube, pero menos:

$$C = C_0 + cY, \quad C_0 > 0, \ c \in (0,1)$$

– *Multiplicador*: un aumento del gasto autónomo eleva la renta en $1/(1-c) > 1$ veces. La causalidad va de la demanda a la oferta, la política fiscal es muy eficaz y redistribuir estimula el consumo. El efecto es tanto mayor cuanto mayor es c, y los hogares de menor renta tienen mayor propensión marginal a consumir.
– *Paradoja del ahorro*: más ahorro autónomo reduce la renta y no eleva necesariamente el ahorro agregado. Ataca así la defensa clásica del ahorro como fuente del crecimiento.

Las primeras contrastaciones parecían corroborarla, pero la evidencia de Kuznets, con una propensión media constante a largo plazo, obligó a considerar la renta a largo plazo y la riqueza: la renta relativa y el ciclo vital de la síntesis y la renta permanente de Friedman. El ahorro keynesiano es un excedente de ingreso, muy poco influido por el tipo de interés.

Inversión: Keynes distingue la inversión *planeada*, parte de la demanda, de la *efectiva*, que difiere en la variación no planeada de existencias. Invierte la relación clásica: la inversión, movida por las expectativas, determina el ahorro. Quien invierte compra un flujo de retornos esperados, y se invierte mientras la **eficiencia marginal del capital**, la tasa que iguala el valor actual de los rendimientos esperados al coste de reemplazo, supere al tipo de interés. Como la EMK decrece con la inversión, porque caen los rendimientos esperados y sube el coste de reemplazo, el óptimo es EMK $= r$, y $I = I_0 - vr$. I_0, los *animal spirits*, es alto, y v, bajo (el **primer eslabón débil**).

Expectativas a largo plazo: guían la inversión y, como apenas se sabe nada del futuro lejano, dependen del *estado de la confianza* y de una *convención*, la de suponer que lo presente continuará. La bolsa la vuelve frágil, porque cada especulador anticipa lo que la opinión media espera de la opinión media (*concurso de belleza*), y la EMK y la preferencia por la liquidez se vuelven inestables.

Curva IS: los pares renta-tipo de interés que vacían el mercado de bienes:

$$Y = \frac{1}{1-c}\left[A_0 - vr\right]$$

Es decreciente, porque un tipo menor eleva la inversión y, por el multiplicador, la renta. Es tanto más inclinada cuanto menor es v, más para los keynesianos que para los monetaristas, y el gasto autónomo la desplaza. El término $1/(1-c)$ es el multiplicador, y $A_0 = C_0 + I_0 + G_0$ agrupa los componentes autónomos del gasto, de modo que, con r dado:

$$\Delta Y = \frac{1}{1-c}\,\Delta A_0$$

Preferencia por la liquidez: la teoría cuantitativa es solo un caso límite, porque con desempleo el dinero altera sobre todo el empleo y solo en pleno empleo los precios varían en proporción. Se demanda dinero por *transacción*, proporcional a la renta; por *precaución*, que se asimila al anterior, y por *especulación*, la novedad, que liga el dinero al tipo de interés. Como el precio de los bonos varía inversamente con el tipo y cada individuo espera que este vuelva a un valor crítico, i^e, mantiene toda su riqueza, W, en bonos o en dinero según el tipo:

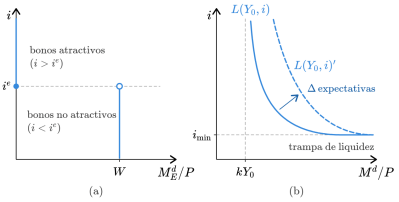
$$\frac{M_E^d}{P} = 0 \ \text{ si } i > i^e, \qquad \frac{M_E^d}{P} = W \ \text{ si } i < i^e$$

Al agregar individuos con valores críticos dispares, la demanda especulativa es continua, decreciente y convexa, y la demanda de dinero queda:

$$\frac{M^d}{P} = L(Y^+, i^-) = kY - hi$$

El término kY recoge la demanda por transacción y precaución, y hi, la especulativa, gobernada por expectativas volátiles, de modo que h es alta y la demanda de dinero, inestable. El interés no recompensa la abstinencia, sino la **renuncia a la liquidez**.

Trampa de liquidez: es el **segundo eslabón débil**. Si el tipo es tan bajo que todos esperan que suba, la demanda de dinero es perfectamente elástica, $h \to \infty$, y cualquier aumento de la cantidad de dinero se mantiene en efectivo, sin destinarse a bonos ni rebajar el tipo de interés. Por eso la política monetaria es inefectiva y la fiscal, muy potente. Keynes la presentó como un caso límite sin ejemplos conocidos.

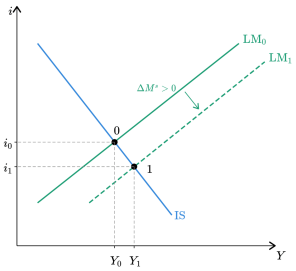


Demanda de dinero keynesiana.

Curva LM: con oferta monetaria exógena, los pares que vacían el mercado de dinero, y por la ley de Walras el de bonos, cumplen $i = (kY - M^s/P)/h$. Es creciente, con pendiente k/h, porque más renta eleva la demanda por transacción y el tipo sube para racionarla. El tipo de interés lo fijan fuerzas monetarias, no los fondos prestables. Es más plana cuanto mayor es h, horizontal en la trampa y casi vertical para los monetaristas, con $h \approx 0$.

Equilibrio: el cruce de la IS y la LM fija la renta y el tipo de interés por el lado de la demanda.

Estática comparativa: una expansión monetaria desplaza la LM a la derecha, baja el tipo y eleva la renta, más cuanto menor es h y mayor v. Una fiscal desplaza la IS a la derecha y eleva renta y tipo, con expulsión parcial de inversión, de modo que, salvo en la trampa de liquidez, su efecto sobre la renta es menor que el del multiplicador simple, $1/(1-c)$. Es más eficaz cuanto mayor es h y menor v.



Equilibrio IS-LM y política monetaria expansiva.

Dinámica: en el **aspa keynesiana**, la demanda planeada, de pendiente c, corta a la recta de 45° en un equilibrio único y estable, que no tiene por qué ser de pleno empleo. Fuera de él se ajustan las existencias: si la producción supera la demanda, se acumulan existencias no deseadas, $(1-c)(Y_1 - Y^*)$, y se recorta la producción. En el mercado de dinero, en cambio, el tipo de interés sí ajusta, aunque en conjunción con las expectativas.

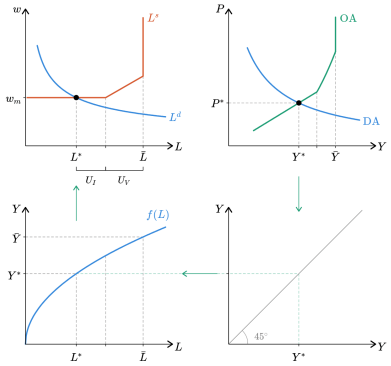
2.4. Macromodelo keynesiano OA-DA

Oferta agregada: se obtiene, como en los neoclásicos, llevando el empleo a la función de producción, pero con causalidad inversa, porque las empresas ajustan la producción a la demanda. Tiene tres tramos: en el primero, los precios elevan el empleo pero no el salario nominal, por la ilusión monetaria; en el segundo, esta se diluye y el salario sube; el tercero, vertical, es el pleno empleo. Hasta $\bar{Y}$ es creciente, porque una subida de precios abarata el salario real y eleva el empleo, y el desempleo involuntario surge en el primer tramo, por una demanda agregada insuficiente. Para Keynes, la cuestión no es asignar recursos, sino explicar por qué la renta puede quedar persistentemente por debajo del pleno empleo. Según la síntesis, la OA sería vertical a largo plazo, pero Keynes ya había advertido en 1923, frente a la teoría cuantitativa, que el largo plazo es una guía engañosa para los asuntos corrientes, porque "a largo plazo todos estaremos muertos".

Demanda agregada: proyecta los equilibrios IS-LM y es decreciente porque un nivel de precios más alto reduce la oferta real de dinero y desplaza la LM a la izquierda.

Equilibrio: la DA corta a la OA con recursos ociosos, con empleo $L^* < \bar{L}$ y paro involuntario y voluntario. Una expansión fiscal o monetaria desplaza la DA, según las pendientes de la IS y la LM, y eleva producción, empleo y precios mientras haya recursos ociosos, más la producción cuanto más plana sea la OA, y solo precios en el tramo vertical.

Una subida de w_m o una caída de la productividad desplazan la OA a la izquierda, con lo que suben los precios y caen la producción y el empleo.



Macromodelo keynesiano OA-DA: con salario rígido en w_m, la OA es creciente hasta el pleno empleo y el equilibrio deja paro involuntario.

Política de Keynes: un uso activo del presupuesto, que supla la demanda que falte, y tipos bajos que animen la inversión productiva mientras el capital sea escaso, hasta la "eutanasia del rentista". En el capítulo 24 pide además dirigir la propensión a consumir con los impuestos y una "socialización bastante completa de la inversión", con un Estado rector de su volumen, sin socialismo de Estado, porque lo importante no es la propiedad de los medios de producción, sino el volumen de inversión.

Política de la síntesis: dominante hasta los años setenta, da la razón a Keynes a corto plazo, con política de demanda discrecional, sobre todo fiscal, la más eficaz por los dos eslabones débiles, y a los neoclásicos a largo, con presupuesto equilibrado y una política monetaria que dé liquidez.

En suma, Keynes explica el desempleo involuntario de los años treinta, rompe la dicotomía con efectos reales del dinero y justifica la intervención pública. Pero no formalizó muchas de sus ideas, de ahí los debates sobre "qué quería decir Keynes", y no considera los retardos de la política económica.

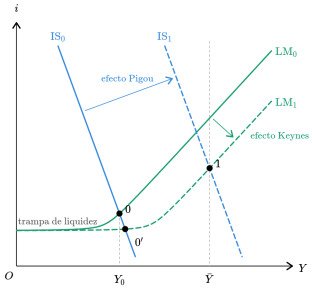
3. Reacciones y desarrollos posteriores

La obra de Keynes y su difusión por la síntesis sitúan la macroeconomía en el centro de la teoría y suscitan varias reacciones.

3.1. El debate sobre los saldos reales

Efecto Keynes (1936): con dinero nominal constante, la deflación eleva los saldos reales, baja el tipo de interés, sube la inversión y, por el multiplicador, la renta: desplaza la LM. Pero el canal es el tipo de interés, que falla en la trampa de liquidez o con inversión insensible, y la deflación daña deudas y expectativas. Por eso Keynes defendía salarios nominales estables, prefería la expansión monetaria a la bajada de salarios y pedía estímulos fiscales ante la deflación. La flexibilidad salarial no logra nada que la política monetaria no consiga con menos costes. Su eficacia depende, además, de las pendientes de la IS y de la LM.

Efecto Pigou (1943): el consumo depende de la riqueza real neta, $W = M^s/P + B/P$, que la deflación revaloriza: desplaza la IS. El tipo de interés deja de ser el único puente entre finanzas y bienes, y los dos eslabones débiles ya no impiden la vuelta automática al pleno empleo. Para los neoclásicos era una oportunidad y un reto, porque rompía la dicotomía. Su desplazamiento de la IS se suma al de la LM del efecto Keynes, y hace a la economía más estable de lo que sugería Keynes.



Efectos Keynes y Pigou en el IS-LM.

- *Metzler* (1951): con el capital en la riqueza, importa su composición. Una compra de títulos del banco central baja permanentemente el tipo real, que deja de depender solo de factores reales, de modo que la compra no es neutral.
- *Patinkin* (1956): reconciliar el efecto Pigou con la neutralidad exige violar la dicotomía.
- *Barro* (1974): con equivalencia ricardiana, los bonos no son riqueza

neta y el efecto se reduce a los saldos monetarios.

En teoría, el equilibrio con desempleo exige alguna rigidez. Pero la deflación necesaria es enorme y, como objetó Kalecki (1944), solo opera sobre el dinero externo, porque con deudas privadas lo que ganan los acreedores lo pierden los deudores, con riesgo de quiebras.

3.2. La síntesis neoclásica

Keynes rompió la dicotomía y abrió canales de transmisión por la demanda, pero su microfundamentación era escasa. La síntesis microfundamenta cada componente de la demanda, con la renta relativa y el ciclo vital en el consumo, el acelerador y la q de Tobin en la inversión y los modelos de Baumol-Tobin y de carteras en la demanda de dinero, y añade un canal riqueza que desplaza la IS. Por el lado de la oferta, la **curva de Phillips** (1958), leída por Samuelson y Solow (1960) como un **menú de política**, y la ley de Okun (1962) parecen permitir el ajuste fino. Pero la curva solo funciona en cualquier horizonte si los trabajadores tienen ilusión monetaria, y la estanflación de los setenta la deshizo y abrió paso al monetarismo.

3.3. Poskeynesianismo

Origen: los discípulos de Keynes en Cambridge, Robinson y Kaldor, con Kalecki, y después Minsky y Davidson en EE. UU., rechazan el "keynesianismo bastardo", como llamó Robinson a la lectura de la síntesis. Le reprochan reducir a Keynes a un caso particular de rigideces.

Idea: la demanda efectiva determina la producción y el empleo también a largo plazo, y la economía capitalista es intrínsecamente inestable, en tiempo histórico irreversible y con incertidumbre no probabilizable, de ahí el rechazo del equilibrio walrasiano. La vertiente británica pone el acento en la distribución, el crecimiento y el valor. Kaldor (1956) y Pasinetti (1962) hacen depender el ahorro, y con él el crecimiento, del reparto entre salarios y beneficios. Robinson y Sraffa sostienen frente a Samuelson y Solow la controversia del capital, y Sraffa (1960) propone una teoría del valor neorricardiana. La norteamericana subraya el papel del dinero y la **inestabilidad financiera** de Minsky, según la cual en el auge la financiación pasa de posiciones cubiertas a especulativas y Ponzi, de modo que la estabilidad genera su propia fragilidad.

Recupera la incertidumbre, la fragilidad financiera y la distribución, y ofrece una explicación endógena de la inestabilidad, pero carece de un modelo canónico, lo que explica su escasa influencia en la corriente principal pese al interés por Minsky tras la crisis *subprime*.

3.4. Keynesianos del desequilibrio

Clower y Leijonhufvud rompen con el IS-LM: para entender a Keynes hay que dejar a Walras y volver a Marshall. A su juicio, la teoría renta-gasto nada tiene que ver con lo que Keynes escribió.

3.4.1. Rechazo de la ley de Walras (Clower, 1965)

Idea: el equilibrio walrasiano es el caso particular, sin racionamiento, de equilibrios no walrasianos con transacciones a precios de no vaciado. La ley de Walras se cumple para las demandas **nocionales**, pero solo se extiende a las **efectivas** con un subastador que coordine los planes. Para Clower, Keynes pensaba en términos de decisiones efectivas.

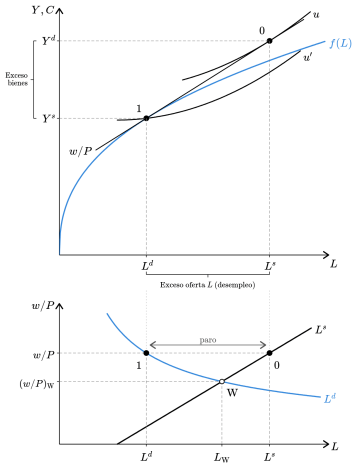
Decisión dual: frente a la decisión unitaria walrasiana, los agentes toman primero decisiones nocionales y después efectivas, sujetas también a **restricciones de cantidades**. Lo que cada uno intercambia depende de la renta efectivamente obtenida. A precios de no vaciado, las decisiones nocionales revelan los lados cortos y largos de los excesos, y los intercambios efectivos son las únicas señales fiables que emiten los mercados.

Equilibrio con racionamiento (Barro y Grossman, 1971; Malinvaud, 1977): rige el intercambio voluntario, hay fallos de coordinación y domina el **lado corto**, $x_i^* = \min\{x_i^d, x_i^s\}$, mientras el largo determina el grado de racionamiento. Los *spillovers* entre mercados hacen el desempleo persistente y subóptimo. Gráficamente, con salarios nominales altos y precios bajos, en el mundo walrasiano hay exceso de oferta de trabajo y de demanda de bienes, que se compensan en valor y el subastador elimina. Sin subastador, el lado corto de cada mercado fija un equilibrio con racionamiento. Los hogares, racionados en el trabajo, ajustan su demanda de bienes a la renta obtenida. Este equilibrio es estable, porque en una economía monetaria el trabajo no vendido no aporta renta con la que comprar. El paro involuntario convive así con agentes que optimizan.

Microfundamenta el multiplicador y el paro involuntario sin ilusión monetaria, pero toma los precios de no vaciado como dados, sin explicarlos, y deja sin fundamentar la rigidez que sostiene el resultado.

3.4.2. Retroalimentación (Leijonhufvud, 1968)

Idea: Leijonhufvud distingue la economía de Keynes de la "economía keynesiana". Para él, la aportación esencial de Keynes es invertir las velocidades de ajuste: sin subastador, las cantidades se ajustan antes que los precios, lo que amplifica las perturbaciones. Una caída de la inversión, que en un contexto walrasiano solo generaría paro friccional, eliminado al instante por el subastador, lleva a las empresas a vender y contratar menos, y a los hogares a ingresar y comprar menos, hasta un



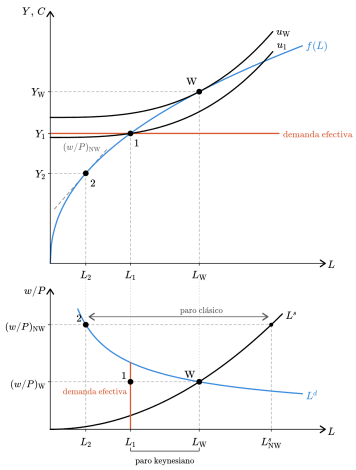
Decisión dual (Clower, 1965): la demanda efectiva de bienes depende de la renta obtenida en el mercado de trabajo; abajo, el paro al salario real dado.

paro mayor y persistente. La incertidumbre eleva además la preferencia por la liquidez y el tipo a largo, justo lo contrario de lo que haría falta. Leijonhufvud defiende así el multiplicador y las políticas de demanda. Desplaza la explicación del paro de la rigidez salarial a los fallos de información y coordinación, pero es una reinterpretación verbal, sin modelo formal manejable.

3.4.3. Política en el mercado de trabajo (Malinvaud, 1977)

Idea: una subida del salario real encarece los costes y estimula la demanda. Su efecto depende del tipo de paro. En el equilibrio walrasiano W, el salario real de vaciado, $(w/P)_{\mathrm{W}}$, iguala la productividad marginal del trabajo y la RMS entre consumo y trabajo.

- *Paro clásico*: salario real superior al de vaciado, por ejemplo por un sindicato con poder, con exceso de demanda de bienes y de oferta de trabajo, de modo que consumidores y trabajadores están racionados. La empresa produce donde $f'(L_2) = (w/P)_{\mathrm{NW}}$, con $L_2 < L_{\mathrm{W}}$. Se corrige bajando el salario real, y el estímulo de demanda es ineficaz.
- *Paro keynesiano*: exceso de oferta en ambos mercados por demanda insuficiente en la fase depresiva, con un salario real inferior, igual o superior al de vaciado, de modo que empresas y trabajadores están racionados. La restricción de demanda efectiva limita las ventas a $Y_1 < Y_{\mathrm{W}}$ y el empleo a $L_1 < L_{\mathrm{W}}$. Ayudan una subida limitada del salario real y el impulso fiscal.
- *Otros*: la inflación reprimida, con excesos de demanda en ambos mercados, típicamente por controles de precios y salarios, y el subconsumo, infrecuente, con exceso de demanda de trabajo y de oferta de bienes.



Paro keynesiano (1) y clásico (2) en una economía de Robinson Crusoe frente al equilibrio walrasiano W; abajo, el mercado de trabajo.

No hay políticas válidas para todos los casos, pero identificar el régimen es difícil y la tipología es estática, sin explicar cómo se pasa de un régimen a otro.

En suma, la escuela aporta la decisión dual y los fallos de coordinación, y la nueva economía keynesiana toma el relevo de microfundamentar las rigideces. Se le objeta un aparato menos elegante que el IS-LM; que el blanco de Keynes era Pigou y no Walras; que Keynes no daba por fijos los precios, sino solo rígidos a la baja los salarios nominales; que

no aclara cuánto duran las rigideces, y que la inconsistencia dinámica y la equivalencia ricardiana de los nuevos clásicos alcanzan a sus recetas.

1. La síntesis neoclásica

La exposición de la síntesis neoclásica recorre sus principales rasgos, el análisis del lado de la demanda y el de la oferta.

1.1. Principales rasgos

Definición: la **síntesis neoclásica**, SNC, es la corriente que, entre los años cuarenta y sesenta, se convirtió en la macroeconomía dominante al reconciliar la *Teoría general* de Keynes (1936) con la tradición neoclásica. El nombre se debe a Samuelson, en la tercera edición de *Economics* (1955).

Arquitectos: fueron tres.

- Hicks formaliza a Keynes en el modelo IS-LM (*Mr. Keynes and the Classics*, 1937).
- Hansen lo difunde, de ahí el nombre de modelo de Hicks-Hansen.
- Modigliani (1944) le añade un mercado de trabajo y muestra que, salvo en la trampa de liquidez, el desempleo involuntario descansa en la rigidez de los salarios nominales.

Idea: a corto plazo la economía se comporta de forma keynesiana, con rigideces, fallos de mercado y posible desempleo; a largo plazo, de forma neoclásica, con flexibilidad, vaciado de los mercados y pleno empleo.

Metodología: su gran aportación es formalizar a Keynes, cuya obra era esencialmente verbal, con el IS-LM con mercado de trabajo, sobre el que se articulan el OA-DA y la curva de Phillips. Combina inducción y deducción: estima las grandes funciones de comportamiento y los primeros macromodelos econométricos, como el de Klein, y hereda el método optimizador para microfundamentarlas.

Supuestos: el rasgo distintivo es su carácter dual.

- *A corto plazo* admite fallos de mercado. El principal es la rigidez a la baja de los salarios nominales, que hace posible el desempleo involuntario. Se rompe la dicotomía clásica y el dinero no es neutral. Entre los fallos figuran también la información imperfecta y la incertidumbre, que dan un papel central a las expectativas, aunque la SNC no ofrece un modelo explícito de su formación.
- *A largo plazo* rige el marco neoclásico, y el efecto saldos reales, o efecto Pigou-Patinkin, devuelve la economía al pleno empleo.

Agentes: optimizan, pero los trabajadores padecen *ilusión monetaria*: atienden al salario nominal y no al real, se resisten a recortes nominales y aceptan caídas del salario real vía inflación.

Ajuste: a corto plazo, por cantidades, con las existencias como variable de ajuste. La SNC invierte la ley de Say y admite un equilibrio de subempleo. A largo plazo, el ajuste es por precios. La flexibilidad de precios y salarios restablece el pleno empleo y vuelve a regir la ley de Say.

Política económica: un intervencionismo activo de gestión de la demanda para cerrar la brecha de producción. Esa confianza en la *sintonía fina* y en la curva de Phillips como disyuntiva explotable será el blanco del monetarismo.

1.2. Análisis del lado de la demanda

El modelo supone corto plazo, economía cerrada y tres mercados: bienes, dinero y bonos. Con precios fijos, la oferta agregada se adapta pasivamente a la demanda y el tipo real coincide con el nominal. El lado de la demanda determina el *output*, con el aspa keynesiana, y el tipo de interés, en el mercado de dinero. El de la oferta introduce el mercado de trabajo y la relación entre inflación y desempleo y relaja el supuesto de precios fijos.

1.2.1. Modelo IS-LM

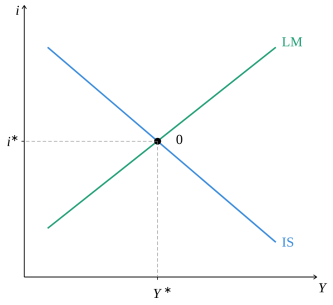
Curvas IS y LM: la IS es el equilibrio en el mercado de bienes, con la producción igual al gasto agregado, y la LM, el del mercado de dinero y bonos, con la oferta monetaria real igual a la demanda de dinero, que depende de la renta y del tipo de interés:

$$\text{IS}: \; Y = C_0 + cY^d + I_0 - vr + G_0, \qquad c \in (0,1)$$
$$\text{LM}: \; \frac{M^s}{P} = kY - hi \;\rightarrow\; i = \frac{1}{h}\left(kY - \frac{M^s}{P}\right)$$

- $c \in (0,1)$ es la propensión marginal a consumir, $Y^d = Y - T_0$ la renta disponible tras impuestos de suma fija, I_0 la inversión autónoma, que recoge los *animal spirits*, v la sensibilidad de la inversión al tipo de interés real y G_0 el gasto público.
- La IS es decreciente, porque un tipo de interés más bajo eleva la inversión y, por el multiplicador, la renta de equilibrio en el aspa keynesiana.
- La LM es creciente, porque más renta eleva la demanda de dinero por transacción y el tipo sube para equilibrar el mercado. En forma general, $M^s/P = L(Y^+, i^-)$, y la forma lineal no recoge el tramo asintóticamente horizontal con que suele dibujarse.

Equilibrio: el corte de ambas curvas da la renta y el tipo de interés de equilibrio, Y^* e i^*, con los dos mercados a la vez en equilibrio. El modelo es de equilibrio general y condensa en dos ecuaciones el comportamiento de una economía cerrada.

Desarrollo gráfico: en el aspa keynesiana, la demanda agregada para el tipo i_0 corta la recta de 45° con renta Y_0, y con un tipo más bajo, i_1, la inversión es mayor y la renta, Y_1. Llevar esos pares al plano renta-tipo de interés da la IS. En el mercado de dinero, con la oferta real vertical, una renta mayor desplaza la demanda de dinero a la derecha y eleva el tipo de i_0 a i_1, lo que da la LM. En el punto 0 se cortan ambas curvas.



Equilibrio simultáneo en los mercados de bienes y de dinero en el modelo IS-LM.

Estática comparativa: con las formas lineales, y $A_0 \equiv C_0 - cT_0 + I_0 + G_0$, la renta de equilibrio y los multiplicadores monetario y fiscal son:

$$Y^* = \frac{h\,A_0 + v\,M^s/P}{h(1-c) + v\,k}$$
$$\frac{\partial Y^*}{\partial (M^s/P)} = \frac{v}{h(1-c)+vk}, \qquad \frac{\partial Y^*}{\partial G_0} = \frac{h}{h(1-c)+vk}$$

- *Política monetaria*: más M^s/P desplaza la LM a la derecha, eleva la renta y reduce el tipo.
- *Política fiscal*: más G_0 desplaza la IS a la derecha y eleva renta y tipo. Su multiplicador es menor que el del aspa keynesiana, $1/(1-c)$, porque la subida del tipo expulsa inversión privada (el **efecto expulsión**, o *crowding out*).

Pendientes: la IS, de pendiente $-(1-c)/v$, es más vertical cuanto menor es v, y la LM, de pendiente k/h, más plana cuanto mayor es h. La política fiscal gana eficacia cuanto más vertical es la IS y más plana la LM, y la monetaria en el caso contrario. Para la SNC, con v bajo y h alto, la política fiscal es la más eficaz. En los extremos, con $v = 0$ o en la trampa de liquidez, $h \to \infty$, el multiplicador fiscal alcanza $1/(1-c)$ y la política monetaria no afecta a la renta.

1.2.2. Demanda de consumo

Renta absoluta (Keynes, 1936): la SNC construye una función de consumo lineal y estable en la renta disponible del periodo, $C = C_0 + c\,(Y - T_0)$, con $c \in (0,1)$. El consumo no depende del tipo de interés y el ahorro es un residuo. En una economía cerrada con impuestos de suma fija y tipo de interés dado, los multiplicadores del gasto público y de los impuestos son:

$$\frac{\partial Y}{\partial G_0} = \frac{1}{1-c} > 1, \qquad \frac{\partial Y}{\partial T_0} = -\frac{c}{1-c} < 0$$

- La política fiscal es muy efectiva. El gasto público lo es más que una rebaja impositiva equivalente, que se reparte entre consumo y ahorro, y la redistribución estimula el consumo.
- La política monetaria apenas afecta a la renta, por la baja sensibilidad de la inversión al tipo, el *eslabón débil*.

Renta relativa (Duesenberry, 1949): la utilidad depende del consumo del entorno, efecto demostración, y del consumo pasado, efecto trinquete. Como el trinquete hace el consumo rígido a la baja, en una recesión cae menos que la renta, la propensión media al consumo, PMeC, sube y se amortigua la caída de la demanda.

Ciclo vital (Modigliani y Brumberg, 1954; Ando y Modigliani, 1963): frente a una renta en forma de joroba, los hogares suavizan el consumo con el ahorro. La PMeC oscila a corto plazo y es estable a largo, como observó Kuznets, y la política fiscal transitoria pierde eficacia. Con la riqueza constante, la PMeC baja en las fases alcistas y sube en las bajistas, porque el aumento de renta se reparte entre varios periodos, y una variación transitoria de la renta disponible apenas altera el consumo corriente.

1.2.3. Demanda de inversión

Eficiencia marginal del capital (Keynes, 1936): se invierte mientras la EMK, la tasa que iguala el valor actual de los rendimientos esperados con el coste de reposición, supere al tipo de interés: $I = I_0 - v\,r$. Al depender de los *animal spirits*, la inversión es volátil y poco sensible al tipo, y la IS es casi vertical.

Teoría de la q (Tobin, 1969): se invierte cuando $q > 1$, con q el valor de mercado del capital instalado respecto de su coste de reposición,

$I_t = f(q_t)$, lo que liga la inversión real a las valoraciones de los activos. La q relevante es la marginal, el precio sombra relativo del capital, que en la práctica se aproxima por la q media, y la teoría formaliza la importancia de la estabilidad de los mercados financieros, ya apuntada por Keynes.

1.2.4. Demanda de dinero

La SNC microfundamenta cada uno de los tres motivos de Keynes.

- **Transacción** (Baumol, 1952; Tobin, 1956): el dinero es un inventario, y su gestión enfrenta el coste de transacción y el coste de oportunidad. La regla de la raíz cuadrada da $M^d/P = k(b^+, i^-) \cdot Y^{1/2}$, menos que proporcional a la renta por economías de escala en la tesorería. Por ello, un aumento de la cantidad de dinero desplaza más la LM, porque a tipo de interés dado la renta que equilibra el mercado de dinero responde de forma amplificada, $Y \propto (M^s/P)^2$, aunque el efecto final sobre la renta depende de la pendiente de la IS.
- **Precaución** (Whalen, 1966): el saldo equilibra el coste de oportunidad con el coste esperado de las rupturas de liquidez y sigue una regla de la raíz cúbica, $M^d = (2\sigma^2 b/i)^{1/3}$, con σ^2 la varianza de los pagos netos y b el coste de cada ruptura. Como la varianza crece con la renta, $M^d = f(b^+, i^-, Y^+)$.
- **Especulación** (Tobin, 1958): los agentes diversifican entre dinero y bonos, y la demanda crece con la riqueza y el riesgo de los bonos y es ambigua en su rentabilidad. La política monetaria actúa sobre la composición de las carteras, con menos efecto sobre la renta cuanto más sensible al tipo sea la demanda de dinero. En símbolos, $M^d = f(W^+, \sigma_B^+, r_B^\pm)$, con W la riqueza, σ_B el riesgo de los bonos y r_B su rentabilidad.

1.2.5. Conclusiones

Aspectos positivos: dota a la demanda de microfundamentos, resuelve la paradoja de Kuznets y ofrece un marco coherente de gestión de la demanda.

Limitaciones: carece de una teoría de las expectativas, descansa en un eslabón débil frágil, segmenta de forma *ad hoc* la demanda de dinero y separa las decisiones de consumo-ahorro e inversión. Los monetaristas rebatirán el eslabón débil con una inversión sensible al tipo de interés y una demanda de dinero estable, y Friedman tratará el dinero como un activo más, con una demanda unificada.

1.3. Análisis del lado de la oferta

La determinación de la inflación y su relación con el mercado de trabajo era el punto débil del keynesianismo.

1.3.1. Curva de Phillips (Phillips, 1958; Lipsey, 1960)

Origen: Phillips, en la LSE, comprobó con datos del Reino Unido entre 1861 y 1957 que los salarios nominales crecían más despacio cuanto mayor era el paro, con una relación no lineal y aparentemente estable. De ahí surgen dos líneas de investigación: una empírica, que busca confirmar su estabilidad, y otra teórica, que busca explicarla a partir del funcionamiento del mercado de trabajo. Phillips añadió que la relación es muy convexa, porque los salarios nominales apenas bajan aunque el paro sea alto, y que la inflación salarial se acelera si el paro cae deprisa.

Idea: inflación y paro varían en sentido opuesto, de modo que la política económica puede elegir entre más inflación y menos paro, o a la inversa.

Razonamiento: Lipsey (1960) le dio fundamento: cuanto más paro, menor la demanda excedente de trabajo y más despacio suben los salarios. Con productividad constante, los precios siguen a los salarios (Samuelson y Solow, 1960). La **curva de Phillips** liga de forma inversa y convexa inflación y desempleo:

$$\pi_t = f(u_t), \qquad f' < 0,\ f'' > 0$$

Casos polares: vertical, y a la tasa natural puede corresponder cualquier inflación, fenómeno de demanda, monetario (suele asociarse a la visión clásica); horizontal, y a cualquier paro corresponde una única inflación, fijada por los costes (suele asociarse a la keynesiana). Ambos son simplificaciones, pues ninguna escuela sostuvo una forma exactamente polar.

Menú de política: Samuelson y Solow (1960) la conciben como un *menú*: mantener el paro bajo a cambio de una inflación continuada con políticas monetarias expansivas sostenidas, un "ajuste fino" del mercado laboral, aunque advierten de que podía no mantener su forma a largo plazo. Las políticas de demanda mueven la economía a lo largo de la curva, sin desplazarla, porque sus variables no aparecen en ella. Lo que desplazan es la DA, con un movimiento a lo largo de la OA en el plano renta-precios.

Conclusiones: condensaba en una sola relación el nexo entre trabajo e inflación, sin microfundamentar las rigideces laborales, y fue durante los sesenta el instrumento central de la síntesis en el análisis de la oferta. Pero no resultó estable, porque con la estanflación subieron a la vez inflación y paro, y carecía de microfundamentos en sentido estricto (agentes que optimizan) y exigía ilusión monetaria permanente.

1.3.2. Ley de Okun (Okun, 1962)

Idea: con datos de Estados Unidos entre 1947 y 1960, un punto más de paro se asociaba a en torno a un 3 % menos de producción, mucho más de lo que predice la productividad marginal decreciente, porque se suman horas, participación y productividad. En primeras diferencias:

$$u_t - u_{t-1} = -\frac{1}{\phi}\left(g_{Y,t} - \bar{g}_Y\right)$$

- La producción debe crecer a su tasa tendencial, $\bar{g}_Y$, solo para mantener el paro constante, y ϕ ronda 3. Un crecimiento superior al tendencial reduce el paro.
- En brechas, $y_t - \bar{y}_t = -\phi\,(u_t - \bar{u})$, equivalente si $\bar{u}$ es constante. Lleva la curva de Phillips lineal al plano producción-inflación, $\pi_t = \kappa\,(y_t - \bar{y}_t)$ con $\kappa = \kappa_u/\phi$, y cierra el IS-LM por el lado de la oferta.

Conclusiones: guio en EE. UU. la estimación oficial de la producción potencial y es una de las regularidades empíricas más sólidas. Pero Okun sabía que sus ecuaciones simplificaban interacciones complejas, con riesgo de sobreestimación, y su coeficiente varía entre países y se rompe hacia 1973-1974, aunque la ley resistió mejor que la curva de Phillips.

1.3.3. Conclusiones

Aspectos positivos: las dos relaciones cubrieron el punto débil del keynesianismo, la determinación de la inflación y su relación con el mercado de trabajo, y completaron el IS-LM con un análisis de la oferta.

Limitaciones: a finales de los sesenta ese análisis entra en crisis por razones teóricas y empíricas. Desde 1970 aparece la **estanflación**, alto paro con alta inflación, una inflación de oferta que la curva original no podía recoger. Bruno y Sachs (1985) la atribuyen al encarecimiento del petróleo (1973-1974 y 1979) y de las materias primas, a la caída del crecimiento de la productividad total de los factores, a la rigidez del salario real, sobre todo en Europa, y a la contracción de la demanda con la que se combatió la inflación.

En suma, la SNC dio a la macroeconomía un instrumental común y reconcilió a Keynes con el equilibrio neoclásico, pero postuló las rigideces sin microfundamentarlas y sin una teoría de las expectativas, y la estanflación la dejó sin respuesta. Ese doble fallo, teórico y empírico, propició la contrarrevolución monetarista y, más tarde, la nueva macroeconomía clásica.

2. Monetaristas

La crítica monetarista tuvo dos etapas: de mediados de los cincuenta a finales de los sesenta, por el lado de la demanda; desde finales de los sesenta, por el de la oferta. Su prolongación en la economía abierta es el enfoque monetario de la balanza de pagos (Johnson, 1958, 1972; desarrollo en B12 §2.1).

2.1. Principales rasgos

Definición: el **monetarismo**, término que acuñó Brunner en 1968, es la doctrina del círculo de Chicago en torno a Friedman, cuyo núcleo es la teoría cuantitativa del dinero.

Metodología: usa el instrumental IS-LM, pero Friedman sostiene que el realismo de los supuestos es irrelevante, porque una teoría se juzga por sus predicciones (**tesis de irrelevancia de los supuestos**, 1953).

Supuestos: conserva postulados neoclásicos, como la racionalidad y la flexibilidad de precios, pero el dinero es neutral a largo plazo y no a corto (*money matters*). Los ciclos se deben sobre todo a las oscilaciones de la oferta monetaria, y la causalidad va del dinero a la demanda.

Agentes: para suplir la falta de una teoría de las expectativas en la SNC, populariza la **hipótesis de expectativas adaptativas**, HEA, anticipada por Fisher (1911) y formalizada por Cagan (1956) y Nerlove (1958). Suele usarse la versión de retardos distribuidos con pesos geométricamente decrecientes, de modo que cada periodo la expectativa se revisa en una fracción $\lambda \in (0, 1]$ del último error de previsión:

$$\pi^e_{t+1} = \pi^e_t + \lambda(\pi_t - \pi^e_t), \qquad \lambda \in (0, 1]$$

- En forma de retardos distribuidos, con X_{t-j} los valores pasados observados, la HEA sostiene la crítica a la curva de Phillips y la estimación de la renta permanente:

$$X^e_{t+1} = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j X_{t-j}$$

Ajuste: Friedman rompe la separación keynesiana entre consumo-ahorro e inversión y abre un canal directo de transmisión a través de la riqueza. Frente a las dos variables de ajuste keynesianas, las existencias en el mercado de bienes y el tipo de interés en el monetario, ese canal opera vía consumo y demanda de dinero.

Política económica: sus recomendaciones chocan punto por punto con las de la SNC. Como el consumo, guiado por la renta permanente, y la demanda de dinero son estables, la fiscal es poco eficaz y expulsa gasto privado. Ningún instrumento iguala a la monetaria, que debe aplicarse con reglas, como la **regla del k %** (Friedman, 1960), y no a discreción.

2.2. La crítica por el lado de la demanda

Friedman reconstruye las tres funciones de la síntesis, consumo, inversión y demanda de dinero, y deduce sus consecuencias de política.

2.2.1. Renta permanente (Friedman, 1957)

Idea: el consumo no depende de la renta corriente, sino de la **renta permanente**, el rendimiento de la riqueza total, humana y no humana:

$$Y_t^P = rW_t, \qquad W_t = \frac{B_t}{r} + M_t + K_t + A_t + W_H$$

- La renta corriente suma la permanente y una **renta transitoria** aleatoria, sin correlación entre sí. Como la permanente no se observa, los consumidores la estiman con expectativas adaptativas. En símbolos, $Y_t = Y_t^P + Y_t^T$. En la riqueza, B_t/r son los bonos perpetuos, M_t los saldos reales, K_t los activos físicos, A_t las acciones y W_H la riqueza humana, valor descontado de las rentas laborales esperadas, $W_H = \sum_{s=t}^{\infty} y_s^e/(1+r)^{s-t}$.

Consumo: el permanente es proporcional a la renta permanente, $C_t^P = k_P Y_t^P$, con $k_P = f(r^-, (W_H/W)^-, z^{\pm})$: un tipo más alto encarece el consumo presente, y la riqueza humana es más arriesgada y peor aval ante un prestamista; z recoge gustos, demografía e incertidumbre, y k_P no depende de cuánta renta permanente se tenga. El transitorio no guarda correlación con la renta transitoria, de modo que la **propensión a consumir la renta transitoria es nula**: un *shock* positivo se ahorra y uno negativo se absorbe con desahorro. Formalmente, $\rho(C_t^T, Y_t^T) = 0$. La proporcionalidad entre consumo y renta permanentes exige una utilidad intertemporal homotética.

Horizontes: la función de consumo agregada depende del horizonte:

$$\text{LP}: \; C_t = k_P Y_t^P$$
$$\text{CP}: \; C_t = C_0 + c\,Y_t, \qquad C_0 > 0, \; c < k_P$$

- A largo plazo, los *shocks* transitorios se cancelan en el agregado y la PMeC es constante, como observó Kuznets. En el plano renta-consumo, la función parte del origen con pendiente k_P.
- A corto plazo, como la renta permanente se estima como un retardo distribuido de la corriente, la función se aproxima con una keynesiana, más plana y con ordenada positiva. Estimada por grupos de renta, el de renta alta tiene una renta transitoria media positiva y el de renta baja, negativa, de modo que este consume por encima de lo que indica su renta observada y financia el exceso con desahorro. Así, la PMeC de los hogares de renta baja es mayor a corto plazo y la PMeC agregada es anticíclica.

Extensiones:

- *Política económica*: un aumento monetario no anticipado y transitorio eleva el consumo por efectos riqueza. Si es permanente, solo genera inflación. Un cambio fiscal transitorio se ahorra íntegramente. Uno permanente sí eleva el consumo, y que lo compensen los impuestos futuros no es un resultado de Friedman, sino de Barro (1974), la equivalencia ricardiana. En la política monetaria, el efecto riqueza es directo, vía saldos monetarios, e indirecto, vía tipo de interés, y este último es siempre positivo ante una bajada del tipo, porque k_P decrece con r.

Conclusiones: encaja con el ciclo vital y con Kuznets, pero supone mercados de capitales perfectos, combina una mirada *forward looking* con expectativas adaptativas, que miran atrás, y es difícil de contrastar porque la renta y el consumo permanentes no se observan.

2.2.2. Demanda de inversión

Friedman acepta en lo esencial la EMK, y el debate gira en torno a v. Para la síntesis, la inversión sigue a los *animal spirits*, apenas responde al tipo y la IS es casi vertical. Para los monetaristas, los agentes reparten su riqueza entre muchos activos según su rentabilidad, la inversión reacciona con fuerza a r y desaparece el eslabón débil. Como además el consumo depende de la riqueza, cuyo valor cambia con el tipo, la IS es **elástica** al tipo de interés.

2.2.3. Demanda de dinero (Friedman, 1956)

Idea: Friedman reformula la teoría cuantitativa como una teoría de la demanda de dinero, que trata como **un bien de capital más**, con muchos sustitutivos imperfectos: bonos, acciones, bienes duraderos e inmuebles. Keynes, en cambio, le atribuía pocos sustitutos, aunque muy próximos. Friedman trata a la vez el consumo-ahorro y la cartera, sin segmentar la demanda en motivos. El dinero es factor de producción para las empresas y fuente de poder adquisitivo para los consumidores. Con un enfoque de cartera, como Tobin, la demanda real depende de la renta permanente, del peso de la riqueza humana, W_H/W, de las rentabilidades nominales de bonos y acciones, r_B y r_E, y de la inflación esperada. Agrupando las rentabilidades de los activos, resulta una ecuación de Cambridge revisada:

$$M^d/P = f((Y^P)^+, (W_H/W)^+, r_B^-, r_E^-, \pi^{e-})$$
$$\rightarrow \; M^d/P = k(r^-, \pi^{e-}) \cdot Y^P$$

- Con muchos sustitutivos y dependiente de la renta permanente, la demanda es **estable** y poco sensible al tipo, $h \approx 0$.
- La inflación esperada, como rentabilidad, liga la demanda de dinero con la de activos reales y financieros, y es el centro de la transmisión.
- La especificación puede incluir la rentabilidad nominal del dinero, r_m, como los intereses de los depósitos, con $\partial M^d/\partial r_m > 0$.

Conclusiones: recupera la estabilidad de la demanda de dinero con base en la elección de cartera, pero postula sus argumentos sin un programa de optimización explícito, y esa estabilidad se quebró con la innovación financiera de los ochenta.

2.2.4. Implicaciones para la política de demanda

Teoría cuantitativa revisada: parte de la ecuación de cambio, $MV = PY$. La velocidad, $V = Y/(kY^P)$, es una función estable de pocas variables y, por tanto, predecible. En tasas de crecimiento:

$$g_M + g_V = g_P + g_Y \; \rightarrow \; \Delta g_P = \Delta g_M$$

- A corto plazo el dinero no es neutral y afecta a renta y empleo. A largo plazo, con $g_V \approx 0$ y la renta en su tasa tendencial, cualquier aumento del crecimiento monetario se traslada íntegro a la inflación, de modo que el dinero es neutral (superneutralidad del dinero respecto del *output*, que fijan factores reales). Friedman duda de que la oferta monetaria sea exógena, por la amplitud del agregado, pero lo acepta como simplificación, con el equilibrio instantáneo $M^s = M^d$.

Mecanismo de transmisión: surge al incluir los saldos reales en el consumo y en la demanda de dinero.

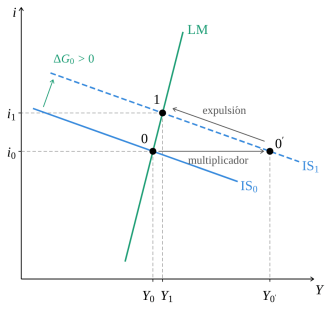
- *Efecto riqueza directo*: los cambios en el dinero y la deuda pública en manos del público desplazan la IS. Formalmente, $\partial C/\partial W \cdot \partial W/\partial M^s$ o $\partial C/\partial W \cdot \partial W/\partial B$.
- *Efecto riqueza indirecto*: vía tipo de interés, que altera el valor de los bonos y hace la IS más elástica, y vía precios, el efecto Pigou.
- *Reajuste de cartera*: dinero y bonos no son sustitutivos perfectos, y emitir deuda puede desplazar la LM a la izquierda.
- *Ley de Walras*: en el IS-LM, $M^s - M^d + B^s - B^d = 0$, y las carteras solo se relacionan de forma indirecta con el mercado de bienes. Friedman añade el sector real, $M^s - M^d + B^s - B^d + (Y^s - Y^d) = 0$, de modo que un exceso en el mercado monetario no exige uno compensatorio en el de bonos, que puede producirse en el sector real.

Política monetaria: más oferta monetaria desplaza la LM a la derecha, con más renta y menos tipo. Supera con mucho a la fiscal en eficacia.

Política fiscal: con una IS casi horizontal y una LM casi vertical, el multiplicador (0-0) es menor que en la SNC y el **efecto expulsión** (0 -1) es elevado. Si $h = 0$, la expulsión es total, y los multiplicadores del IS-LM dan el caso opuesto a la síntesis:

$$\lim_{h\to 0} \frac{\partial Y^*}{\partial G_0} = 0, \qquad \lim_{h\to 0} \frac{\partial Y^*}{\partial (M^s/P)} = \frac{1}{k}$$

Con $h = 0$ se recupera el supuesto clásico, en que el dinero solo se demanda para transacciones, en una proporción fija de la renta. La sensibilidad de la demanda de dinero a la renta, k, es baja para los keynesianos, porque domina el tipo de interés, y alta para los monetaristas, porque domina la renta permanente.



Política fiscal expansiva con IS muy elástica y LM muy rígida: el efecto multiplicador (0-0) se pierde casi entero por la expulsión (0 -1).

Extensiones:

- *Gran Depresión*: para Friedman y Schwartz (*A Monetary History of the United States, 1867-1960*, 1963), la caída de más de un tercio de la cantidad de dinero entre 1929 y 1933, por las quiebras bancarias que la Reserva Federal no frenó, fue su origen principal. El proteccionismo agravó la crisis, pero su tesis, estrictamente monetaria, lo deja fuera.
- *Diseño*: el objetivo es la cantidad de dinero, no el tipo de interés, controlada con operaciones de mercado abierto.
- *Desinflación*: gradual y preanunciada, aunque ante inflaciones muy altas Friedman aconsejó un "tratamiento de choque", como en Chile en 1975. La gradualidad da tiempo a los agentes para revisar sus expectativas y limita el coste en producción y empleo, aunque la transición implica un periodo de menor actividad.

– *Retardos*: largos y variables, sobre todo en la política fiscal, de modo que la estabilización activa puede ser procíclica. Friedman (1960) estimó que los máximos del crecimiento monetario se adelantaban en promedio unos 16 meses a los del ciclo y los mínimos unos 12.

Política reglada: la política monetaria, por poderosa, debe quedar a salvo de presiones políticas, tener la inflación como meta y renunciar a la discrecionalidad. A corto plazo debe hacerlo porque nadie sabe qué parte de una sacudida monetaria irá a precios y qué parte a renta; a largo, porque con una tasa natural de paro solo puede estabilizar los precios.

Reglas: la del k % (Friedman, 1960), un crecimiento constante de la masa monetaria ajustado a las necesidades de liquidez, y la de inflación óptima (Friedman, 1969): el óptimo de Pareto exige igualar el coste privado de mantener dinero, el tipo nominal, a su coste social, casi nulo, con una deflación igual al tipo real:

$$\pi = i - r \;\rightarrow\; \pi = -r$$

2.2.5. Ecuación perdida y evidencia empírica

Ecuación perdida: la teoría cuantitativa determina la renta nominal, pero no su reparto entre precios y renta real. Friedman la cierra con la tasa natural y la HEA. A corto plazo el reparto depende del error de previsión y, a largo, todo recae sobre los precios.

Evidencia: el análisis histórico de Friedman y Schwartz (1963), según el cual los impulsos monetarios nacen en el propio sector monetario y van seguidos de ciclos reales, el contraste de Friedman y Meiselman (1963) y la ecuación de St. Louis (Andersen y Jordan, 1968), con un efecto del dinero sobre la renta nominal mayor, más rápido y más predecible que el de la política fiscal. Los críticos objetaron una posible causalidad inversa.

Conclusiones: devolvió al dinero un papel central en la inflación y el ciclo, y ayudó a que la estabilidad de precios fuera el objetivo de la política monetaria. Pero desde los ochenta la innovación financiera quebró la estabilidad de la velocidad, y los bancos centrales pasaron a los tipos de interés y a los objetivos directos de inflación.

2.3. La crítica por el lado de la oferta

Idea: en *The Role of Monetary Policy* (1968), Friedman sostiene que existe una **tasa natural de paro**, la única compatible con una inflación estable, de modo que la curva de Phillips es vertical a largo plazo. La inflación es "siempre y en todas partes un fenómeno monetario" (1963). Una inflación persistente tiene, por tanto, un origen de demanda, monetario, y nunca de costes, pues solo surge cuando el dinero crece más deprisa que la producción. El nombre evoca el tipo de interés natural de Wicksell.

2.3.1. Inestabilidad de la curva de Phillips

La curva original es inestable porque deja fuera las expectativas: supone ilusión monetaria y expectativas estáticas. A los agentes les importa lo que su salario permite comprar, y al pactar una cifra nominal la ajustan a los precios que prevén. Con la HEA, la curva se desplaza al cambiar las expectativas. La ilusión monetaria es plausible, pero no para siempre. El parámetro λ mide lo rápido que aprenden los agentes, y las variables nominales no pueden alterar de forma permanente el desempleo.

2.3.2. Curva ampliada (Friedman, 1968; Phelps, 1967)

Definición: la **tasa natural de paro** es la de equilibrio a largo plazo, a la que llega la economía por el comportamiento optimizador de los agentes. Incluye el paro friccional y el estructural y la fijan factores reales. Su análoga en el modelo de salarios y precios, la NAIRU, crece con el margen de las empresas y con los factores institucionales que presionan al alza los salarios.

Curva ampliada por las expectativas: la inflación depende de la esperada y del desempleo cíclico, y la disyuntiva solo existe mientras las expectativas no se cumplen:

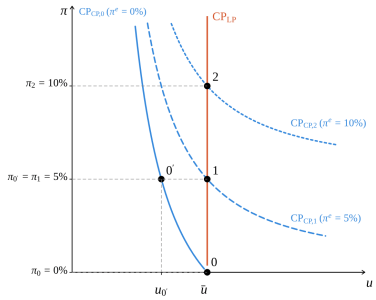
$$\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u}), \quad \kappa_u > 0$$

– A corto plazo hay una curva por cada inflación esperada y, a largo, las expectativas recogen la inflación efectiva y la curva es vertical en $\bar{u}$.
– Despejando el paro, $u_t = \bar{u} - \frac{1}{\kappa_u}(\pi_t - \pi_t^e)$, que solo se aparta de la tasa natural mientras las expectativas no se cumplen.
– Con la ley de Okun en brechas, la oferta agregada ampliada, $\pi_t = \pi_t^e + \kappa\,(y_t - \bar{y}_t)$, tiene pendiente positiva a corto plazo y es vertical en $\bar{y}_t$ a largo. Con la teoría cuantitativa y la HEA, reparte el crecimiento de la renta nominal entre inflación y crecimiento real, que es la ecuación perdida.

Desarrollo gráfico: en el plano desempleo-inflación, cada curva a corto plazo, $\mathrm{CP}_{\mathrm{CP},0}$, $\mathrm{CP}_{\mathrm{CP},1}$ y $\mathrm{CP}_{\mathrm{CP},2}$, corresponde a una inflación esperada del 0 %, el 5 % y el 10 % y tiene pendiente negativa. La $\mathrm{CP}_{\mathrm{LP}}$ es vertical en $\bar{u}$, donde la inflación observada iguala a la esperada, y la economía parte del punto 0, con expectativas nulas y en la tasa natural.

Dinámica: de ahí procede la **teoría aceleracionista** (Friedman, 1968).

1. El Gobierno estimula la demanda y, con los agentes aún sin revisar sus expectativas, el paro cae por debajo de $\bar{u}$, en el punto 0 .
2. Un periodo después revisan sus previsiones, perciben que su salario real ha empeorado, la curva sube a $\mathrm{CP}_{\mathrm{CP},1}$ y el aumento del paro lleva la economía al punto 1. Quienes entraron en el mercado de trabajo creyendo que su salario había aumentado salen de él.
3. Si el Gobierno insiste, el proceso se repite y desata una espiral inflacionista, infructuosa cuando los agentes anticipan también la aceleración. En el extremo, la espiral puede privar al dinero de su función de medio de cambio.



Curva de Phillips ampliada por las expectativas: disyuntiva a corto plazo y curva vertical a largo plazo en la tasa natural de paro.

Extensiones: mantener el paro bajo la tasa natural exige acelerar sin fin el crecimiento monetario, para que los agentes sigan equivocándose. A largo plazo, solo las políticas de oferta reducen el desempleo, como la reducción de las asimetrías informativas, los incentivos a la innovación o la formación de capital humano.

Conclusiones: explica, antes de la estanflación, por qué la disyuntiva es transitoria y separa los determinantes reales del paro de los monetarios de la inflación. Pero descansa en expectativas adaptativas, que permiten engañar sistemáticamente a los agentes, y la tasa natural no es observable.

En suma, el monetarismo devolvió al centro del análisis la oferta agregada y el mercado monetario, y dejó tres legados: la tasa natural como guía para medir el paro cíclico, las expectativas como elemento central de todo modelo y los retardos de la política económica. Sus recetas fueron costosas. La Reserva Federal (1979-1982) logró la desinflación con tipos muy volátiles y un paro del 10,8 % a finales de 1982. Poole (1970) mostró que, si predominan los *shocks* monetarios, fijar el tipo de interés estabiliza mejor la economía real que fijar el agregado, y la regla del k % cedió ante la de Taylor (1993). En lo teórico, la renta permanente sufre exceso de sensibilidad (Flavin, 1981), Tobin (1970) advirtió de que la correlación entre dinero y renta nominal podía reflejar una causalidad inversa, con un dinero endógeno que se adapta a la renta, y Lucas rechazó que la HEA permitiera engañar una y otra vez a los agentes. Esta última objeción, la de más peso, surgió dentro del propio pensamiento neoclásico. Aun así, el núcleo de la renta permanente, compartido con el ciclo vital, sigue siendo la base de la teoría moderna del consumo.

3. Otros desarrollos posteriores

La controversia dio paso a tres desarrollos posteriores, que se exponen en sus temas propios.

3.1. La nueva macroeconomía clásica

La crítica de Lucas a las expectativas adaptativas enlaza con la **nueva macroeconomía clásica**, NMC, que sustituye la HEA por la hipótesis de expectativas racionales y explica el ciclo con mercados que se vacían continuamente. Con expectativas racionales, una política monetaria anticipada no tiene efectos reales ni siquiera a corto plazo.

3.2. Nueva economía keynesiana de 1.ª generación

La **nueva economía keynesiana de primera generación**, NEK1G, responde a la NMC microfundamentando las rigideces reales y nominales que la síntesis neoclásica postulaba sin justificar, como en el modelo de la NAIRU de Layard y Nickell.

3.3. Nueva economía keynesiana de 2.ª generación

La **nueva economía keynesiana de segunda generación**, NEK2G, completa la secuencia al combinar las rigideces nominales con el método de equilibrio general dinámico de la NMC, en lo que se conoce como **nueva síntesis neoclásica**.

A6. La nueva macroeconomía clásica. La hipótesis de las expectativas racionales; la crítica de Lucas; el surgimiento de los modelos dinámicos estocásticos de equilibrio general

1. Fundamentos del pensamiento de la NMC

La aportación de la nueva macroeconomía clásica, NMC, supuso un cambio metodológico que se asienta en dos ideas, la crítica de Lucas y la microfundamentación de los modelos macroeconómicos, y en un supuesto sobre los agentes, las expectativas racionales.

1.1. Principales rasgos

Metodología: en gran parte a raíz de la crítica de Lucas, la NMC exige **microfundamentar** los modelos, es decir, derivar las relaciones agregadas del comportamiento optimizador de los agentes.

Supuestos: un mundo neoclásico, con competencia perfecta y plena flexibilidad de precios, salarios y tipo de interés, pero con información imperfecta, por la que las perturbaciones nominales no anticipadas pueden tener efectos reales transitorios sin rigideces nominales.

Agentes: es el ámbito más novedoso, con dos ideas.

- *Racionalidad dinámica*: los agentes optimizan en el tiempo, con sustitución intertemporal del trabajo y el ocio entre periodos (Lucas y Rapping, 1969), y tienen una conducta previsora, *forward-looking*.
- *Expectativas racionales* (Muth, 1961): los agentes usan la información disponible de la mejor manera posible y no cometen errores sistemáticos.

Ajuste y política: estos elementos llevan más allá el pensamiento clásico y garantizan el **vaciado continuo** de los mercados: el equilibrio se alcanza en cada instante. De ahí se deduce la defensa del no intervencionismo.

1.2. Expectativas racionales (Muth, 1961)

Hipótesis de expectativas racionales, HER: los agentes no ajustan sus expectativas gradualmente, sino que usan la información disponible de la mejor manera posible. La expectativa es la esperanza condicionada a esa información, Ω_t, y el error de predicción es de media cero y ortogonal a ella:

$$X^e_{t+i} \equiv \mathbb{E}_t[X_{t+i} \mid \Omega_t], \qquad X_{t+i} - X^e_{t+i} = \varepsilon_{t+i}, \quad \mathbb{E}_t[\varepsilon_{t+i}] = 0$$

La HER admite errores, pero no sistemáticos, y no debe confundirse con las expectativas adaptativas, que corrigen la previsión con los errores pasados.

Características: son cuatro.

- *Información*: no exige información completa ni conocer la estructura de la economía. Si es incompleta, los agentes usan de forma óptima la disponible. Si es completa, actúan como si conocieran el modelo.
- *Posibilidad de error*: el error puede ser grande con poca información o perturbaciones imprevisibles y solo desaparece con previsión perfecta. Lo que se exige es que no afecte a lo deducible con la información disponible.
- *Conocimiento de reglas*: no hace falta conocer las reglas de política futura, pero las creencias sobre ellas no pueden diferir de las reales en lo deducible.
- *Expectativas iteradas*: la predicción de hoy de la predicción futura coincide con la predicción actual, $\mathbb{E}_t[\mathbb{E}_{t+1}(X_{t+2})] = \mathbb{E}_t(X_{t+2})$. Si se esperara que la predicción fuera a cambiar, sería porque hay información no incorporada a la actual.

Error: la ortogonalidad exige dos propiedades, esperanza nula y ausencia de autocorrelación de los errores:

$$\mathbb{E}[\varepsilon_t] = 0, \qquad \mathbb{E}[(X_t - X^e_t)(X_{t-1} - X^e_{t-1})] = 0$$

- La homocedasticidad y la normalidad no son requisitos, sino supuestos simplificadores. La normalidad permite, por ejemplo, la extracción de señal lineal del modelo de Lucas.

1.3. La crítica de Lucas (1976)

Origen: en *Econometric Policy Evaluation: A Critique* (1976), Lucas cuestiona la práctica de evaluar políticas simulándolas en grandes modelos macroeconométricos keynesianos, construidos sobre el IS-LM y estimados con datos históricos.

Idea: la estructura de esos modelos está formada por las reglas de decisión óptimas de los agentes, que cambian cuando cambia la propia política. Cualquier cambio de política altera, por tanto, la estructura del modelo.

Razonamiento: hay que distinguir dos tipos de parámetros.

- *Profundos o estructurales*: preferencias, tecnología y restricciones, invariantes ante la política.
- *De las reglas de decisión*: los coeficientes en forma reducida, como los de una función de consumo o una curva de Phillips, que incorporan las expectativas sobre la política y varían con el régimen.

Tratar los coeficientes en forma reducida como estructurales impide predecir los efectos de un cambio de política, porque la relación estimada bajo un régimen deja de valer justo cuando se adopta otro. Lucas lo ilustra con el consumo, la inversión y, sobre todo, la curva de Phillips.

Consecuencias metodológicas: son tres.

- Se evalúan **reglas** estables y conocidas, no medidas aisladas. De ahí procede el énfasis posterior en la credibilidad.
- Las expectativas deben ser endógenas al régimen, lo que consagra la HER.
- Solo sirven los modelos formulados con parámetros profundos, que desembocan en los modelos de equilibrio general dinámico y estocástico, DSGE.

La inefectividad de la política anticipada de Sargent y Wallace aplica la misma lógica. Si la regla es conocida, su parte sistemática ya está descontada en las expectativas.

Conclusiones: fue un punto de inflexión, a menudo descrito como un cambio de paradigma en el sentido de Kuhn, pero su relevancia empírica ha sido menor de lo temido, pues los contrastes de superexogeneidad (Favero y Hendry, 1992), como en la demanda de dinero, apenas hallan la inestabilidad que predice. Esos contrastes comprueban si los coeficientes de una ecuación se mantienen estables cuando cambia el proceso que siguen las variables de política.

En suma, la NMC conserva el mundo neoclásico y renueva la modelización de los agentes, y la crítica de Lucas convierte esos supuestos en una exigencia de método.

2. Aplicaciones: estudio del ciclo walrasiano

Para Lucas, un **ciclo** es un movimiento autocorrelacionado de las principales variables alrededor de una tendencia, descrito por una ecuación con perturbaciones estocásticas. La NMC ofrece dos explicaciones, ambas como respuesta óptima de agentes racionales en mercados que se vacían: el ciclo nominal, por sorpresas monetarias, y el ciclo real, por perturbaciones tecnológicas.

2.1. OA-DA con información imperfecta

Origen: Lucas (1972; 1973; 1975) elabora el modelo de las islas sobre dos trabajos previos: la *parábola de las islas* de Phelps (1969), base del **problema de extracción de señales**, y el modelo de Lucas y Rapping (1969), que destaca la elasticidad de sustitución intertemporal del trabajo.

Idea: la diferencia con el modelo clásico está en la oferta. La de corto y la de largo plazo difieren por las percepciones transitorias erróneas sobre los precios, que crean a corto plazo una relación positiva entre precios y producción. La producción responde a la diferencia entre los precios observados y los esperados.

2.1.1. Supuestos

Entorno: mundo neoclásico con información imperfecta. La economía son muchas "islas". Cada productor observa el precio de su mercado, p_{it}, pero no el nivel general, p_t, y no sabe si una subida responde a un *shock* específico, ante el que conviene producir más, o general, ante el que no. Las variables van en logaritmos.

Agentes: cada productor vende un único bien en su mercado y compra en otros, sobre cuyos precios forma una expectativa racional, $p^e_t \equiv \mathbb{E}_{t-1}[p_t]$, con la información previa a observar su precio. Optimiza intertemporalmente, de modo que un precio relativo alto lleva a producir más, con más demanda de trabajo de las empresas y más oferta de los hogares, por la sustitución intertemporal del ocio. En cuanto a la autoridad monetaria, el dinero tiene un componente previsible o sistemático y otro no previsible, ε_t, que recoge las sorpresas monetarias o el error de control:

$$m_t = \mathbb{E}_{t-1}[m_t \mid \Omega_{t-1}] + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma^2_m)$$

2.1.2. Modelo

Extracción de señales: el productor pondera la información con $\vartheta \in [0,1]$, la fracción de la varianza de los precios debida a *shocks* relativos, o probabilidad de que el *shock* sea específico:

$$\vartheta = \frac{\sigma^2_i}{\sigma^2_i + \sigma^2_p}$$

- Con inflación alta y volátil, $\vartheta \to 0$ y el productor apenas reacciona, mientras que con precios estables $\vartheta \to 1$.

Planteamiento: el modelo combina la oferta de Lucas, derivada de la extracción de señales, y una demanda de teoría cuantitativa:

$$\text{OA: } y_t = \bar{y}_t + \zeta\,(p_t - p^e_t) + \varepsilon^s_t$$
$$\text{DA: } y_t = m_t - p_t + v_t$$

- *Oferta*: la de cada mercado suma un componente tendencial, $\bar{y}_t$, ligado a la tasa natural de paro, y uno cíclico, $y_{it} = \bar{y}_t + \zeta\,(p_{it} - p^e_t)$, con $\zeta \equiv \vartheta\eta > 0$ y η la sensibilidad de la producción a los *shocks* de precios de su mercado. El productor solo responde a la parte del diferencial que atribuye a un *shock* relativo. La OA agrega mercados iguales, en los que el componente específico promedia a cero, y añade los *shocks*

de oferta, ε_t^s. Salvo por estos, la producción solo se desvía de su nivel natural si hay **sorpresa de precios**, y no lo hace con $\vartheta = 0$, como en una hiperinflación, con $\eta = 0$ ni con la inflación anticipada, $p_t = p_t^e$.
– *Demanda*: parte de $m_t + v_t = p_t + y_t$, con v_t un *shock* de velocidad.

Equilibrio: igualando OA y DA con la HER, todo lo anticipable se cancela y la brecha de producción depende solo de los *shocks*:

$$\tilde{y}_t = \frac{\zeta}{1+\zeta}(\varepsilon_t + v_t - \varepsilon_t^s) + \varepsilon_t^s$$

Estática comparativa: es la lectura principal del modelo, con los efectos de cada *shock* sobre la producción:

$$\frac{\partial y_t}{\partial \varepsilon_t} = \frac{\partial y_t}{\partial v_t} = \frac{\zeta}{1+\zeta} > 0, \qquad \frac{\partial y_t}{\partial \varepsilon_t^s} = \frac{1}{1+\zeta} > 0$$

– Un cambio monetario o de demanda anticipado se traslada íntegro a los precios. Solo su parte no anticipada altera la producción, con efecto $\zeta/(1+\zeta)$, hasta que los agentes infieren el nivel de precios.
– Los *shocks* de oferta son reales y mueven la producción directamente, con un efecto amortiguado, $1/(1+\zeta)$.
– Donde la demanda es más volátil, la producción responde menos a las sorpresas (Lucas, 1973): la curva de Phillips a corto plazo es más vertical y no hay disyuntiva explotable de forma sistemática.
– Con la ley de Okun, $\tilde{y}_t = -\phi\,(u_t - \bar{u})$, la OA de Lucas da una curva de Phillips con HER, vertical a largo plazo, cuando $\pi_t = \pi_t^e$:

$$\pi_t = \pi_t^e - \kappa_u\,(u_t - \bar{u}) - \frac{\varepsilon_t^s}{\zeta}, \qquad \kappa_u = \phi/\zeta > 0$$

2.1.3. Conclusiones

Aspectos positivos: deriva la oferta del comportamiento optimizador, en coherencia con la crítica de Lucas, y es la primera explicación microfundamentada de la no neutralidad del dinero, sin rigideces nominales. Su predicción, dinero anticipado neutral y no anticipado real, es nítida y contrastable. Concilia, además, la relación positiva entre precios y producción a corto plazo con la tasa natural y la neutralidad a largo plazo, con una OA_{LP} vertical.

Limitaciones: con estadísticas de precios rápidas y baratas, cuesta creer en una confusión persistente; las sorpresas monetarias explican poca variabilidad del producto y el modelo no tiene **mecanismo de propagación**; exige una elasticidad intertemporal del trabajo contraria a la evidencia; y el vaciado continuo excluye el desempleo involuntario. Los hechos estilizados lo contradicen, porque la mayor parte del ajuste del empleo se produce por flujos entre empleo y desempleo y no por las horas de los ocupados.

Evidencia: Barro (1977; 1978) halló que solo el dinero no anticipado afectaba a paro y producción, pero con efectos de unos dos años, difíciles de atribuir a confusión informativa. Mishkin (1982) encontró que el anticipado también los tenía. La escuela conservó el método y cambió el impulso monetario por uno tecnológico.

2.2. Ciclos por *shocks* reales no anticipados

Origen: Kydland y Prescott (1982) y Long y Plosser (1983) fundan la **teoría del ciclo real**, TCR, replicando la economía de EE. UU. entre 1950 y 1979. Kydland y Prescott propagan los *shocks* con un plazo de construcción del capital (*time to build*) y resuelven numéricamente, y Long y Plosser obtienen el caso analítico.

Idea: los ciclos son la respuesta óptima de los agentes a perturbaciones reales, tecnológicas, estocásticas y autocorrelacionadas, sin fricciones informativas.

2.2.1. Supuestos

Entorno: economía cerrada, información perfecta, tiempo discreto y horizonte infinito, competencia perfecta, mercados completos y plena flexibilidad.

Agentes: muchos agentes idénticos, consumidores, productores y trabajadores a la vez, con vida infinita y expectativas racionales. Su único activo es el capital y su utilidad, separable en el tiempo, depende del consumo y del ocio. No hay sector público.

Tecnología: Cobb-Douglas con rendimientos constantes, progreso neutral en el sentido de Harrod y perturbaciones autocorrelacionadas. El capital se deprecia a una tasa $\delta \in (0, 1]$.

2.2.2. Modelo

Cuatro ecuaciones de comportamiento de hogares y empresas, tres de tecnología y una condición de equilibrio competitivo permiten estudiar la dinámica del *output*.

Hogares: el hogar representativo maximiza la utilidad esperada descontada sujeta a su restricción presupuestaria intertemporal, en la que el capital es el único activo y se remunera a su coste de uso bruto:

$$\text{máx } \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, 1-l_t) \quad \text{s.a.:} \begin{cases} c_t + k_{t+1} - (1-\delta)k_t = w_t l_t + (r_t + \delta)k_t \\ k_0 \text{ dado} \end{cases}, \quad \beta \in (0,1)$$

En las CPO, la condición intratemporal iguala la RMS entre ocio y consumo al salario, como en el caso estático, y la **ecuación de Euler** estocástica iguala la utilidad marginal del consumo presente a la esperada de esa unidad un periodo después. Si no, convendría desplazar consumo entre periodos. La concavidad estricta de la utilidad y de la tecnología, con la condición de transversalidad, las hace necesarias y suficientes:

$$\frac{u'(1-l_t)}{u'(c_t)} = w_t, \qquad u'(c_t) = \beta\,\mathbb{E}_t[u'(c_{t+1})\,(1+r_{t+1})]$$

Empresas: maximizan su beneficio y remuneran cada factor por su productividad marginal, $f'(l_t) = w_t$ y $f'(k_t) = r_t + \delta$, como en el caso estático. Su programa es $\text{máx}_{l_t,k_t}\,\Pi_t = y_t - w_t l_t - (r_t + \delta)k_t$, y la Cobb-Douglas, cóncava, garantiza que las CPO caracterizan el máximo.

Tecnología: la producción es $y_t = k_t^{\alpha}(A_t l_t)^{1-\alpha}$, y la tecnología crece en torno a una tendencia a un ritmo exógeno, g_A, con un componente aleatorio AR(1), $a_t = \rho_A a_{t-1} + \phi_t$. En logaritmos, $\ln A_t = \ln \bar{A} + g_A\,t + a_t$, con ϕ_t un ruido blanco y $\rho_A \in (-1, 1)$.

Equilibrio: con vaciado de los mercados, $c_t + i_t = y_t$, donde $i_t = k_{t+1} - (1-\delta)k_t$ es la inversión bruta. En el caso que se resuelve analíticamente, con $\delta = 1$ y utilidad logarítmica, $u(c_t, 1-l_t) = \ln c_t + b \ln(1-l_t)$, donde b es el peso del ocio, el empleo y la tasa de ahorro son constantes, porque solo dependen de parámetros estructurales:

$$l^* = \frac{1-\alpha}{(1-\alpha) + b(1-\alpha\beta)}, \qquad s = \alpha\beta$$

– El consumo y el ahorro, en cambio, responden a los *shocks*. No es una propiedad general de la TCR, sino de esta versión.

Dinámica: la desviación del *output* respecto de su tendencia sigue un AR(2), y la duración del ciclo crece con la persistencia de la perturbación, ρ_A:

$$\tilde{y}_t = (\alpha + \rho_A)\,\tilde{y}_{t-1} - \alpha\rho_A\,\tilde{y}_{t-2} + (1-\alpha)\,\phi_t$$

Bienestar: en una economía walrasiana, las fluctuaciones son la respuesta **Pareto-óptima** a los *shocks*, y toda intervención para mitigarlas reduce el bienestar. Si los *shocks* son permanentes, $\rho_A = 1$, alteran la propia tendencia (Nelson y Plosser, 1982), y explicarían a la vez el ciclo y el crecimiento.

2.2.3. Conclusiones

Aspectos positivos: con modelos sencillos reproducía buena parte de los hechos estilizados. Los *shocks* tecnológicos explicarían en torno al 70 % de la volatilidad del PIB de posguerra en EE. UU. (Kydland y Prescott, 1991), y las versiones calibradas recogen la suavización del consumo y la volatilidad de la inversión. La TCR fue una revolución metodológica que adoptó después la NEK.

Limitaciones:

– *Impulso*: el progreso técnico apenas cambia a corto plazo, y cuesta explicar una recesión con una regresión tecnológica. Además, la propagación por acumulación de capital es débil.
– *Ahorro y trabajo constantes*: en la versión básica, inversión y consumo serían igual de volátiles, cuando en los datos la inversión lo es unas tres veces más que la producción.
– *Mercado de trabajo*: el salario real es mucho más procíclico que en los datos, y la oferta de trabajo necesaria es un orden de magnitud más elástica que la estimada con microdatos.
– *Neutralidad del dinero*: la evidencia, como la de los modelos VAR, muestra que no siempre se cumple a corto plazo.

2.3. Comparación y legado: los modelos DSGE

Comparación: ambos modelos comparten agentes racionales que optimizan intertemporalmente y mercados que se vacían, de modo que el ciclo es una respuesta óptima y no un fallo. Difieren en el impulso: en Lucas, una sorpresa nominal con efectos reales transitorios; en la TCR, una perturbación tecnológica con dinero neutral incluso a corto plazo. La TCR aporta además persistencia, por la acumulación de capital y la autocorrelación de los *shocks*, aunque su propagación interna es débil.

Legado: el más duradero de la TCR fue su método, del que surgen los **modelos DSGE**: equilibrio general con hogares y empresas que optimizan con expectativas racionales, perturbaciones aleatorias y parámetros profundos, en principio robustos a la crítica de Lucas. Kydland y Prescott no estimaban los parámetros, sino que los **calibraban** con estudios microeconómicos y promedios de largo plazo. El modelo se juzga entonces por su capacidad para reproducir las volatilidades y correlaciones del ciclo.

Evolución: desde los noventa, el programa incorporó las rigideces de la NEK, con competencia monopolística y precios *à la* Calvo, y Christiano, Eichenbaum y Evans (2005) reprodujeron la inercia de la inflación y la persistencia del producto tras una perturbación monetaria, con rigideces nominales de precios y salarios y rigideces reales: hábitos de consumo, costes de ajuste de la inversión y utilización variable del capital. Smets

y Wouters los estimaron con técnicas bayesianas para la zona del euro (2003) y EE. UU. (2007), y el de EE. UU. obtuvo previsiones fuera de la muestra comparables a las de los VAR bayesianos.

Uso y límites: los bancos centrales los usan desde mediados de los dos mil, como el FRB/EDO de la Reserva Federal, el NAWM del BCE y el COMPASS del Banco de Inglaterra. La Gran Recesión expuso sus límites: sin sector financiero, con agente representativo, linealizados y con mal tratamiento del límite inferior de los tipos. La respuesta ha sido ampliarlos con fricciones financieras y agentes heterogéneos, como los modelos HANK (Kaplan et al., 2018). En ellos, la política monetaria actúa sobre el consumo sobre todo a través de efectos indirectos sobre la renta.

En suma, el ciclo es la respuesta óptima a perturbaciones, monetarias o tecnológicas, y el método de la TCR, más duradero que su explicación, dio lugar a los DSGE. Si el ciclo es óptimo y solo las sorpresas tienen efectos reales, la estabilización es inútil o perjudicial.

3. Implicaciones del pensamiento de la NMC

De ese marco se derivan implicaciones de política monetaria y fiscal, las críticas a la escuela y su convergencia hacia la nueva síntesis.

3.1. Implicaciones de política monetaria

3.1.1. Inefectividad (Sargent y Wallace, 1975; 1976)

Idea: las políticas que responden a información pública son ineficaces para estabilizar la producción.

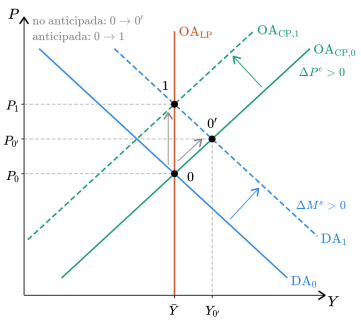
Razonamiento:

- *Política reglada y creíble*: los agentes la anticipan y se traslada íntegra a los precios, $\partial \mathbb{E}_{t-1}[p_t]/\partial \mathbb{E}_{t-1}[m_t] = 1$, sin efectos reales.
- *Política no creíble*: si los agentes esperan una expansión distinta de la anunciada, la parte "anticipada" puede tener efectos reales.

Aun así, la NMC no recomienda sorpresas monetarias, porque restan credibilidad y aumentan la incertidumbre.

Desarrollo gráfico: la OA_{LP} es vertical en $\bar{Y}$, porque sin sorpresa de precios la producción no se desvía de su nivel natural, la OA_{CP} es la oferta de Lucas para unos precios esperados dados, creciente, y la DA es decreciente, porque una subida de los precios reduce los saldos reales. Las tres se cortan en el punto 0.

Estática comparativa: una expansión monetaria desplaza la DA a la derecha. Si no es anticipada, los precios esperados no cambian y la economía pasa de 0 a 0 , con más producción. Si es anticipada, la oferta a corto plazo se desplaza a $\text{OA}_{\text{CP},1}$ y la economía va directamente de 0 a 1, con la producción natural y más precios.



Modelo de la NMC: solo la expansión monetaria no anticipada (0-0) eleva la producción; la anticipada (0-1) solo eleva los precios.

Conclusiones: distingue política anticipada y no anticipada y refuerza las reglas sencillas y predecibles, pero exige precios plenamente flexibles, porque con contratos nominales escalonados la política anticipada tiene efectos reales, y los contrastes no confirman el resultado.

3.1.2. Inconsistencia dinámica

Origen: Kydland y Prescott (*Rules Rather than Discretion: The Inconsistency of Optimal Plans*, 1977), en el marco de la tasa natural, y Barro y Gordon (1983a), en el de la teoría de juegos, se preguntan si mantener una política monetaria sorpresiva es coherente con la HER.

Definición: un plan es **dinámicamente inconsistente** si, sin cambiar las preferencias ni llegar información nueva, deja de ser óptimo cuando llega el momento de ejecutarlo. Una política anunciada lo es si los agentes anticipan que la autoridad romperá su compromiso, y entonces no es creíble.

Idea: la autoridad sabe que a corto plazo hay una relación positiva entre *output* e inflación y tiene incentivos para romper sus compromisos. Pero los agentes, conscientes de ello, lo anticipan, de modo que la política **discrecional** no logra efectos reales ni siquiera a corto plazo y genera una pérdida de bienestar. Así se introduce la importancia de la credibilidad.

Supuestos: información completa, juego único y contratos nominales fijados antes de que la autoridad decida la inflación. La sociedad forma expectativas racionales, $\pi_t^e = \mathbb{E}_{t-1}[\pi_t]$. El Gobierno tiene un objetivo de paro inferior a la tasa natural, $u^T < \bar{u}$, y da un peso b a la inflación: alto en un Gobierno duro, como el mandato único del BCE, y bajo en uno blando, como el dual de la Reserva Federal.

Planteamiento: el Gobierno minimiza la pérdida social sujeto a la curva de Phillips, con $\gamma = 1/\kappa_u$ la inversa del valor absoluto de su pendiente:

$$\min_{\pi_t} L_t = (u_t - u^T)^2 + b(\pi_t - \pi^T)^2 \quad \text{s.a.: } u_t = \bar{u} - \gamma(\pi_t - \pi_t^e)$$

CPO: iguala el coste marginal de desviar la inflación de su objetivo con el beneficio marginal de reducir el paro a lo largo de la curva de Phillips, y la convexidad estricta de L_t en π_t la hace suficiente:

$$-2\gamma(\bar{u} - \gamma(\pi_t - \pi_t^e) - u^T) + 2b(\pi_t - \pi^T) = 0$$

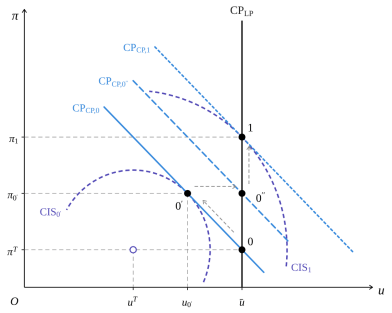
Equilibrio: con agentes racionales, $\pi_t^e = \pi_t$, la inflación supera el objetivo en el **sesgo inflacionista**, y el paro queda en su tasa natural, sin ganancia alguna:

$$\pi_1 = \pi^T + \frac{\gamma}{b}(\bar{u} - u^T) > \pi^T$$

Desarrollo gráfico: en el plano (u, π), las curvas de Phillips a corto plazo, una por cada inflación esperada y con pendiente $-\kappa_u$, son la restricción del Gobierno, y las curvas de indiferencia social son curvas de nivel de la pérdida en torno a (u^T, π^T). El punto 1 es un equilibrio de Nash, porque la curva de Phillips es tangente a la de indiferencia y la inflación coincide con la esperada, y es peor que el inicial, con el mismo paro y más inflación.

Estática comparativa: el sesgo disminuye con la aversión a la inflación, b, en lo que se funda la delegación, y crece con la ambición del objetivo de empleo, $\bar{u} - u^T$, y con una curva de Phillips más plana (mayor γ), que tienta más a la sorpresa. En los tres casos solo cambia la inflación: el paro de equilibrio sigue siendo $\bar{u}$.

Dinámica: desde el compromiso, en 0, a la autoridad le conviene una inflación sorpresa que reduce el salario real y el paro, en 0 . La sociedad revisa sus expectativas y la curva de Phillips se desplaza, hasta el equilibrio de Nash en 1. En el juego único, los agentes lo anticipan y la economía va directamente a 1.



Modelo de Barro y Gordon (1983a): del compromiso (0) al equilibrio discrecional (1), con sesgo inflacionista y sin ganancia de empleo.

Extensiones: las soluciones al sesgo inflacionista son cuatro.

1. *Reglas* (Kydland y Prescott, 1977), como la de Taylor (1993), $i_t = \bar{i}_t + \alpha_\pi \tilde{\pi}_t + \alpha_y \tilde{y}_t$.
2. *Reputación* (Barro y Gordon, 1983b): en un juego repetido, con estrategias de castigo, la autoridad sacrifica el corto plazo para forjarse una reputación de halcón; en economía abierta, con un cambio fijo o una caja de conversión.
3. *Delegación* (Rogoff, 1985) en una autoridad independiente que dé más peso a la estabilidad de precios que la sociedad. En las economías avanzadas la independencia se correlaciona negativamente con la inflación, aunque correlación no es causalidad, porque la independencia y la baja inflación pueden tener una causa común, la oposición a la inflación del sector financiero (Posen, 1993). Además, la respuesta a las perturbaciones puede dejar de ser óptima.
4. *Contratos* (Walsh, 1995) entre Gobierno y banco central, que añaden un coste a su función de pérdida y eliminan el sesgo sin renunciar a estabilizar. Svensson (1997) muestra que equivalen a un objetivo de inflación inferior al socialmente óptimo.

Conclusiones: explica el sesgo inflacionista sin irracionalidad y fundamenta las reglas, la independencia del banco central y los objetivos de inflación, pero supone una autoridad que busca un paro inferior al natural, en un marco estático y de información completa.

3.2. Implicaciones de política fiscal

3.2.1. Equivalencia ricardiana (Barro, 1974)

Idea: en *Are Government Bonds Net Wealth?*, Barro recupera una tesis de Ricardo: los bonos no son riqueza neta y es irrelevante financiar un gasto dado con impuestos o con deuda.

Supuestos: mercados de capitales perfectos, sin restricciones de liquidez; agentes con HER y altruismo intergeneracional perfecto; impuestos de suma fija y una senda de gasto dada.

Razonamiento: al combinar las restricciones intertemporales de hogares y Gobierno, los bonos desaparecen y solo importa el valor actual del gasto. Sin ilusión de deuda, una rebaja de impuestos financiada con deuda solo cambia la composición del ahorro: el desahorro público se compensa con ahorro privado. La política fiscal sigue teniendo efectos reales, pero no dependen de su financiación.

Extensiones: una consolidación puede tener efectos no keynesianos, expansivos: si el recorte del gasto se cree permanente, los agentes anticipan menos impuestos, su riqueza crece y su incertidumbre cae, el **efecto atracción**, y bajan las primas de riesgo de la deuda. En economía abierta, la deuda la absorbe el ahorro adicional de los hogares, sin cambio en la cuenta corriente.

Conclusiones: es un valioso punto de partida teórico, pero mala aproximación a la realidad, porque sus supuestos rara vez se cumplen.

- *Mercados*: hay restricciones de liquidez, y el Estado se endeuda a un tipo menor que el sector privado, de modo que la deuda sí es riqueza.
- *Agentes*: con ilusión fiscal (Buchanan y Wagner, 1977) infravaloran la carga fiscal, y con altruismo imperfecto la carga se traslada a generaciones futuras.
- *Estado*: los impuestos rara vez son de suma fija.

3.2.2. Suavización impositiva (Barro, 1979)

Origen: con impuestos distorsionantes, la financiación deja de ser irrelevante. Barro plantea la política fiscal óptima como un programa intertemporal: dada la senda de gasto, elegir la de impuestos y deuda que minimiza la pérdida de bienestar, cumpliendo su restricción presupuestaria intertemporal y las condiciones de optimalidad del sector privado.

Idea: como el coste marginal de la distorsión crece con el tipo, lo óptimo es un tipo impositivo **constante en lo esperado**, *tax smoothing*, que con incertidumbre sigue un paseo aleatorio, que minimiza el exceso de gravamen. Es preferible un impuesto bajo durante muchos años a uno alto durante uno, y dejar actuar a los estabilizadores automáticos, con déficit en recesión y superávit en expansión. La deuda amortigua los *shocks* y hace posible la suavización. El déficit resulta así contracíclico y el ahorro público, procíclico.

Conclusiones: es una teoría de la deuda óptima que justifica los déficits cíclicos y financiar con deuda el gasto transitorio, pero supone un Gobierno benevolente y no explica la acumulación persistente de deuda.

3.2.3. Dominancia fiscal (Sargent y Wallace, 1981)

Idea: hay **aritmética monetarista desagradable** cuando una política fiscal expansiva financiada con deuda impide a la política monetaria controlar la inflación.

Razonamiento: si la autoridad monetaria no monetiza hoy, el déficit se financia con deuda.

1. Si $r > g_Y$, la ratio deuda/PIB crece de forma explosiva.
2. Llega un límite en que el mercado no absorbe más deuda o exige primas crecientes.
3. Con dominancia fiscal, el banco central tiene que monetizar el déficit y los intereses acumulados, con más inflación que si hubiera monetizado al principio.

Con la HER, los agentes anticipan la monetización y la inflación sube ya hoy, por lo que la inflación es más un fenómeno fiscal que monetario. Una política monetaria restrictiva puede resultar así, paradójicamente, más inflacionista que una acomodaticia, y de ahí lo desagradable de la aritmética.

Conclusiones: liga ambas políticas por la restricción intertemporal del sector público, de modo que controlar la inflación exige disciplina fiscal. La teoría fiscal del nivel de precios generaliza el razonamiento y apoya las reglas fiscales. Pero depende de que $r > g_Y$, de un límite a la deuda y de que la autoridad fiscal no ajuste, y supone, en lugar de explicar, cuál de las dos cede.

3.3. Críticas al pensamiento de la NMC

A la HER: ignora las asimetrías de información; supone que la información no cuesta; y supone riesgo y no incertidumbre en el sentido de Knight, central en la crisis financiera global, que la NMC no ha explicado de forma satisfactoria.

Al vaciado continuo: la NMC descansa en la plena flexibilidad de precios, pero en la realidad hay rigideces de precios y salarios. A comienzos de los ochenta, la **nueva economía keynesiana** introdujo en el mercado de trabajo y en el de bienes la competencia imperfecta y fricciones y rigideces nominales y reales.

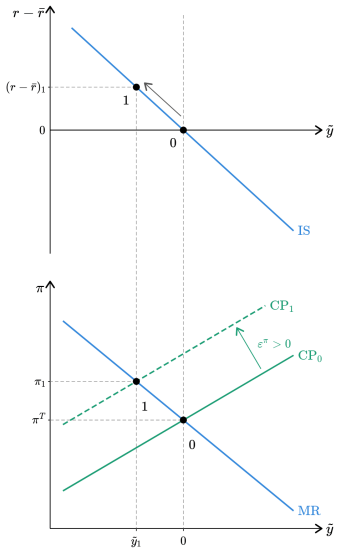
3.4. La nueva síntesis neoclásica

Convergencia: según Woodford (2009), desde finales de los noventa las opiniones convergen y se difumina la frontera entre micro y macroeconomía. Siguiendo a Goodfriend y King (1997), se habla de una **nueva síntesis neoclásica** o nuevo consenso.

Principios:

- *Metodología*: modelos de equilibrio general estocástico intertemporal, que funden el ciclo real de la NMC con las rigideces de la NEK; robustos a la crítica de Lucas; y con expectativas endógenas, racionales, sin aceptar por ello la inefectividad de la política anticipada.
- *Impulso del ciclo*: los *shocks* reales son una fuente importante de fluctuaciones, y los monetarios explican solo una pequeña parte de la varianza del PIB.
- *Política monetaria*: es efectiva, en especial para controlar la inflación, y se abandona la neutralidad a corto plazo. Tras la Gran Recesión se reconoce también un papel estabilizador a la política fiscal, sobre todo en el límite inferior de los tipos.

Desarrollo gráfico: la IS es decreciente, porque un tipo real superior al natural reduce el gasto, la CP es creciente, porque más actividad eleva costes y precios, y la curva de regla monetaria, MR, es decreciente, porque ante más inflación el banco central sube el tipo real y reduce la producción. La economía parte del punto 0, con la inflación en el objetivo y brecha nula. Ante un *shock* tecnológico positivo, si el tipo de interés no sigue la caída del natural, la MR se desplaza y la economía pasa de 0 a 1. Con acomodación plena, la brecha no se abre. Ante un *shock* de costes, la curva de Phillips se desplaza de CP_0 a CP_1 y el banco central acepta una brecha de producción negativa para devolver la inflación al objetivo.



Modelo IS-PC-MR ante un *shock* de costes: el banco central acepta una brecha de producción negativa para devolver la inflación al objetivo.

1. Fundamentos del pensamiento neokeynesiano

Los autores neokeynesianos son heterogéneos, pero sus modelos coinciden en justificar equilibrios subóptimos a corto plazo que requieren la intervención de las autoridades.

1.1. Principales rasgos

Metodología: ambas generaciones adoptan, en general, la hipótesis de expectativas racionales, HER, lo que las separa del poskeynesianismo. La primera (Fischer, 1977; Taylor, 1980; Mankiw, 1985) injerta fricciones sobre el equilibrio parcial y el IS-LM. La crítica de Lucas impulsó la microfundamentación en equilibrio general, y la segunda la completa con modelos DSGE, en la nueva síntesis neoclásica (Woodford, 2003; Galí, 2008).

Supuestos:

- *Entorno*: la NEK subraya las instituciones y los contratos que restringen la conducta individual. Defiende la microfundamentación de la macroeconomía, pero también la macrofundamentación de la microeconomía. De ahí los fallos de mercado:
  - *Poder de mercado*, generalmente competencia monopolística.
  - *Información imperfecta*, sobre todo asimétrica, como en los mercados financieros o los contratos.
  - *Rigideces nominales*: resistencia al ajuste de precios y salarios nominales por fricciones, como los costes de menú, que con pérdidas privadas de segundo orden generan efectos agregados de primer orden.
  - *Rigideces reales*: baja sensibilidad del precio o salario real deseado, el óptimo de la empresa, a la actividad agregada. Surgen de la aversión al riesgo vía contratos implícitos, los salarios de eficiencia o la curvatura de la demanda, y amplifican la rigidez nominal, porque, si el óptimo apenas se mueve, la pérdida de no ajustar el precio nominal es pequeña.
  - *Fallos de coordinación*: sin plena flexibilidad puede haber equilibrios múltiples subóptimos, por la complementariedad estratégica en la fijación de precios (Cooper y John, 1988), con "profecías autocumplidas".
- *Agentes*: el grueso de la NEK mantiene la racionalidad plena, aunque una parte admite desviaciones menores, como la *near-rationality* de Akerlof y Yellen (1985).

Ajuste: no hay vaciado continuo, sino un ajuste lento y preferentemente por cantidades, porque con los precios rígidos la producción y el empleo absorben las perturbaciones.

Política económica: intervención para corregir los fallos de mercado. Como las rigideces nominales rompen la neutralidad, la política monetaria sistemática tiene efectos reales, de ahí el énfasis en las reglas de tipos y en el principio de Taylor (1993).

1.2. Teorías sobre rigideces reales

Estas teorías explican por qué el salario o el precio real deseado apenas responde a la actividad.

1.2.1. Mercado de trabajo

Contratos implícitos (Baily, 1974; Gordon, 1974; Azariadis, 1975): ante relaciones laborales duraderas, con trabajadores aversos al riesgo y empresas neutrales, los contratos de largo plazo fijan rentas estables. La renta salarial actúa como un seguro frente a una demanda de trabajo fluctuante y se desvincula de la productividad corriente. Explica la estabilidad salarial sin irracionalidad, pero no el desempleo involuntario, porque el empleo puede seguir siendo eficiente.

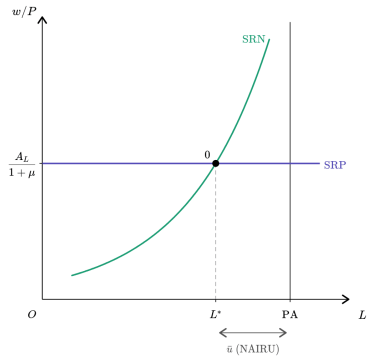
Salarios de eficiencia (Solow, 1979; Shapiro y Stiglitz, 1984): el salario afecta al esfuerzo, que en Shapiro y Stiglitz no es observable, y a las empresas les interesa incentivarlo pagando más que el de vaciado. Explica el desempleo involuntario como equilibrio, aunque supone que no hay otros incentivos, como salarios diferidos.

Insiders* y *outsiders (Lindbeck y Snower, 1988): con trabajadores heterogéneos en su peso negociador, los empleados negocian en su interés, sin atender a los desempleados, y el salario supera al de vaciado.

- El resultado es triple: dualidad, con *outsiders* sin influencia; rigidez salarial, porque los *insiders* que quedan tras los despidos atienden a su propio empleo; e histéresis, con efectos permanentes de perturbaciones transitorias.
- Exige costes de rotación elevados, por lo que pesa más con sindicatos fuertes y alta protección del empleo.

Negociación (Layard y Nickell, 1985): con este macromodelo, completado por Carlin y Soskice, la NEK europea estudia el desempleo involuntario y el poder de negociación de los setenta. Sindicatos y empresas con poder de mercado se reparten la renta negociando en términos reales, y el reparto depende de factores institucionales.

- *Salario real negociado*, SRN: lo fijan los sindicatos en términos nominales según los precios esperados. Decrece con el paro y crece con los factores institucionales.
- *Salario real pagado*, SRP: es el margen de las empresas sobre el coste laboral unitario, $w/P = A_L/(1+\mu)$, independiente del empleo. Supone rendimientos constantes, con $Y = A_L L$, y disminuye con el poder de mercado y con las cotizaciones a cargo de la empresa.
- Dado el salario nominal, la demanda agregada no altera los precios, sino las existencias o la producción, lo que respalda el ajuste keynesiano por cantidades.
- Con SRN = SRP y $P = P^e$ se obtiene un desempleo de equilibrio único y estable, la **NAIRU**, compatible con el desempleo involuntario, el estructural y el friccional. Sube con los factores institucionales, z, como el poder sindical, que desplazan la SRN hacia arriba, y con el poder de mercado, μ, que desplaza la SRP hacia abajo. El efecto es mayor cuanto más plana es la SRN. Una mayor elasticidad de la demanda de trabajo reduce el salario negociado y la NAIRU.



Determinación de la NAIRU (Layard y Nickell, 1985).

- Sin imponer $P = P^e$ se obtiene una curva de Phillips de primera generación, con expectativas adaptativas, $\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u})$.
- Es más rico que la tasa natural friedmaniana y fundamenta las políticas estructurales, pero la NAIRU no es observable, sus estimaciones son inestables y el modelo, estático, no explica la persistencia del paro europeo, que motiva la histéresis.

1.2.2. Mercado de bienes

Como los costes marginales caen mucho en las recesiones, la rigidez exige que no caigan o que la elasticidad de la demanda caiga mucho ante una perturbación negativa.

Mercados de clientela (Okun, 1981): traslada al mercado de bienes los contratos implícitos. Los costes de cambio y de búsqueda fidelizan al cliente y hacen la demanda menos elástica a corto plazo, y la elasticidad es procíclica. Las empresas no suben precios en las expansiones y sacrifican beneficios a corto plazo para conservar a sus clientes, con un margen anticíclico. El argumento es sobre todo verbal y difícil de contrastar.

Economías de aglomeración (Diamond, 1982): en un mercado denso es más fácil intercambiar, de modo que el coste marginal cae en los auges y sube en las recesiones. Puede haber equilibrios de alta y baja actividad, y los fallos de coordinación pueden atrapar a la economía en el bajo, lo que fundamenta las políticas de demanda, pero no se sabe cuál se alcanza.

Guerras de precios en los auges (Rotemberg y Saloner, 1986): en un oligopolio que sostiene la colusión tácita en un juego repetido, con demanda alta crece la ganancia inmediata de desviarse frente al castigo futuro, y el precio colusivo sostenible y el margen son anticíclicos. Exige perturbaciones observables y transitorias, y Green y Porter (1984), con información imperfecta, obtienen lo contrario.

Costes marginales constantes (Stigler, 1939): por la capacidad ociosa que las empresas mantienen como reserva, el coste marginal apenas cae en las recesiones y el precio deseado tampoco. Las perturbaciones se absorben por cantidades y la rigidez real amplifica la de los costes de menú. Esta explicación es la más sencilla, aunque supone la forma de la función de costes.

1.2.3. Conclusiones

Fundamentan la rigidez real con información asimétrica y competencia imperfecta, sin irracionalidad. Pero, como señalan Ball y Romer (1990), la **rigidez real no implica rigidez nominal**. La amplifica, pero no la genera, y sin fricciones nominales el dinero sigue siendo neutral.

1.3. Teorías sobre rigideces nominales

1.3.1. Cuasirracionalidad (Akerlof y Yellen, 1985)

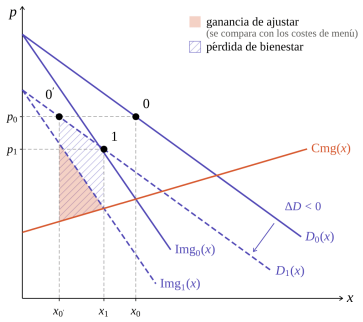
Idea: en competencia imperfecta, una proporción de empresas cuasirracionales sigue reglas prácticas, *rules of thumb*, y no ajusta sus precios, porque la pérdida de no hacerlo es de **segundo orden**, pequeña, aunque por agregación es de primer orden.

Conclusiones: exige desviaciones mínimas de la racionalidad, pero no es condición suficiente para la rigidez nominal, porque muchas otras políticas de precios dan pérdidas pequeñas.

1.3.2. Costes de menú (Mankiw, 1985)

Idea: a la empresa no le compensa cambiar el precio cuando la pérdida de mantenerlo es menor que los **costes de menú**: administrativos, físicos y de información, o no administrativos, como el daño a la imagen. Mankiw llega a la conclusión de Akerlof y Yellen sin renunciar a la racionalidad.

Desarrollo gráfico: con demanda decreciente, ingreso marginal por debajo de ella y coste marginal creciente, el competidor monopolista parte del punto 0, donde el ingreso marginal iguala al coste marginal. Si la demanda cae de D_0 a D_1, puede mantener el precio, en 0 , con ajuste íntegro por cantidades, o ajustarlo, en 1, optimizando. Solo lo cambia si el triángulo de beneficios del ajuste supera los costes de menú. Si no, la producción queda por debajo de la óptima y hay pérdida de bienestar.



Costes de menú (Mankiw, 1985): la empresa mantiene el precio (0) si la ganancia de ajustarlo (1) no supera los costes de menú.

Extensiones: un coste marginal más plano, por mayor rigidez del salario real, reduce el triángulo, y la pendiente del ingreso marginal, reflejo del poder de mercado, también condiciona su tamaño. Los factores reales acentúan o atenúan la rigidez nominal.

Conclusiones: fricciones pequeñas bastan para que la demanda tenga efectos reales. Pero el modelo es estático: si la perturbación es permanente, las pérdidas se acumulan y el ajuste es más rápido. Admite equilibrios múltiples e ignora los costes de ajustar las cantidades. Ball, Mankiw y Romer (1988) hallan que la curva de Phillips es más vertical donde la inflación media es mayor, como predicen los costes de menú.

1.3.3. Externalidades de demanda (Blanchard y Kiyotaki, 1987)

Idea: aceptan la cuasirracionalidad y los costes de menú y amplían el análisis al equilibrio general con competencia monopolística. La demanda de cada empresa depende del *stock* real de dinero y de su precio relativo. El precio de cada una afecta a la demanda de las demás, una **externalidad de demanda agregada** que ninguna internaliza, y las rigideces individuales se acumulan en una inercia nominal que hace fluctuar la producción con un nivel de precios constante.

Conclusiones: la producción de equilibrio es inferior a la eficiente, y una expansión monetaria con precios rígidos puede mejorar el bienestar de todos, lo que fundamenta la política de demanda.

1.3.4. Conclusiones

Las fricciones nominales justifican la no neutralidad, pero no bastan para explicar el tamaño y la persistencia de los efectos monetarios, lo que abrió el estudio de cómo se fijan los precios.

1.4. Análisis dinámico de las rigideces

Si todas las empresas revisan sus precios a la vez, los *shocks* solo tienen efectos transitorios, mientras que, si las revisiones se **escalonan**, los efectos reales duran mucho más. La fijación puede ser en función del tiempo o dependiente del estado.

1.4.1. Ajuste en función del tiempo (Calvo, 1983)

Origen: Fischer (1977) considera precios o salarios de t fijados en $t-j$ con la información disponible. Si no aciertan, el mercado no se vacía, a diferencia de Lucas, e incluso la política sistemática tiene efectos reales a muy corto plazo, porque reacciona a información posterior a los contratos. Pero no explica la persistencia. Taylor (1980) introduce contratos escalonados y solapados, con los que el nivel de precios hereda inercia.

Idea: Calvo supone que en cada periodo solo una fracción de las empresas puede revisar su precio, de modo que el nivel de precios se ajusta gradualmente.

Supuestos: competencia monopolística, horizonte infinito, HER y un coste marginal procíclico, que con precios rígidos hace el margen contracíclico. Cada empresa mantiene su precio con probabilidad ξ, que mide la rigidez nominal, con independencia del tiempo transcurrido:

$$p_t = \xi\, p_{t-1} + (1-\xi)\, p_t^*, \qquad \xi \in [0,1)$$

Planteamiento: sin rigideces, la empresa fijaría un margen constante sobre el coste marginal, $P_{i,t} = (1+\mu)\,\mathrm{Cmg}_t$. Con ellas, maximiza el valor actual esperado de sus beneficios mientras su precio siga vigente, con $\gamma_{t,t+j}$ el factor de descuento estocástico, sujeta a la demanda de su variedad:

$$\max_{P_t^*} V_t^e = \sum_{j=0}^{\infty} \xi^j\, \mathbb{E}_t\left[\gamma_{t,t+j}(P_t^*\, Y_{t+j} - \mathrm{CT}_{t+j}(Y_{t+j}))\right], \qquad \xi \in [0,1)$$

$$\text{s.a.: } Y_{t+j} = (P_t^*/P_{t+j})^{-\sigma} C_{t+j}$$

CPO: simplificada, da $p_t^* = f(\xi, \gamma, \mu^+, (\mathbb{E}_t[\pi_{t+j}])^+, (\mathbb{E}_t[\mathrm{Cmg}_{t+j}])^+)$. La rigidez aparece elevada a j porque la empresa contempla seguir sin revisar en el futuro, y con un descuento constante cada periodo pesa $(\beta\xi)^j$. Fija así su margen deseado sobre un promedio de su coste marginal corriente y esperado, teniendo en cuenta la inflación esperada, la rigidez y su impaciencia. El efecto de ξ es ambiguo, porque con inflación esperada positiva más rigidez eleva el precio de reajuste.

CSO: el valor esperado es estrictamente cuasicóncavo en P_t^*, porque la demanda es isoelástica con $\sigma > 1$ y el coste marginal no decrece, de modo que la CPO caracteriza el máximo.

Solución: la curva de Phillips de la NEK, en la que la inflación depende de la esperada y de la desviación del coste marginal real respecto de su nivel sin fricciones. Como el coste marginal no es observable, suele usarse la brecha de producción, proporcional a él porque es procíclico, $\widehat{\mathrm{cmg}}_t = \eta\, \tilde{y}_t$, o el desempleo cíclico:

$$\pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] + \kappa_{\mathrm{cm}}\, \widehat{\mathrm{cmg}}_t$$

$$\kappa_{\mathrm{cm}} = \frac{(1-\xi)(1-\beta\xi)}{\xi} \cdot \frac{1-\alpha}{1-\alpha+\alpha\sigma}, \qquad \xi \in [0,1)$$

$$\pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] + \kappa\, \tilde{y}_t, \qquad \kappa = \eta\, \kappa_{\mathrm{cm}}$$

$$\pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] - \kappa_u\, (u_t - \bar{u})$$

- La pendiente κ_{cm} decrece con la rigidez, porque menos empresas ajustan sus precios; con la sustituibilidad entre variedades y con la participación del capital, porque la empresa que sube su precio vende menos y reduce su coste marginal; y con el factor de descuento.
- En la pendiente en brecha, $\eta = \theta + \frac{\varphi+\alpha}{1-\alpha} > 0$, con φ la inversa de la elasticidad de Frisch.
- Frente a la de primera generación, es **prospectiva**, *forward-looking*, y por eso mismo no explica la persistencia de la inflación.

1.4.2. Ajuste en función del estado

Origen: Caplin y Spulber (1987) plantean los modelos *Ss*, en los que la empresa revisa el precio cuando su desviación del óptimo rebasa una banda.

Idea: la empresa paga el coste de menú solo cuando la pérdida acumulada lo justifica, lo que hace endógena la frecuencia de ajuste, exógena en las reglas en función del tiempo.

Razonamiento: Caplin y Spulber muestran que, con una distribución uniforme de precios en la banda e inflación positiva, ajustan justo las empresas necesarias para que los precios se muevan uno a uno con el dinero, de modo que la rigidez individual no implica rigidez agregada. Blanchard y Fischer (1989) matizan que, con *shocks* de ambos signos, el dinero recupera efectos reales, aunque menores.

En suma, Ball y Romer (1990) mostraron que la interacción entre rigideces reales y fricciones nominales permite romper la dicotomía clásica. Pero la primera generación no frenó a la NMC, por una metodología que quedaba obsoleta, y la segunda aceptó las reglas del ciclo real incorporando estas rigideces.

2. Aplicaciones: el estudio del ciclo con rigideces

Las aportaciones de la NEK se condensan hoy en un modelo DSGE que permite estudiar el ciclo y la política económica en un marco de rigideces.

2.1. Modelo IS-PC-MR

Origen: a finales de los noventa, Clarida, Galí, Gertler y Woodford dan lugar al **modelo canónico de la NEK**, que resume la economía en tres ecuaciones sobre la metodología del ciclo real, con competencia monopolística y precios *à la* Calvo. La política monetaria tiene efectos reales a corto plazo, pero no a largo.

2.1.1. Supuestos

Competencia monopolística y precios *à la* Calvo en el mercado de bienes, mercado de trabajo competitivo, HER y cuatro agentes: hogares, empresas, sector público y banco central. Por simplicidad no hay capital ni inversión. Los hogares tienen utilidad $u = \frac{c^{1-\theta}}{1-\theta} - \frac{l^{1+\varphi}}{1+\varphi}$, con θ la aversión relativa al riesgo, y ofrecen trabajo hasta $c^\theta l^\varphi = w/P$. Las empresas producen $y_i = A\, l_i^{1-\alpha}$ y, sin rigideces, fijan el margen $1+\mu = \sigma/(\sigma-1)$, con σ la elasticidad de sustitución entre variedades.

2.1.2. Modelo

Ecuaciones: la curva IS dinámica procede de la ecuación de Euler y del equilibrio en el mercado de bienes, log-linealizados, la curva de Phillips, de la fijación de precios *à la* Calvo, y la MR, de llevar a la IS la regla de Taylor del banco central. En términos de brecha, con $r_t = i_t - \mathbb{E}_t[\pi_{t+1}]$

y $\tilde{\pi}_t = \pi_t - \pi^T$:

$$\text{IS: } \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(r_t - \bar{r}_t)$$
$$\text{CP: } \pi_t = \beta\,\mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t$$
$$\text{MR: } \tilde{y}_t = \frac{\theta\,\mathbb{E}_t[\tilde{y}_{t+1}] + (\mathbb{E}_t[\pi_{t+1}] - \pi^T) - \alpha_\pi\,\tilde{\pi}_t}{\theta + \alpha_y}$$

Curva IS dinámica: la demanda crece con la brecha esperada y decrece con el tipo real, con una sensibilidad $1/\theta$, la inversa de la aversión relativa al riesgo. A diferencia de la IS estática, recoge la naturaleza previsora de los agentes: el efecto riqueza, la impaciencia y el alisamiento del consumo. Cuanto mayor es el deseo de alisar, menos reacciona el gasto a los tipos.

- En niveles, con el gasto público no productivo, g_t, el consumo esperado, $\mathbb{E}_t[c_{t+1}] = \mathbb{E}_t[y_{t+1} - g_{t+1}]$, y la tasa de descuento, ρ, la IS es la siguiente, y la forma en brecha corresponde al caso sin sector público:

$$y_t = \mathbb{E}_t[c_{t+1}] - \frac{1}{\theta}(i_t - \mathbb{E}_t[\pi_{t+1}] - \rho) + g_t$$

- Si el tipo real supera al natural, la brecha se vuelve negativa (el canal de la política monetaria). Iterando hacia delante, la brecha depende de toda la senda esperada de desviaciones del tipo real:

$$\tilde{y}_t = -\frac{1}{\theta}\sum_{k=0}^{\infty}\mathbb{E}_t[r_{t+k} - \bar{r}_{t+k}]$$

- El **tipo de interés real natural**, en línea con Wicksell, es el que fijan los factores reales y, sin *shocks* de costes, garantiza la estabilidad de precios: $\bar{r}_t = \rho + \theta\,\mathbb{E}_t[\Delta\bar{y}_{t+1}]$. Crece con la impaciencia, porque los hogares ahorran menos y escasean los fondos prestables, y cae con los *shocks* tecnológicos transitorios, que reducen el crecimiento esperado del *output* natural. Sin *shocks* de costes ni rigidez salarial real, es eficiente.

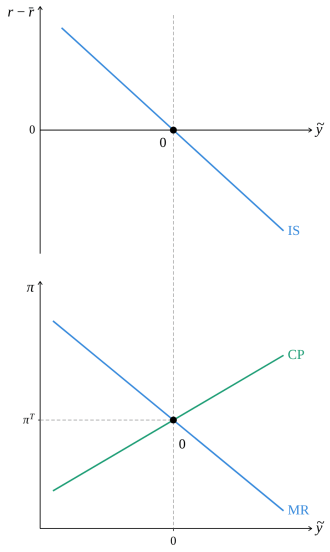
Curva de Phillips de la NEK: determina la inflación dada la senda de la brecha, que fijan a su vez los tipos y el gasto público. El gasto actúa sobre la brecha moviendo tanto la producción, vía la IS, como la potencial, porque su efecto riqueza cambia la oferta de trabajo.

Regla de política monetaria: una regla de Taylor, con $\bar{i}_t = \bar{r}_t + \pi^T$ el tipo neutral:

$$i_t = \bar{i}_t + \alpha_\pi\tilde{\pi}_t + \alpha_y\tilde{y}_t, \qquad \alpha_\pi > 1,\ \alpha_y \geq 0$$

- Con ese intercepto, la regla acomoda por completo los cambios del tipo natural. Si el banco central usa una estimación que se revisa con retraso o un intercepto constante, $\rho + \pi^T$, esos cambios desplazan la MR.
- Por el **principio de Taylor**, $\alpha_\pi > 1$: si la inflación supera el objetivo, el tipo nominal sube más que ella para elevar el tipo real. Se relaciona con la condición de Blanchard-Kahn, que garantiza un equilibrio estable y único.
- Llevada a la IS, la regla da la **MR**, una relación negativa entre brecha y desviación de la inflación. El compromiso con la estabilidad de precios implica menos *output* ante más inflación.

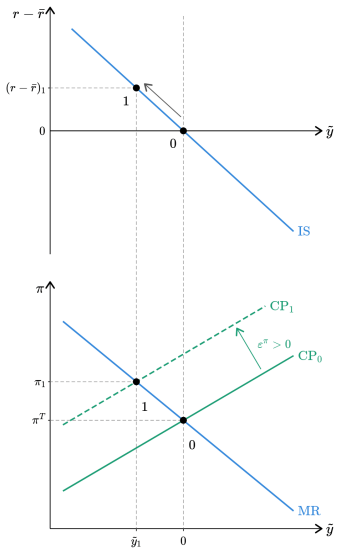
Desarrollo gráfico: arriba, la IS relaciona la brecha con la desviación del tipo real respecto del natural. Abajo, la CP crece con la brecha y la MR decrece. En el punto 0, la inflación está en el objetivo y la producción en su nivel natural.



Equilibrio del modelo IS-PC-MR: IS dinámica arriba; curva de Phillips y MR abajo, con la economía en el punto 0.

Estática comparativa:

- *Shock tecnológico positivo*: eleva el *output* natural y reduce el tipo natural. Si el banco central traslada la caída a su tipo, la brecha no se abre. Si no, la MR se desplaza y la economía va de 0 a 1, con brechas negativas de producción e inflación.
- *Shock de costes*: desplaza la CP hacia arriba y el banco central acepta transitoriamente una brecha negativa para devolver la inflación al objetivo, a lo largo de la MR.



Modelo IS-PC-MR ante un *shock* de costes: el banco central acepta una brecha de producción negativa para devolver la inflación al objetivo.

2.1.3. Conclusiones

Aspectos positivos: microfundamentación rigurosa, en principio robusta a la crítica de Lucas; tres ecuaciones tratables que permiten analizar el bienestar y lo hacen el modelo de referencia de los bancos centrales; y un marco coherente para la estabilidad de precios, la regla de Taylor y los objetivos de inflación.

Limitaciones: por ser prospectivo, no explica la persistencia de la inflación, lo que obliga a versiones híbridas; descansa en un agente representativo y omite capital, sector financiero, heterogeneidad y desempleo involuntario; y la log-linealización le impide tratar grandes perturbaciones o el límite inferior de los tipos, ZLB.

2.2. Extensiones al modelo

Economía abierta (Galí y Monacelli, 2005): en un país pequeño, con Marshall-Lerner y mercados completos, se mantienen las tres ecuaciones, pero cada una recoge la apertura, α, el peso de los importados en el consumo (con $\alpha = 0$ se recupera el caso cerrado). Con $p_{o,t}$ el precio relativo de los importados y $\tilde{E}_t$ la desviación del tipo de cambio respecto

de su nivel natural (desarrollo en A32 §3.3):

$$\text{IS: } \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta_\alpha}(r_t - \bar{r}_t) + \Gamma_y\, p_{o,t} + \varepsilon_t^d$$

$$\text{CP: } \pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] + \kappa_\alpha\, \tilde{y}_t + \Gamma_\pi\, p_{o,t} + \varepsilon_t^\pi$$

$$\text{Regla: } i_t = \bar{i}_t + \alpha_\pi\, \tilde{\pi}_t + \alpha_y\, \tilde{y}_t + \alpha_E\, \tilde{E}_t + \varepsilon_t^m$$

- El tipo natural depende también de la renta exterior. Γ_y tiene signo indeterminado, porque el efecto sustitución y el efecto renta de un encarecimiento de los importados se oponen; $\Gamma_\pi\, p_{o,t}$, con $\Gamma_\pi > 0$, es la inflación importada, y α_E va de la flotación pura ($\alpha_E = 0$) al tipo fijo ($\alpha_E \to \infty$). Parte del ajuste se canaliza por el tipo de cambio real, de modo que la estabilización ya no recae solo en el banco central y el tipo de interés.
- El modelo ordena por su bienestar los regímenes monetarios y, en un caso particular, muestra que estabilizar la inflación nacional es óptimo y la paridad fija, subóptima. Pero el país pequeño y los mercados completos dejan fuera el efecto sobre el resto del mundo, las primas de riesgo y la dinámica de la deuda externa.

Blanchard y Galí: la **rigidez salarial real** (2007) genera un término de costes endógeno y una verdadera disyuntiva entre estabilizar la inflación y la actividad. Las **fricciones de búsqueda** (2010), *à la* Diamond-Mortensen-Pissarides, introducen el desempleo de equilibrio.

Curva de Phillips salarial (Galí, 2011): la inflación salarial depende de la desviación del paro respecto de la NAWRU, lo que reintroduce el desempleo y reconecta con la relación original de Phillips:

$$\pi_t^w = \beta\, \mathbb{E}_t[\pi_{t+1}^w] - \kappa_w\, (u_t - \bar{u}_t)$$

En suma, el modelo canónico recupera la no neutralidad monetaria a corto plazo sin renunciar al rigor de los nuevos clásicos, y admite extensiones sin alterar su estructura de tres ecuaciones.

3. Implicaciones del pensamiento de la NEK

La política monetaria puede tener efectos importantes y persistentes sobre las variables reales y nominales.

3.1. Implicaciones de política monetaria

Niveles naturales: el *output* y el tipo de interés naturales, los de precios flexibles, son la referencia del banco central, que no puede desviarse persistentemente de ellos sin presiones inflacionistas o deflacionistas. Como no son observables, antes se aproximaban con tendencias estadísticas y hoy con un modelo del ciclo real sin fricciones.

Divina coincidencia (Blanchard y Galí, 2007): sin imperfecciones reales no triviales, estabilizar la inflación estabiliza también la brecha respecto del nivel eficiente. Con rigidez salarial real, la coincidencia se rompe y conviene tolerar algo más de inflación para evitar caídas del *output* ineficientemente grandes. Otros autores la rompen con un *shock* de costes en la CP, ε_t^π, por impuestos distorsionantes, salarios deseados o márgenes:

$$\pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] + \kappa\, \tilde{y}_t + \varepsilon_t^\pi, \qquad \kappa > 0,\ \beta \in (0,1)$$

- La COVID-19, en su componente de demanda, es una caída de ρ, más ahorro por precaución, que reduce el tipo natural. Si la regla de Taylor no la sigue, la MR se desplaza y aparecen brechas negativas de producción e inflación. La caída de ρ equivale a un mayor factor de descuento, $\beta = 1/(1+\rho)$.

Forward guidance (Campbell et al., 2012): la inflación depende no solo de la brecha, sino también de la senda anticipada de tipos, y gestionar las expectativas del sector privado es decisivo para la eficacia de la política monetaria. Desde la Gran Recesión, el banco central comunica su senda futura como previsión, délfica, o como compromiso, odiseica, por ejemplo estado-contingente, lo que ayuda a anclar las expectativas.

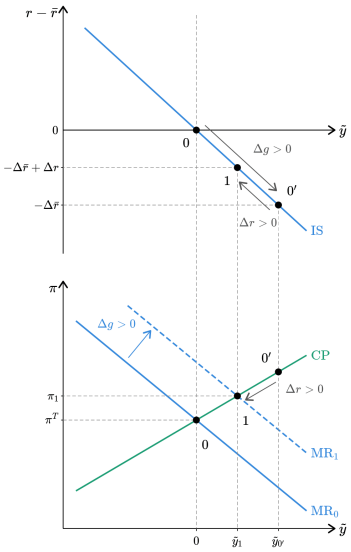
Overshooting (Dornbusch, 1976): con precios rígidos y tipo de cambio flexible, una expansión monetaria no anticipada baja el tipo de interés. Por la paridad descubierta, debe esperarse una apreciación, y el tipo de cambio se deprecia hoy por encima de su nivel de largo plazo. A medida que los precios suben, el tipo de interés se recupera y el tipo de cambio se aprecia gradualmente. Explica la volatilidad cambiaria sin irracionalidad, pero la evidencia apunta a una sobrerreacción retardada y la paridad descubierta excluye la prima de riesgo cambiario.

3.2. Implicaciones de política fiscal

Restricciones de liquidez (Galí, López-Salido y Vallés, 2007): junto a agentes ricardianos hay agentes no ricardianos que consumen toda su renta corriente y no sufren el efecto riqueza negativo del gasto público. Con suficientes, el multiplicador puede superar la unidad.

Respuesta monetaria (Woodford, 2011; Eggertsson, 2011): el multiplicador depende del banco central.

- Con tipo real constante, es igual a la unidad.
- Con objetivo de inflación o regla de Taylor, es inferior a la unidad, porque la subida de tipos desplaza gasto privado.
- En el ZLB es mayor que la unidad, porque el gasto eleva las expectativas de inflación y reduce el tipo real. Eggertsson muestra que, ante un aumento del ahorro precautorio, la espiral deflacionista hace el multiplicador elevado.
- Gráficamente, $\Delta g > 0$ eleva el tipo natural y desplaza la MR. A tipo real constante, la economía iría por la IS y la CP hasta 0 , y con la regla de Taylor el banco central sube el tipo real y la deja en 1, con menos producción.



Expansión fiscal en el IS-PC-MR: a tipo real constante, de 0 a 0 ; con la respuesta del banco central, a 1, con multiplicador inferior a la unidad.

Unión monetaria (Galí y Monacelli, 2008): cada país conserva su política fiscal, mientras que la monetaria, común, debe estabilizar la inflación de la unión. Sin política monetaria propia y con precios rígidos, la fiscal nacional gana un papel estabilizador ante perturbaciones propias, más allá de proveer bienes públicos. Como el banco central no responde a la inflación de un país pequeño, el efecto del gasto nacional no depende de la acomodación monetaria.

3.3. Política de empleo (Layard y Nickell, 1985)

El modelo de Layard y Nickell muestra que las políticas de demanda tienen margen a corto plazo, pero a costa de inflación. La escuela subraya por ello las políticas estructurales y de rentas que reducen la NAIRU desplazando la SRN y la SRP: menos cuña fiscal, gasto en formación y reforma del subsidio de desempleo. La formación eleva la productividad y reduce el poder de negociación de los trabajadores cualificados.

3.4. Críticas al pensamiento de la NEK

A la CP-NEK: por ser prospectiva, predice un ajuste inmediato cuando la desinflación es gradual (Mankiw, 2001), no reproduce la inercia de la inflación y contradice las encuestas, en las que las empresas atienden al pasado y al futuro (Fabiani et al., 2006).

- *Curvas híbridas*: Gordon (1982) introduce retardos en su modelo del triángulo; Galí y Gertler (1999) añaden un retardo de la inflación, y Christiano, Eichenbaum y Evans (2005), la indexación a la inflación pasada. Gordon combina la persistencia extrínseca, heredada de la demanda y la oferta, y la intrínseca, de la propia inflación, y en Galí y Gertler el retardo procede de una fracción de empresas que siguen reglas sencillas.
- *Información rígida* (Mankiw y Reis, 2002): los agentes actualizan sus expectativas de forma intermitente. Si los precios cambian con rapidez, solo una fracción reacciona con información al día.

Sistema financiero (Cúrdia y Woodford, 2010): su ausencia es una deficiencia. Los modelos con acelerador financiero (Bernanke, Gertler y Gilchrist, 1999) muestran cómo las imperfecciones del crédito amplifican las perturbaciones.

Política fiscal (Solow, 2004): el consenso de políticas fiscales neutrales, que no interfirieran con la monetaria, subestimó su papel anticíclico, salvo ante grandes *shocks* que llevaran a la trampa de liquidez.

Metodología (Blanchard, 2008): pide cierta simplificación y más contacto con los datos y, como Solow, observa que se ha atendido más a la política monetaria que a la fiscal. Los debates posteriores a la crisis se han alejado en cierta medida de los planteamientos de la NEK.

3.5. La nueva síntesis neoclásica

La escuela desemboca en la **nueva síntesis neoclásica** (Goodfriend y King, 1997), que Woodford (2009) describe como un consenso en tres

planos: modelos DSGE con expectativas racionales y robustos a la crítica de Lucas; *shocks* reales como fuente importante del ciclo, sin excluir los nominales; y una política monetaria efectiva, sobre todo para controlar la inflación. Las críticas marcan la agenda posterior a la Gran Recesión: fricciones financieras, heterogeneidad de agentes y más papel para la política fiscal.

A8. Teoría de la demanda del consumidor (I). Axiomas sobre las preferencias, función de utilidad y función de demanda marshalliana. La teoría de la preferencia revelada. Precios hedónicos

1. Fundamentos de la demanda del consumidor

La teoría neoclásica parte de un consumidor racional que maximiza su utilidad con una renta exógena, en un marco estático y de certidumbre.

1.1. Conjunto de elección y preferencias

1.1.1. Conjunto de elección y relaciones binarias

Definición: el **conjunto de elección**, CE, reúne todas las cestas físicamente consumibles, vectores de n bienes no negativos y divisibles. Es no vacío, cerrado y acotado inferiormente.

Sobre él se definen tres relaciones binarias: la indiferencia, $\sim$, la preferencia estricta, $\succ$, y la preferencia débil, $\succsim$, que se lee "al menos tan preferido como" y de la que se derivan las otras dos.

Para que la elección tenga sentido económico, esas relaciones deben cumplir axiomas de orden y de regularidad.

1.1.2. Axiomas de orden

Completitud: ante dos cestas cualesquiera, el consumidor siempre puede ordenarlas: $\mathbf{x}^0 \succsim \mathbf{x}^1$, $\mathbf{x}^1 \succsim \mathbf{x}^0$ o ambas, y entonces le son indiferentes.

Reflexividad: toda cesta es al menos tan buena como ella misma, $\mathbf{x}^0 \succsim \mathbf{x}^0$.

Transitividad: si $\mathbf{x}^0 \succsim \mathbf{x}^1$ y $\mathbf{x}^1 \succsim \mathbf{x}^2$, entonces $\mathbf{x}^0 \succsim \mathbf{x}^2$ (coherencia lógica entre comparaciones).

Los tres dan un **preorden completo débil**: todas las cestas se ordenan, con empates, y la elección es racional. No basta, sin embargo, para asignar un número a cada cesta (las preferencias lexicográficas son el contraejemplo clásico).

1.1.3. Axiomas de regularidad

Continuidad: pequeños cambios en las cantidades no provocan saltos en las preferencias, es decir, los conjuntos de contorno superior e inferior son cerrados.

No saturación local: en cualquier entorno de una cesta hay otra estrictamente preferida, de modo que no hay zonas de saciedad y el óptimo agota la renta.

Monotonía: más es mejor. Una cesta con al menos tanto de todo y más de algún bien es estrictamente preferida. Es más fuerte que la no saturación local, porque excluye los males.

Convexidad: si dos cestas son indiferentes, cualquier combinación convexa de ellas es al menos tan buena como cada una:

$$\mathbf{x}^0 \sim \mathbf{x}^1 \to \theta\mathbf{x}^0 + (1-\theta)\mathbf{x}^1 \succsim \mathbf{x}^0, \quad \theta \in (0,1)$$

- Recoge la preferencia por la variedad. Con convexidad estricta, la utilidad es estrictamente cuasicóncava, la RMS estrictamente decreciente y el óptimo único.

Diferenciabilidad: la utilidad es al menos dos veces diferenciable, sin vértices en las curvas de indiferencia, lo que permite usar las condiciones de Kuhn-Tucker.

1.2. Utilidad y curvas de indiferencia

1.2.1. Función de utilidad

Definición: una **función de utilidad** asigna a cada cesta un índice que respeta el orden de las preferencias: $\mathbf{x}^0 \succsim \mathbf{x}^1 \leftrightarrow u(\mathbf{x}^0) \geq u(\mathbf{x}^1)$.

Existencia: por el teorema de representación de Debreu, los axiomas de orden y el de continuidad bastan para que exista una utilidad continua. Los demás axiomas le dan propiedades analíticas convenientes.

Propiedades: es continua y dos veces diferenciable, estrictamente creciente en cada bien, con utilidades marginales positivas por monotonía, y estrictamente cuasicóncava por convexidad estricta.

Es **ordinal**: solo importa el orden, no la magnitud, y cualquier transformación $v = f(u)$ con $f' > 0$ representa las mismas preferencias. Además, no es observable, porque solo lo son las elecciones.

1.2.2. Curvas de indiferencia

Definición: una **curva de indiferencia** es el lugar geométrico de las cestas que dan un mismo nivel de utilidad, es decir, un corte a utilidad constante de la superficie de utilidad.

Propiedades: cada una procede de un axioma.

- Por transitividad, y no por monotonía, no se cortan, porque si lo hicieran, cestas de niveles distintos de utilidad tendrían que ser indiferentes.
- Por monotonía, tienen pendiente negativa, y las más alejadas del origen representan más utilidad.
- Por convexidad, son convexas respecto al origen. Solo lo son en sentido débil si tienen tramos rectos, como con sustitutivos o complementarios perfectos.

1.2.3. Pendiente: la RMS

Definición: la **relación marginal de sustitución** del bien j por el bien i, el precio relativo subjetivo de los dos bienes, mide a cuántas unidades de j renuncia el consumidor por una unidad más de i sin cambiar su utilidad ni el consumo de los demás bienes.

Se obtiene de la diferencial total de la utilidad igualada a cero y, con dos bienes, 1 y 2, es el valor absoluto de la pendiente de la curva de indiferencia:

$$\mathrm{RMS}_{ji} = -\left.\frac{dx_j}{dx_i}\right|_{du=0} = \frac{u'_i}{u'_j} \geq 0, \qquad \mathrm{RMS}_{21} = \frac{u'_1}{u'_2}$$

Como es un cociente de utilidades marginales, no depende de cómo se mida la utilidad, a diferencia de la utilidad marginal decreciente.

Con convexidad estricta la **RMS es decreciente**: cuanto más bien 1 tiene el consumidor, menos bien 2 está dispuesto a ceder por otra unidad. Con preferencias de "todo o nada", las curvas serían cóncavas y el óptimo, de esquina.

1.2.4. Forma: la elasticidad de sustitución

Definición: la **elasticidad de sustitución** de Hicks (1932) mide la facilidad del intercambio, es decir, cuánto varía en porcentaje la proporción consumida cuando la RMS varía un uno por ciento a lo largo de una curva de indiferencia, con el consumo de los demás bienes constante:

$$\sigma_{ji} = \frac{\partial \ln(x_j/x_i)}{\partial \ln \mathrm{RMS}_{ji}} \geq 0$$

Clasificación: con $\sigma_{ji} = \infty$ hay sustitutivos perfectos y curvas lineales; con $\sigma_{ji} = 0$, complementarios perfectos y curvas angulares, tipo Leontief; entre ambos, sustitutivos imperfectos, como la Cobb-Douglas, con $\sigma_{ji} = 1$.

Gráficamente, mide la curvatura: cuanto mayor es σ, más plana es la curva. Con más de dos bienes, la elasticidad de sustitución de Allen (1938), $\sigma_{ij}^{\text{Allen}}$, deja que se ajuste el consumo de los demás.

1.3. Restricción presupuestaria

Definición: las preferencias dicen qué desea el consumidor. Unos precios positivos, $\mathbf{p}$, y una renta fija, M_0, dicen qué puede pagar. El **conjunto asequible**, $\mathrm{CA} = \{\mathbf{x} \in \mathrm{CE} : \mathbf{p} \cdot \mathbf{x} \leq M_0\}$, es no vacío, cerrado, acotado y convexo, y hace el problema genuinamente económico. Es acotado porque los precios positivos y la renta finita limitan el consumo de cada bien, y convexo porque es la intersección de un semiplano con el ortante no negativo.

Recta de balance: por monotonía se agota la renta y la restricción se cumple con igualdad. Con dos bienes:

$$x_2 = \frac{M_0}{p_2} - \frac{p_1}{p_2}x_1$$

- Corta los ejes en M_0/p_1 y M_0/p_2, y su pendiente, $-p_1/p_2$, es el precio relativo objetivo (lo que el mercado exige del bien 2 por cada unidad del bien 1).
- Como una de las tres variables sobra, puede escribirse en términos reales, tomando un bien o la renta como numerario. Con el bien 2 como numerario, $\frac{p_1}{p_2}x_1 + x_2 = \frac{M_0}{p_2}$, y con la renta, $\frac{p_1}{M_0}x_1 + \frac{p_2}{M_0}x_2 = 1$.

Bien compuesto: si los precios de un grupo de bienes se mueven a la par, su gasto, G, se trata como un solo bien, $p_1x_1 + G \leq M_0$, y el problema de n bienes se reduce a dos dimensiones.

1.4. Decisión de consumo y demanda individual

1.4.1. Equilibrio del consumidor

Planteamiento: el consumidor elige la cesta que maximiza su utilidad dentro de su presupuesto y sin consumos negativos:

$$\max_{\mathbf{x}} \; u(\mathbf{x}) \quad \text{s.a.:} \begin{cases} \sum_{i=1}^{n} p_i x_i \leq M_0 \\ x_i \geq 0 \; \forall i \end{cases}$$

- En el lagrangiano, $\mathcal{L} = u(\mathbf{x}) + \lambda(M_0 - \sum_i p_i x_i)$, el multiplicador λ es el precio sombra de la renta, es decir, la utilidad marginal de la renta en el óptimo. Formalmente, $\lambda = \partial u^*/\partial M \equiv u'_M$.

CPO: las condiciones de Kuhn-Tucker exigen, para cada bien:

$$u'_i - \lambda p_i \leq 0, \quad x_i \geq 0, \quad x_i\,(u'_i - \lambda p_i) = 0$$

- Por la holgura complementaria, o el bien se consume y la condición se cumple con igualdad, o no se consume. Por monotonía, $\lambda > 0$ y se agota la renta.

CSO: como la restricción es lineal, las CPO son suficientes si la utilidad es cuasicóncava, y el máximo es único si lo es estrictamente.

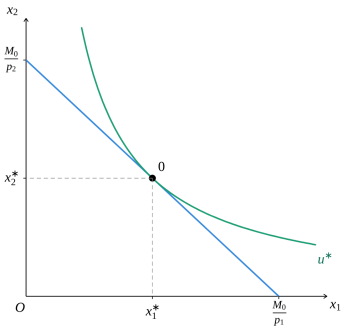
Solución interior: la RMS, que es subjetiva, iguala el precio relativo, que es objetivo. Dicho de otro modo, la utilidad marginal por euro gastado es la misma en todos los bienes e igual a λ, que es la segunda ley de Gossen. Para cada par de bienes i, j:

$$\mathrm{RMS}_{ji} = \frac{p_i}{p_j} \;\Leftrightarrow\; \frac{u'_i}{p_i} = \frac{u'_j}{p_j} = \lambda$$

- Si la igualdad no se cumpliera, reasignar el gasto de un bien a otro elevaría la utilidad.

Solución de esquina: con dos bienes, si $x_2^* = 0$, la valoración subjetiva del bien 1 es al menos su precio relativo, $\text{RMS}_{21} \geq p_1/p_2$, y toda la renta va al bien 1. Si $x_1^* = 0$, la desigualdad se invierte.

Desarrollo gráfico: el equilibrio 0 es la tangencia entre la curva de indiferencia más alta alcanzable, u^*, y la recta de balance, en la cesta (x_1^*, x_2^*).



Equilibrio del consumidor: tangencia de la curva de indiferencia más alta con la recta de balance.

1.4.2. Demanda marshalliana

Definición: resolviendo las CPO, por el teorema de la función implícita, la cantidad óptima de cada bien depende solo de los precios y la renta (**demandas marshallianas**, ordinarias o walrasianas):

$$x_i^m = d_i(\mathbf{p}, M) = d_i(k\mathbf{p}, kM)$$

- La solución existe por el teorema de Weierstrass, porque el conjunto asequible es compacto y la utilidad continua, y es única por la cuasiconcavidad estricta.
- Con una Cobb-Douglas, $u = x_1^\alpha x_2^{1-\alpha}$, la solución es interior, la tangencia da $p_2 x_2 = \frac{1-\alpha}{\alpha} p_1 x_1$ y cada bien absorbe una fracción constante del gasto:

$$x_1^m = \frac{\alpha M}{p_1}, \qquad x_2^m = \frac{(1-\alpha)M}{p_2}$$

Propiedades:

- Son continuas, por el teorema del máximo, y diferenciables si la utilidad es regular, lo que permite la estática comparativa.
- Son **homogéneas de grado cero** en precios y renta: multiplicar todo por k no cambia nada, porque el consumidor no padece ilusión monetaria. Tampoco cambian ante transformaciones monótonas de la utilidad, porque la cesta óptima depende de las preferencias y no de su representación.
- Agotan la renta, $\mathbf{p} \cdot \mathbf{x}^* = M$, que en el agregado es la ley de Walras.
- Coinciden con las hicksianas (las que minimizan el gasto para una utilidad dada) evaluadas en la utilidad máxima, $d_i(\mathbf{p}, M) = h_i(\mathbf{p}, v(\mathbf{p}, M))$, y, si su matriz de Slutsky es simétrica y semidefinida negativa, de ellas pueden recuperarse las preferencias que las generan, propiedad de integrabilidad.
- Respecto al precio propio pueden tener cualquier pendiente, porque combinan el efecto renta y el efecto sustitución. Respecto a la renta, crecen para los bienes normales y decrecen para los inferiores.
- Son observables, a diferencia de la utilidad.

1.4.3. Función indirecta de utilidad

Definición: llevando las demandas a la utilidad se obtiene la **función indirecta de utilidad**, la utilidad máxima alcanzable con cada combinación de precios y renta:

$$v(\mathbf{p}, M) = u(d_1(\mathbf{p}, M), \dots, d_n(\mathbf{p}, M))$$

De ella se recuperan las demandas con la identidad de Roy, $d_i = -(\partial v/\partial p_i)/(\partial v/\partial M)$. Con la Cobb-Douglas y $\alpha = 1/2$, $v = M/(2p_1^{1/2}p_2^{1/2})$.

En suma, con pocos axiomas la teoría obtiene una utilidad ordinal y demandas con propiedades contrastables. Sus límites son unos axiomas de regularidad elegidos por conveniencia analítica, un marco estático y una utilidad que no se observa.

2. Estática comparativa

Conocida la demanda marshalliana, la estática comparativa estudia cómo responde a cambios en la renta y en los precios, para identificar leyes de demanda y construir la curva de demanda individual y, por agregación, la de mercado.

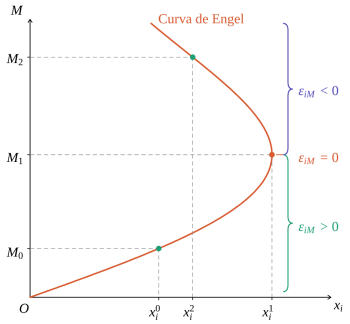
2.1. Cambios en la renta

2.1.1. Curva renta-consumo y curva de Engel

Un cambio de la renta, con precios y preferencias constantes, desplaza en paralelo la recta de balance, porque su pendiente no varía, y el nuevo equilibrio está en una curva de indiferencia más alejada del origen, o más próxima si la renta cae. Su efecto es $\partial d_i/\partial M$, cuyo signo no fijan los axiomas.

Curva renta-consumo: une las cestas óptimas al variar la renta. Tiene pendiente positiva si ambos bienes son normales, se inclina hacia el eje del bien normal si el otro es inferior y, aunque su forma puede ser irregular, es una recta desde el origen con preferencias homotéticas.

Curva de Engel: proyecta esa curva sobre el plano cantidad-renta, $x_i^* = f_i(M)$. Es creciente para un bien normal, más que proporcionalmente si es de lujo, decreciente para uno inferior y una recta desde el origen con preferencias homotéticas.



Curva de Engel: el bien es normal hasta M_1 e inferior a partir de ese nivel.

2.1.2. Elasticidad-renta

Definición: la **elasticidad-renta** es la variación porcentual de la cantidad demandada ante una variación porcentual de la renta, a precios constantes. Gráficamente, es la elasticidad de la curva de Engel en cada punto:

$$\varepsilon_{iM} = \frac{\partial x_i}{\partial M} \cdot \frac{M}{x_i}$$

2.1.3. Clasificación de los bienes

Clasificación: un bien es inferior si $\varepsilon_{iM} < 0$, como las marcas blancas; frontera si es nula, y normal si es positiva. Entre los normales, es de primera necesidad si está entre cero y uno, como el pan; de elasticidad unitaria si vale uno, con una participación en el gasto que no varía, y de lujo si supera la unidad. Si todos los bienes tienen elasticidad unitaria, las preferencias son homotéticas.

La clasificación no es del bien, sino de su relación con el consumidor. El transporte público puede ser normal para una renta e inferior para otra.

Ley de Engel (1857): al crecer la renta, cae la fracción que se dedica a alimentos, y por eso esa fracción sirve como indicador del nivel de vida.

En política económica, la elasticidad-renta explica cómo cambia la estructura del consumo con el crecimiento y la progresividad de los impuestos indirectos: gravar los bienes de primera necesidad es regresivo y gravar los de lujo, progresivo.

2.2. Cambios en los precios

Si todos los precios varían en la misma proporción, la homogeneidad hace que equivalga a un cambio de renta de signo contrario. Si varía un solo precio, cambian a la vez el precio relativo y el poder adquisitivo de la renta.

2.2.1. Efectos sustitución y renta: Slutsky

Para separar los dos efectos hay que definir la renta real constante.

- Para Slutsky (1915), es la que permite comprar la cesta inicial a los nuevos precios. Es observable y no exige conocer las preferencias.
- Para Hicks (Hicks y Allen, 1934), es la que permite alcanzar la utilidad inicial. Exige conocer las preferencias, pero aísla el efecto sustitución puro. Ambas coinciden para cambios infinitesimales.

La definición de Hicks requiere la **demanda hicksiana o compensada**, $h_i(\mathbf{p}, u_0)$: la cantidad que minimiza el gasto necesario para alcanzar la utilidad u_0.

Ecuación de Slutsky: el efecto total de p_j sobre el bien i se descompone en un efecto sustitución y un efecto renta:

$$\underbrace{\frac{\partial x_i^m}{\partial p_j}}_{\text{ET}} = \underbrace{\frac{\partial h_i}{\partial p_j}}_{\text{ES}} \underbrace{- \frac{\partial x_i^m}{\partial M} x_j}_{\text{ER}}$$

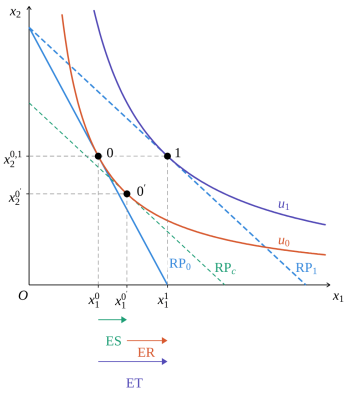
- El término x_j mide cuánto poder adquisitivo se pierde por cada unidad que sube p_j.
- Si la renta procede de vender una dotación inicial, ω_j, aparece un efecto renta-dotación de signo contrario al ordinario, y el efecto renta depende de la demanda neta, $z_j = x_j^m - \omega_j$. Así, el vendedor neto de un bien normal puede demandar más de él cuando sube su precio sin que el bien sea Giffen:

$$\frac{\partial x_i^m}{\partial p_j} = \frac{\partial h_i}{\partial p_j} - \frac{\partial x_i^m}{\partial M} z_j$$

Efecto sustitución: es el movimiento a lo largo de la curva de indiferencia inicial. El propio es **siempre no positivo** (ley de la demanda compensada), por la concavidad de la función de gasto, y solo es nulo con complementarios perfectos. El cruzado tiene signo indeterminado.

Efecto renta: es el paso a otra curva de indiferencia. El propio refuerza el efecto sustitución si el bien es normal y lo contrarresta si es inferior, hasta poder dominarlo.

Gráficamente, al bajar p_1, la recta pivota sobre su corte con el eje de x_2. Una recta auxiliar con la nueva pendiente y tangente a u_0 da la cesta 0': de 0 a 0' va el efecto sustitución y de 0' a 1, ya en u_1, el efecto renta. Solo el efecto total, de 0 a 1, es observable.



Ecuación de Slutsky: efecto sustitución ($0 \to 0'$) y efecto renta ($0' \to 1$) de una bajada de p_1.

2.2.2. Elasticidad-precio

Definición: la elasticidad-precio propia, ε_{ii}, y la cruzada, ε_{ij}, miden la respuesta porcentual de la cantidad al precio del propio bien o al de otro, *ceteris paribus*:

$$\varepsilon_{ij} = \frac{\partial x_i}{\partial p_j} \cdot \frac{p_j}{x_i}$$

Las marshallianas recogen el efecto total y las compensadas, ε_{ij}^c, solo el de sustitución. Se relacionan por la ecuación de Slutsky en elasticidades, $\varepsilon_{ij} = \varepsilon_{ij}^c - s_j \varepsilon_{iM}$, donde s_j es la participación del bien j en el gasto.

2.2.3. Clasificación de los bienes

Según el efecto propio: los bienes normales y los frontera son siempre ordinarios. Un bien inferior es ordinario si domina el efecto sustitución, rígido si ambos se igualan y **Giffen** si domina el efecto renta, de modo que su cantidad sube con el precio.

- El bien Giffen no debe confundirse con el bien Veblen, cuya demanda sube con el precio por consumo ostentoso. La ley de la demanda es, así, un teorema condicionado y no un axioma, porque el caso Giffen es posible, pero raro.

Según la elasticidad propia: la demanda es inelástica si $\varepsilon_{ii} \in (-1, 0)$, unitaria si vale -1 y elástica si es menor que -1, con los extremos de la demanda perfectamente inelástica y perfectamente elástica.

Relaciones brutas: con el efecto total cruzado, dos bienes son sustitutivos brutos si $\partial x_i/\partial p_j > 0$, complementarios brutos si es negativo e independientes si es nulo. Son observables, pero asimétricas, porque los efectos renta difieren entre bienes.

Relaciones netas: con el efecto sustitución, son sustitutivos netos si $\partial h_i/\partial p_j > 0$ y complementarios netos si es negativo. No son observables, pero sí simétricas, $\partial h_i/\partial p_j = \partial h_j/\partial p_i$. Con dos bienes y RMS decreciente, ambos son siempre sustitutivos netos. La razón es la homogeneidad de grado cero de la demanda hicksiana, $p_1\,\partial h_1/\partial p_1 + p_2\,\partial h_1/\partial p_2 = 0$, que con un efecto sustitución propio negativo exige uno cruzado positivo.

2.3. Curva de demanda

2.3.1. Curva precio-consumo

Curva precio-consumo: haciendo pivotar la recta de balance para distintos precios del bien 1, *ceteris paribus*, se obtiene el lugar de las cestas óptimas. Es creciente si la demanda es inelástica, horizontal si es unitaria y decreciente si es elástica. Con demanda inelástica, el gasto en x_1 aumenta al subir su precio, y la relación entre la elasticidad y la forma de la curva se sigue de la condición de Cournot.

Proyectada sobre el plano cantidad-precio, da la curva de demanda precio-propio del consumidor.

2.3.2. Demanda individual

Definición: la **curva de demanda individual** es la sección de la demanda marshalliana en su propio precio, $x_i^* = d_i(p_i)$, con los otros precios, la renta nominal y las preferencias constantes. Un cambio del precio propio es un movimiento a lo largo de la curva, mientras que cualquier otro determinante la desplaza.

Hereda la continuidad y la homogeneidad de grado cero, y su pendiente es el efecto total propio de Slutsky, negativa salvo en los bienes rígidos y Giffen. Con una Cobb-Douglas es una hipérbola rectangular de elasticidad unitaria.

Las demandas compensadas solo recogen el efecto sustitución y cumplen siempre la ley de la demanda, pero solo la hicksiana no es observable. Para un bien normal, la marshalliana es más elástica que la hicksiana.

2.3.3. Demanda de mercado

Definición: la **demanda de mercado** es la suma horizontal de las demandas individuales de los H consumidores: $D_i(p_i) = \sum_{h=1}^{H} d_i^h(p_i, \mathbf{p}_{-i}^0, M_0^h)$.

Hereda la continuidad, con posibles quiebros, la homogeneidad y la ley de Walras agregada, y es decreciente si lo son las individuales. En general, en cambio, no hereda la simetría ni la ley de la demanda compensada. La pendiente negativa de las individuales es condición suficiente, no necesaria, y la de mercado puede decrecer aunque algún consumidor tenga una demanda Giffen.

Es más plana que cada demanda individual, aunque no necesariamente más elástica, y se desplaza con la renta y su distribución, los demás precios, las preferencias y el número de consumidores. Su elasticidad, a un precio dado, es la media de las individuales ponderada por la participación de cada consumidor en la cantidad total.

En conjunto, la estática comparativa deja cinco leyes: ausencia de ilusión monetaria, ley de Walras, efecto sustitución propio negativo, efecto sustitución cruzado simétrico y bienes normales siempre ordinarios. Su límite es que ni las preferencias ni la utilidad son observables y que los datos de mercado suman a muchos consumidores.

3. Extensiones

La teoría ordinal se completa con la agregación y las restricciones de los sistemas de demanda, la preferencia revelada, las teorías de Lancaster y de Becker y los precios hedónicos.

3.1. Agregación y restricciones de la demanda

3.1.1. Agregación de bienes

Bien compuesto (Hicks, 1939b; Leontief, 1936): si los precios de un grupo varían siempre en la misma proporción, el gasto del grupo se comporta como la demanda de un solo bien.

Separabilidad débil: si la RMS entre dos bienes de un grupo no depende del consumo de los demás grupos, el consumidor decide en dos etapas: reparte la renta entre grupos y después, dentro de cada uno, atiende solo a sus precios.

Ambas vías justifican los agregados, los deflactores y la estimación por etapas. Su incumplimiento sesga las elasticidades estimadas del grupo.

3.1.2. Agregación de consumidores

Condición de Gorman (1953): la demanda de mercado no depende de la distribución de la renta si las curvas de Engel de todos los consumidores son rectas paralelas, como con preferencias homotéticas e idénticas o cuasilineales.

Esta condición es el fundamento, frágil, del agente representativo. Sin ella, la demanda de mercado pierde la simetría y la negatividad de Slutsky.

3.1.3. Restricciones del sistema de demanda

La homogeneidad y la ley de Walras imponen tres restricciones en elasticidades a todo sistema completo de demanda, con $s_i = p_i x_i / M$ la participación de cada bien en el gasto:

$$\sum_j \varepsilon_{ij} + \varepsilon_{iM} = 0$$

$$\sum_i s_i\, \varepsilon_{iM} = 1$$

$$\sum_i s_i\, \varepsilon_{ij} = -s_j$$

- *Homogeneidad*: los efectos de todos los precios y de la renta sobre cada bien se compensan exactamente.
- *Agregación de Engel*: la media ponderada de las elasticidades-renta es uno, de modo que no todos los bienes pueden ser inferiores y los de lujo conviven con otros por debajo de la unidad.
- *Agregación de Cournot*: el encarecimiento de la cesta al subir p_j se absorbe con cambios en todas las cantidades. Con dos bienes, el 1 es sustitutivo bruto del 2 si y solo si la demanda del 2 es elástica.

Las tres son condiciones necesarias, no suficientes: que el sistema proceda de maximizar una utilidad exige además una matriz de Slutsky simétrica y semidefinida negativa.

3.1.4. Conclusiones

Las restricciones de homogeneidad, Engel y Cournot hacen contrastable la teoría, pero las condiciones de agregación exacta son muy restrictivas.

3.2. Preferencia revelada

Idea: como la utilidad no se observa, la preferencia revelada invierte el enfoque y es inductiva: parte de las cestas elegidas a unos precios y una

renta y se pregunta qué revelan de las preferencias. Supone una cesta única en cada situación y que se agota la renta.

3.2.1. ADPR (Samuelson, 1938)

ADPR: si a los precios $\mathbf{p}^0$ el consumidor elige $\mathbf{x}^0$ pudiendo comprar $\mathbf{x}^1$, la primera se revela directamente preferida, y la segunda nunca podrá revelarse preferida a ella. Cuando elija $\mathbf{x}^1$, por tanto, $\mathbf{x}^0$ tiene que ser inasequible:

$$\mathbf{p}^0 \cdot \mathbf{x}^0 \geq \mathbf{p}^0 \cdot \mathbf{x}^1 \rightarrow \mathbf{p}^1 \cdot \mathbf{x}^1 < \mathbf{p}^1 \cdot \mathbf{x}^0$$

Del ADPR se sigue, en primer lugar, la estabilidad: ante las mismas opciones, el consumidor elige siempre lo mismo. Además, sin función de utilidad, basta para el efecto sustitución propio no positivo y para la pendiente negativa de la demanda de los bienes normales.

Con dos bienes, el ADPR garantiza también la transitividad, pero no con más de dos, y por eso es condición necesaria pero no suficiente de racionalidad.

3.2.2. AFPR (Houthakker, 1950)

Puede ocurrir que dos cestas no sean comparables directamente, porque ninguna es asequible cuando se elige la otra. Se define entonces la preferencia revelada indirecta, una cadena de preferencias directas. En símbolos, con $\succ_d$ una elección observada y $\succ_i$ una inferencia a partir de una cadena de elecciones:

$$\mathbf{x}^0 \succ_d \mathbf{x}^1 \text{ y } \mathbf{x}^1 \succ_d \mathbf{x}^2 \rightarrow \mathbf{x}^0 \succ_i \mathbf{x}^2 \rightarrow \neg(\mathbf{x}^2 \succ_d \mathbf{x}^0)$$

AFPR: si $\mathbf{x}^0$ se revela preferida a $\mathbf{x}^2$, directamente o a través de una cadena de elecciones, $\mathbf{x}^2$ nunca podrá revelarse preferida a $\mathbf{x}^0$. Añade la transitividad y es **condición necesaria y suficiente** de racionalidad.

Teorema de Afriat (1967): si un conjunto finito de observaciones cumple el AFPR generalizado, GARP, existe una utilidad continua, monótona y cóncava que las racionaliza. El teorema ofrece un contraste no paramétrico, sin forma funcional.

3.2.3. Conclusiones

Con las cestas observadas se construyen el índice de Laspeyres, que sobreestima la subida del coste de la vida por ignorar la sustitución, y el de Paasche, que la subestima, además de contrastes de racionalidad con encuestas y experimentos.

Pero exige muchas elecciones con preferencias estables, sus contrastes tienen poca potencia y, por Afriat, no relaja la racionalidad: la expresa en términos observables.

3.3. Restricciones tecnológicas

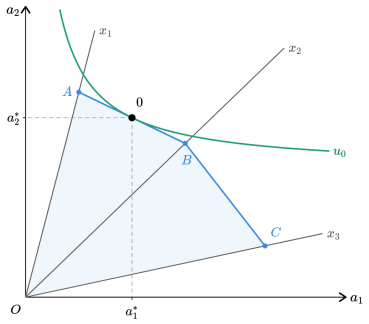
En las dos teorías que siguen, la utilidad no procede de los bienes en sí, sino de lo que el consumidor produce con ellos.

3.3.1. Demanda de características (Lancaster, 1966)

Idea: ante la incapacidad de la teoría con bienes homogéneos para explicar las variedades, Lancaster sostiene que se demandan las **características** de los bienes, que estos combinan en proporciones distintas.

Tecnología: es lineal y objetiva, $a_j = \sum_i \beta_{ji} x_i$, donde β_{ji} es la cantidad de la característica j que aporta una unidad del bien i. Cada bien es un rayo en el plano de características.

Frontera eficiente de consumo: une los puntos que se alcanzan gastando toda la renta en cada bien. Es cóncava, y los bienes que quedan dentro son dominados y no se consumen. El óptimo es la tangencia con la curva de indiferencia más alta, en un tramo, combinando dos bienes, o en un vértice.



Demanda de características: tres bienes, frontera eficiente quebrada A-B-C y equilibrio en el tramo AB.

Explica la diferenciación de producto como fenómeno de demanda. Sus límites son la tecnología lineal, las características objetivas y la falta de una teoría de los precios de las variedades.

3.3.2. Producción familiar (Becker, 1965)

Idea: el hogar produce actividades o bienes finales con bienes de mercado y tiempo, $a_j = f_j(x_j, t_j)$, y el coste de oportunidad del tiempo es el salario.

Planteamiento: al unir las restricciones de tiempo y de presupuesto se obtiene la de **renta completa**, en la que $\pi_j = p_j b_j + w\tau_j$ es el precio sombra de cada actividad, con b_j y τ_j los bienes y el tiempo por unidad de actividad:

$$\max_{\mathbf{a}} \; u(\mathbf{a}) \quad \text{s.a.:} \quad \sum_j \pi_j a_j = wT_0 + M_0$$

CPO: la habitual, con precios sombra en lugar de precios de mercado:

$$\frac{\partial u/\partial a_j}{\partial u/\partial a_k} = \frac{\pi_j}{\pi_k}$$

Estática comparativa: si sube el salario, se encarecen las actividades intensivas en tiempo, que se sustituyen por otras intensivas en bienes, como la comida preparada o la guardería. La teoría funda la economía de la familia, aunque ignora las relaciones de poder dentro del hogar.

3.4. Precios hedónicos (Court, 1939; Griliches, 1961; Rosen, 1974)

Origen: Court valoró automóviles por sus atributos, y Griliches y Rosen formalizaron el marco, versión empírica de Lancaster que recupera de los precios observados el precio implícito de cada característica.

3.4.1. Supuestos

Cada variedad es un vector de características observables, $\mathbf{z}$, como la superficie o la antigüedad de una vivienda. El mercado es competitivo, y consumidores y productores son precio-aceptantes ante la función de precios hedónicos, $p(\mathbf{z})$, en general no lineal.

3.4.2. Modelo

Planteamiento: el consumidor elige las características de la variedad, $\mathbf{z}$, y el gasto en el resto de bienes, y:

$$\max_{\mathbf{z},\, y} \; u(\mathbf{z}, y) \quad \text{s.a.:} \quad y + p(\mathbf{z}) = M_0$$

CPO: la RMS entre cada característica y el resto de bienes iguala su **precio implícito**, es decir, lo que el mercado cobra por una unidad más de esa característica:

$$\frac{\partial u/\partial z_j}{\partial u/\partial y} = \frac{\partial p(\mathbf{z})}{\partial z_j}$$

CSO: la CPO basta si la utilidad es cuasicóncava y $p(\mathbf{z})$ es convexa en las características. Si $p(\mathbf{z})$ es cóncava, la elección se desplaza a los extremos.

Equilibrio: Rosen muestra que $p(\mathbf{z})$ es la envolvente de las tangencias entre las curvas de oferta de las empresas y las de puja de los consumidores. Se estima por MCO, con especificación lineal, semilogarítmica o doble logarítmica. En esta última, cada coeficiente es directamente la elasticidad del precio respecto a la característica.

3.4.3. Conclusiones

Su aplicación principal son los índices de precios ajustados por calidad: un ordenador que cuesta lo mismo que hace cinco años con el doble de capacidad ha bajado de precio en términos ajustados por calidad. Institutos como el BLS o Eurostat usan estos métodos en los índices de precios de consumo y de vivienda.

Se aplica también a la vivienda y a la valoración de bienes ambientales, como el ruido, a través del mercado inmobiliario.

Sus límites son que los precios implícitos son propios de un mercado concreto, que el segundo paso de Rosen plantea problemas de identificación y que solo admite características cuantificables.

En suma, la sección delimita cuándo la demanda agregada conserva las propiedades individuales, recupera las preferencias desde la conducta y traslada la utilidad a las características, las actividades y el tiempo.

A9. Teoría de la demanda del consumidor (II). Dualidad e integrabilidad de las preferencias. Sistemas de demanda utilizados en estudios empíricos. Medidas de cambio en el bienestar

1. Fundamentos del análisis dual: el problema primal

La teoría neoclásica trata al consumidor como un agente racional que persigue la mayor utilidad que le permite su renta, exógena, en un marco estático y de certidumbre, y la utilidad nace de comprar bienes de consumo final.

1.1. El problema primal: maximizar la utilidad

Planteamiento: dados precios y renta, exógenos, el consumidor elige en el conjunto asequible, $\text{CA} = \{\mathbf{x} \in \mathbb{R}^n_+ : \mathbf{p}\cdot\mathbf{x} \leq M_0\}$, la cesta de máxima utilidad. Este problema primal es el punto de partida, y el dual se presentará como su imagen especular:

$$\max_{\mathbf{x}} \; u(\mathbf{x}) \quad \text{s.a.:} \begin{cases} \sum p_i x_i \leq M_0 \\ x_i \geq 0 \; \forall i \end{cases}$$

- El lagrangiano es $\mathcal{L} = u(\mathbf{x}) + \lambda\,(M_0 - \sum p_i x_i)$, con λ la utilidad marginal de la renta. Por la envolvente, $\lambda = \partial v/\partial M$, el precio sombra de la renta.

CPO: con preferencias monótonas y solución interior, el consumidor agota su renta e iguala la RMS al precio relativo, $\text{RMS}_{ji} \equiv u_i'/u_j' = p_i/p_j$, es decir, la tasa subjetiva de intercambio iguala a la objetiva y, en general, $u_i'/p_i = \lambda$. La utilidad marginal por euro gastado es así la misma en todos los bienes e igual a la utilidad marginal de la renta.

CSO: con utilidad estrictamente cuasicóncava, es decir, curvas de indiferencia estrictamente convexas, la tangencia es el máximo y la solución es única.

Solución: las demandas marshallianas, $d_i(\mathbf{p}, M)$, y, como función de valor óptimo, la función indirecta de utilidad, $v(\mathbf{p}, M) = u(\mathbf{d}(\mathbf{p}, M))$, ambas con réplica en el problema dual. En el caso Cobb-Douglas, $u = x_1^{1/2}x_2^{1/2}$, el gasto se reparte a partes iguales y:

$$x_1^m = \frac{M}{2p_1}, \qquad v(\mathbf{p}, M) = \frac{M}{2p_1^{1/2}p_2^{1/2}}$$

1.2. Demanda marshalliana

Definición: resolviendo las CPO y la restricción para cualesquiera precios y renta se obtienen las demandas ordinarias, **marshallianas** o walrasianas, $x_i^m = d_i(\mathbf{p}, M)$. Son *no compensadas*, porque ante una variación de los precios la renta monetaria permanece fija y la utilidad alcanzada cambia.

Propiedades:

- *Continuidad y diferenciabilidad*: son continuas por el teorema del máximo, y diferenciables si u es C^2 con curvatura estrictamente positiva (hessiano orlado no singular), lo que permite la estática comparativa.
- *Transformaciones*: son homogéneas de grado cero en precios y renta, sin ilusión monetaria, e invariantes ante transformaciones monótonas de u. En símbolos, $d_i(k\mathbf{p}, kM) = d_i(\mathbf{p}, M)$, y dependen de las preferencias y no de su representación.
- *Inclinación*: en el precio propio, pendiente indeterminada, porque suman efecto renta y sustitución, negativa salvo en los bienes Giffen; en la renta, crecientes para los normales y decrecientes para los inferiores.
- *Forma*: la teoría no restringe su curvatura; la restricción de segundo orden recae sobre la matriz de Slutsky, simétrica y semidefinida negativa.
- *Otras*: agotan la renta, $\sum_i p_i d_i(\mathbf{p}, M) = M$, la ley de Walras; se obtienen de la FIU por la identidad de Roy; coinciden con las hicksianas evaluadas en la utilidad máxima, $d_i(\mathbf{p}, M) = h_i(\mathbf{p}, v(\mathbf{p}, M))$; de ellas pueden recuperarse la función de gasto y las preferencias si cumplen las condiciones del teorema de la integrabilidad; y son **observables**, estimables con datos de precios, rentas y cantidades.

1.3. Función indirecta de utilidad

Definición: la **función indirecta de utilidad**, FIU, es la utilidad máxima alcanzable para cada par de precios y renta. Se obtiene sustituyendo las demandas marshallianas en la función objetivo:

$$v(\mathbf{p}, M) = u(\mathbf{d}(\mathbf{p}, M)) = \max_{\mathbf{x}} \; u(\mathbf{x}) \quad \text{s.a.: } \mathbf{p}\cdot\mathbf{x} = M$$

Propiedades:

- *Continuidad*: es continua y diferenciable cuando lo son las demandas, lo que permite aplicar el teorema de la envolvente.
- *Transformaciones*: es homogénea de grado cero en precios y renta. Como la utilidad es ordinal, toda transformación monótona creciente de v es otra FIU, lo que permite una representación monetaria. En símbolos, $v(k\mathbf{p}, kM) = v(\mathbf{p}, M)$.
- *Inclinación*: estrictamente creciente en la renta, $\partial v/\partial M = \lambda > 0$, y no creciente en los precios, $\partial v/\partial p_i = -\lambda x_i \leq 0$. Estas derivadas dan, por el teorema de la envolvente, la identidad de Roy.
- *Forma*: es cuasiconvexa en precios: un vector de precios promedio de dos situaciones no es mejor que al menos una de ellas.

Identidad de Roy (1942): el cociente de sus derivadas, cambiado de signo, recupera la demanda marshalliana:

$$d_i(\mathbf{p}, M) = -\frac{\partial v/\partial p_i}{\partial v/\partial M}$$

Dualidad: evaluada en el gasto mínimo, devuelve la utilidad exigida, $v(\mathbf{p}, e(\mathbf{p}, u_0)) = u_0$, de modo que invirtiéndola respecto a la renta se obtiene la función de gasto. No es observable, porque la utilidad no lo es. En símbolos, $v(\mathbf{p}, M) = e^{-1}(\mathbf{p}, M)$.

En suma, el primal da unas demandas observables y homogéneas de grado cero, con pendiente indeterminada en el precio propio, y una FIU no observable cuya inversión abre el problema dual, que intercambia los papeles de la renta y la utilidad.

2. Análisis de dualidad

El problema dual se desarrolla con la misma estructura que el primal, problema, demanda y función de valor, y después se comparan ambos.

2.1. El problema dual: minimizar el gasto

Idea: el dual invierte los papeles de la función objetivo y de la restricción. Dados los precios y un nivel de utilidad exigido, el consumidor busca la cesta que lo alcanza con el mínimo gasto. La renta pasa a ser el resultado y la utilidad, el dato. Su primer producto es la misma cesta óptima que la del primal. Al generalizar para cualesquiera precios y utilidad, se obtienen además la demanda hicksiana y la función de gasto.

Planteamiento: suponiendo solución interior:

$$\min_{\mathbf{x}} \; G = \mathbf{p}\cdot\mathbf{x} \quad \text{s.a.:} \begin{cases} u(\mathbf{x}) \geq u_0 \\ x_i \geq 0 \; \forall i \end{cases}$$

- En el lagrangiano, $\mathcal{L} = \mathbf{p}\cdot\mathbf{x} + \mu\,(u_0 - u(\mathbf{x}))$, el multiplicador $\mu = 1/\lambda = \partial G^*/\partial u$ es el **coste marginal de la utilidad**, el gasto adicional para una unidad más de utilidad: el precio sombra de la utilidad, simétrico del de la renta en el primal. En el caso Cobb-Douglas, $\mu = 2p_1^{1/2}p_2^{1/2} = 1/\lambda$.

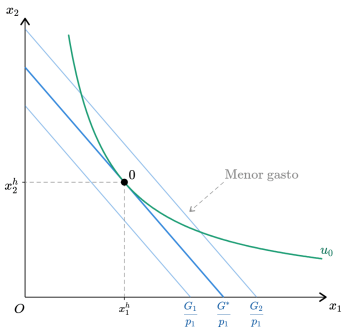
CPO: el consumidor alcanza exactamente la utilidad exigida, sin exceso, e iguala el precio de cada bien a su utilidad marginal valorada al coste marginal de la utilidad, $u_i' = p_i/\mu$. Resulta la misma tangencia que en el primal, $\text{RMS}_{21} = p_1/p_2$. Formalmente:

$$u(\mathbf{x}^*) = u_0, \qquad u_i' = \frac{p_i}{\mu} \;\Rightarrow\; \text{RMS}_{ji} = \frac{p_i}{p_j}$$

CSO: la tangencia es un mínimo si la utilidad es cuasicóncava, la misma condición que en el primal, porque ahora el contorno superior de u actúa de restricción. El conjunto de cestas que alcanzan u_0 es entonces convexo.

Solución: las demandas hicksianas y, como función de valor óptimo, la función de gasto.

Desarrollo gráfico: en el plano (x_1, x_2), la curva de indiferencia objetivo, u_0, es la restricción y permanece fija. Las rectas isogasto tienen pendiente $-p_1/p_2$ y el gasto disminuye al acercarse al origen. El equilibrio, en el punto 0, es la tangencia entre u_0 y la isogasto más baja compatible con ella, G^*. Una isogasto menor no alcanza u_0 y una mayor la alcanza con más gasto. Cada isogasto corta el eje horizontal en G/p_1, y en 0 se elige la cesta (x_1^h, x_2^h).



El problema dual: tangencia de la curva de indiferencia objetivo, u_0, con la isogasto más baja compatible con ella.

2.2. Demanda hicksiana

Definición: resolviendo las CPO y la restricción de utilidad para cualesquiera precios y utilidad se obtienen las **demandas hicksianas**, $x_i^h = h_i(\mathbf{p}, u_0)$ (la cesta que alcanza una utilidad dada con el mínimo

gasto). Son **compensadas**: ante un cambio de precios el gasto no permanece fijo, sino que se reajusta para mantener la utilidad. Mantienen así constante la renta real en el sentido de Hicks mediante una compensación hipotética de renta.

Propiedades:

- *Continuidad y diferenciabilidad*: son continuas por el teorema del máximo y diferenciables bajo los supuestos habituales, lo que permite la estática comparativa.
- *Transformaciones*: son homogéneas de grado cero en precios: si todos suben en la misma proporción, el gasto sube igual y la cesta no cambia. En símbolos, $h_i(k\mathbf{p}, u_0) = h_i(\mathbf{p}, u_0)$.
- *Inclinación*: solo recogen el efecto sustitución, por lo que **nunca crecen con el precio propio**, la ley de la demanda compensada, y no admiten bienes Giffen. Para un bien normal son más inclinadas, menos elásticas, que las marshallianas, porque no incorporan el efecto renta.
- *Forma*: la teoría no restringe su curvatura. La restricción de segundo orden recae sobre la función de gasto, cóncava en precios, y por eso las pendientes de las hicksianas forman una matriz simétrica y semidefinida negativa. Así, la concavidad de e en precios es la pendiente negativa de la demanda compensada, $\partial^2 e/\partial p_i^2 = \partial h_i/\partial p_i \leq 0$.
- *Otras*: alcanzan exactamente u_0; son el gradiente de la función de gasto, por el lema de Shephard; coinciden con las marshallianas evaluadas en el gasto mínimo; y no son observables, porque la utilidad no lo es, de modo que solo se recuperan indirectamente por dualidad. En símbolos, $u(\mathbf{h}(\mathbf{p}, u_0)) = u_0$ y $h_i = \partial e/\partial p_i$.

2.3. Función de gasto

Definición: sustituyendo las demandas hicksianas en el objetivo se obtiene la **función de gasto**, FG, el mínimo gasto necesario para cada vector de precios y nivel de utilidad:

$$e(\mathbf{p}, u_0) = \mathbf{p} \cdot \mathbf{h}(\mathbf{p}, u_0) = \min_{\mathbf{x}} \ \mathbf{p} \cdot \mathbf{x} \quad \text{s.a.: } u(\mathbf{x}) = u_0$$

- Da el gasto mínimo para cualesquiera precios y utilidad sin repetir el problema, y sus propiedades son análogas a las de la función de costes de la empresa. En el óptimo, $G^* = e(\mathbf{p}, u_0)$.

Propiedades:

- *Continuidad*: es continua en precios y utilidad y diferenciable en precios, lo que permite aplicar el teorema de la envolvente.
- *Transformaciones*: es **homogénea de grado uno** en precios: un cambio equiproporcional deja la cesta igual y multiplica el gasto en la misma proporción. En símbolos, $e(k\mathbf{p}, u_0) = k \cdot e(\mathbf{p}, u_0)$.
- *Inclinación*: estrictamente creciente en la utilidad, $\partial e/\partial u = \mu > 0$, y no decreciente en precios, $\partial e/\partial p_i = h_i \geq 0$.
- *Forma*: es **cóncava en precios**, es decir, al subir un precio, el gasto mínimo aumenta menos que si la cesta se mantuviera fija, porque se puede sustituir. Por eso la matriz de Slutsky es semidefinida negativa. El consumo compensado del bien encarecido no aumenta, pero puede caer también el de sus complementarios netos: si sube la electricidad, se sustituye por gas, y puede caer la compra de electrodomésticos eléctricos.

Lema de Shephard (1953): la derivada de la función de gasto respecto a un precio es la demanda hicksiana:

$$\frac{\partial e(\mathbf{p}, u_0)}{\partial p_i} = h_i(\mathbf{p}, u_0)$$

- Al subir marginalmente un precio, el gasto mínimo aumenta en la cantidad que se consumía, porque el reajuste de la cesta es de segundo orden. El lema es el análogo de la identidad de Roy, pero más simple, porque el objetivo ya está en dinero. Con $u = x_1^{1/2} x_2^{1/2}$, la tangencia da $p_1 x_1 = p_2 x_2 = G^*/2$ y, sustituyendo en la restricción de utilidad, se obtiene la función de gasto, inversa de la FIU respecto a la renta, y de ella la hicksiana:

$$e(\mathbf{p}, u_0) = 2u_0\, p_1^{1/2} p_2^{1/2} \quad \rightarrow \quad x_1^h = \frac{\partial e}{\partial p_1} = u_0 \left(\frac{p_2}{p_1}\right)^{1/2}$$

Dualidad e integrabilidad: evaluada en la utilidad máxima que permite una renta, devuelve esa renta, $e(\mathbf{p}, v(\mathbf{p}, M)) = M$, de modo que es la inversa de la FIU, $e(\mathbf{p}, u_0) = v^{-1}(\mathbf{p}, u_0)$. Recíprocamente, toda función continua, homogénea de grado uno, no decreciente y cóncava en precios, y creciente en la utilidad, es la función de gasto de unas preferencias de buen comportamiento. No es observable, pero al medirse en dinero es la base de las medidas del bienestar.

2.4. Relaciones de dualidad y pendientes

Entre la maximización de la utilidad, PMU, y la minimización del gasto, PMG, que tienen los mismos componentes, hay tres relaciones de dualidad, a las que se añade la comparación de las pendientes de las demandas.

2.4.1. Teorema básico de la dualidad

Las cestas óptimas de ambos problemas son idénticas con el debido intercambio de información.

- *Del primal al dual*: si $\mathbf{x}^*$ resuelve el primal con renta M_0, resuelve el dual cuando la utilidad exigida es la obtenida, $u_0 = v(\mathbf{p}, M_0)$. Entonces $G^* = M_0$.
- *Del dual al primal*: si $\mathbf{x}^*$ resuelve el dual con utilidad u_0, resuelve el primal cuando la renta es el gasto mínimo, $M_0 = e(\mathbf{p}, u_0)$. Entonces $u^* = u_0$.

El resultado se refiere a las cestas, porque las demandas marshallianas y hicksianas dependen de variables distintas.

2.4.2. Dualidad entre funciones de demanda

Hallada la demanda de un problema, la del otro se obtiene con la función de valor de este último: la marshalliana, de la hicksiana y la FIU; la hicksiana, de la marshalliana y la función de gasto. Esta relación extiende el teorema básico a las funciones de demanda:

$$d_i(\mathbf{p}, M) = h_i(\mathbf{p}, v(\mathbf{p}, M)), \qquad h_i(\mathbf{p}, u_0) = d_i(\mathbf{p}, e(\mathbf{p}, u_0))$$

2.4.3. Dualidad entre funciones de valor

La FIU es la inversa de la función de gasto respecto a su segundo argumento, y viceversa, de modo que basta resolver un problema:

$$e(\mathbf{p}, v(\mathbf{p}, M)) = M, \qquad v(\mathbf{p}, e(\mathbf{p}, u_0)) = u_0$$

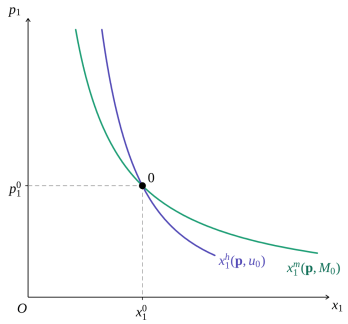
2.4.4. Pendientes de las curvas de demanda

Idea: la marshalliana recoge el efecto total y la hicksiana solo el de sustitución, de modo que su diferencia es el **efecto renta**. En elasticidades:

$$\varepsilon_{ii}^m = \varepsilon_{ii}^c - s_i\, \varepsilon_{iM}$$

- Para un bien normal, la marshalliana es más elástica que la hicksiana; para un inferior no Giffen, menos; para un bien frontera, coinciden. Aquí ε_{ii}^c es la elasticidad compensada, s_i la participación del bien en el gasto y ε_{iM} su elasticidad-renta.

Desarrollo gráfico: en el plano (x_1, p_1), la marshalliana añade, ante una bajada del precio, el salto a una curva de indiferencia superior. La hicksiana mantiene fija la curva de indiferencia y solo recoge el movimiento a lo largo de ella. Ambas se cortan en el punto 0, al precio inicial. Allí la utilidad exigida en el dual, u_0, es la máxima que permite la renta M_0 en el primal, y la recta de balance coincide con la isogasto.



Demanda marshalliana y hicksiana: la compensada recoge solo el efecto sustitución y es menos elástica para un bien normal.

En suma, el dual fija la utilidad y minimiza el gasto. Sus demandas hicksianas, no observables, nunca crecen con el precio propio, y la función de gasto, cóncava y homogénea de grado uno en precios, las reúne. Las relaciones de dualidad permiten resolver un solo problema y obtener todas las funciones del otro.

3. Extensiones y aplicaciones

Conocidas las herramientas, la dualidad tiene implicaciones teóricas, fundamenta los sistemas de demanda empíricos y proporciona las medidas de bienestar.

3.1. Ecuación de Slutsky e integrabilidad

La dualidad tiene dos implicaciones en direcciones opuestas: hacia delante, de las preferencias a la conducta, la ecuación de Slutsky; hacia atrás, de la conducta a las preferencias, la integrabilidad.

3.1.1. Ecuación de Slutsky: derivación formal

Derivación: se parte de la dualidad entre demandas, $h_i(\mathbf{p}, u_0) = d_i(\mathbf{p}, e(\mathbf{p}, u_0))$. Derivando respecto a p_j por la regla de la cadena, sustituyendo $\partial e/\partial p_j = x_j$ por el lema de Shephard y teniendo en cuenta que en el óptimo $M = G^*$:

$$\frac{\partial x_i^m}{\partial p_j} = \frac{\partial h_i}{\partial p_j} - \frac{\partial x_i^m}{\partial M} \cdot x_j$$

- El primer término es el **efecto sustitución** y el segundo, con su signo, el **efecto renta**. Su suma es el efecto total.

– En elasticidades, $\varepsilon_{ij}^m = \varepsilon_{ij}^c - s_j \cdot \varepsilon_{iM}$, con s_j la participación del bien j en el gasto. Se obtiene multiplicando por p_j/x_i, y para $i = j$ es la comparación de pendientes de las dos demandas.

Consecuencias: el efecto sustitución queda definido formalmente por las derivadas de la demanda hicksiana. El propio es **no positivo** en todo caso, por la concavidad de la función de gasto. El cruzado tiene signo indeterminado y clasifica los bienes en sustitutivos netos, si es positivo, complementarios netos, si es negativo, e independientes, si es nulo. Las clasificaciones por efectos totales, como ordinarios, Giffen o sustitutivos brutos, proceden del efecto total. En símbolos, $\partial h_i/\partial p_i \leq 0$, y el bien j es sustitutivo neto de i si $\partial h_i/\partial p_j > 0$.

3.1.2. Matriz de Slutsky

Definición: con cambios en varios precios, los efectos sustitución se ordenan en la **matriz de Slutsky**, la hessiana de la función de gasto. Por Shephard, $\partial e/\partial p_i = h_i$, y derivando de nuevo:

$$S_{ij} = \frac{\partial^2 e(\mathbf{p}, u_0)}{\partial p_i \, \partial p_j} = \frac{\partial h_i(\mathbf{p}, u_0)}{\partial p_j}$$

– Los propios están en la diagonal y los cruzados fuera de ella, y el efecto de cualquier vector de cambios de precios sobre las demandas compensadas es $d\mathbf{h} = \mathbf{S}\, d\mathbf{p}$. Para dos bienes:

$$\mathbf{S}_2 = \begin{pmatrix} \partial h_1/\partial p_1 & \partial h_1/\partial p_2 \\ \partial h_2/\partial p_1 & \partial h_2/\partial p_2 \end{pmatrix}$$

– Por la ecuación de Slutsky, cada elemento se calcula con magnitudes observables, $S_{ij} = \partial x_i^m/\partial p_j + (\partial x_i^m/\partial M)\, x_j$, lo que hace la matriz contrastable.

Propiedades:

– *Simétrica*: $\partial h_i/\partial p_j = \partial h_j/\partial p_i$, lo que en general no ocurre con las demandas ordinarias, salvo, por ejemplo, con preferencias homotéticas. Las relaciones netas son simétricas, pero las brutas no, porque el efecto renta introduce asimetría: Coca-Cola puede ser sustitutivo bruto de Pepsi y no al revés. Además, no todos los bienes pueden ser complementarios netos a la vez.
– **Semidefinida negativa**, por la concavidad de la función de gasto: los efectos propios compensados son no positivos, y las hicksianas cumplen la ley de la demanda compensada.

3.1.3. Integrabilidad (Antonelli, 1886)

Problema: solo se observan las demandas marshallianas. La cuestión es si un sistema completo de demandas estimado es compatible con unas preferencias de buen comportamiento y si estas pueden reconstruirse. Antonelli lo planteó y resolvió en 1886, y Hurwicz y Uzawa (1971) le dieron su formulación moderna.

Teorema: la solución pasa por la función de gasto, en tres pasos. La función de gasto contiene toda la información sobre las preferencias, y sus derivadas son observables a través de la ecuación de Slutsky.

1. De las demandas observadas se calcula la matriz de Slutsky, sin conocer las hicksianas.
2. Como $\partial e/\partial p_i = d_i(\mathbf{p}, e(\mathbf{p}, u_0))$, la función de gasto se obtiene integrando ese sistema, lo que exige tres condiciones: **ley de Walras**, que con la simetría implica homogeneidad; **simetría**, la condición de integrabilidad propiamente dicha (teorema de Frobenius); y **semidefinición negativa**, que garantiza una función de gasto cóncava. Esa función procede, así, de una minimización del gasto.
3. De la función de gasto se recupera la FIU por inversión y, con ella, las preferencias.

Las tres condiciones son **necesarias y suficientes**: toda demanda que procede de maximizar una utilidad de buen comportamiento las cumple, y todo sistema completo que las cumple procede de alguna utilidad. Las condiciones de homogeneidad, Engel y Cournot, consecuencias de la ley de Walras y la homogeneidad, son solo necesarias, porque un sistema puede cumplirlas sin proceder de ninguna utilidad. La dualidad añade la simetría y la negatividad, y con ellas el conjunto basta.

Consecuencias: permite contrastar la teoría sin especificar ninguna utilidad, comprobando o imponiendo las condiciones en las demandas estimadas, y garantiza que de ellas se recupera la función de gasto, la métrica del bienestar, lo que cierra el círculo del análisis dual.

3.2. Sistemas de demanda empíricos

Definición: un **sistema completo de demanda** especifica conjuntamente las n ecuaciones de demanda, o de participaciones en el gasto, $s_i = p_i x_i/M$, y las estima simultáneamente. Su fundamento es la dualidad: se parte de una función de gasto o una FIU y se obtienen las demandas por Shephard o Roy, de modo que las restricciones teóricas se convierten en restricciones sobre los parámetros, que pueden imponerse o contrastarse. Agregación, homogeneidad y simetría se traducen en restricciones lineales, y la negatividad, en desigualdades.

3.2.1. Sistema lineal de gasto (Stone, 1954)

Origen: Klein y Rubin (1948) propusieron la forma funcional, Geary (1950) derivó la utilidad que la genera y Stone (1954) lo estimó con datos británicos como sistema completo, el primero no homotético estimado.

Planteamiento: parte de la utilidad Stone-Geary, $u = \sum_i \gamma_i \ln(x_i - \bar{x}_i)$, con $\sum_i \gamma_i = 1$ y consumos de *subsistencia* $\bar{x}_i$ cubiertos con independencia de precios y renta. Su función de gasto es de la forma polar de Gorman, $e(\mathbf{p}, u) = a(\mathbf{p}) + b(\mathbf{p})\, u$, donde $a(\mathbf{p}) = \sum_k p_k \bar{x}_k$ es el coste de la subsistencia, y el gasto en cada bien es lineal en la renta, de ahí el nombre:

$$p_i x_i = p_i \bar{x}_i + \gamma_i \left(M - \sum_k p_k \bar{x}_k \right)$$

– Cada bien recibe una proporción γ_i de la renta supernumeraria. Las restricciones se cumplen por construcción, tiene pocos parámetros, $2n - 1$, y sus curvas de Engel son rectas que no pasan por el origen, lo que, si los parámetros son comunes a los consumidores, permite el consumidor representativo de Gorman. Con el índice de utilidad $\prod_k (x_k - \bar{x}_k)^{\gamma_k}$, $b(\mathbf{p}) = \prod_k (p_k/\gamma_k)^{\gamma_k}$, y la elasticidad-precio propia es $\varepsilon_{ii}^m = -1 + (1 - \gamma_i)\bar{x}_i/x_i$, mayor que -1 si $\bar{x}_i > 0$.

Conclusiones: es parsimonioso y coherente con la teoría, pero rígido: todos los bienes son normales, con $\varepsilon_{iM} = \gamma_i M/p_i x_i > 0$, sustitutivos netos y complementarios brutos, y la demanda propia es inelástica si $\bar{x}_i > 0$. Al no admitir bienes inferiores ni complementarios netos, impone la teoría en lugar de contrastarla.

3.2.2. Modelo de Rotterdam (Barten, 1964; Theil, 1965)

Idea: en vez de partir de una utilidad, especifica el sistema en diferenciales logarítmicas y aplica la ecuación de Slutsky, con un efecto renta real y efectos precio compensados:

$$s_i \, d\ln x_i = b_i \, d\ln Q + \sum_j c_{ij} \, d\ln p_j$$

– $d\ln Q = \sum_k s_k \, d\ln x_k$ es el índice de volumen de Divisia, la variación de la renta real, y $b_i = p_i \, \partial x_i/\partial M$, la participación marginal en el gasto, positiva si el bien es normal y negativa si es inferior. Los coeficientes de Slutsky, $c_{ij} = (p_i p_j/M)\, \partial h_i/\partial p_j$, constantes en el modelo, cumplen $c_{ii} < 0$, y los cruzados son positivos entre sustitutivos netos y negativos entre complementarios netos.
– La agregación, $\sum_i b_i = 1$ y $\sum_i c_{ij} = 0$, la homogeneidad, $\sum_j c_{ij} = 0$, y la simetría, $c_{ij} = c_{ji}$, son restricciones lineales que pueden imponerse o contrastarse con tests convencionales. La negatividad, que exige una matriz $[c_{ij}]$ semidefinida negativa, no es lineal.

Conclusiones: fue el primero en contrastar sistemáticamente homogeneidad y simetría, con rechazos frecuentes de la homogeneidad, pero sus coeficientes constantes solo son compatibles globalmente con una Cobb-Douglas, así que es una aproximación local.

3.2.3. AIDS (Deaton y Muellbauer, 1980a)

Idea: el sistema casi ideal reúne las ventajas de los anteriores: procede de una función de gasto, como el SLG, con restricciones lineales contrastables, como Rotterdam; aproxima en primer orden cualquier sistema y agrega exactamente sobre consumidores. Parte de una función de gasto **PIGLOG**, logarítmicamente lineal en la utilidad, $\ln e = (1 - u)\ln a(\mathbf{p}) + u \ln b(\mathbf{p})$, con $\ln a(\mathbf{p})$ translogarítmica, y, por el lema de Shephard, las participaciones son:

$$s_i = \alpha_i + \sum_j \gamma_{ij} \ln p_j + \beta_i \ln \left(\frac{M}{P} \right)$$

– Las restricciones de agregación, $\sum_i \alpha_i = 1$ y $\sum_i \gamma_{ij} = \sum_i \beta_i = 0$, de homogeneidad, $\sum_j \gamma_{ij} = 0$, y de simetría, $\gamma_{ij} = \gamma_{ji}$, son lineales, y la negatividad se comprueba en la muestra. La elasticidad-renta es $\varepsilon_{iM} = 1 + \beta_i/s_i$, de modo que $\beta_i > 0$ identifica los bienes de lujo y $\beta_i < 0$, los de primera necesidad. Las curvas de Engel en participaciones son lineales en el logaritmo de la renta, la forma de Working-Leser, lo que preserva la agregación exacta de Muellbauer con un índice de distribución de la renta.
– Con el índice de Stone, $\ln P^* = \sum_k s_k \ln p_k$, el sistema es lineal en los parámetros, LA-AIDS, y su generalización cuadrática, QUAIDS (Banks et al., 1997), admite curvas de Engel en U invertida. El índice de precios exacto, que el de Stone aproxima, es:

$$\ln P = \alpha_0 + \sum_k \alpha_k \ln p_k + \tfrac{1}{2} \sum_k \sum_j \gamma_{kj} \ln p_k \ln p_j$$

Conclusiones: satisface los axiomas, agrega, contrasta la teoría y se ajusta a los presupuestos familiares, de ahí su uso generalizado. Pero tiene muchos parámetros, y el índice de Stone sesga las elasticidades.

3.2.4. Aplicaciones empíricas

Los sistemas de demanda llevan la teoría del consumidor a los datos en cuatro usos.

1. *Elasticidades*: precio, cruzadas y renta por grupos de bienes, con encuestas de presupuestos familiares o series agregadas, que alimentan los modelos de equilibrio general aplicado y las previsiones.
2. *Contraste de la teoría*: los tests de homogeneidad y simetría. El rechazo frecuente de la homogeneidad en series temporales se atribuye a la agregación, la dinámica omitida y los errores de medida, más que a la irracionalidad.
3. *Evaluación de políticas*: con la función de gasto estimada se calculan la VC y la VE de una reforma de la imposición indirecta o de una subida de precios regulados, y su distribución por hogares.
4. *Coste de la vida*: el sesgo de sustitución del IPC se cuantifica comparando Laspeyres con el índice verdadero implícito en el sistema.

3.3. Medidas de cambio de bienestar

La pregunta normativa es cómo medir el cambio de bienestar de un consumidor cuando los precios pasan de $\mathbf{p}^0$ a $\mathbf{p}^1$, con renta M_0 y utilidades u_0 y u_1. El cambio puede deberse a la inflación, al poder de mercado o a impuestos.

3.3.1. Utilidad métrica monetaria (Samuelson, 1974)

Definición: como la utilidad es ordinal, la función de gasto permite elegir la representación que la mide en dinero. La **función de utilidad métrica monetaria** es el gasto mínimo, a unos precios de referencia $\mathbf{q}$, para alcanzar la utilidad de una cesta, $m(\mathbf{q}, \mathbf{x}) = e(\mathbf{q}, u(\mathbf{x}))$. Su versión indirecta, $e(\mathbf{q}, v(\mathbf{p}, M))$, mide en dinero, a los precios $\mathbf{q}$, la utilidad máxima que permiten los precios $\mathbf{p}$ y la renta M:

$$\tilde{m}(\mathbf{q}; \mathbf{p}, M) = e(\mathbf{q}, v(\mathbf{p}, M))$$

- Ambas son transformaciones monótonas crecientes de la utilidad, válidas para representar las preferencias, pero en unidades monetarias. Heredan las propiedades de u y v: la directa es creciente y cuasicóncava en las cantidades; la indirecta, homogénea de grado cero, creciente en la renta y cuasiconvexa en precios. Su continuidad y diferenciabilidad son las de la función de gasto y las de u o de v.
- Con precios de referencia iguales a los vigentes, la indirecta es la renta. No son observables, porque exigen la función de gasto, pero son la base de las medidas exactas del bienestar. En símbolos, $\tilde{m}(\mathbf{p}; \mathbf{p}, M) = e(\mathbf{p}, v(\mathbf{p}, M)) = M$.

3.3.2. VC, VE y VEC

Variación compensatoria (Hicks, 1939b): es la diferencia entre la renta y la que, a los precios finales, permitiría alcanzar la utilidad inicial. Responde a ¿cuánto dinero hay que dar o quitar al consumidor después del cambio para restaurar su bienestar inicial?

$$\mathrm{VC} = M_0 - e(\mathbf{p}^1, u_0) = \int_{p_i^1}^{p_i^0} h_i(\mathbf{p}, u_0)\, dp_i$$

- Es la diferencia de utilidad métrica monetaria con referencia $\mathbf{p}^1$. Al valorar el cambio a la utilidad de partida, mide lo que habría que pagar o cobrar a los afectados para dejarlos como estaban, por lo que es la medida natural para compensar a quienes soportan un cambio. En símbolos, $\mathrm{VC} = e(\mathbf{p}^1, u_1) - e(\mathbf{p}^1, u_0)$, porque a los precios finales $M_0 = e(\mathbf{p}^1, u_1)$.

Variación equivalente: es la diferencia entre la renta que, a los precios iniciales, permitiría alcanzar la utilidad final y la renta. Responde a ¿cuánto dinero hay que dar o quitar antes del cambio para alterar el bienestar en la misma cuantía?

$$\mathrm{VE} = e(\mathbf{p}^0, u_1) - M_0 = \int_{p_i^1}^{p_i^0} h_i(\mathbf{p}, u_1)\, dp_i$$

- Es la diferencia de utilidad métrica monetaria con referencia $\mathbf{p}^0$. Al usar unos precios fijos, los iniciales, cualesquiera que sean las alternativas, es la adecuada para comparar y ordenar de forma coherente varias políticas. En símbolos, $\mathrm{VE} = e(\mathbf{p}^0, u_1) - e(\mathbf{p}^0, u_0)$, porque a los precios iniciales $M_0 = e(\mathbf{p}^0, u_0)$.

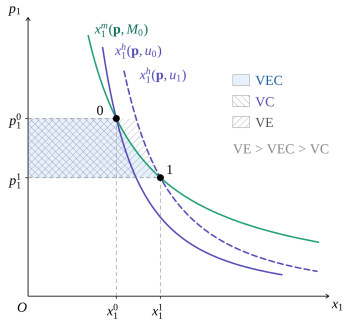
Variación del excedente del consumidor (Dupuit, 1844; Marshall, 1890): a diferencia de las anteriores, usa la demanda ordinaria. El excedente es la diferencia entre la disposición máxima a pagar y el precio pagado, y su variación es el área bajo la demanda marshalliana entre ambos precios, $\int_{p_i^1}^{p_i^0} d_i(p_i, M_0)\, dp_i$. Gráficamente, es el área entre la curva de demanda ordinaria y las cotas de precios.

Comparación: las tres tienen el signo del cambio de bienestar, pero difieren según el efecto renta. Para un bien normal:

$$\begin{aligned} &\text{bajada de } p_i: && \mathrm{VE} > \mathrm{VEC} > \mathrm{VC} \\ &\text{subida de } p_i: && |\mathrm{VC}| > |\mathrm{VEC}| > |\mathrm{VE}| \end{aligned}$$

- La razón es que la hicksiana a la utilidad final está a la derecha de la hicksiana a la inicial y la marshalliana, entre ambas. Con un bien inferior ocurre lo contrario en cada caso y, sin efecto renta, las tres coinciden.

Desarrollo gráfico: en el plano (x_1, p_1), una bajada de p_1^0 a p_1^1 lleva al consumidor de 0 a 1. Cada medida es el área a la izquierda de una curva entre ambos precios: la marshalliana, que pasa por 0 y 1, da la VEC; la hicksiana a u_0, que pasa por 0, la VC; y la hicksiana a u_1, que pasa por 1, la VE. La hicksiana a u_1 queda a la derecha de la de u_0 porque, para un bien normal, una utilidad mayor exige más cantidad a cada precio.



Medidas monetarias del bienestar ante una bajada de precio: VC, VEC y VE como áreas a la izquierda de cada curva de demanda.

Conclusiones: VC y VE, diferencias de utilidad métrica monetaria, jerarquizan rigurosamente las situaciones, pero exigen conocer la función de gasto, que se estima con un sistema de demanda. La VEC solo es exacta sin efecto renta, con preferencias cuasilineales o bienes frontera, y por ello se usa como aproximación en equilibrio parcial.

3.3.3. Números índice

Cuando no hay demandas hicksianas, pero sí cestas y precios, el bienestar se aproxima con **números índice**.

- *Índice de variación compensatoria*: es el factor de actualización de la renta que mantiene constante la utilidad inicial, $e(\mathbf{p}^1, u_0)/e(\mathbf{p}^0, u_0) \le P_L$. El índice de **Laspeyres** (1871) valora la cesta base a los nuevos precios. Como esa cesta sigue siendo asequible con la renta $P_L \cdot M_0$, pero el consumidor puede sustituir, sobrestima la subida del coste de la vida. Con $u = x_1 x_2^2$, $\mathbf{p}^0 = (10, 20)$, $M_0 = 300$ y $\mathbf{p}^1 = (15, 40)$, el índice verdadero es $545{,}1/300 \approx 1{,}817$ y el de Laspeyres, $550/300 \approx 1{,}833$.
- *Índice de variación equivalente*: mantiene constante la utilidad final, $e(\mathbf{p}^1, u_1)/e(\mathbf{p}^0, u_1) \ge P_P$. El de **Paasche** (1874) valora la cesta corriente a los precios de ambos periodos y, por el argumento simétrico, la subestima. Los dos índices son:

$$P_L = \frac{\sum_i p_i^1 x_i^0}{\sum_i p_i^0 x_i^0}, \qquad P_P = \frac{\sum_i p_i^1 x_i^1}{\sum_i p_i^0 x_i^1}$$

Índice de Konüs (1924): el índice verdadero del coste de la vida, $P_K(u) = e(\mathbf{p}^1, u)/e(\mathbf{p}^0, u)$, es el de variación compensatoria con $u = u_0$ y el de variación equivalente con $u = u_1$. Con preferencias homotéticas no depende del nivel de utilidad y queda entre Paasche y Laspeyres, y, en general, lo está para una utilidad intermedia entre u_0 y u_1. El índice ideal de Fisher (1922), su media geométrica, es exacto con preferencias homotéticas cuadráticas. En símbolos, $P_P \le P_K \le P_L$ con preferencias homotéticas, y $P_F = \sqrt{P_L\, P_P}$.

Alcance: el análisis vale para un individuo aislado. Con utilidad ordinal, la utilidad marginal de la renta no es comparable entre agentes, y el bienestar colectivo solo es medible en casos restringidos, como preferencias idénticas, cuasilineales o un agente representativo. El análisis del bienestar social trata de cubrir estas lagunas con criterios de deseabilidad colectiva.

En suma, la dualidad conecta la función de gasto y la demanda hicksiana, no observables, con las demandas marshallianas observables: la ecuación de Slutsky descompone los efectos precio, la integrabilidad recupera las preferencias, los sistemas de demanda convierten la teoría en restricciones estimables y las medidas de bienestar usan la función de gasto como métrica monetaria.

1. Riesgo: teoría de la utilidad esperada

En la **paradoja de San Petersburgo** (Nicolaus Bernoulli, 1713), se tira una moneda hasta que sale cara y el premio se duplica con cada lanzamiento. El valor esperado es infinito, pero nadie paga mucho por jugar. Daniel Bernoulli (1738) la resolvió: a los individuos no les importan los premios, sino su utilidad.

1.1. Conjunto de elección y preferencias

Idea: la **teoría de la utilidad esperada** (TUE), formulada axiomáticamente por von Neumann y Morgenstern (1944) y extendida a probabilidades subjetivas por Savage (1954), obtiene una función de utilidad a partir de axiomas sobre las preferencias entre loterías. Parte de dos premisas: no hay certeza sobre los sucesos, pero se conoce la distribución de probabilidades de los estados de la naturaleza.

Conjunto de elección: el consumidor ordena un conjunto $\mathcal{L}$, no vacío y compacto, de **loterías simples**, distribuciones de probabilidad objetivas sobre un vector fijo de pagos $(x_1, \dots, x_n)$:

$$\mathcal{L} = \left\{ \mathbf{l} = (p_1, p_2, \dots, p_n) \in \mathbb{R}^n_+ \;\middle|\; \sum_{i=1}^{n} p_i = 1 \right\}$$

Por el axioma de reducción de loterías compuestas, la premisa "consecuencialista", una lotería sobre otras loterías equivale a la lotería simple que asigna a cada resultado su probabilidad final, y basta analizar $\mathcal{L}$.

El agente no elige cestas de utilidad cierta, sino alternativas de riesgo, y sobre ellas define la indiferencia, $\sim$, la preferencia estricta, $\succ$, y la preferencia débil, $\succsim$.

1.1.1. Axiomas de orden

Completitud: ante dos loterías cualesquiera, el consumidor siempre puede ordenarlas: $\mathbf{l}_1 \succsim \mathbf{l}_2$, $\mathbf{l}_2 \succsim \mathbf{l}_1$ o ambas, y entonces le son indiferentes.

Reflexividad: toda lotería es comparable consigo misma, $\mathbf{l}_0 \succsim \mathbf{l}_0$.

Transitividad: si $\mathbf{l}_1 \succsim \mathbf{l}_2$ y $\mathbf{l}_2 \succsim \mathbf{l}_3$, entonces $\mathbf{l}_1 \succsim \mathbf{l}_3$ (coherencia entre comparaciones).

1.1.2. Axiomas de regularidad

Monotonía: la preferencia por una lotería aumenta cuando aumenta la probabilidad del suceso más favorable.

Continuidad: pequeños cambios en las probabilidades no alteran la ordenación entre dos loterías, lo que excluye las preferencias lexicográficas.

Independencia: si dos loterías se mezclan con una tercera, la ordenación entre las mezclas no depende de la tercera lotería:

$$\mathbf{l}_1 \succsim \mathbf{l}_2 \leftrightarrow \alpha\mathbf{l}_1 + (1-\alpha)\mathbf{l}_3 \succsim \alpha\mathbf{l}_2 + (1-\alpha)\mathbf{l}_3, \quad \alpha \in (0,1]$$

El axioma de independencia, el más controvertido, es el corazón de la teoría. Como solo uno de los estados llegará a ocurrir, los pagos en estados distintos, los bienes contingentes, no pueden ser sustitutivos ni complementarios en la utilidad, es decir, la utilidad marginal del pago en un estado no depende del pago en otro.

Diferenciabilidad, opcional: evita puntos angulares en la función de utilidad.

1.2. Utilidad esperada y perfiles de riesgo

1.2.1. Función de utilidad esperada

Representación: por el teorema de Debreu (1954), la racionalidad y la continuidad garantizan una utilidad continua que preserva el orden. Pero en riesgo hace falta una forma que recoja el "valor moral" de la lotería (la utilidad de los pagos, no los pagos).

Teorema de la utilidad esperada (von Neumann y Morgenstern, 1944): si a la racionalidad y la continuidad se une la independencia, la ordenación se representa con una **función de utilidad esperada**, donde $u(x_i)$ es la utilidad de Bernoulli de cada pago y p_i su probabilidad objetiva:

$$U(\mathbf{l}) \equiv u^e(\mathbf{l}) = \mathbb{E}[u(\mathbf{l})] = \sum_{i=1}^{n} p_i \cdot u(x_i)$$

Propiedades:

- Es continua en las probabilidades, y la utilidad de Bernoulli se supone dos veces diferenciable en los pagos.
- Es única salvo transformaciones afines positivas, las únicas que conservan la forma de esperanza, por lo que la utilidad esperada es **cardinal**.
- Es creciente en la probabilidad del suceso más favorable y en los pagos, y lineal en las probabilidades. La curvatura de $u(x)$ queda libre y caracteriza la actitud ante el riesgo.

San Petersburgo: con una utilidad de Bernoulli suficientemente cóncava, como la logarítmica que propuso Bernoulli, la esperanza de $u(x)$ es finita aunque la de x no lo sea.

1.2.2. Perfiles de riesgo

El individuo elige la lotería que maximiza su utilidad esperada, y su conducta depende de su **actitud frente al riesgo**: del distinto efecto que la renta ganada y la perdida tienen sobre su utilidad. Se caracteriza con tres claves: (i) la lotería degenerada, $\mathbf{l}_d$, que paga con certeza la esperanza, frente a la lotería $\mathbf{l}$, y $u(\mathbb{E}(\mathbf{l}))$ frente a $\mathbb{E}[u(\mathbf{l})]$; (ii) el signo de $u''(x)$, y (iii) la prima de riesgo.

Prima de riesgo: $\rho = \mathbb{E}(\mathbf{l}) - x_c$, la diferencia entre la esperanza y el **equivalente cierto**, el pago cuya utilidad iguala la utilidad esperada de la lotería. Esta diferencia es el sobreprecio máximo que el agente pagaría, sobre una prima actuarialmente justa, por eliminar el riesgo. El equivalente cierto resuelve $u(x_c) = \mathbb{E}[u(\mathbf{l})]$, de modo que $x_c < \mathbb{E}(\mathbf{l})$ con aversión y $x_c > \mathbb{E}(\mathbf{l})$ con propensión.

Averso al riesgo: prefiere el valor esperado cierto a la lotería y paga por evitar el riesgo:

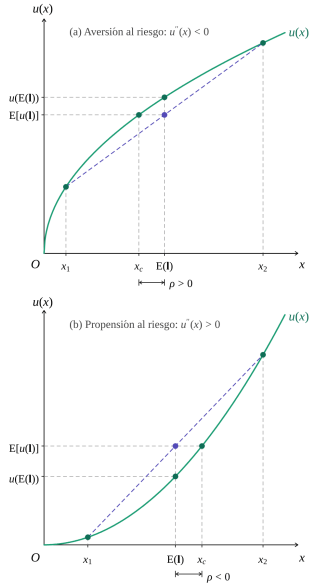
$$\mathbf{l}_d \succ \mathbf{l} \leftrightarrow u(\mathbb{E}(\mathbf{l})) > \mathbb{E}[u(\mathbf{l})], \qquad u''(x) < 0, \qquad \rho > 0$$

Su utilidad de Bernoulli es estrictamente cóncava, de modo que, con utilidad marginal decreciente, ante variaciones iguales al alza y a la baja, la renta perdida resta más utilidad de la que suma la ganada.

Neutral al riesgo: le es indiferente la lotería degenerada o cualquier otra de igual esperanza, con $u''(x) = 0$ y $\rho = 0$. Su utilidad es una recta, y ni paga por evitar el riesgo ni exige compensación por asumirlo.

Propenso al riesgo: prefiere la lotería, con $u''(x) > 0$ y $\rho < 0$. Su utilidad es estrictamente convexa, el equivalente cierto supera la esperanza y exige una compensación por renunciar al riesgo.

Desarrollo gráfico: la altura de la cuerda entre x_1 y x_2 sobre $\mathbb{E}(\mathbf{l})$ es la utilidad esperada, por debajo de $u(\mathbb{E}(\mathbf{l}))$ con aversión y por encima con propensión. La distancia entre x_c y $\mathbb{E}(\mathbf{l})$ es la prima de riesgo.



Perfiles de riesgo: utilidad cóncava con aversión (a) y convexa con propensión (b).

1.2.3. Coeficientes de Arrow-Pratt

Motivo: $u''(x)$ da el signo del perfil, pero no es invariante ante transformaciones afines y no permite cuantificar ni ordenar las actitudes. Los coeficientes toman como referencia el perfil averso y suponen preferencias monótonas.

Aversión absoluta (Pratt, 1964; Arrow, 1965): el **CAA** es la curvatura relativa de la utilidad de Bernoulli. Dividir por $u'(x)$ lo hace invariante, y el signo menos, creciente con la aversión:

$$\text{CAA} = -\frac{u''(x)}{u'(x)}$$

- Su signo caracteriza al agente, y su valor ordena la aversión entre individuos y entre niveles de renta. La evidencia sugiere que es decreciente (Guiso y Paiella, 2008): con más riqueza se asume más fácilmente un riesgo de cuantía absoluta dada.
- Aproxima, con un desarrollo de Taylor de segundo orden, la prima de riesgo, proporcional a la aversión por la varianza de los pagos: $\rho \approx \frac{1}{2} \cdot \text{CAA} \cdot \sigma_x^2$. Con $u(x) = x^{1/2}$ y una lotería que paga 36 o 100 con probabilidad 1/2, $\mathbb{E}(\mathbf{l}) = 68$ y $\mathbb{E}[u(\mathbf{l})] = 8$, de modo que $x_c = 64$ y $\rho = 4$. Con $\text{CAA} = 1/(2x)$, que en $x = 68$ vale 1/136, y $\sigma_x^2 = 1024$, la aproximación da $\rho \approx 3{,}76$.

Aversión relativa: el CAA depende de las unidades monetarias. El **CAR** tiene en cuenta la riqueza y es la elasticidad de la utilidad marginal respecto a los pagos, cambiada de signo:

$$\text{CAR} = -\frac{u''(x)}{u'(x)} \cdot x$$

Si es constante, la proporción de la riqueza en activos con riesgo no varía con la riqueza. Con una utilidad CRRA, el CAR es constante, θ, y, en un contexto intertemporal, su inversa, $1/\theta$, es la elasticidad de sustitución intertemporal.

1.3. Aplicaciones financieras

La TUE se aplica a dos decisiones financieras simétricas. En la de **aseguramiento**, el agente parte de una exposición al riesgo, fuera de la línea de certeza, y la reduce con un seguro. En la de **cartera**, parte de la certeza y aumenta su exposición comprando acciones.

1.3.1. Demanda de seguros

Idea: el agente elige la cuantía asegurada igualando la tasa subjetiva a la que intercambia renta entre dos estados de la naturaleza con la tasa objetiva que ofrece el mercado de seguros.

Supuestos: hay información imperfecta y simétrica y dos estados: sin pérdida, con la riqueza inicial M_0 y probabilidad p_1, y con pérdida, con $M_0 - L$ y probabilidad p_2. Un agente averso al riesgo elige la cuantía asegurada, M_s, y paga una prima γ por unidad asegurada, de modo que obtiene $x_1 = M_0 - \gamma M_s$ sin pérdida y $x_2 = x_1 - L + M_s$ con ella.

Recta de aseguramiento: despejando M_s y sustituyendo, su pendiente, $-(1-\gamma)/\gamma$, es la tasa objetiva a la que el mercado intercambia renta entre estados:

$$x_2 = \frac{M_0}{\gamma} - L - \frac{1-\gamma}{\gamma}\, x_1$$

Planteamiento: el agente elige la cuantía asegurada que maximiza su utilidad esperada:

$$\max_{M_s}\ u^e = p_1 \cdot u(x_1) + p_2 \cdot u(x_2) \quad \text{s.a.:} \begin{cases} x_1 = M_0 - \gamma M_s \\ x_2 = x_1 - L + M_s \\ M_s \geq 0 \end{cases}$$

CPO: la **relación marginal de sustitución estocástica**, tasa subjetiva de intercambio de renta entre estados, iguala el precio relativo de la renta entre ambos, la indemnización neta por unidad de prima:

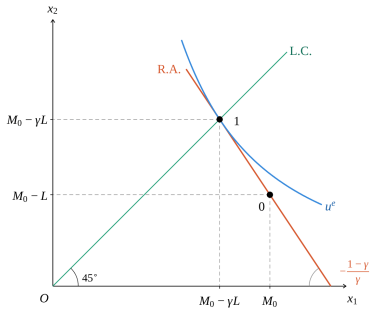
$$\text{RMSE}_{21} = \frac{p_1 \cdot u'(x_1)}{p_2 \cdot u'(x_2)} = \frac{1-\gamma}{\gamma}$$

CSO: la segunda derivada respecto a M_s, $\gamma^2 p_1 u''(x_1) + (1-\gamma)^2 p_2 u''(x_2)$, es negativa por la concavidad de la utilidad de Bernoulli, y la CPO da un máximo global.

Solución: depende de la prima frente a la probabilidad de pérdida, es decir, de la competencia y de los costes de gestión del mercado:

1. Con prima **actuarialmente justa**, $\gamma = p_2$, propia de competencia perfecta sin costes de gestión, el agente se asegura por toda la pérdida, $M_s^* = L$, y queda sobre la línea de certeza, $x_1 = x_2$.
2. Con prima desfavorable, $\gamma > p_2$, por costes de gestión o poder de mercado, se infraasegura, $M_s^* < L$. La recta de aseguramiento es menos inclinada que la curva de indiferencia sobre la línea de certeza, y $x_1 > x_2$.
3. Con prima favorable, $\gamma < p_2$, se sobreasegura, $M_s^* > L$. La recta es más inclinada, y $x_1 < x_2$.

Desarrollo gráfico: en el plano de rentas contingentes, la recta de aseguramiento parte del punto sin seguro, $(M_0, M_0 - L)$. Con prima justa, la curva de indiferencia más alta le es tangente sobre la línea de certeza, y el agente obtiene $M_0 - \gamma L$ en ambos estados.



Demanda de seguro: con prima justa, el averso se asegura por completo.

Estática comparativa: por el teorema de la función implícita, la demanda de aseguramiento depende de la aversión y de los parámetros de la lotería:

$$M_s^* = d_s(\underset{+}{\text{CAA}},\ \underset{-}{p_1},\ \underset{+}{p_2},\ \underset{\pm}{M_0},\ \underset{+}{L},\ \underset{\pm}{\gamma})$$

Extensiones: que la competencia lleve la prima a la justa exige una aseguradora casi neutral al riesgo, gracias a la **agrupación de riesgos**, que con n riesgos independientes reduce la varianza de la pérdida media a σ^2/n, aunque solo elimina el riesgo idiosincrásico, y al **reparto de riesgos**, que entre muchos partícipes lleva la prima de riesgo total hacia cero, como en el argumento de Arrow y Lind (1970).

Conclusiones: el modelo da los fundamentos de la demanda de seguros, sus determinantes y su estática comparativa. Su extensión más relevante es la información asimétrica, *ex ante*, la selección adversa, y *ex post*, el riesgo moral.

1.3.2. Decisión de cartera

Origen: Markowitz (1952) formaliza la selección de carteras con el enfoque media-varianza, Tobin (1958) añade un activo sin riesgo y Sharpe (1963) la simplifica con el modelo de índice único. En utilidad esperada, se busca qué empuja a un averso a abandonar la línea de certeza.

Supuestos: el agente tiene su riqueza M_0 en un bono de rentabilidad cierta r y puede invertir una proporción α en una acción que rinde r_a con probabilidad $1-p$ y r_b con probabilidad p, con $r_a > r > r_b$.

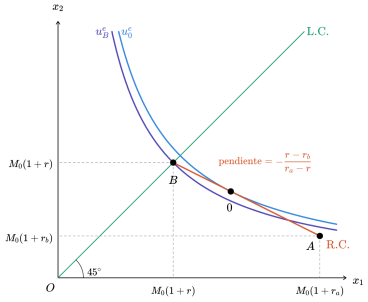
Planteamiento: elige la proporción α que maximiza su utilidad esperada:

$$\max_{\alpha}\ u^e = (1-p)\cdot u(x_1) + p \cdot u(x_2) \quad \text{s.a.:} \begin{cases} x_1 = M_0[1 + r + \alpha(r_a - r)] \\ x_2 = M_0[1 + r - \alpha(r - r_b)] \end{cases}$$

CPO: el óptimo es la tangencia entre la curva de indiferencia más alta y la recta de cartera, cuyos extremos son las carteras solo de bonos y solo de acciones. La RMSE iguala la tasa objetiva a la que la cartera intercambia renta entre estados:

$$\text{RMSE}_{21} = \frac{(1-p)\cdot u'(x_1)}{p \cdot u'(x_2)} = \frac{r - r_b}{r_a - r}$$

El averso abandona la línea de certeza si el rendimiento esperado de la acción supera a r.



Decisión de cartera.

CSO: la aversión garantiza el máximo. Con neutralidad o propensión, la solución es de esquina, $\alpha^* \in \{0, 1\}$.

Solución: la demanda de activo con riesgo cae con la aversión y con la probabilidad del resultado desfavorable, es ambigua con la riqueza y depende de la ganancia y la pérdida frente al bono:

$$\alpha^* = d(\underset{-}{\text{CAA}},\ \underset{-}{p},\ \underset{\pm}{M_0},\ \underset{+}{r_a},\ \underset{+}{r_b},\ \underset{-}{r})$$

Solo el signo de r_b es inequívoco porque, al subir, reduce la pérdida en el estado desfavorable, y los efectos sustitución y renta actúan en el mismo sentido. El de r_a exige CAR ≤ 1, y el de r, CAA constante o creciente. Según los tres rendimientos, la cartera va toda al bono, se diversifica o va toda a la acción.

Conclusiones: explica la demanda de activos con riesgo a partir de la aversión y la riqueza, y enlaza la TUE con la selección de carteras. Pero, con dos estados y un solo activo arriesgado, no recoge la correlación entre rendimientos, de la que dependen las ganancias de diversificar.

1.4. Dominancia y preferencia por estados

1.4.1. Dominancia estocástica

Idea: los coeficientes de Arrow-Pratt caracterizan al individuo, mientras que la **dominancia estocástica** caracteriza las loterías y las ordena parcialmente sin conocer la utilidad concreta. Con F y G las funciones de distribución de dos loterías:

- F domina a G en **primer orden** si $F(x) \leq G(x)$ para todo x: la probabilidad de no alcanzar cualquier pago nunca es mayor con F, y todo agente con utilidad creciente la prefiere.
- F domina a G en segundo orden si $\int_{-\infty}^{x} [G(t) - F(t)]\, dt \geq 0$ para todo x: todo averso al riesgo la prefiere, y, con igual media, G es una versión de F con más dispersión (Rothschild y Stiglitz, 1970).

La de segundo orden formaliza que, a igual rendimiento esperado, el averso elige la lotería menos arriesgada. El criterio media-varianza es un caso particular, con distribuciones normales o utilidad cuadrática.

1.4.2. Preferencia por estados

Idea: Arrow (1953) y Debreu (1959) tratan el consumo en cada estado de la naturaleza como un bien distinto.

Planteamiento: con dos estados, el individuo maximiza $p_1u(x_1) + p_2u(x_2)$ sujeto a $q_1x_1 + q_2x_2 = M_0$, donde q_s es el precio de un **activo Arrow**, que paga una unidad si ocurre el estado s y nada en otro caso.

CPO: la RMS entre el consumo de los dos estados, ponderada por sus probabilidades, iguala el precio relativo de los activos contingentes. Es suficiente por la concavidad de u y la restricción lineal:

$$\frac{p_1u'(x_1)}{p_2u'(x_2)} = \frac{q_1}{q_2}$$

Solución: con precios actuarialmente justos, $q_1/q_2 = p_1/p_2$, el averso iguala el consumo en ambos estados y se asegura plenamente, como en el modelo de seguro.

Conclusiones: convierte la elección con riesgo en un problema ordinario de demanda y muestra que un sistema completo de mercados contingentes asigna el riesgo de forma eficiente. Pero ese sistema no existe, porque crear un mercado por estado es costoso, y los estados deben ser observables y verificables, lo que excluye la información asimétrica.

En suma, con un aparato mínimo la TUE representa las preferencias sobre loterías como una esperanza de utilidades, mide la actitud ante el riesgo y resuelve el seguro y la cartera. Sus límites son la independencia, el axioma más contestado, la inestabilidad de los perfiles y el supuesto de probabilidades objetivas.

2. Críticas y teoría de la perspectiva

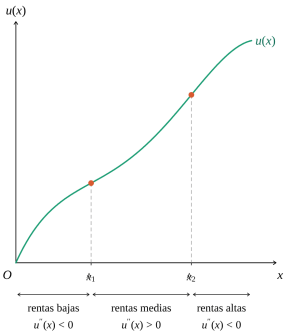
Los perfiles de riesgo pueden cambiar con distintas variables, lo que ha motivado críticas teóricas, sobre la forma de la utilidad, y empíricas, sobre los axiomas. Todas mantienen probabilidades objetivas, es decir, riesgo.

2.1. Críticas teóricas: forma de la utilidad

2.1.1. Friedman y Savage (1948)

Origen: un mismo agente compra seguros, como un averso, y participa en juegos de azar desfavorables, como un propenso, algo incoherente con los perfiles de la TUE.

Idea: una utilidad de Bernoulli cóncava, convexa y de nuevo cóncava, con dos puntos de inflexión, $\hat{x}_1$ y $\hat{x}_2$: hay aversión en las rentas bajas y altas y propensión en las intermedias, y la actitud cambia con la renta.



Utilidad de Friedman y Savage: tramos cóncavo, convexo y cóncavo.

Conclusiones: permitía entender conductas observadas, pero, como señaló Markowitz (1952), sus implicaciones chocan con la evidencia. Los ricos, en el último tramo cóncavo, no aceptarían apuestas desfavorables, y no existirían los casinos de Montecarlo. Las rentas intermedias desearían apuestas simétricas de gran cuantía.

2.1.2. Utilidad de la riqueza (Markowitz, 1952)

Idea: Markowitz cuestiona que las loterías se evalúen por la renta final. Importan los cambios, $x = M_0 + \Delta x$, respecto a la riqueza habitual, M_0, que actúa como punto de referencia o "ancla" del comportamiento.

Forma: la utilidad tiene tres puntos de inflexión, el central en la riqueza habitual. Es convexa ante ganancias pequeñas y cóncava ante las grandes, cóncava ante pérdidas pequeñas y convexa ante las grandes.

Conclusiones: ofrece más matices que la de Friedman y Savage e influyó en los trabajos de Kahneman y Tversky sobre la aversión a la pérdida. Pero es difícil postular funciones de este tipo y puede haber incoherencias entre perfiles.

2.2. Críticas empíricas: los axiomas

Idea: el contraste experimental revela violaciones sistemáticas de la TUE, que se ordenan según el supuesto que incumplen.

Independencia (Allais, 1953): la **paradoja de Allais** muestra que cambiar un componente común a dos loterías, que no debería alterar la elección, sí la altera. Con dos pares de loterías sobre los mismos resultados, la mayoría prefiere $\mathbf{l}_1^1$ en el primero y $\mathbf{l}_2^2$ en el segundo, aunque los pares solo difieren en una consecuencia común. MacCrimmon y Larsson (1979) distinguen dos manifestaciones:

- *Efecto consecuencia común*: la preferencia cambia al modificar un premio que ambas loterías comparten con la misma probabilidad. Cuanto mayor es el pago común, más averso se vuelve el agente ante la parte incierta.
- *Efecto razón común*, o efecto certeza (Kahneman y Tversky, 1979): la preferencia cambia al reducir en la misma proporción todas las probabilidades de ganar, porque se sobrevaloran los resultados ciertos frente a los meramente probables.

Invarianza: la TUE supone que la elección depende de las loterías y no de cómo se presentan ni de cómo se pregunta por ellas. Lo contradicen dos resultados:

- *Efecto marco* (Tversky y Kahneman, 1981): describir un mismo problema en ganancias o en pérdidas lleva a elecciones distintas, por sesgos cognitivos.
- *Inversiones de preferencia* (Lichtenstein y Slovic, 1971): la lotería elegida entre dos no siempre es la mejor valorada por separado, lo que contradice la transitividad y la invarianza del procedimiento.

2.3. Perspectiva (Kahneman y Tversky, 1979)

Origen: esas regularidades han guiado enfoques que rompen la independencia, unos preservando la dominancia estocástica de primer orden, como la utilidad dependiente del rango (Quiggin, 1982), y otros transformando las probabilidades en pesos de decisión.

Idea: el más influyente es la **teoría de la perspectiva** de Daniel Kahneman (Nobel 2002) y Amos Tversky, formulada en 1979 para el riesgo. Los agentes valoran las loterías como ganancias y pérdidas respecto a un punto de referencia, con una función de valor y una función de ponderación que sustituyen a la utilidad de Bernoulli y a las probabilidades objetivas.

2.3.1. Supuestos

Entorno: loterías con probabilidades objetivas, como en la TUE, pero cuyos resultados se codifican como ganancias y pérdidas, Δx_i, respecto a un punto de referencia (habitualmente la riqueza actual, aunque puede ser una expectativa o una aspiración). Sigue siendo riesgo, con pesos de decisión sobre probabilidades conocidas.

Agentes: el agente evalúa la lotería en dos fases y elige la de mayor valor ponderado.

- *Edición*: simplifica y codifica el problema: fija el punto de referencia, clasifica los resultados en ganancias y pérdidas y elimina los componentes comunes, lo que hace posibles los efectos marco.
- *Evaluación*: valora cada lotería con una función de valor, $v(\cdot)$, sobre ganancias y pérdidas, y una función de ponderación, $\pi(\cdot)$, sobre las probabilidades.

2.3.2. Modelo

Planteamiento: en el caso discreto, con $v(\cdot)$ creciente, $v(0) = 0$ y definida sobre $\Delta x_i = x_i - x_0$, y $\pi(\cdot)$ no lineal, las loterías se evalúan según:

$$V(\mathbf{l}) = \sum_{i=1}^{n} \pi(p_i) \cdot v(\Delta x_i)$$

Como los pesos se aplican a cada probabilidad por separado, puede elegirse una lotería dominada en primer orden. Kahneman y Tversky lo evitan suponiendo que en la edición el agente descarta las alternativas dominadas.

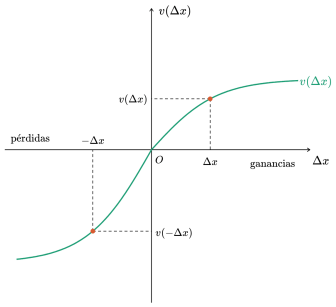
Función de valor: tiene tres rasgos.

- *Dependencia del punto de referencia*: se define sobre ganancias y pérdidas, porque el sistema perceptivo responde más a cambios que a magnitudes absolutas.
- *Aversión a la pérdida*: los agentes son más sensibles a las pérdidas que a ganancias de igual magnitud, y la función es más inclinada en las pérdidas: $-v(-\Delta x) > v(\Delta x)$ para $\Delta x > 0$.
- *Sensibilidad decreciente*: es cóncava en ganancias y convexa en pérdidas, lo que implica aversión al riesgo en las ganancias y propensión en las pérdidas, para probabilidades moderadas. Gráficamente, $v(\Delta x)$ pasa por el origen, que es el punto de referencia.

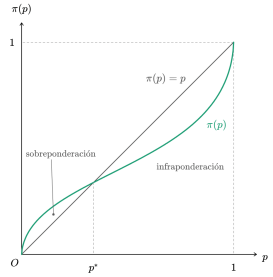
Función de ponderación: $\pi(\cdot)$ sobrepondera las probabilidades bajas e infrapondera las altas, en forma de S invertida, de modo que pesan más las colas. No son creencias erróneas, sino pesos de decisión. Con la función de valor explica el **patrón cuádruple**: aversión al riesgo ante ganancias probables y pérdidas improbables, y propensión ante ganancias improbables, como las loterías, y pérdidas probables.

2.3.3. Conclusiones

Aspectos positivos: recoge las violaciones de la independencia y de la invarianza con un aparato formal comparable al de la TUE. Camerer (2000) reúne diez regularidades de campo contrarias a la utilidad esperada que así se explican, y conceptos derivados, como el efecto dotación o la contabilidad mental (Thaler), son hoy el núcleo de la economía conductual.



Función de valor: cóncava en ganancias, convexa y más inclinada en pérdidas.



Función de ponderación de probabilidades (Tversky y Kahneman, 1992).

Limitaciones: es difícil de aplicar. En una decisión de cartera, no es evidente si el punto de referencia es la riqueza total, el valor de la cartera o cada activo. Además, la fase de edición no está formalizada, y la teoría es descriptiva, no normativa.

En suma, la teoría de la perspectiva reúne las críticas teóricas y empíricas con el punto de referencia de Markowitz, la aversión a la pérdida y la ponderación no lineal, aunque en su versión de 1979 sigue limitada a probabilidades objetivas.

3. Elección en incertidumbre

Rara vez las situaciones se reducen a probabilidades objetivas. Lo habitual es la incertidumbre en sentido de Knight, en la que la "probabilidad" es una estimación subjetiva.

3.1. Probabilidad subjetiva y acumulativa

Las dos teorías del riesgo se extienden a la incertidumbre: la utilidad esperada, con la teoría de la probabilidad subjetiva, y la perspectiva, con la perspectiva acumulativa.

3.1.1. Probabilidad subjetiva (Anscombe y Aumann, 1963)

Origen: las formulaciones de Savage (1954) y de Anscombe y Aumann (1963) buscan determinar a la vez una utilidad, los gustos, y una medida de probabilidad, las creencias, a partir de las preferencias. La de Anscombe y Aumann es la más próxima a la TUE: el agente se comporta *como si* estableciera probabilidades subjetivas.

Supuestos: hay un conjunto finito de estados, $j = 1, \dots, s$, sin probabilidades dadas, y cada alternativa asigna a cada estado una lotería objetiva F_j sobre los pagos. El agente cumple los axiomas de la TUE, con la independencia ampliada a valoraciones estado-dependientes.

Razonamiento: la representación se obtiene en tres pasos.

1. Por el teorema de la utilidad esperada ampliada, cada alternativa se evalúa con valoraciones estado-dependientes, $\sum_j \int u_j(x)\, dF_j(x)$. Con pagos ciertos se reduce a $\sum_j u_j(x_j)$.
2. Para separar las probabilidades de la utilidad, se imponen preferencias **estado-uniformes**: la valoración de una consecuencia no depende del estado en que se logre. En símbolos, $\succsim_j = \succsim_{j'}$ para todo j, j'.
3. Por el teorema de la utilidad esperada subjetiva, existen probabilidades π_j positivas, que suman uno, y una utilidad $u(\cdot)$ que representan las preferencias.

Utilidad esperada subjetiva: la diferencia con la TUE está en π_j, probabilidades subjetivas reveladas por las elecciones:

$$(x_1, \dots, x_s) \succsim (x'_1, \dots, x'_s) \leftrightarrow \sum_j \pi_j \cdot u(x_j) \geq \sum_j \pi_j \cdot u(x'_j)$$

Conclusiones: unifica riesgo e incertidumbre y hace de las creencias objeto de análisis, porque se revelan en las elecciones y se actualizan con información. Pero hereda las críticas a la TUE y tiene problemas propios.

Paradoja de Ellsberg (1961): los individuos prefieren apostar con una urna de probabilidades conocidas antes que con otra de probabilidades desconocidas, un patrón que ninguna asignación de probabilidades subjetivas justifica, conocido como **aversión a la ambigüedad**. La incorporan la utilidad esperada *maxmin* (Gilboa y Schmeidler, 1989), que evalúa el peor caso de un conjunto de probabilidades, y las probabilidades no aditivas (Schmeidler, 1989).

3.1.2. Perspectiva acumulativa (Tversky y Kahneman, 1992)

Idea: Tversky y Kahneman (1992) reformulan la versión de 1979, limitada al riesgo y expuesta a violar la dominancia de primer orden, con la ponderación dependiente del rango de Quiggin (1982) y las probabilidades no aditivas de Schmeidler (1989). La función de valor no cambia, pero el peso de cada resultado depende de su posición en la ordenación y se obtiene de las probabilidades acumuladas, por separado para ganancias y pérdidas.

Razonamiento: con incertidumbre, el agente no asigna probabilidades aditivas, sino capacidades, pesos no aditivos que dependen de la fuente de la incertidumbre, y así acomoda la aversión a la ambigüedad.

Conclusiones: respeta la dominancia de primer orden, unifica riesgo e incertidumbre y reproduce con pocos parámetros el patrón cuádruple. Hereda la dificultad de especificar el punto de referencia y el carácter descriptivo, y exige estimar más parámetros.

3.2. Análisis bayesiano (Bayes, 1763)

Idea: es la contrapartida operativa de la probabilidad subjetiva. Muestra cómo formar las creencias y revisarlas con información experimental.

Probabilidad total: la probabilidad de un suceso B se calcula con las probabilidades condicionadas a los sucesos A_i de una partición del espacio muestral:

$$P(B) = \sum_i P(B \mid A_i) \cdot P(A_i)$$

Teorema de Bayes (1763): da la probabilidad de un suceso condicionada a otro cuando se conoce la del condicionado inverso:

$$P(A_i \mid B) = \frac{P(B \mid A_i) \cdot P(A_i)}{P(B)}$$

Árboles de decisión: representan el problema como una sucesión de nodos de decisión, donde elige el agente, y de azar, donde actúa la naturaleza, y se resuelven hacia atrás. En cada nodo de azar se calcula el valor esperado, o la utilidad esperada si el agente no es neutral al riesgo, y en cada nodo de decisión se elige la rama de mayor valor.

El teorema de Bayes actualiza las probabilidades con información experimental, como un ensayo o una encuesta, y permite valorarla: el **valor esperado de la información** es la diferencia entre el valor del problema con ella y sin ella. Si la demanda de un producto nuevo es alta, A_1, con probabilidad 0,4 y un estudio da un informe favorable, B, con probabilidad 0,8 si es alta y 0,3 si es baja, $P(B) = 0{,}32 + 0{,}18 = 0{,}50$ y $P(A_1 \mid B) = 0{,}32/0{,}50 = 0{,}64$.

3.3. Criterios de decisión en incertidumbre

Idea: si el agente no está dispuesto a asignar probabilidades, ni siquiera subjetivas, decide con reglas definidas sobre la matriz de resultados de cada acción en cada estado:

1. *Wald*, pesimista: el mejor de los peores resultados, *maximin* en ganancias o *minimax* en costes. Asegura el mejor resultado posible en el peor escenario.
2. *Optimista*, *maximax*: el mejor de los mejores resultados.
3. *Hurwicz*: combina ambos con un índice de optimismo $\alpha \in [0, 1]$, valora cada acción como α por su mejor resultado más $1 - \alpha$ por el peor y elige la de mayor valor. Wald y *maximax* son los casos $\alpha = 0$ y $\alpha = 1$. En símbolos, cada acción vale:

$$\alpha \cdot (\text{mejor resultado}) + (1 - \alpha) \cdot (\text{peor resultado})$$

4. *Savage*: calcula el coste de oportunidad de cada acción en cada estado, el mejor resultado del estado menos el obtenido, y elige la acción que minimiza el máximo, el *minimax* del arrepentimiento.
5. *Laplace*: por el principio de razón insuficiente, da a todos los estados el peso $1/s$ y elige la acción de mayor resultado esperado, lo que devuelve el problema a la probabilidad subjetiva.

A11. Teoría de la producción. Caracterización de la tecnología de la empresa a corto y largo plazo. El conjunto de posibilidades de producción. La función de producción. Rendimientos locales y globales a escala. Elasticidad de sustitución. Producción conjunta

1. Fundamentos de la teoría de la producción

La teoría de la producción concibe la empresa como un agente racional que procura obtener, con una tecnología dada, el mayor *output* posible de los *inputs* que emplea. Excluye, por ello, los precios y los costes, y se centra en las restricciones técnicas que condicionan físicamente la producción.

1.1. Planes de producción y tecnología

El conjunto de elección se delimita en tres pasos: del espacio de todos los planes de producción al subconjunto de los técnicamente posibles y, dentro de este, a su frontera, formada por los técnicamente eficientes.

1.1.1. Espacio de planes de producción

Definición: un **plan de producción** recoge las combinaciones de *output* e *inputs* de un proyecto de producción. El **espacio de planes de producción** (EPP) reúne todos los planes, vectores $\mathbf{y} \in \mathbb{R}^n$ en los que los componentes positivos son *outputs* y los negativos, *inputs*. En símbolos:

$$\text{EPP} = \{\mathbf{y} \in \mathbb{R}^n \mid y_i > 0 \to x_i\text{: output; } y_i < 0 \to z_i\text{: input}\}$$

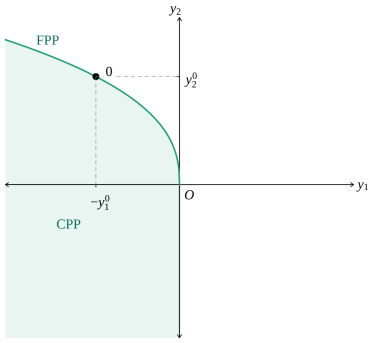
Pero los recursos son limitados. La empresa solo tiene a su alcance los planes que permite el conocimiento técnico y que su dotación de recursos le deja ejecutar, los físicamente posibles, que forman el **conjunto de posibilidades de producción**, CPP $\subseteq$ EPP.

1.1.2. CPP: axiomas de delimitación

Propiedades: el CPP se delimita con seis axiomas.

- *No vacío*: existe al menos un plan a disposición de la empresa.
- *Cerrado*: el límite de una secuencia de planes posibles es posible, de modo que el CPP incluye su frontera y los planes eficientes son posibles.
- *Libre disposición*: si un plan es posible, también lo es cualquier otro que emplee los mismos *inputs* o más y obtenga el mismo *output* o menos. La capacidad productiva puede derrocharse sin coste, y la frontera tiene pendiente no positiva. Formalmente, $\mathbf{y}_0 \in$ CPP, $\mathbf{y}_1 \leq \mathbf{y}_0 \Rightarrow \mathbf{y}_1 \in$ CPP, y, con la inacción, deja dentro del CPP el ortante negativo.
- *Imposibilidad de producción gratuita* (*no free lunch*): no puede obtenerse *output* sin emplear algún *input*, CPP $\cap \mathbb{R}^n_+ \subseteq \{\mathbf{0}\}$. Gráficamente, el CPP no tiene ningún punto en el ortante positivo salvo, a lo sumo, el origen.
- *Posibilidad de inacción*: el plan nulo es posible, $\mathbf{0} \in$ CPP. Puede incumplirse a corto plazo, con factores fijos o costes irrecuperables. Con el axioma anterior, el origen queda sobre la frontera.
- *Irreversibilidad*: si un plan no nulo es posible, su inverso no lo es: no se pueden recuperar los *inputs* a partir del *output*. Formalmente, $\mathbf{y} \in$ CPP, $\mathbf{y} \neq \mathbf{0} \Rightarrow -\mathbf{y} \notin$ CPP.

Gráficamente, en el plano *input-output*, (y_1, y_2), con el *input* en negativo, el CPP es la región que reúne los planes posibles, y la FPP, su límite superior, una curva que pasa por el origen. Un plan eficiente, como el 0, emplea y_1^0 unidades de *input* para obtener y_2^0 de *output*.



Conjunto (CPP) y frontera (FPP) de posibilidades de producción: el conjunto reúne los planes factibles y la frontera, los eficientes.

1.1.3. FPP: eficiencia técnica y rendimientos

Eficiencia técnica: en la **frontera de posibilidades de producción** (FPP), la restricción tecnológica efectiva de la empresa, no es posible obtener más *output* con los mismos *inputs* ni el mismo *output* con menos.

Axiomas de regularidad: la forma de la frontera depende de tres axiomas sobre la escala.

- *Aditividad*, o libre entrada: sumar dos planes posibles da otro posible, porque pueden ejecutarse a la vez en unidades independientes. Implica rendimientos no decrecientes. Garantiza que duplicar un plan es posible, pero nada impide planes más productivos que esa duplicación.
- *Proporcionalidad*, o divisibilidad: todo plan posible puede reducirse a escala en cualquier proporción. Implica rendimientos no crecientes. Formalmente, $\mathbf{y}_0 \in \text{CPP} \Rightarrow \lambda \mathbf{y}_0 \in \text{CPP}$ para $\lambda \in [0, 1]$.
- *Convexidad*: toda combinación de dos planes posibles es posible. Las mezclas equilibradas de factores son al menos tan productivas como las extremas. En la producción simple equivale a una función de producción cóncava. Formalmente, $\alpha \mathbf{y}_0 + (1-\alpha)\mathbf{y}_1 \in$ CPP para todo $\alpha \in [0, 1]$, y la convexidad estricta excluye los rendimientos crecientes y, por tanto, los mixtos.

Rendimientos: solo con rendimientos constantes se cumplen los tres a la vez.

1. Con proporcionalidad y aditividad, el CPP es un cono convexo, los rendimientos son **constantes** y la frontera es una recta desde el origen. Todo plan admite cualquier cambio de escala.
2. Un CPP estrictamente convexo admite reducciones proporcionales, pero no la suma de planes, de modo que los rendimientos son **decrecientes** y la frontera, estrictamente cóncava.
3. Con aditividad y sin proporcionalidad, el CPP deja de ser convexo, un plan grande rinde más que varios pequeños, los rendimientos son **crecientes** y la frontera, convexa. La teoría neoclásica excluye este caso, que recupera la organización industrial.

1.2. La función de producción

1.2.1. Función de producción y propiedades

Definición: la **función de producción** es una relación exclusivamente técnica que da, para cada vector de *inputs*, el *output* máximo que permite la tecnología. Describe la frontera del CPP:

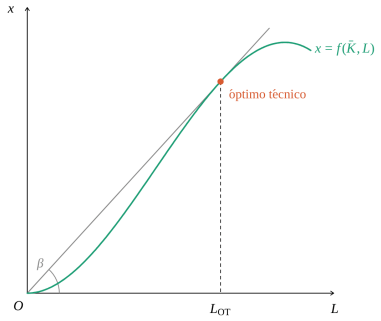
$$x = f(\mathbf{z}), \qquad f : \mathbb{R}^m_+ \to \mathbb{R}_+$$

En la versión dual, la tecnología se describe con el conjunto de requerimientos de *inputs*, $V(x)$, las combinaciones que obtienen al menos x, cuya frontera es la isocuanta. Con varios *outputs*, se usa una función de transformación implícita, $F(\mathbf{x}, \mathbf{z}) = 0$. El conjunto de requerimientos es $V(x) = \{\mathbf{z} \in \mathbb{R}^m_+ : f(\mathbf{z}) \geq x\}$.

Propiedades:

- Es continua y dos veces diferenciable, lo que permite definir productividades marginales y relaciones de sustitución. Supone la perfecta divisibilidad de factores y producto.
- Es **cardinal**, porque sus variables son observables y medibles. A diferencia de la utilidad, no admite transformaciones que alteren sus valores.
- Es no decreciente en cada factor, por la libre disposición, con productividades marginales positivas en la región económica, y cuasicóncava, con isocuantas convexas que recogen la sustituibilidad decreciente entre factores.
- Comparar $f(\lambda \mathbf{z})$ con $\lambda f(\mathbf{z})$ ante un cambio equiproporcional de todos los factores define los rendimientos a escala: si es homogénea de grado h, crecientes, constantes o decrecientes según h sea mayor, igual o menor que uno.
- Si es homogénea de grado h, por el teorema de Euler, $\sum_i z_i \, \partial f / \partial z_i = h \cdot f(\mathbf{z})$: con rendimientos constantes, pagar a cada factor su productividad marginal agota exactamente el producto; con crecientes, los pagos lo superan, y con decrecientes, no lo agotan.
- No es directamente observable, porque las empresas no siempre producen sobre la frontera, pero puede estimarse con datos de factores y de producto, separando la tecnología de la ineficiencia de cada empresa.

Gráficamente, con un *output*, trabajo, L, y capital, K, la función es una colina en $\mathbb{R}^3$. Fijando $\bar{K}$, el corte vertical es la función a corto plazo, de la que salen las productividades. Fijando el *output*, el corte horizontal es la isocuanta, de la que salen la RMST y la elasticidad de sustitución. Con rendimientos mixtos, la función a corto plazo es primero convexa y luego cóncava, y en el óptimo técnico, L_{OT}, el radio-vector le es tangente y la productividad media es máxima.



Función de producción a corto plazo: en el óptimo técnico, la productividad media es máxima.

1.2.2. Productividad marginal y media

Productividad marginal: el aumento del producto ante un aumento infinitesimal del factor variable, *ceteris paribus*. Es la pendiente de la función de producción:

$$\text{Pmg}_L \equiv f'_L = \frac{\partial f(K,L)}{\partial L}$$

Productividad media: el producto por unidad de factor variable, $\text{Pme}_L = x/L$. Es la pendiente del radio-vector desde el origen, $\tan\beta$. Crece hasta que el radio-vector es tangente a la función de producción. La marginal, pendiente de la función, es creciente en su tramo convexo, decreciente en el cóncavo y nula en el máximo técnico.

Si la productividad marginal supera la media, cada trabajador adicional produce más que la media y esta sube. Si es inferior, la media cae. Ambas coinciden, por tanto, en el máximo de la productividad media, el **óptimo técnico**.

1.3. Las curvas isocuantas

Definición: una **isocuanta** (curva de nivel de la función de producción, proyectada sobre el plano de los factores) recoge los modos de combinar los factores que dan un mismo volumen de producto.

Propiedades: no se cortan, las más alejadas del origen reflejan más producto y, con eficiencia técnica, tienen pendiente negativa. Pueden ser lineales, con sustituibilidad perfecta; angulares o Leontief, con complementariedad perfecta; quebradas, con sustituibilidad limitada, como en la programación lineal, o continuas y convexas, con sustituibilidad continua e imperfecta.

1.3.1. Pendiente: la RMST

Definición: la **relación marginal de sustitución técnica** es la tasa a la que un factor sustituye a otro con la producción constante (cuántas unidades de capital libera la empresa al contratar una unidad más de trabajo). Se obtiene igualando a cero la diferencial total de la función; para dos factores i y j y, en particular, para el capital y el trabajo:

$$\text{RMST}_{ji} = -\left.\frac{dz_j}{dz_i}\right|_{x_0} = \frac{f'_i}{f'_j}, \qquad \text{RMST}_{KL} = -\left.\frac{dK}{dL}\right|_{x_0} = \frac{f'_L}{f'_K}$$

Si la Pmg_L es elevada en relación con la Pmg_K, una unidad de trabajo sustituye muchas unidades de capital y la RMST es alta. Al sustituir capital por trabajo, la Pmg_L tiende a caer y la Pmg_K a subir, y la cuasiconcavidad garantiza una RMST decreciente. Equivale a la RMS del consumidor, pero, como la función es cardinal, la magnitud de las contribuciones marginales sí tiene significado. Su limitación es que depende de las unidades en que se mide cada factor. Gráficamente, la isocuanta es convexa respecto al origen porque la RMST es decreciente, y su pendiente en cada punto, cambiada de signo, es la RMST.

1.3.2. Forma: elasticidad de sustitución (Hicks, 1932)

Definición: la **elasticidad de sustitución técnica** evita el problema de las unidades, porque relaciona dos variaciones porcentuales, la de la mezcla factorial y la de la RMST. Para dos factores i y j, $\sigma_{ji} = \partial \ln(z_j/z_i)/\partial \ln \text{RMST}_{ji}$; con el capital y el trabajo:

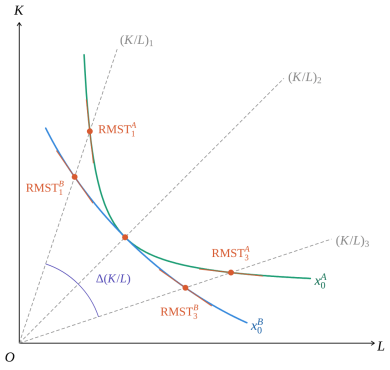
$$\sigma_{KL} = \left.\frac{\Delta\,\%(K/L)}{\Delta\,\%(\text{RMST}_{KL})}\right|_{x_0} = \frac{\partial \ln(K/L)}{\partial \ln \text{RMST}_{KL}}$$
$$= \frac{d(K/L)}{d\,\text{RMST}_{KL}} \cdot \frac{\text{RMST}_{KL}}{K/L} \geq 0$$

Si es alta, los factores se sustituyen con facilidad y la RMST, el cociente de productividades marginales, apenas se mueve aunque cambie la mezcla. Con dos factores y minimización de costes, $\text{RMST} = w/r$, mide la respuesta de la mezcla al precio relativo de los factores. Con más de dos factores, se generaliza con la elasticidad parcial de Allen (1938).

Casos: la función lineal tiene $\sigma = \infty$, sustituibilidad perfecta; la Leontief, $\sigma = 0$, complementariedad perfecta, y la Cobb-Douglas, $\sigma = 1$. En esta última, la ratio K/L varía en la misma proporción que la RMST, y por eso, con minimización de costes, las participaciones relativas de los factores en el coste se mantienen constantes. En la CES, con $\alpha, \beta > 0$, la elasticidad es constante:

$$x = (\alpha K^\rho + \beta L^\rho)^{1/\rho} \;\rightarrow\; \sigma = \frac{1}{1-\rho}$$

Desarrollo gráfico: para un mismo cambio de la proporción de factores, la RMST varía más en la isocuanta más convexa, de elasticidad baja, que en la más recta. Los movimientos sobre un radio-vector son cambios de escala, y el paso de uno a otro, $\Delta(K/L)$, un cambio en las proporciones de factores.

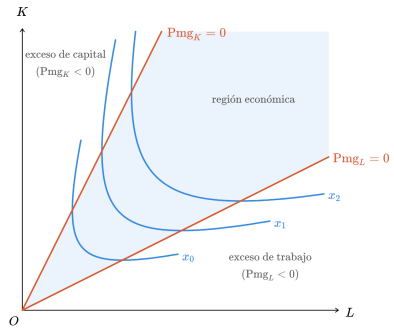


Elasticidad de sustitución: la RMST varía más en la isocuanta más convexa (σ^A baja).

1.3.3. Eficiencia técnica y región económica

Una empresa racional nunca emplea un factor con productividad marginal negativa, porque reduciéndolo obtendría más *output* con menos *input*. Solo interesa, por eso, el tramo de las isocuantas con pendiente negativa ($\text{RMST} \geq 0$).

Región económica: en ella ambos factores tienen productividad marginal positiva, las isocuantas son decrecientes y hay eficiencia técnica. La limitan las **líneas de contorno** (*ridge lines*), sobre las que la productividad marginal de un factor vale cero. Son isoclinas rectas si la función es homotética e irregulares en otro caso. Fuera, uno de los factores tiene productividad marginal negativa y las isocuantas son crecientes, zona que solo existe si se prescinde de la libre disposición, que excluye cualquier productividad marginal negativa.



Región económica.

Relación con las etapas: con rendimientos constantes, la región económica corresponde a la etapa II del corto plazo, la región exterior superior, con exceso de capital, a la etapa I, y la inferior, con exceso de trabajo, a la III. En general, solo es exacta la correspondencia entre la región inferior y la etapa III.

1.4. De la tecnología a la decisión

La empresa se estudia aquí solo como unidad técnica. El conjunto de elección son los planes técnicamente posibles, y su ordenación, la función de producción y sus isocuantas. Para decidir hacen falta dos tipos de consideraciones más.

Económicas: los precios de los factores y el presupuesto definen la recta isocoste, y el equilibrio resulta de combinarla con la tecnología, minimizando el coste de un nivel de producción o maximizando el *output* para un coste dado.

De mercado: decidir cuánto producir y a qué precio exige conocer el mercado del producto y su precio.

En suma, la teoría caracteriza la empresa por sus restricciones técnicas: los axiomas delimitan los planes posibles y la frontera, los de escala fijan los rendimientos, y la función de producción da las productividades, la RMST y la elasticidad de sustitución. No basta para decidir cuánto producir, pero ordena la respuesta del *output* cuando varían un factor, todos o la tecnología.

2. Leyes de producción

La empresa puede aumentar su producción alterando la combinación de factores, pero con restricciones temporales, porque con poco tiempo o costes de ajuste la mezcla es difícil de modificar. Se distinguen cuatro horizontes:

1. En el **muy corto plazo**, la producción ya está hecha y la cantidad ofrecida es fija (periodo de mercado).
2. En el **corto plazo**, al menos un factor es fijo, y rige la ley de las proporciones variables.

3. En el **largo plazo**, todos los factores son variables, y lo relevante son los rendimientos a escala.
4. En el **muy largo plazo**, también cambia la tecnología (el progreso técnico).

2.1. Corto plazo: proporciones variables

Ley: con un factor fijo y la tecnología dada, solo cabe aumentar la producción con más factor variable. Como cambia la proporción entre factores, llegado cierto punto la contribución marginal del factor variable es decreciente y, llevada al extremo, negativa.

2.1.1. Elasticidad-producto

Definición: la **elasticidad-producto** mide la respuesta porcentual del *output* al factor variable y es el cociente entre las productividades marginal y media:

$$\varepsilon_{xL} = \frac{\partial x}{\partial L} \cdot \frac{L}{x} = \frac{\text{Pmg}_L}{\text{Pme}_L}$$

Desempeña en el corto plazo el papel de la elasticidad de escala en el largo. Indica la etapa de la producción y, con el precio del factor dado, fija la forma de las curvas de costes a corto plazo. Es mayor que uno mientras la productividad media crece, menor cuando decrece e igual a uno en el óptimo técnico. En la Cobb-Douglas, es constante e igual al exponente del factor.

2.1.2. Etapas de la producción

Etapa I: llega hasta el margen extensivo, el óptimo técnico, L_{OT}, con $\varepsilon_{xL} > 1$. Sobra factor fijo respecto al variable, el capital no se aprovecha bien y, con rendimientos constantes, su productividad marginal es incluso negativa, de modo que conviene seguir contratando trabajadores. Formalmente, $\varepsilon_{xL} > 1$ para todo $L \in (0, L_{\text{OT}})$ y, con rendimientos constantes, $\text{Pmg}_K < 0$ en toda la etapa, fuera de la región económica.

– La función es primero creciente y convexa, con rendimientos marginales crecientes que pueden deberse a indivisibilidades, y después, desde el punto de inflexión, creciente y cóncava. En símbolos, $f'_L > 0$ y $f''_{LL} > 0$ en el tramo convexo, y $f'_L > 0$ y $f''_{LL} < 0$ en el cóncavo.

Ley de rendimientos marginales decrecientes (Turgot, 1767): si se aumenta un factor variable con el resto constante, a partir de cierto punto el producto crece, $f' > 0$, cada vez menos, $f'' < 0$. Observada en múltiples industrias, explica la forma de las curvas de costes, $\text{Cmg} = w/\text{Pmg}_L$, de la demanda de factores y de la oferta a corto plazo.

Etapa II: va del margen extensivo al intensivo, el máximo técnico, L_{MT}, con $\varepsilon_{xL} \in [0, 1]$, y en ella se altera la intensidad de uso. Con rendimientos constantes coincide con la región económica, de modo que ambas productividades marginales son positivas. La empresa debe producir en esta etapa, pero el punto exacto depende de los costes y, como la teoría de la producción no puede determinarlo por sí sola, hace falta la teoría de costes.

Etapa III: es la de sobreexplotación, con $\varepsilon_{xL} < 0$ a partir del máximo técnico, donde $\text{Pmg}_L = 0$. El factor variable es tan abundante que cada trabajador adicional reduce el producto total, y ninguna empresa racional produce aquí. Con jornadas laborales, el producto crece en proporción a las horas hasta unas 49 semanales, a ritmo decreciente a partir de ahí y, por encima de unas 55, prácticamente deja de aumentar (Pencavel, 2014).

Gráficamente, la Pmg_L alcanza su máximo en el punto de inflexión, corta a la Pme_L en el máximo de esta, en L_{OT}, y se anula en L_{MT}, donde el producto total es máximo. Así, la elasticidad-producto delimita las tres etapas:

$$\begin{aligned} \text{Etapa I:} \quad & \varepsilon_{xL} > 1, && L \in (0, L_{\text{OT}}) \\ \text{Etapa II:} \quad & \varepsilon_{xL} \in [0, 1], && L \in [L_{\text{OT}}, L_{\text{MT}}] \\ \text{Etapa III:} \quad & \varepsilon_{xL} < 0, && L > L_{\text{MT}} \end{aligned}$$

Los rendimientos marginales decrecientes parecen independientes del tipo de rendimientos a escala. Si la función es homogénea de grado $h \leq 1$ y cuasicóncava, es cóncava y la productividad marginal es no creciente, mientras que con rendimientos crecientes puede ser creciente en algún tramo.

2.2. Largo plazo: rendimientos a escala

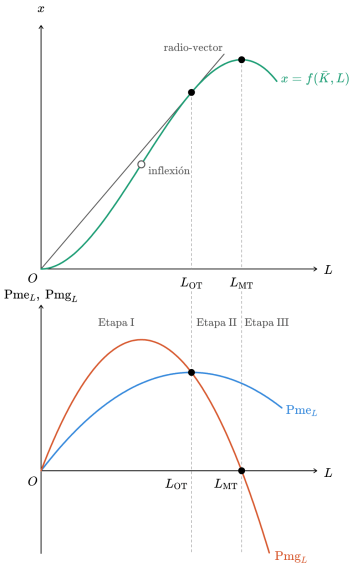
2.2.1. Elasticidad de escala

Definición: la **elasticidad de escala** es la variación porcentual del *output* ante una variación porcentual de todos los factores, con su proporción constante, y es la suma de las elasticidades-producto:

$$\varepsilon_{x\lambda} = \sum_{i=1}^{m} \varepsilon_{xz_i} = \varepsilon_{xK} + \varepsilon_{xL}$$

Clasificación: los rendimientos pueden ser de tres tipos.

1. *Crecientes*, $\varepsilon_{x\lambda} > 1$: por indivisibilidades (Wicksell, Edgeworth, Kaldor), por la especialización y la división del trabajo (Smith, 1776) o por razones físicas, como el oleoducto que, al duplicar el diámetro, cuadruplica el volumen con el doble de material. Con la especialización, cada factor se dedica a una etapa concreta del proceso, lo que aumenta la eficiencia global.
2. *Constantes*, $\varepsilon_{x\lambda} = 1$: por la posibilidad de replicar los procesos y de reducirlos a escala (aditividad y proporcionalidad).
3. *Decrecientes*, $\varepsilon_{x\lambda} < 1$: por capacidades rígidas y problemas de organización, como la pérdida de control al pasar la información por cada nivel de la jerarquía (Williamson, 1967), la burocracia, los fallos de coordinación y los costes de transacción.



Etapas de la producción a corto plazo ($\bar{K}$ fijo).

Funciones homogéneas: si $f(\lambda K, \lambda L) = \lambda^h f(K, L)$, la elasticidad de escala es constante e igual al grado, h, que en la Cobb-Douglas es $\alpha + \beta$. La homogeneidad implica homoteticidad: la RMST depende solo de la proporción de factores, las isocuantas son expansiones radiales unas de otras y la senda de expansión es una recta por el origen. Con $h = 1$, las productividades dependen solo de esa proporción. Con rendimientos crecientes o decrecientes, en cambio, la función no tiene por qué ser homogénea, porque la elasticidad de escala puede variar con la producción.

2.2.2. Rendimientos locales y globales

Rendimientos locales: los de un rango restringido de producción, el entorno de una combinación concreta de factores, $\mathbf{z}$. Se miden con la elasticidad de escala en ese punto, que informa de lo que ocurre si, desde ahí, todos aumentan en la misma proporción en una cuantía infinitesimal:

$$\varepsilon_{x\lambda}(\mathbf{z}) = \frac{df(\lambda \mathbf{z})}{d\lambda} \cdot \frac{\lambda}{f(\lambda \mathbf{z})}\bigg|_{\lambda=1}$$

Intuitivamente, indican si, con el tamaño actual de la empresa, un aumento del 1 % de todos los factores eleva el *output* en más, en igual o en menos del 1 %. Son localmente crecientes, constantes o decrecientes en $\mathbf{z}$ según $\varepsilon_{x\lambda}(\mathbf{z})$ sea mayor, igual o menor que uno.

Rendimientos globales: los de todo el rango de producción, para cualquier combinación de factores y cualquier escala. Son una propiedad de toda la tecnología, como los axiomas de escala, y se comprueban comparando $f(\lambda \mathbf{z})$ con $\lambda f(\mathbf{z})$. Son crecientes si $f(\lambda \mathbf{z}) > \lambda f(\mathbf{z})$ para todo $\mathbf{z}$ y todo $\lambda > 1$, y constantes o decrecientes con la igualdad o la desigualdad contraria. Multiplicar todos los factores por cualquier $\lambda > 1$ multiplica siempre el *output* por más de λ, por λ o por menos de λ.

Relación: a lo largo de cada radio-vector, la elasticidad de escala es la derivada logarítmica del *output* respecto a la escala:

$$\frac{d \ln f(\lambda \mathbf{z})}{d \ln \lambda} = \varepsilon_{x\lambda}(\lambda \mathbf{z})$$

Los globales resultan, por tanto, de acumular los locales. Si la elasticidad de escala es mayor que uno en todos los puntos, al integrar se obtiene $f(\lambda \mathbf{z}) > \lambda f(\mathbf{z})$ para todo $\lambda > 1$, y análogamente en los otros dos casos. En una función homogénea, la elasticidad vale h en todos los puntos y ambos coinciden. Pero, en general, la elasticidad cambia con la producción y puede no haber rendimientos globales de ningún tipo, como con rendimientos mixtos. La producción tiene forma de S sobre el radio-vector, creciente a pequeña escala y decreciente a gran escala, la tecnología que subyace al coste medio a largo plazo en U. Así ocurre con $x = KL/(1 + KL)$, con rendimientos localmente crecientes si $x < 1/2$, constantes en $x = 1/2$ y decrecientes si $x > 1/2$, porque cada

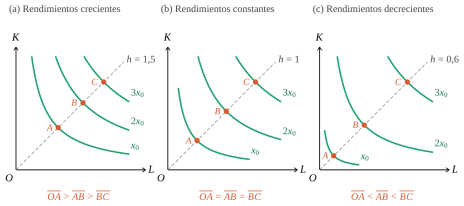
elasticidad-producto vale $1/(1+KL)$ y:

$$\varepsilon_{x\lambda} = \varepsilon_{xK} + \varepsilon_{xL} = \frac{2}{1+KL} = 2(1-x)$$

Costes: si la empresa minimiza costes, la elasticidad del coste respecto al *output* es la inversa de la de escala, $\varepsilon_{Cx} = 1/\varepsilon_{x\lambda}$, de modo que, con precios de los factores dados, los rendimientos localmente crecientes equivalen a **economías de escala**, y los decrecientes, a deseconomías.

2.2.3. Análisis gráfico

Con una función homogénea, los rendimientos se ven en la distancia, sobre un radio-vector, entre isocuantas de niveles x_0, $2x_0$ y $3x_0$. Dependen de la posición de las isocuantas, no de su forma, que fija la elasticidad de sustitución. Con rendimientos crecientes, las isocuantas sucesivas están cada vez más juntas; con constantes, equidistantes, y con decrecientes, cada vez más separadas. Si las isocuantas x_0, $2x_0$ y $3x_0$ cortan el radio-vector en A, B y C, $\overline{OA} > \overline{AB} > \overline{BC}$ con rendimientos crecientes, $\overline{OA} = \overline{AB} = \overline{BC}$ con constantes y $\overline{OA} < \overline{AB} < \overline{BC}$ con decrecientes.



Rendimientos a escala: crecientes (a), constantes (b) y decrecientes (c).

2.3. Muy largo plazo: progreso técnico

2.3.1. Definición y efectos

Definición: en el muy largo plazo, la función de producción incluye un factor que recoge las mejoras de los métodos de producción. El **progreso técnico** se manifiesta como innovación de producto, que crea variedades o bienes nuevos, de proceso, que permite producir más con los mismos *inputs*, o mejora de la calidad.

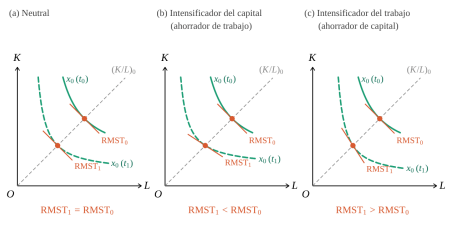
Efectos: se estudian en tres planos, con $x = A \cdot f(K, L)$ y un parámetro de eficiencia que pasa de A_0 a $A_1 > A_0$. La curva de productividad marginal se desplaza hacia fuera, de $\text{Pmg}_L(L; A_0)$ a $\text{Pmg}_L(L; A_1)$, la función a corto plazo pivota hacia arriba, de $A_0 f(\bar{K}, L)$ a $A_1 f(\bar{K}, L)$, y la isocuanta se desplaza hacia el origen. Los tres planos expresan la misma idea: un producto dado se obtiene con menos factores y, si los factores se mantienen, la producción es mayor. El progreso es neutral o sesgado según el criterio con que se mida su efecto sobre los factores.

2.3.2. Clasificación del progreso técnico

Su clasificación importa porque sus efectos alcanzan al reparto de la renta: si el progreso eleva la productividad marginal de un factor más que la del otro, con una proporción de factores dada, ese factor gana remuneración relativa y peso en la renta. Hicks (1932) y Harrod (1937; 1948) propusieron los dos criterios más usados.

Hicks (1932): compara la RMST, f'_L/f'_K, sobre un radio-vector dado antes y después del progreso.

- *Neutral*: el cociente de productividades marginales no varía, ambas suben en la misma proporción y la isocuanta se desplaza radialmente: $x_t = A(t) \cdot f(K_t, L_t)$, donde $A(t)$ recoge lo que influye en x sin ser K ni L, como el saber empresarial o la cualificación.
- *Intensificador del capital*, o ahorrador de trabajo: eleva más la productividad marginal del capital, la RMST disminuye y la empresa sustituye trabajo por capital. La isocuanta se hace más plana sobre el radio-vector.
- *Intensificador del trabajo*, o ahorrador de capital: eleva más la del trabajo, la RMST aumenta y la empresa sustituye capital por trabajo. La isocuanta se hace más inclinada sobre el radio-vector.



Clasificación de Hicks: neutral (a), ahorrador de trabajo (b) y de capital (c).

Harrod (1937; 1948): el progreso es neutral si las participaciones de los factores en la renta no cambian con una relación capital-producto dada. Como mostraron Robinson (1938) y Uzawa (1961), exige que el progreso aumente la productividad del trabajo, de modo que, con una misma cantidad de capital, se precisa cada vez menos trabajo para obtener la misma producción. Uzawa demostró además que es la condición necesaria y suficiente para un estado estacionario. Por eso es la hipótesis estándar de los modelos de tipo Solow, ya que es la única forma de progreso compatible con una senda de crecimiento equilibrado.

Solow (1960): es neutral si las participaciones no cambian con una relación trabajo-producto dada.

Síntesis: cada criterio de neutralidad corresponde a una forma de la función, según $A(t)$ multiplique a toda la función, al trabajo o al capital:

$$\begin{aligned} \text{Hicks: } x_t &= A(t) \cdot f(K_t, L_t) \\ \text{Harrod: } x_t &= f(K_t, A(t) \cdot L_t) \\ \text{Solow: } x_t &= f(A(t) \cdot K_t, L_t) \end{aligned}$$

Con una Cobb-Douglas, los tres tipos de progreso neutral son equivalentes.

Conclusiones: el progreso técnico es un determinante fundamental del crecimiento a largo plazo, y la clasificación distingue sus efectos distributivos: el ahorrador de trabajo puede generar desempleo tecnológico, mientras que el neutral de Harrod favorece un crecimiento equilibrado. Pero, como con una Cobb-Douglas las tres neutralidades coinciden, se limita la capacidad del modelo para captar los efectos asimétricos observados en la práctica.

En suma, a corto plazo la elasticidad-producto sitúa a la empresa en una etapa, y solo la II es racional; a largo plazo, la elasticidad de escala clasifica los rendimientos, y a muy largo plazo el progreso técnico desplaza la función y su sesgo altera el reparto de la renta.

3. Extensiones

Las leyes suponen un único *output* y producción sobre la frontera. Las extensiones relajan ambos supuestos: la producción de varios bienes con recursos dados, paso previo al equilibrio general, y la medición de la distancia de las empresas reales a la frontera.

3.1. Producción conjunta o multiproducto

Con un solo producto, bastaba situarse en el tramo decreciente de la isocuanta para la eficiencia técnica, pero con varios, ya no.

3.1.1. Planteamiento

Se producen dos bienes, x e y, con dos factores, mediante funciones f_x y f_y, sujetas a la dotación de la economía, $L_0 = L_x + L_y$ y $K_0 = K_x + K_y$. La eficiencia técnica exige que reasignar los factores no permita producir más de un bien sin producir menos del otro, es decir, igual RMST en ambas producciones:

$$\text{RMST}^x_{KL} = \text{RMST}^y_{KL}$$

Esta condición es la de eficiencia productiva del equilibrio general, donde un precio relativo de los factores común, w/r, garantiza la igualdad de forma descentralizada. En productividades, $\text{Pmg}^x_L/\text{Pmg}^x_K = \text{Pmg}^y_L/\text{Pmg}^y_K$, sin necesidad de precios. El caso equivale a una economía con dos productores, uno por bien, o a un único productor que obtiene los dos.

3.1.2. Caja de Edgeworth-Bowley

En la caja, de dimensiones iguales a las dotaciones, todo punto interior es una asignación de pleno empleo, aunque no necesariamente eficiente. Las isocuantas de cada empresa tienen mayor nivel cuanto más lejos de su origen, pendiente negativa y son convexas si la sustituibilidad es imperfecta.

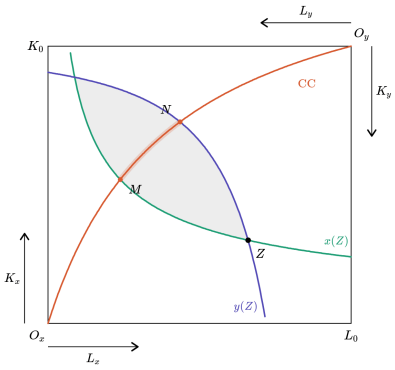
Superioridad de Pareto: desde una asignación, los puntos de la lente que forman las isocuantas que pasan por ella, el "área del excedente distribuible" de Allais (1943), son superiores en el sentido de Pareto. Lo son también las de su frontera, salvo sus dos vértices, la asignación de partida y el otro corte de ambas isocuantas.

Curva de contratos: las asignaciones en que ninguna empresa puede producir más sin que la otra produzca menos son eficientes en el sentido de Pareto, y su lugar geométrico nace de la tangencia de las isocuantas. No son comparables entre sí, porque pasar de una a otra aumenta un bien a costa del otro. En todos sus puntos se cumple $\text{RMST}^x_{KL} = \text{RMST}^y_{KL}$.

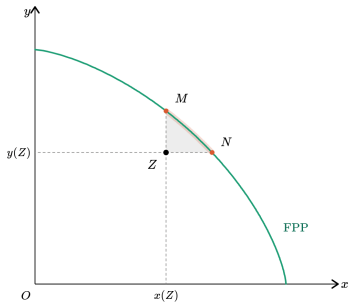
La curva no es necesariamente una isoclina e incluye los vértices de la caja. Su forma depende de la intensidad factorial: se curva por encima de la diagonal si x es más intensivo en capital, por debajo si lo es menos y es recta si la intensidad es idéntica. En los vértices, una empresa usa todos los factores y la otra, ninguno.

3.1.3. Curva de transformación

Definición: proyectando la curva de contratos sobre el plano de bienes se obtiene la **curva de transformación**, la máxima cantidad de x producible para cada cantidad de y, dadas la tecnología y las dotaciones. Todos sus puntos son óptimos de Pareto. Los exteriores no son alcanzables, y los interiores son ineficientes, por recursos ociosos o mal repartidos. Desde un punto interior cabría mantener el *output* de un bien y aumentar el del otro, y el progreso técnico podría poner a su alcance los exteriores.



(a) Caja de Edgeworth de la producción



(b) Frontera de posibilidades de producción

Caja de Edgeworth-Bowley de la producción y curva de transformación.

Pendiente: sin externalidades y con productividades marginales positivas, producir más de un bien exige producir menos del otro. La **relación marginal de transformación** mide ese coste de oportunidad; entre dos bienes i y j, $\text{RMT}_{ji} = -dx_j/dx_i$, y con los bienes x e y:

$$\text{RMT}_{yx} = -\left.\frac{dy}{dx}\right|_{K_0,L_0} = \frac{\text{Pmg}_L^y}{\text{Pmg}_L^x} = \frac{\text{Pmg}_K^y}{\text{Pmg}_K^x}$$

Desde la teoría de costes, la RMT es el cociente de costes marginales, $\text{Cmg}_x/\text{Cmg}_y$: al producir más de x y menos de y, el Cmg_x sube y el Cmg_y baja, y la RMT aumenta a lo largo de la curva.

Forma: es cóncava, por tres razones.

- Rendimientos decrecientes a escala.
- Especialización de factores, más adecuados para un bien que para otro, con sustituibilidad imperfecta entre ellos.
- Distinta intensidad factorial de los bienes, incluso con rendimientos constantes. Con factores homogéneos y rendimientos constantes, basta que cada bien requiera proporciones de factores distintas.

Con tecnología convexa y preferencias estrictamente convexas, la tangencia con la curva de indiferencia más alta, $\text{RMT}_{yx} = \text{RMS}_{yx}$, es el equilibrio general de una economía de un solo agente, tipo Robinson Crusoe, o representada por una curva de indiferencia social.

3.1.4. Producción conjunta en sentido técnico

Si un mismo proceso da varios productos, como lana y carne, la tecnología es $F(x, y, \mathbf{z}) = 0$, no separable. Con factores dados, la empresa precio-aceptante maximiza $p_x x + p_y y$ sobre su curva de transformación y elige la mezcla donde $\text{RMT}_{yx} = p_x/p_y$. Con proporciones fijas, $y = \kappa x$, solo elige la escala, y produce hasta que $p_x + \kappa\, p_y$ iguala el coste marginal del proceso (oferta conjunta de Marshall).

3.2. Fronteras de eficiencia

Idea: en la práctica, empresas con la misma tecnología obtienen *outputs* distintos con los mismos *inputs*. El análisis empírico mide la **eficiencia productiva**, la distancia de cada empresa a una frontera no observable, estimada a partir de la mejor práctica.

Farrell (1957): distingue dos componentes, cuyo producto es la eficiencia económica, $\text{EE} = \text{ET} \cdot \text{EA}$, el cociente entre el coste mínimo y el observado.

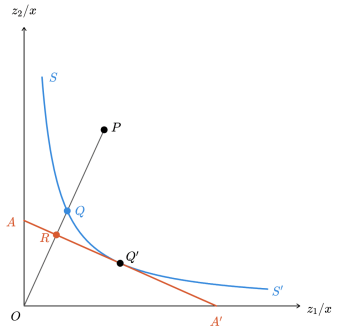
- *Eficiencia técnica*: la máxima reducción radial de los *inputs* compatible con el *output* observado, el menor $\theta \in (0, 1]$ tal que $\theta\mathbf{z} \in V(x)$. Vale uno sobre la isocuanta y no depende de las unidades. Puede orientarse también al *output*, y ambas orientaciones coinciden con rendimientos constantes. En símbolos:

$$\text{ET} = \min\{\theta \in (0, 1] : \theta\mathbf{z} \in V(x)\}$$

- *Eficiencia asignativa*: si la mezcla de factores minimiza el coste a los precios vigentes. Vale uno en la tangencia de la isocuanta con la isocoste.

Desarrollo gráfico: con rendimientos constantes, la tecnología se resume en la isocuanta unitaria SS', en el plano $(z_1/x, z_2/x)$. El radio-vector de la empresa observada, OP, corta la isocuanta en Q y la isocoste AA', tangente en Q', en R:

$$\text{ET} = \frac{OQ}{OP}, \qquad \text{EA} = \frac{OR}{OQ}, \qquad \text{EE} = \frac{OR}{OP} = \text{ET} \cdot \text{EA}$$



Medición de la eficiencia (Farrell, 1957).

Análisis envolvente de datos, DEA (Charnes, Cooper y Rhodes, 1978): método no paramétrico que toma como frontera la menor tecnología que contiene todas las observaciones y cumple libre disposición, convexidad y rendimientos constantes. Un programa lineal da la ET de cada empresa, la menor fracción θ de sus *inputs* con la que una combinación de empresas observadas obtiene al menos su *output*. Si se admiten rendimientos variables (Banker, Charnes y Cooper, 1984), el cociente entre ambas medidas aísla la eficiencia de escala. No exige forma funcional y admite varios *outputs*, pero es determinista, porque atribuye a ineficiencia cualquier desviación, y sensible a los atípicos.

Fronteras estocásticas, SFA (Aigner et al., 1977; Meeusen y van den Broeck, 1977): método paramétrico con un error compuesto de ruido simétrico, v_i, e ineficiencia, $u_i \geq 0$, estimado por máxima verosimilitud:

$$\ln x_i = \ln f(\mathbf{z}_i; \beta) + v_i - u_i$$

La eficiencia técnica de cada empresa, $\exp(-u_i)$, se infiere de la distribución de u_i condicionada al residuo. A diferencia del DEA, separa ruido e ineficiencia y permite la inferencia, a costa de imponer una forma funcional y una distribución a la ineficiencia. La función suele ser Cobb-Douglas o translog, y u_i sigue una distribución asimétrica, seminormal o exponencial.

Con datos de panel, el índice de Malmquist descompone la variación de la productividad en cambio de eficiencia y cambio técnico. Ambos métodos se aplican a hospitales, centros educativos y bancos, y a la competencia por comparación. La ineficiencia-X ofrece una explicación de la distancia a la frontera.

1. Fundamentos de la teoría de costes

La **teoría de costes** estudia cómo se traduce la tecnología en términos monetarios: dados un nivel de producción y unos precios de los factores, qué combinación de factores elige una empresa racional y cuánto le cuesta producir.

1.1. Conjunto de elección y premisas

Coste: la economía lo entiende de forma más amplia que la contabilidad. El coste contable recoge los costes explícitos, el implícito es el coste de oportunidad de los recursos propios (el valor de la mejor alternativa a la que se renuncia) y el **coste económico** suma ambos.

Conjunto de elección: es el mismo que el de la teoría de la producción, pero, como se fija el nivel de producción, se restringe al **conjunto de requerimientos de inputs**, las combinaciones de factores que obtienen al menos x, cuya frontera es la isocuanta:

$$V(x) = \{\mathbf{z} \in \mathbb{R}^m_+ \mid f(\mathbf{z}) \geq x\}$$

Supuestos: se pueden resumir en cuatro premisas.

1. *Entorno*: un único producto, en un marco estático y con certidumbre.
2. *Agentes*: la empresa minimiza el coste de cada nivel de producción, condición necesaria para maximizar el beneficio.
3. *Tecnología*: de buen comportamiento, lo que garantiza isocuantas estrictamente convexas y, si los factores son esenciales, soluciones interiores.
4. *Mercado de factores*: la empresa es precio-aceptante, con el vector $\mathbf{p}^z$ dado.

1.2. Gasto en factores e isocostes

Función objetivo: la empresa ordena sus alternativas por el **gasto en factores**, $C(\mathbf{z}) = \mathbf{p}^z \cdot \mathbf{z}$. Es lineal, homogéneo de grado uno en cantidades y en precios, estrictamente creciente en las cantidades, cardinal, medido en unidades monetarias, y observable.

Rectas isocoste: sus curvas de nivel reúnen las combinaciones de igual gasto, $\sum_i p_i^z z_i = C_0$. Con dos factores, su pendiente es el precio relativo, la tasa externa u objetiva a la que el mercado permite sustituir un factor por otro sin alterar el gasto:

$$\left.\frac{dz_2}{dz_1}\right|_{C_0} = -\frac{p_1^z}{p_2^z}$$

Es constante a lo largo de la recta. Un aumento del gasto la desplaza en paralelo hacia fuera, y un cambio del precio relativo la gira en torno a su corte con el eje del factor cuyo precio no varía.

1.3. Tecnología y eficiencia técnica

Restricción: para producir x, la empresa debe elegir una combinación del conjunto de requerimientos, $f(\mathbf{z}) \geq x$. Las isocuantas reúnen las combinaciones **técnicamente eficientes**, y la empresa se sitúa en su tramo decreciente, la región económica, con productividades marginales positivas.

Eficiencia en costes: la eficiencia técnica deja abierta la decisión. En qué isocuanta producir lo decide la teoría de los mercados, y en qué punto de ella, la teoría de costes con los precios de los factores. Al productor le interesan los puntos que usan más del factor relativamente barato, cuyo precio relativo es menor que su productividad marginal relativa. La eficiencia en costes pide una asignación técnicamente eficiente y, además, de coste mínimo.

1.4. Equilibrio y función de costes

La decisión se plantea como un problema primal, la máxima producción con un gasto dado, o como su dual, el mínimo coste de una producción dada. Sus soluciones, relacionadas por dualidad, dan las demandas de factores y la función de costes.

1.4.1. Primal: maximización de la producción

Planteamiento: la empresa maximiza la producción con un gasto dado en factores:

$$\max_{\mathbf{z}} \; x = f(\mathbf{z}) \quad \text{s.a.:} \quad \sum p_i^z z_i = C_0$$

– El lagrangiano es $\mathcal{L} = f(\mathbf{z}) + \lambda(C_0 - \sum p_i^z z_i)$, y, por el teorema de la envolvente, el multiplicador, $\lambda = \partial\mathcal{L}/\partial C_0 = df^*/dC$, es el producto marginal del coste, la producción adicional de un euro más gastado en factores. Si ambos problemas coinciden, $\lambda = 1/\text{Cmg}$.

CPO: la tasa interna a la que la tecnología permite intercambiar los factores, la RMST, iguala la tasa externa, el precio relativo. Reordenando, la productividad marginal por euro es la misma en todos los factores e igual a λ:

$$\text{RMST}_{ji} = \frac{f_i'}{f_j'} = \frac{p_i^z}{p_j^z} \;\Leftrightarrow\; \frac{f_i'}{p_i^z} = \lambda \quad \forall i$$

CSO: la CPO es suficiente si las isocuantas son estrictamente convexas, porque la restricción es lineal.

Solución: las **demandas marshallianas de factores**, $z_i^m = d_i(\mathbf{p}^z, C)$, son homogéneas de grado cero en precios y coste, agotan el presupuesto, $\sum_i p_i^z z_i^m = C$, y son observables. Respecto a su precio tienen pendiente indeterminada, porque recogen el efecto producción y el efecto sustitución. El valor óptimo del primal es la **función indirecta de producción**, $x^* = \phi(\mathbf{p}^z, C)$, cardinal, creciente en el coste y no creciente en los precios, de la que las demandas se obtienen con una identidad análoga a la de Roy:

$$z_i^m = -\frac{\partial\phi(\mathbf{p}^z, C)/\partial p_i^z}{\partial\phi(\mathbf{p}^z, C)/\partial C}$$

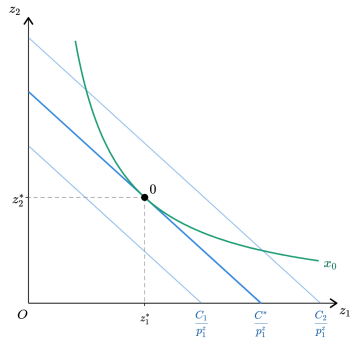
1.4.2. Dual: minimización del coste

Planteamiento: la empresa minimiza el coste de obtener un nivel de producción dado:

$$\min_{\mathbf{z}} \; C = \sum p_i^z z_i \quad \text{s.a.:} \quad f(\mathbf{z}) = x_0$$

CPO y CSO: el multiplicador, $\mu = 1/\lambda$, es el coste marginal. De nuevo, la RMST iguala el precio relativo de los factores, y la isocuanta x_0 es tangente a la isocoste más baja. Cualquier otra combinación técnicamente eficiente sería más cara. La convexidad de las isocuantas garantiza un mínimo. Formalmente:

$$x_0 = f(\mathbf{z}^*), \qquad f_i' = \frac{p_i^z}{\mu} \;\Rightarrow\; \text{RMST}_{ji} = \frac{p_i^z}{p_j^z}$$



Equilibrio de la empresa: tangencia de la isocuanta con la isocoste más baja.

Solución: las **demandas condicionadas de factores**, $z_i^c = d_i^c(\mathbf{p}^z, x_0)$, y, como valor óptimo, la función de costes, $c(\mathbf{p}^z, x_0)$. Las demandas tienen estas propiedades:

– Son homogéneas de grado cero en los precios: una subida equiproporcional no altera la combinación de mínimo coste.
– Su efecto sustitución propio es no positivo, y el cruzado, indeterminado. Al ser condicionadas, no recogen el efecto producción. Con la producción crecen para los factores normales y decrecen para los inferiores.
– Son el gradiente de la función de costes, por el lema de Shephard, y observables, porque la producción que las condiciona es medible, a diferencia de la utilidad en las hicksianas.

1.4.3. Relaciones de dualidad

Los planes óptimos de ambos problemas son idénticos si se intercambia la información, $C_0 = C^*$ o $x_0 = x^*$. Las demandas marshallianas coinciden con las condicionadas evaluadas en la producción máxima, y estas con aquellas evaluadas en el coste mínimo. La función de costes y la indirecta de producción son inversas. En símbolos:

$$d_i^c(\mathbf{p}^z, x_0) = d_i(\mathbf{p}^z, c(\mathbf{p}^z, x_0)), \qquad \phi(\mathbf{p}^z, C) = c^{-1}(\mathbf{p}^z, C)$$

1.4.4. Función de costes y sus curvas

Definición: la **función de costes**, valor del dual, da el coste mínimo de producir x a los precios dados. Análoga a la función de gasto del consumidor, resume toda la información relevante de la tecnología:

$$C^* = c(\mathbf{p}^z, x) = \sum_{i=1}^{m} p_i^z \cdot d_i^c(\mathbf{p}^z, x)$$

Propiedades:

– Es homogénea de grado uno en los precios, estrictamente creciente en la producción y no decreciente en los precios. En símbolos, $c(\lambda\mathbf{p}^z, x) = \lambda \cdot c(\mathbf{p}^z, x)$, y el Cmg y el Cme son también homogéneos de grado uno en precios.
– Es **cóncava en los precios**, es decir, si sube un precio, el coste aumenta menos que con cantidades fijas, por la sustitución. Por eso, la matriz de sustitución es semidefinida negativa y simétrica.
– Por el **lema de Shephard**, sus derivadas respecto a los precios son las demandas condicionadas, $z_i^c = \partial c(\mathbf{p}^z, x)/\partial p_i^z$.
– Por integrabilidad, toda función continua, homogénea de grado uno, no decreciente y cóncava en precios es la función de costes de alguna tecnología. Además, es observable, a diferencia de la función de gasto, que depende de una utilidad no observable.

Coste marginal y medio: el Cmg es la pendiente del coste total, y el Cme, la del rayo desde el origen. El marginal corta al medio en su mínimo.

Economías de escala: son caídas del coste medio al aumentar la escala. Según Marshall, las internas dependen del tamaño de la empresa y dan forma a su curva de coste medio a largo plazo, y las externas dependen del tamaño de la industria y la desplazan. Pueden ser estáticas o dinámicas, y reales o pecuniarias. Las mide la elasticidad-producto del coste, inversa de la elasticidad de escala en la combinación de mínimo coste:

$$\varepsilon_{Cx} = \frac{\text{Cmg}}{\text{Cme}} = \frac{1}{\varepsilon_{x\lambda}}$$

Si es menor que uno, el marginal está por debajo del medio y hay economías de escala; si es mayor, deseconomías. Con precios dados, los rendimientos crecientes equivalen localmente a economías de escala, aunque, si la función no es homotética, puede haber economías de escala sin rendimientos crecientes sobre un radio-vector fijo, porque la empresa cambia la mezcla de factores al crecer.

Precios: la elasticidad del coste respecto al precio de un factor es su participación en el coste, $s_i = p_i^z z_i^c / c$. La **elasticidad de sustitución parcial de Allen** (1938), $\sigma_{ij}^P = c \cdot c_{ij}/(c_i \cdot c_j) = \varepsilon_{z_i p_j}/s_j$, la elasticidad cruzada de la demanda condicionada dividida por la participación de j, mide la respuesta de la demanda de un factor al precio de otro y permite que varíe el empleo de los demás. Es simétrica y clasifica los factores en sustitutivos, si es positiva, y complementarios, si es negativa, caso solo posible con más de dos factores. Con dos, coincide con la de Hicks y es no negativa, $\sigma_{KL}^P = \Delta\%(K/L)/\Delta\%(w/r) \geq 0$.

En suma, la empresa iguala la RMST al precio relativo de los factores tanto si maximiza la producción como si minimiza el coste, y ambos problemas llevan a la misma combinación. La función de costes reúne la información de la tecnología que importa a la empresa.

2. Teoría neoclásica de los costes

El horizonte temporal determina qué factores pueden alterarse: a corto plazo hay algún factor fijo; a largo, todos son variables, y a muy largo, también la tecnología. La estática comparativa separa los cambios del objetivo de producción, que trazan la senda de expansión, de los de los precios de los factores.

2.1. Corto plazo: proporciones variables

Con al menos un factor fijo, hay **costes fijos**, independientes del volumen y, si son irrecuperables, hundidos, y **costes variables**, que dependen de él.

2.1.1. Análisis gráfico

Costes totales: el fijo es una recta horizontal. El variable tiene, con costes cúbicos, forma de S tumbada e invertida, y crece primero a ritmo decreciente y, desde el punto de inflexión, donde la productividad marginal es máxima, a ritmo creciente. El total es el variable desplazado hacia arriba en el coste fijo.

Costes unitarios:

- El coste fijo medio decrece, porque el coste fijo se diluye con la producción.
- El **coste marginal** es inversamente proporcional a la productividad marginal del factor variable, $\text{Cmg} = w/\text{Pmg}_L$, y tiene forma de U: decrece mientras los rendimientos marginales son crecientes y crece después, pasando por los mínimos del coste variable medio y del coste medio.
- El coste variable medio decrece hasta el **mínimo de explotación**, donde se alcanza el óptimo técnico, la productividad media máxima.
- El coste medio, suma del variable medio y del fijo medio, tiene forma de U, y su mínimo es el **óptimo de explotación**, que, por la dilución del coste fijo, está a la derecha del mínimo de explotación.

2.1.2. Limitaciones a la optimalidad

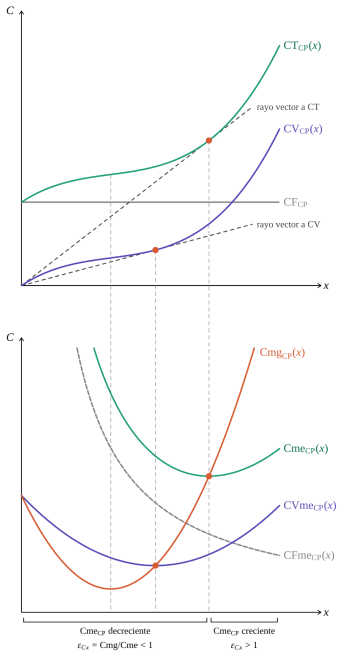
El factor fijo impide cumplir $\text{RMST}_{KL} = w/r$ salvo en la producción para la que la planta es óptima, y la empresa usa combinaciones subóptimas, con exceso o insuficiencia de capital. Los efectos dependen del contrato: con un **arrendamiento**, paga solo las unidades que usa, si el factor es divisible, y mientras no agote la capacidad elige la mezcla como a largo plazo, y con la **propiedad** paga con independencia del uso, de modo que $r\bar{K}$ es un coste fijo.

2.1.3. Estática comparativa

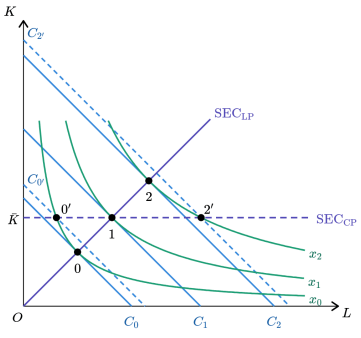
Producción: con el capital fijo, la senda de expansión a corto plazo es la recta $K = \bar{K}$, que solo corta a la de largo plazo donde la planta es óptima. En los demás niveles, el coste a corto plazo supera al de largo. Un cambio del objetivo de producción es un movimiento a lo largo de las curvas. El coste a corto plazo, con $L(x;\bar{K})$ el trabajo que exige la isocuanta de cada nivel, es:

$$\text{CT}_{\text{CP}}(x) = w\,L(x;\bar{K}) + r\bar{K}$$

Precios: los efectos dependen del factor encarecido.



Costes a corto plazo: totales (arriba) y unitarios (abajo).



Factor fijo y condición de tangencia.

- Si suben todos los precios, las curvas se desplazan proporcionalmente y la mezcla no cambia.
- Si sube un **coste fijo** no evitable, como un alquiler o un impuesto de suma fija, suben el coste fijo medio y el medio, pero no el variable medio ni el marginal, de modo que la oferta a corto plazo no cambia y solo cae el beneficio. Un impuesto sobre los beneficios tampoco la altera.
- Si sube un **coste variable**, como el salario o un impuesto específico, suben el variable medio, el medio y el marginal, y la empresa produce menos. Con arrendamiento y capacidad sin agotar, la empresa sustituye trabajo por capital, las curvas suben menos que en proporción y la mezcla cambia según la curvatura de la isocuanta. Con propiedad, el coste variable medio y el marginal se multiplican por w_1/w_0 y la mezcla no cambia.

2.2. Largo plazo: dimensión óptima de planta

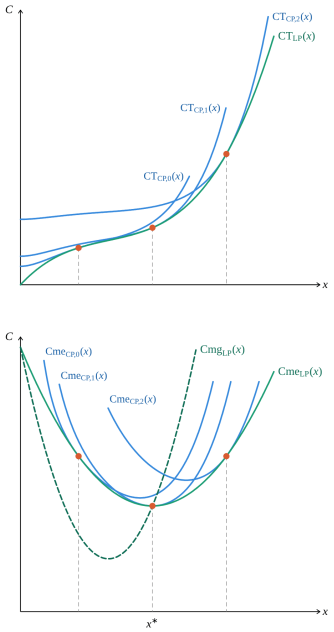
Con todos los factores ajustables, la empresa elige la planta de menor coste. Mientras no ha invertido, los costes a largo plazo son de planificación, *ex ante*, y no hay costes fijos.

2.2.1. Análisis gráfico

Envolventes: como el coste a corto plazo nunca es inferior al de largo, el coste total a largo plazo es la envolvente de los de corto. El coste medio a largo plazo es también la envolvente de los de corto, pero no pasa por sus mínimos salvo en el suyo propio, porque, con su forma de U, cada curva de corto lo toca cuando aún decrece, a la izquierda del mínimo, o cuando ya crece, a la derecha. El marginal a largo plazo no es envolvente, porque el Cmg_{CP} puede ser mayor o menor que él. Se obtiene proyectando en vertical las tangencias entre cada Cme_{CP} y el Cme_{LP}.

Dimensión óptima de planta: se alcanza donde $\text{Cme}_{\text{CP}} = \text{Cmg}_{\text{CP}} = \text{Cmg}_{\text{LP}} = \text{Cme}_{\text{LP}}$, el tamaño que lleva el coste medio a largo plazo a su mínimo, sin economías de escala por aprovechar. Las plantas menores tienen capacidad ociosa, y las mayores trabajan sobrecargadas, con deseconomías de escala. En ella $\varepsilon_{Cx} = 1$,

y la sobrecarga se traduce en horas extra, velocidad antieconómica y averías.



Costes a largo plazo: envolvente y dimensión óptima de planta.

2.2.2. Estática comparativa a largo plazo

Producción: la **senda de expansión del coste** une las tangencias entre isocuantas e isocostes al crecer la producción, con precios relativos constantes. Recorrerla es recorrer la función de costes a largo plazo. Es una recta por el origen si la función es homotética y tiene pendiente negativa si algún factor es regresivo. Un factor es normal si su demanda condicionada crece con la producción e inferior, o regresivo, si decrece, pero no todos pueden ser inferiores, porque el coste mínimo crece con la producción. El factor es normal si $\varepsilon_{z_i x} = (\partial z_i^c/\partial x)(x/z_i^c) > 0$, y con tecnología homotética todas estas elasticidades son iguales a ε_{Cx}. La restricción sobre los inferiores es:

$$\sum_i p_i^z \frac{\partial z_i^c}{\partial x} = \text{Cmg} > 0$$

Precios: si sube el precio de un factor, la isocoste gira y la tangencia se desplaza a lo largo de la isocuanta hacia menos de ese factor, y el coste sube según su participación. Los efectos cruzados son simétricos. Con dos factores, estos son siempre sustitutivos. Un aumento permanente de un coste cuasifijo, independiente del volumen, eleva la curva de coste medio y traslada su mínimo a la derecha; el de un coste variable la eleva con efecto indeterminado sobre su mínimo.

Teorema de Le Châtelier-Samuelson (1947): en la producción para la que la planta es óptima, el coste marginal a corto plazo tiene más pendiente que el de largo, y la oferta a largo plazo es más elástica, porque es tanto más rígida cuantos más factores se mantienen fijos.

2.3. Muy largo plazo: progreso técnico

Idea: las mejoras tecnológicas desplazan las curvas de costes hacia abajo. Con rendimientos constantes y progreso neutral en el sentido de Hicks, $x = A(t)\, f(K, L)$, cada unidad de producto exige $1/A(t)$ unidades de f, y el coste medio cae al ritmo del progreso técnico, $g_{\text{Cme}} = -g_A$:

$$\text{Cme}_{\text{LP}}(t) = \text{Cmg}_{\text{LP}}(t) = \frac{c(w, r)}{A(t)}$$

2.4. Extensiones neoclásicas

Frente a los rendimientos mixtos, los casos polares, con tecnología homogénea de grado h y $c(\mathbf{p}^z, x) = x^{1/h} c(\mathbf{p}^z, 1)$, cambian el perfil del coste a largo plazo, y las economías externas lo trasladan.

Rendimientos constantes: el coste total es una recta por el origen, y el medio y el marginal, una misma recta horizontal, sin economías ni deseconomías de escala, $\varepsilon_{Cx} = 1$.

Rendimientos decrecientes: el coste total es convexo, y el medio y el marginal son crecientes, con el marginal siempre por encima y deseconomías de escala, $\varepsilon_{Cx} > 1$.

Rendimientos crecientes: es el caso que Sraffa (1926) opuso a la forma de U. El coste medio y el marginal decrecen, la regla $p = \text{Cmg}$ genera pérdidas y no se cumple la CSO del problema competitivo, por lo que los rendimientos crecientes son incompatibles con la competencia perfecta y conducen al monopolio natural. El coste total es cóncavo, el medio queda siempre por encima del marginal y $\varepsilon_{Cx} < 1$.

Economías externas (Marshall, 1890): si la oferta de factores no es perfectamente elástica o hay efectos externos tecnológicos, la producción de la industria altera los costes de cada empresa.

- Con **deseconomías externas**, como un factor escaso o la congestión, las curvas suben y la oferta de la industria a largo plazo es creciente.
- Con **economías externas**, como la difusión de conocimientos entre empresas próximas, bajan y la oferta es decreciente. Fundamentan la protección de la industria naciente y la promoción de clústeres.

Permiten una oferta de la industria creciente o decreciente sin renunciar a curvas en U en cada empresa, pero son difíciles de identificar y medir, y Sraffa (1926) las consideró poco frecuentes.

En suma, con rendimientos mixtos hay curvas en U y una dimensión óptima de planta. Pero Sraffa (1926) señaló la incoherencia del esquema con el equilibrio parcial en competencia perfecta: las economías de escala internas solo son compatibles con la competencia imperfecta, en la que la demanda limita el tamaño, respuesta que llegó con Chamberlin (1933) y Robinson (1933).

3. Desarrollos posteriores

Las aportaciones posteriores critican la forma de U y la minimización, relacionan el coste con la empresa y la estructura del mercado, introducen el tiempo y estiman las curvas.

3.1. Críticas a la curva neoclásica

3.1.1. Reserva de capacidad (Stigler, 1939)

Idea: George Stigler cuestiona una dimensión óptima única. La empresa mantiene una **reserva de capacidad** ante imprevistos, como las oscilaciones de la demanda, y por razones técnicas, como las indivisibilidades.

Corto plazo: el coste variable medio tiene forma de bañera, con tres tramos que reflejan la productividad del factor variable.

1. Exceso de capacidad, con productividad media creciente y coste variable medio decreciente. Su elasticidad-producto, $\varepsilon_{\text{CV}x} = \text{Cmg}_{\text{CP}}/\text{CVme}_{\text{CP}}$, es menor que uno.
2. *Reserva de capacidad*: coste variable medio y marginal coinciden y son horizontales, con $\varepsilon_{\text{CV}x} = 1$. Es la zona de uso normal, cuya extensión se fija con el **factor de carga** o coeficiente de aprovechamiento, cociente entre la producción efectiva y la de plena capacidad.
3. Sobrecarga, con el factor fijo plenamente utilizado y coste variable medio creciente. Aquí $\varepsilon_{\text{CV}x} > 1$.

Largo plazo: el coste medio deja de ser la envolvente y pierde su mínimo único. Los costes de producción caen de forma continua, y los administrativos se reducen hasta cierta dimensión y, en las plantas muy grandes, repuntan con mucha lentitud. Si las economías de escala no se agotan, el coste medio baja siempre, en forma de L. Si se agotan, como sostiene Joe Bain, baja hasta la **escala mínima eficiente** y luego es plano, con infinitos puntos de coste mínimo.

Conclusiones: estas formas se parecen más a las curvas observadas, y la reserva de capacidad es una decisión optimizadora, no una ineficiencia. Pero, sin mínimo único, el tamaño de la empresa y el número de empresas quedan indeterminados.

3.1.2. Ineficiencia-X (Leibenstein, 1966)

Idea: Harvey Leibenstein cuestiona que la empresa produzca al mínimo coste. Con información incompleta, contratos incompletos, trabajo penoso y competencia imperfecta, unos mismos *inputs* dan *outputs* distintos. Sin presión de la competencia ni de la supervisión, el esfuerzo cae y se pierden recursos, la **ineficiencia-X**, mayor cuanto más débil es la competencia.

Conclusiones: explica costes distintos con igual tecnología y precios, y añade a la pérdida asignativa del monopolio una pérdida en costes. Se le objeta que, si el control se lleva hasta que su coste marginal iguala su beneficio marginal, no existiría.

3.2. Coste, empresa y estructura de mercado

3.2.1. Coase (1937) y Williamson (1975)

Idea: para Coase, la empresa es una estructura de contratos, y recurrir al mercado tiene **costes de transacción**, que Williamson vincula a la racionalidad limitada, el oportunismo y la especificidad de los activos.

Razonamiento: la dimensión óptima de la empresa se alcanza cuando el coste marginal de organizar internamente una transacción, creciente con el tamaño, iguala el de realizarla en el mercado.

Conclusiones: explica los límites de la empresa y la integración vertical, pero el concepto es amplio y difícil de medir.

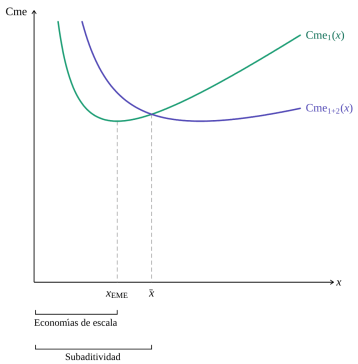
3.2.2. Subaditividad (Baumol, Panzar y Willig, 1982)

Idea: una función de costes es **subaditiva** en x si ningún reparto de x entre varias empresas de igual tecnología es más barato que una sola. Si lo es estrictamente, el mercado es un **monopolio natural**:

$$\text{CT}(x) < \sum_{i=1}^{k} \text{CT}(x_i), \quad \sum_i x_i = x$$

Relación con las economías de escala: en el caso monoproducto, son suficientes, pero no necesarias, y en el multiproducto, ni siquiera suficientes.

- Si el coste medio decrece hasta x, cada parte del reparto tiene un coste medio mayor y la función es subaditiva. En símbolos, $\mathrm{CT}(x) = x \cdot \mathrm{Cme}(x) < \sum_i x_i \cdot \mathrm{Cme}(x_i) = \sum_i \mathrm{CT}(x_i)$.
- Pero puede ser subaditiva aunque el coste medio ya crezca, si la escala mínima eficiente es grande respecto al mercado, porque es una propiedad global, y las economías de escala, local. Gráficamente, el coste medio de dos empresas que se reparten x a partes iguales queda por encima del de una hasta $\bar{x}$, aunque entre x_{EME} y $\bar{x}$ la empresa única esté ya en deseconomías.
- La condición necesaria es alguna no convexidad, como costes fijos o hundidos, indivisibilidades o economías de alcance. Sin ninguna, con costes convexos y $\mathrm{CT}(0) = 0$, la función es superaditiva y la producción se reparte eficientemente entre muchas empresas.



Subaditividad: una empresa es más barata que dos hasta $\bar{x}$.

Conclusiones: define el monopolio natural con una condición exacta, válida para la empresa multiproducto, y da a la regulación un criterio riguroso, pero comprobarla exige conocer el coste de repartos que no se observan.

3.2.3. Economías de alcance (Panzar y Willig, 1977; 1981)

Idea: hay **economías de alcance** cuando producir varios bienes en una empresa cuesta menos que hacerlo en empresas especializadas, y la eficiencia viene de la variedad, no del volumen. Nacen de la dependencia técnica entre productos, de factores indivisibles, de *inputs* públicos internos, como una marca o una red, y de compartir la organización. Su grado mide el ahorro de la producción conjunta:

$$S_C = \frac{\mathrm{CT}(x_1, 0) + \mathrm{CT}(0, x_2) - \mathrm{CT}(x_1, x_2)}{\mathrm{CT}(x_1, x_2)}$$

Ni las economías de escala ni las de alcance implican las otras. Las de alcance son necesarias, pero no suficientes, para la subaditividad multiproducto, y la complementariedad de costes, que aumentar la producción de un bien reduzca el coste marginal del otro, $\partial^2 \mathrm{CT}/\partial x_1 \partial x_2 < 0$, las implica y, con costes fijos comunes y costes marginales propios no crecientes, garantiza la subaditividad. Con $S_C = 0{,}25$, producir por separado costaría un 25 % más.

Conclusiones: explican la diversificación y juzgan la estructura de las industrias reguladas: con S_C alto, separar actividades destruye eficiencia. Pero medirlas exige costes de empresas especializadas que a menudo no existen.

3.2.4. Costes hundidos (Baumol, Panzar y Willig, 1982)

Idea: un coste es **hundido** si no puede recuperarse al abandonar la actividad, como una vía férrea, la publicidad o la I+D, a diferencia del fijo recuperable. Es irrelevante una vez incurrido, pero decisivo antes.

Razonamiento: antes de entrar, es una barrera a la entrada y, para la instalada, de salida. Después, hace creíble el compromiso de permanecer y de luchar. Por eso separa el monopolio natural de la impugnabilidad: sin costes hundidos, la competencia potencial disciplina al monopolio. Con ellos, la regulación recupera su papel.

Conclusiones: precisa qué costes condicionan la entrada y la salida, pero la parte hundida de un coste depende del horizonte y de los mercados de segunda mano, de modo que la distinción es de grado.

3.3. Coste y tiempo

3.3.1. Aprendizaje (Wright, 1936; Henderson, 1968)

Idea: Wright observó en la aeronáutica que las horas por avión caían con la producción acumulada, y Henderson y el Boston Consulting Group lo generalizaron como **curva de experiencia**. Las economías de aprendizaje, o de escala dinámicas, ligan el coste medio a la producción acumulada, con b la elasticidad de aprendizaje:

$$\mathrm{Cme}_t = a \cdot (x_t^{\mathrm{ac}})^{-b}$$

Producir hoy abarata el futuro, por lo que la empresa produce más de lo que igualaría el ingreso y el coste marginales actuales, y el primero en entrar gana una ventaja acumulativa. Explica los precios de penetración y las barreras basadas en la experiencia, el aprendizaje por la práctica de Arrow (1962) en el crecimiento endógeno y el argumento de la industria naciente. Es una regularidad bien documentada, pero empírica: no explica la tasa de aprendizaje y cuesta separarla de las economías de escala estáticas y del progreso técnico. Cada vez que se duplica la producción acumulada, el coste medio cae en la proporción $1 - 2^{-b}$, la tasa de aprendizaje.

3.3.2. Costes de ajuste

Idea: son los costes de alterar la velocidad a la que se incorporan los factores. Eisner y Strotz (1963) introducen los internos, como las mermas de producción al cambiar máquinas, que Lucas (1967) formaliza, y la literatura posterior añade los externos, como las indemnizaciones o la selección de personal.

Forma: la función de costes de ajuste suele suponerse convexa, como $\psi(\Delta z) = \frac{\gamma}{2}(\Delta z)^2$, aunque puede ser asimétrica, lineal o fija. Los factores variables y fijos tradicionales son los casos de costes de ajuste nulos o infinitos, y con costes convexos el ajuste se reparte en varios periodos. Fundamentan la teoría q de la inversión, aunque la evidencia de ajustes a saltos apunta a costes fijos o lineales. La lineal es $\psi(\Delta z) = \gamma \cdot |\Delta z|$, con un ajuste que puede ser inmediato, y la fija, $\psi = \bar{\psi}$ si $\Delta z \neq 0$, como un expediente de regulación de empleo.

3.4. Aplicaciones empíricas

La forma de las curvas de coste es, en último término, una cuestión empírica, de la que dependen el número de empresas que caben en un mercado, la existencia de monopolios naturales y el diseño de la regulación.

3.4.1. La función de costes estimada

Idea: Nerlove (1963) estima una función de costes deducida por dualidad, con los precios de los factores como explicativas, para 145 eléctricas de EE. UU. en 1955. En una industria regulada la producción y los precios son exógenos, y con una Cobb-Douglas el coeficiente de $\ln x$ es $1/h$.

Resultados: con toda la muestra halla rendimientos crecientes, que por grupos de tamaño caen con la escala, de modo que la Cobb-Douglas era demasiado rígida. La translog (Christensen et al., 1973) deja libres los rendimientos y las elasticidades de sustitución, y con ella Christensen y Greene (1976) muestran que en 1970 el grueso de la electricidad se generaba en el tramo plano del coste medio.

3.4.2. Fronteras de costes

Idea: la función estimada supone que todas las empresas minimizan el coste. Las fronteras abandonan ese supuesto, de modo que la función de costes es el mínimo y cada empresa puede quedar por encima. La distancia es la eficiencia económica de Farrell, que mide la ineficiencia-X.

Métodos: son los de la producción, ahora desde el coste. El **SFA** añade a la función de costes el error compuesto, $\ln C_i = \ln c(\mathbf{p}_i^z, x_i) + v_i + u_i$, con la ineficiencia $u_i \geq 0$ sumando, porque encarece. El **DEA** calcula con un programa lineal el coste mínimo de cada empresa a los precios que paga.

Conclusiones: miden cuánto podría ahorrar cada empresa y sirven a la regulación por comparación, pero no siempre ordenan igual a las mismas empresas.

3.4.3. Evidencia y problemas de medición

Evidencia: el coste medio a largo plazo cae con fuerza en escalas pequeñas y luego es casi horizontal, en forma de L más que de U. La escala mínima eficiente suele ser pequeña respecto al mercado, salvo en las industrias de red y de grandes costes fijos.

Cautelas: la **falacia de la regresión** (Friedman, 1955) hace parecer decreciente la curva, porque las empresas que producen por encima de lo normal muestran costes medios bajos. Cada observación es un punto de una curva de corto plazo, y los costes contables omiten el coste de oportunidad del capital propio.

1. Teoría de la agencia: el principal-agente

La **teoría de la agencia** (Ross, 1973; Jensen y Meckling, 1976) estudia las relaciones en las que una parte, el principal, encarga a otra, el agente, que actúe en su nombre cuando sus intereses no coinciden. Su objeto es el diseño del **contrato óptimo**, el que mejor alinea ambos intereses.

1.1. Elementos del problema principal-agente

Agentes: la parte contratante es el principal, y la contratista, el agente. Pueden ser individuos, instituciones o centros de decisión, como Estado y ciudadanía, accionista y directivo, empleador y empleado o paciente y médico.

Negociación: se desarrolla con un reparto dado del poder negociador, en general de la forma "lo tomas o lo dejas", sin contraofertas, o de regateo. En el primero, una parte tiene todo el poder negociador, y en el regateo ambas lo tienen y pueden hacer contraofertas.

Contrato: fija, en cada estado de la naturaleza, qué corresponde hacer a cada parte y qué retribución recibe el agente por su esfuerzo. Solo se basa en variables **verificables** por un árbitro independiente. Si no lo fueran, las partes anticiparían que nadie puede comprobar su cumplimiento y no lo firmarían.

Aceptación: el agente acepta si la utilidad de participar no es inferior a su **utilidad de reserva**, la de sus oportunidades alternativas.

Conflicto de interés: ambos se mueven por el interés propio, y el coste de uno, como el salario, es el ingreso del otro. Da lugar a **costes de agencia**, la destrucción de valor por la falta de alineación entre las partes.

1.2. Modelo base: información simétrica

Idea: en el modelo canónico (Ross, 1973; Holmström, 1979), el pago del agente depende del reparto óptimo del riesgo, que se condensa en el contrato.

1.2.1. Supuestos

Entorno: información imperfecta pero simétrica, es decir, nadie conoce de antemano el estado de la naturaleza, pero ambos observan lo mismo. El resultado es una variable aleatoria verificable, $x_i = f(e, s_i)$, que depende del esfuerzo del agente, también verificable, y del estado de la naturaleza. De forma equivalente, el conjunto de resultados está dado y el esfuerzo actúa sobre sus probabilidades, $p_i(e)$. Un mayor esfuerzo desplaza la distribución hacia los resultados altos. Las probabilidades son positivas y suman uno, $\sum_i p_i(e) = 1$.

Agentes:

- *Principal*: diseña el contrato, el salario para cada resultado, $w(x_i)$, y el esfuerzo exigido, e, y maximiza la utilidad esperada de su beneficio, $\sum_i p_i(e) \cdot \pi[x_i - w(x_i)]$, con $\pi' > 0$ y $\pi'' \leq 0$. El caso $\pi'' = 0$, el del principal neutral al riesgo, sirve de referencia.
- *Agente*: si participa, aporta el esfuerzo fijado y obtiene $\sum_i p_i(e) \cdot u[w(x_i)] - v(e)$, con u creciente y cóncava y una desutilidad del esfuerzo, v, creciente. La desutilidad es convexa, $v'' \geq 0$, y la utilidad es aditivamente separable, de modo que la aversión al riesgo no varía con el esfuerzo.
- *Conflicto*: para el principal, el resultado es su ganancia y el salario, su coste, mientras que para el agente el salario es su ganancia y el esfuerzo, su coste. El contrato es el medio que hace compatibles ambos intereses. El esfuerzo no es un coste para el principal ni el resultado una ganancia para el agente, pero el esfuerzo mejora la probabilidad de un buen resultado.

1.2.2. Modelo

Planteamiento: el principal maximiza la utilidad esperada de su beneficio sujeto a la **condición de participación** del agente:

$$\max_{e,\{w(x_i)\}} \sum_i p_i(e) \cdot \pi[x_i - w(x_i)] \quad \text{s.a.:} \quad \sum_i p_i(e) \cdot u[w(x_i)] - v(e) \geq \underline{u}$$

Equivale a resolverlo en dos etapas, buscando primero, para cada esfuerzo, el contrato aceptable que da al principal mayor utilidad esperada (el más barato, si es neutral al riesgo) y, después, el esfuerzo que maximiza esa utilidad.

CPO: se obtienen tres condiciones.

1. *Participación*: la restricción se satura, y el agente recibe exactamente su utilidad de reserva, porque en caso contrario el principal podría pagarle menos y el agente aún aceptaría. La restricción se cumple con igualdad:
$$\sum_i p_i(e) \cdot u[w(x_i)] - v(e) = \underline{u}$$
2. *Reparto del riesgo*: derivando respecto al salario de cada estado, el cociente de utilidades marginales del principal y del agente es constante en todos los estados e igual al multiplicador de la participación.
3. *Esfuerzo*: el principal lo demanda hasta que su ganancia marginal iguala la compensación que el agente exige por la desutilidad marginal. Como es verificable, lo exige directamente. La condición resulta de derivar respecto al esfuerzo:
$$\sum_i p_i'(e) \cdot \pi[x_i - w(x_i)] + \lambda \left[\sum_i p_i'(e) \cdot u[w(x_i)] - v'(e) \right] = 0$$

Condición de Borch: el cociente es λ en todos los estados y, para dos estados cualesquiera, λ se cancela:

$$\frac{\pi'[x_i - w(x_i)]}{u'[w(x_i)]} = \lambda \quad \forall i$$
$$\frac{\pi'(x_i - w_i)}{\pi'(x_j - w_j)} = \frac{u'(w_i)}{u'(w_j)} \quad \forall i, j$$

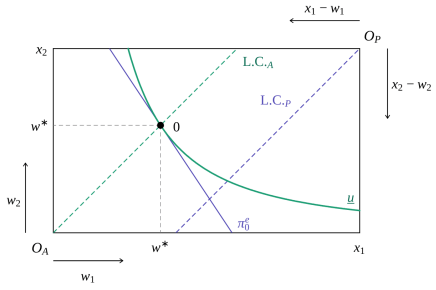
La RMS entre rentas contingentes es, por tanto, la misma para principal y agente. La condición caracteriza las asignaciones Pareto-eficientes bajo incertidumbre, la curva de contratos de la caja de Edgeworth de estados.

CSO: con funciones de beneficio y utilidad cóncavas, una de ellas estrictamente, la solución es única e interior.

Solución: el contrato óptimo, $\{e^*, w^*(x_i)\}$, equivale a la función de demanda del consumidor, y su parámetro decisivo son las actitudes ante el riesgo de las partes. Como la demanda, recoge cómo responde la decisión, aquí el reparto de rentas entre estados, a los parámetros del problema.

- **Caso 1**, principal neutral y agente averso: con π' constante, u' también debe serlo, y el salario no depende del resultado. El principal asume todo el riesgo y paga un **salario fijo**, $w^* = u^{-1}[\underline{u} + v(e^*)]$, que compensa la utilidad de reserva y la desutilidad del esfuerzo.
- **Caso 2**, principal averso y agente neutral: el beneficio del principal debe ser el mismo en todos los estados. El agente asume todo el riesgo en un **contrato de franquicia**, $w^* = x_i - k$: el principal cede su saber hacer y su marca a cambio de un canon fijo, k. El principal recibe así una cantidad fija con independencia del resultado, y el agente se queda con el resultado neto.
- **Caso 3**, ambos aversos: los pagos varían con el resultado según los grados de aversión, $dw^*/dx_i = \mathrm{CAA}_P/(\mathrm{CAA}_P + \mathrm{CAA}_A)$, lineales con utilidades CARA. Ambos soportan riesgo, y el más averso es el mejor cubierto, como en una *joint venture* o una remuneración con parte fija y variable. El esquema es $w^* = a + b \cdot x_i$, con $0 < b < 1$, y los casos 1 y 2 son los extremos en que uno de los coeficientes de aversión absoluta es nulo, con pendiente 0 en el salario fijo y 1 en la franquicia.

Desarrollo gráfico: en la caja de Edgeworth de estados, cuyos lados son los dos resultados posibles, el agente mide sus salarios desde O_A y el principal sus rentas netas desde O_P. Desde cada origen sale su línea de certeza, y la curvatura de las curvas de indiferencia refleja la aversión de cada parte (rectas si es neutral y convexas si es averso). Solo son aceptables los contratos sobre o por encima de la curva de reserva del agente, y el principal elige, entre los contratos que el agente acepta, el de su curva de indiferencia más alejada, tangente a la de reserva del agente. La tangencia está sobre la línea de certeza del averso si el aseguramiento es completo, casos 1 y 2, y entre ambas líneas si es parcial. En el caso 1, que recoge la figura, el óptimo es el punto 0, con $w_1^* = w_2^* = w^*$, un plan de pagos contingente que paga lo mismo con cualquier resultado.



Principal-agente con información simétrica: el principal neutral asegura al agente averso (0).

Estática comparativa: un aumento de la utilidad de reserva desplaza el contrato a lo largo de la curva de contratos, en favor del agente, y una mayor aversión de una parte lleva la tangencia hacia su línea de certeza, con más aseguramiento para ella. La primera perturbación desplaza la restricción de participación hacia O_P, y la segunda aumenta la curvatura de las curvas de indiferencia de esa parte.

1.2.3. Conclusiones

Aspectos positivos: el modelo base es la referencia de eficiencia. El contrato reparte el riesgo según los perfiles de las partes y no hay pérdida de eficiencia, de modo que toda ineficiencia posterior es imputable a la asimetría de información.

Limitaciones: la observabilidad del esfuerzo y de los tipos falla justo en las relaciones relevantes, como el empleo, los seguros, el crédito o la

regulación. El modelo es, por ello, un punto de partida y no una descripción, y el riesgo moral y la selección adversa relajan, respectivamente, la observabilidad del esfuerzo y la de los tipos.

1.3. Aplicaciones

El marco es transversal: aparece cada vez que una parte actúa por cuenta de otra con objetivos distintos y bajo incertidumbre.

- *Asegurador-asegurado*: el asegurador desconoce el riesgo del asegurado y su conducta una vez asegurado.
- *Accionista-directivo*: la separación entre propiedad y control, origen de la teoría (Jensen y Meckling, 1976) y del gobierno corporativo.
- *Regulador-empresa*: la regulación de monopolios naturales cuyos costes el regulador desconoce.
- *Empleador-empleado*: los contratos de trabajo, los incentivos y los salarios de eficiencia.
- *Prestamista-prestatario*: el racionamiento del crédito y los costes de agencia de la deuda.
- **Estado-ciudadano** y organismos internacionales-Estados: la condicionalidad de la ayuda y de los préstamos, y el diseño de las políticas públicas cuando el Estado no observa las características de los beneficiarios.

En todos falla la información simétrica: si el principal no observa lo que el agente **hace**, hay riesgo moral, y si no observa lo que el agente **es**, hay selección adversa.

En suma, toda relación de agencia es un contrato con restricciones de participación y, con información asimétrica, de incentivos. El marco supone contratos verificables y exigibles sin coste, racionalidad y una relación bilateral, supuestos que relajan los contratos incompletos (Grossman-Hart-Moore) y los relacionales.

2. Riesgo moral (Arrow, 1963; Pauly, 1968)

La primera desviación respecto de la simetría surge cuando el principal no observa lo que el agente hace una vez firmado el contrato.

2.1. Concepto de riesgo moral

Origen: el término procede del sector asegurador del siglo XIX, y se populariza en economía con Arrow (1963).

Definición: hay **riesgo moral** cuando, después de firmar el contrato, el esfuerzo del agente no es observable ni verificable. El salario fijo, óptimo con información simétrica en el caso 1, incentiva entonces un esfuerzo bajo, $e_L < e_H$, porque el esfuerzo cuesta y el salario no depende de él. El contrato ya no puede **exigir** el esfuerzo, sino que debe **inducirlo**, haciendo que el esfuerzo que desea el principal coincida con el que el agente elige por su cuenta.

2.2. Modelo: esfuerzo no observable

Origen: Arrow (1963) y Pauly (1968) plantean el problema en los seguros, y Bengt Holmström (1979) formaliza el contrato óptimo con esfuerzo no observable.

2.2.1. Supuestos

Entorno: se mantienen los del modelo base, salvo uno: el principal ya no observa el esfuerzo, solo el resultado. El esfuerzo toma dos valores, alto y bajo, y el principal desea inducir el alto con el esquema de pagos.

2.2.2. Modelo

Planteamiento: el principal elige ahora solo los salarios, con la participación evaluada con el esfuerzo alto y una segunda **restricción de incentivos**, con multiplicador μ:

$$\max_{\{w(x_i)\}} \sum_i p_i(e_H)\cdot\pi[x_i - w(x_i)]$$

$$\text{s.a.:}\begin{cases}\sum_i p_i(e_H)\cdot u[w(x_i)] - v(e_H) \geq \underline{u}\\ \sum_i [p_i(e_H) - p_i(e_L)]\cdot u[w(x_i)] \geq v(e_H) - v(e_L)\end{cases}$$

Esta exige que el esfuerzo alto dé al agente una utilidad esperada no inferior a la del bajo.

CPO: derivando respecto al salario de cada estado se obtiene la **condición de Holmström** (1979): el salario de cada estado depende del cociente de verosimilitud, que informa sobre el esfuerzo a partir del resultado observado:

$$\frac{\pi'[x_i - w(x_i)]}{u'[w(x_i)]} = \lambda + \mu\left(1 - \frac{p_i(e_L)}{p_i(e_H)}\right)$$

Se diferencia de la de Borch en dos elementos.

1. **Coste del riesgo moral**, $\mu > 0$: mide cuánto ganaría el principal si se relajara la restricción de incentivos. En el caso 1 es positivo, porque, si fuera nulo, se volvería al salario fijo, que induce el esfuerzo bajo y viola la restricción. La restricción de incentivos es, por tanto, activa, y el principal obtiene un beneficio menor que con información simétrica.
2. **Cociente de verosimilitud**, $p_i(e_L)/p_i(e_H)$: cuanto menor es, más indica x_i un esfuerzo alto y mayor es el salario en ese estado. El principal condiciona el pago a toda señal informativa sobre el esfuerzo, y el salario crece con el resultado si el cociente es monótono decreciente. En el caso 1 el efecto es inmediato, porque π' es constante y un mayor lado derecho exige una menor u', es decir, un salario mayor.

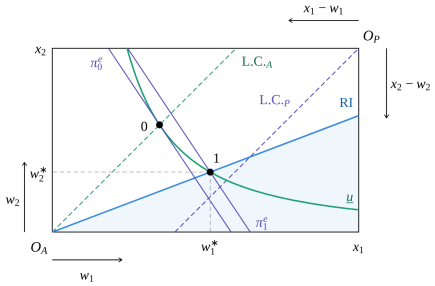
La participación sigue saturada. Como el contrato traslada riesgo al agente, el principal debe pagarle una prima de riesgo, un salario esperado superior al de información simétrica, que no mejora al agente pero reduce el beneficio del principal, y esa caída es el coste del riesgo moral.

CSO: con funciones cóncavas, las CPO caracterizan el óptimo, y el principal compara además el beneficio de inducir el esfuerzo alto con el de conformarse con el bajo a salario fijo y elige el mayor.

Solución: el esfuerzo alto se induce en lugar de exigirse. Si principal y agente persiguieran lo mismo, el contrato de información simétrica ya cumpliría la restricción de incentivos y el riesgo moral no tendría coste. Con intereses enfrentados, los casos cambian así:

- En el caso 1, el salario ya no puede ser fijo. El agente soporta parte del riesgo aunque sea averso, porque solo así el salario transmite incentivos, y de ahí procede el conflicto entre **aseguramiento e incentivos**.
- En el caso 2, la franquicia sigue siendo óptima y no hay pérdida de eficiencia, porque el agente neutral internaliza plenamente los incentivos, siempre que no tenga responsabilidad limitada. Así puede aceptar pagos negativos, y el principal obtiene el mismo beneficio que con información simétrica.
- En el caso 3, el salario combina el reparto del riesgo de Borch con la prima de incentivos y es más sensible al resultado que con información simétrica. Ocurre si el contrato de información simétrica, ya variable, no cumple la restricción de incentivos, y el agente soporta entonces más riesgo que en el modelo base.

Desarrollo gráfico: la restricción de incentivos excluye los contratos con salario igual en ambos estados. El contrato ya no puede situarse en el punto 0, sobre la línea de certeza del agente, y se desplaza por su curva de reserva hasta el punto 1, en la frontera de la restricción, con menor beneficio para el principal. En él, con $w_1^* \neq w_2^*$, el principal renuncia a parte del reparto eficiente del riesgo a cambio de inducir el esfuerzo.



Riesgo moral: la restricción de incentivos lleva el contrato de 0 a 1.

2.2.3. Conclusiones

Aspectos positivos: la ineficiencia no se debe a la asimetría por sí misma, sino a su combinación con el conflicto de interés y la aversión al riesgo, y el modelo da el principio de todo contrato de incentivos: usar cualquier señal informativa sobre el esfuerzo. Con la restricción activa y un agente averso, el pago es variable y se llega a un *second best*. El agente no puede quedar tan asegurado como en el reparto eficiente del riesgo.

Limitaciones: el coste del riesgo moral depende de actitudes ante el riesgo no siempre contrastables, y el modelo es estático y con una tarea. Con relaciones repetidas o varias tareas (Holmström y Milgrom, 1991), el contrato cambia sustancialmente.

2.3. Aplicaciones del riesgo moral

Todas tienen la misma estructura: una parte no observa la conducta de la otra tras el contrato, y el remedio es trasladarle parte del riesgo o condicionar los pagos a resultados observables.

- *Seguros*: el asegurado no tiene incentivos a prevenir el siniestro, y la póliza cubre solo una parte del daño, con franquicias o copagos. Como prescribe el modelo, se traslada así al asegurado parte del riesgo, que paga el primer tramo del daño o recibe solo una fracción del valor del bien.
- *Gobierno corporativo*: los directivos pueden perseguir tamaño o prestigio, y su remuneración se liga a los resultados, con opciones o primas, y a la supervisión. Ocurre cuando propiedad y gestión están separadas, y esos fines se anteponen entonces al valor de la empresa.
- *Trabajo*: si el escaqueo solo se detecta con cierta probabilidad, se paga un salario superior al de vaciado para que el despido sea costoso. El modelo de Shapiro y Stiglitz (1984) explica así el desempleo involuntario como dispositivo de disciplina.

- *Crédito*: tras una subida del tipo de interés, el prestatario tiene incentivos a elegir proyectos más arriesgados, y el prestamista prefiere racionar el crédito (Stiglitz y Weiss, 1981). Esos proyectos rinden más en caso de éxito, con lo que permiten afrontar el mayor coste, pero elevan la morosidad esperada.
- *Rescates*: el país rescatado no tiene incentivos a corregir su política, y la respuesta es la condicionalidad.

2.4. Soluciones al riesgo moral

Todas reducen la asimetría o hacen que el agente soporte las consecuencias de su conducta.

- *Toda la información disponible*: por el resultado de suficiencia estadística (Holmström, 1979), es valiosa cualquier señal que informe sobre el esfuerzo además del resultado, como el estado de la naturaleza, y solo esa, porque permite estimar mejor el esfuerzo y reducir el riesgo que soporta el agente.
- **Competencia por comparación** (*yardstick competition*): los incentivos dependen del rendimiento frente al promedio, lo que filtra las perturbaciones comunes, como en los torneos de promoción. El promedio puede ser el del mercado, el de la propia empresa en otros periodos o el de otras empresas.
- *Supervisión y control*: auditorías e indicadores, hasta que su coste marginal iguala la reducción marginal del coste de agencia. Reducen la asimetría y fomentan que el agente alinee su conducta con los intereses del principal.
- *Castigos*: el despido, la pérdida de reputación o las penalizaciones, eficaces si son creíbles y el agente tiene algo que perder, lo que conecta con los salarios de eficiencia.
- *Regulación*: tras la Gran Recesión, el riesgo moral del *too big to fail* se combate con colchones de capital, *bail-in* y mecanismos de resolución. Las garantías implícitas de rescate público fomentaban la toma excesiva de riesgos, y estos instrumentos devuelven a la entidad y a sus acreedores las consecuencias de sus decisiones.
- *Separación de funciones y condicionalidad*: como la independencia de la autoridad monetaria frente a la monetización del déficit, o la condicionalidad de la ayuda oficial y de los rescates. La financiación en condiciones favorables a países con instituciones débiles genera riesgo moral, y la condicionalidad subordina los desembolsos al cumplimiento de reformas.

En suma, el riesgo moral explica la forma de los contratos reales, como las remuneraciones variables, los copagos o la condicionalidad, y su principio de diseño. Pero el modelo es bilateral y estático. Con terceros, los accionistas pueden remunerar a los directivos para que los competidores produzcan menos o, si compiten en precios, para que los fijen más altos, y el contrato de agencia se vuelve así instrumento de competencia. Con el tiempo, la reputación sustituye en parte a los incentivos explícitos.

3. Selección adversa (Akerlof, 1970)

Si lo que el principal no observa es lo que el agente es, y no lo que hace, la asimetría existe ya antes del contrato.

3.1. Concepto de selección adversa

Definición: hay **selección adversa** cuando el agente sabe, antes de firmar, algo que el principal ignora y que pesa en el resultado del intercambio. Requiere heterogeneidad de la contraparte y conflicto de interés, y nace de que el mercado no permite diferenciar unos bienes de otros o de que una parte oculte lo que sabe.

Consecuencia: la peor calidad desplaza a la mejor, hasta poder hacer desaparecer segmentos del mercado o el mercado entero.

3.2. El mercado de "limones" (Akerlof, 1970)

Origen: Akerlof (*The Market for "Lemons"*, 1970; Nobel 2001) toma el ejemplo de los coches usados: una proporción α es de mala calidad y el resto, $1-\alpha$, de buena.

Idea: quien compra sin observar la calidad solo paga lo que vale un coche medio, y eso expulsa del mercado a los coches buenos hasta que solo se venden los malos, los "limones".

3.2.1. Supuestos

Entorno: la información sobre la calidad está repartida de forma asimétrica: la tiene quien vende y no quien compra.

3.2.2. Modelo

Planteamiento: con calidad observable, cada coche se vendería por su valor para el comprador, p_A el bueno y p_B el malo, con $p_A > p_B$.

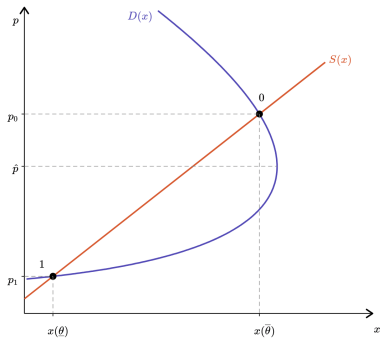
- *Demanda*: como no sabe cuál le venden, el comprador paga solo la media ponderada:

$$p_C = (1-\alpha)\,p_A + \alpha\,p_B$$

- *Oferta*: el dueño de un coche bueno exige un precio mínimo, p_A^V, más alto que el de los dueños de limones. Si supera esa media, $p_A^V > p_C$, no le compensa vender y lo retira, mientras los limones se siguen ofreciendo.

Equilibrio: si $p_A^V > p_C$, el mercado pasa de la situación inicial al equilibrio final en cuatro pasos. Primero, los compradores ofrecen p_C por la calidad media. Después, los dueños de coches buenos los retiran. Los compradores reciben entonces limones y, por la regla de Bayes, revisan a cero la probabilidad de que un coche en venta sea bueno. Por último, el precio cae a p_B, al que solo se venden limones. Los vendedores de limones rebajan así la calidad media percibida y el precio de todos los coches, una externalidad negativa.

Desarrollo gráfico: con un continuo de calidades, la oferta es creciente, porque cada vendedor solo ofrece su coche si el precio supera el valor que le atribuye, y los dueños de coches mejores los valoran más. Al subir el precio entran así coches mejores y sube la calidad media. La demanda tiene dos pendientes y se "dobla" en $\hat{p}$. Por encima domina el efecto precio y la cantidad demandada decrece con el precio, y por debajo domina el efecto calidad y crece. La oferta la corta en dos equilibrios, el 0, al precio p_0, con $x(\overline{\theta})$ coches de calidad media alta, $\overline{\theta}$, y el 1, al precio p_1, con solo $x(\underline{\theta})$ limones, de calidad $\underline{\theta}$. Los dos se sostienen por sí mismos, y cuál se alcanza depende de las expectativas, porque una revisión a la baja de la calidad esperada retira los mejores coches y puede llevar el mercado al 1. La espiral endógena de expulsión es propia del caso de dos calidades.



Mercado de limones: la demanda se dobla en $\hat{p}$ y hay dos equilibrios.

Bienestar: con información simétrica, cada calidad tiene su mercado y el resultado es eficiente. Con asimetría, el equilibrio no es un óptimo de Pareto, porque los coches buenos podrían venderse a un precio entre p_A^V y p_A, y ese excedente perdido mide la ineficiencia. Esos intercambios mutuamente beneficiosos no se realizan porque la asimetría impide que el precio los señale.

3.2.3. Conclusiones

Aspectos positivos: es aplicable al trabajo con trabajadores heterogéneos y a los seguros con clientes heterogéneos.

Limitaciones: supone una calidad exógena, y con garantías, reputación o inspección muchos mercados escapan al colapso total, que es un caso extremo. Supone además que los compradores solo conocen la distribución de las calidades.

3.3. Aplicaciones de la selección adversa

El propio Akerlof (1970) la extendió al seguro médico de los mayores de 65 años y al crédito en los países en desarrollo. Todas siguen la estructura de los limones: un precio basado en la calidad media expulsa a los mejores tipos, y el remedio es hacer observable el tipo o lograr que se revele.

- *Seguros*: con una prima única, salen los de bajo riesgo, la prima sube y el proceso se repite, la **espiral de selección adversa**. La respuesta es el seguro obligatorio, que mantiene a todos los tipos en el mismo fondo, como la responsabilidad civil del automóvil o la sanidad pública, o un menú de pólizas que los separa.
- *Crédito*: si sube el tipo, salen primero los prestatarios con proyectos más seguros, cuyo beneficio esperado con responsabilidad limitada es menor, y empeora la cartera. Este efecto selección se suma al de incentivos y completa el racionamiento del crédito de Stiglitz y Weiss (1981). Las garantías sirven para autoseleccionarse, porque el prestatario de bajo riesgo, que tiene menos probabilidad de perderlas, acepta más colateral a cambio de un tipo menor.
- *Financiación de la empresa*: los directivos solo emiten acciones sobrevaloradas, de modo que una ampliación de capital es una mala noticia, y se prefiere la autofinanciación a la deuda y esta a las acciones (Myers y Majluf, 1984).
- *Trabajo*: el salario de reserva crece con la productividad, y si baja el salario se van primero los más productivos, por lo que el empleador puede pagar salarios de eficiencia por selección adversa. Los periodos de prueba y la educación como señal reducen la asimetría.

3.4. Soluciones: señalización y autoselección

Spence (1973), con la señalización, y Rothschild y Stiglitz (1976), con la autoselección, dan dos soluciones.

1. En la **señalización**, el agente toma la iniciativa y transmite su tipo

con una acción observable y costosa. Funciona como señal cualquier rasgo visible que el agente elige y cuya adquisición le resulta costosa.
2. En la **autoselección** o filtro, el principal diseña un menú de contratos en el que cada agente revela su tipo al elegir. Por el **principio de revelación** de Myerson (1979), basta considerar menús en los que a cada agente le interesa decir la verdad. El principal puede limitarse a ellos sin pérdida de generalidad, y los agentes dicen la verdad cuando es su mejor estrategia.

Ambas vías mejoran la circulación de la información para que las dos partes salgan ganando.

3.4.1. Señalización (Spence, 1973)

Idea: ¿por qué tantas ofertas de empleo piden un título universitario sin especificar cuál? Para Michael Spence (Nobel 2001), la educación demuestra la capacidad de aprendizaje y es una señal de las cualidades del trabajador.

Supuestos: en un mercado de trabajo hay trabajadores de alta productividad, A, y baja, B, igual de numerosos, y el empleador no observa el tipo. Sin señales, paga un salario medio, de modo que A cobra menos de lo que produce y B, más. La educación, h, no altera la productividad, pero es más cara para B. Si la alterase, un trabajador B podría convertirse en A. El coste puede ser financiero, de oportunidad o en desutilidad, y la educación se aproxima por los años de escolarización o los títulos obtenidos.

Planteamiento: cada trabajador elige educarse, $h = h^*$, o no, porque la señal vale lo mismo para cualquier nivel superior al umbral y su coste crece con h. Las empresas tratan como A, con salario w_A, a quien alcanza el umbral, y como B, con $w_B < w_A$, a quien no. La regla salarial, basada en sus creencias previas, es:

$$w(h) = \begin{cases} w_A & \text{si } h \geq h^* \\ w_B & \text{si } h < h^* \end{cases}$$

Restricciones de incentivos: la señal separa si a A le compensa educarse y a B no:

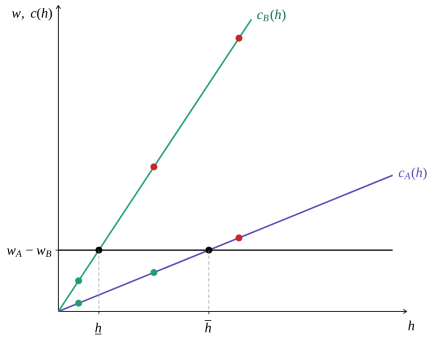
$$w_A - c_A(h^*) \geq w_B$$
$$w_A - c_B(h^*) < w_B$$

La de A pide que su salario neto del coste de educarse no quede por debajo del que obtendría sin educarse, y la de B, que educarse le deje estrictamente menos que no educarse.

Equilibrio: la condición necesaria para el equilibrio separador es $c_A(h^*) < c_B(h^*)$, pero hay equilibrios múltiples.

- **Agrupador con educación**, si el diferencial salarial cubre el coste de B, $w_A - w_B \geq c_B(h^*)$: ambos se educan. La señal no distingue, y con creencias consistentes todos cobrarían el salario medio.
- **Agrupador sin educación**, si no cubre el de A: nadie se educa.
- **Separador**, si $c_B(h^*) > w_A - w_B \geq c_A(h^*)$: solo A se educa, porque la señal es costosa en su justa medida.

Desarrollo gráfico: en el plano educación-salario, el diferencial salarial es una recta horizontal que corta las curvas de coste de B y de A en $\underline{h}$ y $\overline{h}$. El equilibrio es agrupador con educación si $h^* \leq \underline{h}$, separador si $\underline{h} < h^* \leq \overline{h}$ y agrupador sin educación si $h^* > \overline{h}$. Los costes de B superan a los de A para cualquier nivel de educación, y el diferencial es horizontal porque no depende de ella.



Señalización: el diferencial salarial delimita las zonas de equilibrio.

Estática comparativa: un mayor diferencial salarial o una educación más barata desplazan a la derecha $\underline{h}$ y $\overline{h}$. Con h^* dado, la economía puede pasar del agrupador sin educación al separador y de este al agrupador con educación, de modo que, para mantener la separación, el umbral ha de subir. Una mayor diferencia de costes entre tipos amplía la zona separadora y hace la señal más robusta. La razón es que los costes de educarse crecen con h, de modo que todo lo que eleva la recta del diferencial o baja las curvas de costes desplaza los cortes a la derecha.

Bienestar: el agrupador sin educación es eficiente, porque nadie paga una señal improductiva, y el separador es un *second best*. El de menor coste (Riley, 1979), con el umbral mínimo compatible con la separación, es Pareto-superior a los demás separadores. También cabría un óptimo de Pareto si un planificador fijara subsidios cruzados. El agrupador con educación, por su parte, despilfarra recursos, porque todos pagan una señal que no distingue.

Conclusiones: apoya una educación exigente y explica el **efecto diploma**. Pero suponer que la educación no altera la productividad es exigente: para Spence, la productividad causa la educación, mientras que para Becker, la educación causa la productividad. En Spence la educación actúa como señal y permite cobrar más, y en Becker la mayor productividad es la que eleva el salario.

3.4.2. Autoselección (Rothschild y Stiglitz, 1976)

Idea: el principal, que no observa el tipo del agente, ofrece un menú de contratos entre los que cada tipo se autoselecciona. La versión con un único principal procede de Mussa y Rosen (1978).

Razonamiento: en el mercado de trabajo, el tipo A, más productivo, querría hacerse pasar por el tipo B, al que el esfuerzo le pesa más y cuyo contrato pide menos esfuerzo. Para impedirlo, el principal distorsiona el contrato de B, con menos esfuerzo del eficiente, y deja a A una **renta informacional**, una utilidad por encima de la de reserva. El resultado es un separador de *second best*, eficiente para el tipo alto.

Extensiones: el mismo mecanismo explica dos regularidades del mercado de trabajo.

- **Grilletes dorados** (Salop y Salop, 1976): perder a un trabajador formado sale caro, y la empresa paga poco al principio y promete subidas que solo convienen a quien piensa quedarse, con lo que el perfil del salario filtra y retiene.
- **Menú de salario fijo y variable** (Lazear, 1986): el poco productivo elige el fijo y el productivo el variable, que mejora sus ofertas externas, y cada uno revela su tipo, aunque el productivo conserva una renta informacional. La empresa ofrece elegir porque los más productivos tienen mejores alternativas fuera.

Conclusiones: en su versión original, con aseguradoras en competencia y beneficio nulo, no hay agrupador, y el separador da cobertura completa a prima justa a los de alto riesgo y parcial a los de bajo, sin renta informacional, lo que explica las franquicias. Pero impone un coste al tipo bajo, infraasegurado o con un contrato distorsionado, y en seguros el equilibrio puede no existir si abundan los de bajo riesgo. Los refinamientos lo restauran cambiando el concepto de equilibrio.

3.4.3. Regulación (Laffont y Tirole, 1986)

Idea: el regulador no conoce la eficiencia de la empresa, selección adversa, ni su esfuerzo, riesgo moral. La regulación óptima de los monopolios naturales, la aplicación más influyente de la autoselección, es un menú de contratos.

Razonamiento: la empresa eficiente elige un precio fijo, con incentivos plenos y renta informacional, y la ineficiente, más reembolso de costes, con menor potencia de incentivos y sin renta.

Conclusiones: reúne los dos fallos en un marco con una misma respuesta, un menú diseñado con el principio de revelación, a costa de una renta informacional inevitable. Pero es estático y supone un regulador benevolente y comprometido. Con efecto trinquete, la empresa deja de revelar su tipo, y con captura, la regulación deja de perseguir el bienestar. Además, el regulador debe conocer la distribución de los tipos.

En suma, la selección adversa explica por qué se contraen mercados como los seguros, el crédito o el trabajo, y el principio de revelación unifica sus soluciones. Pero estas son de *second best*, porque la señal y la autoselección tienen un coste real, los equilibrios son a menudo múltiples y los modelos suponen tipos fijos y agentes racionales. Con tipos endógenos o racionalidad limitada, la asimetría puede ser más o menos grave de lo que predicen.

1. Fundamentos de la teoría de juegos

Conviene fijar primero los elementos de un juego, sus tipos, la noción de estrategia y los conceptos de solución.

1.1. Definición y elementos de un juego

Definición: un **juego** es la representación formal de una situación en la que varios individuos interactúan en un entorno de interdependencia estratégica. Su *solución* es una predicción de cómo jugarán.

Elementos: todo juego queda definido por cuatro categorías de elementos.

- Jugadores, $N = \{1, 2, \dots, n\}$: individuos, empresas o países, generalmente en número fijo.
- *Reglas*: el conjunto de acciones de cada jugador, A_i; la temporalidad; la información sobre el juego y sobre las acciones ya tomadas, y los compromisos vinculantes. El caso extremo de conflicto son los juegos de suma cero.
- *Resultados*: el desenlace del juego para cada combinación de acciones.
- *Pagos*: las preferencias sobre los resultados, representadas por utilidades cardinales. La función de pagos $u_i(s_i, \mathbf{s}_{-i})$ asigna a cada perfil de estrategias un pago para el jugador i.

1.2. Tipos de juego y representaciones

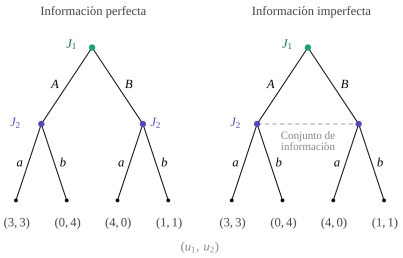
Según el tiempo: en los juegos **estáticos** los jugadores eligen simultáneamente, o sin observar las elecciones de los demás. En los **dinámicos** deciden secuencialmente, y al menos uno observa alguna acción previa de otro. Pueden ser **secuenciales**, jugados una vez por turnos, o **repetidos**, si un mismo juego estático, el juego de etapa, se juega varias veces, con horizonte finito o infinito.

Según la información: se distingue el conocimiento del juego en sí, completo o incompleto, y el de las acciones de los demás, perfecto o imperfecto.

- Con información completa, todos conocen jugadores, estrategias y pagos, y se supone *conocimiento común* (Aumann, 1976), es decir, cada jugador sabe que los otros lo saben, y así *ad infinitum*.
- Con información incompleta, algún jugador desconoce una característica relevante, típicamente los pagos o el "tipo" de algún rival, y se modeliza con la transformación de Harsanyi.
- Con información perfecta, cada jugador conoce al decidir todas las acciones previas, y con imperfecta, no.

Según los acuerdos: en los juegos **no cooperativos** la unidad de análisis es el jugador individual y no hay mecanismos externos que hagan exigibles los acuerdos. En los **cooperativos** la unidad es la coalición, los acuerdos son vinculantes y la pregunta es cómo repartir las ganancias de cooperar.

Representaciones: la **forma normal**, matricial o estratégica, es una matriz con las estrategias y los pagos de cada combinación, y es la natural en los juegos estáticos. La **forma extensiva** es un árbol con la secuencia de decisiones, la información en cada nodo y los pagos finales, y es la natural en los dinámicos, aunque todo juego en forma extensiva admite una forma normal. Los nodos que un jugador no puede distinguir forman un *conjunto de información*.



Forma extensiva: versión secuencial con información perfecta (izquierda) y simultánea con información imperfecta (derecha).

1.3. La estrategia

Acción y estrategia: una acción es cada alternativa disponible en un nodo de decisión. Una **estrategia** es un plan completo que indica qué acción tomar en cada conjunto de información del jugador, incluidos los nodos que ese mismo plan nunca alcanzará. Ambos conceptos coinciden en un juego estático con información completa, pero no en uno dinámico.

1.3.1. Estrategias puras

Definición: una **estrategia pura**, s_i^j, es un plan contingente completo que el jugador sigue con certeza. En un juego finito, el conjunto $S_i = \{s_i^1, s_i^2, \dots, s_i^m\}$ es finito. Su tamaño crece rápido con la estructura del juego, y un jugador con k conjuntos de información y dos acciones en cada uno dispone de 2^k estrategias puras.

Un **perfil de estrategias** reúne una estrategia de cada jugador y se escribe $\mathbf{s} = (s_i, \mathbf{s}_{-i})$ para separar la del jugador i de las de todos los demás.

1.3.2. Estrategias mixtas

Definición: una **estrategia mixta** es una distribución de probabilidad sobre las estrategias puras del jugador:

$$\sigma_i = (p_1, p_2, \dots, p_m), \quad p_j \geq 0, \quad \sum_j p_j = 1$$

Toda estrategia pura es una mixta degenerada, y los pagos de un perfil mixto son pagos esperados, lo que presupone preferencias de von Neumann-Morgenstern.

Interpretación: aleatorización deliberada para no ser previsible, frecuencias en una población de jugadores o incertidumbre de los rivales, según Harsanyi.

1.3.3. Mejor respuesta y dominancia

Mejor respuesta: la estrategia s_i^* es una mejor respuesta a $\mathbf{s}_{-i}$ si ninguna otra da al jugador un pago mayor:

$$u_i(s_i^*, \mathbf{s}_{-i}) \geq u_i(s_i, \mathbf{s}_{-i}) \quad \forall s_i \in S_i$$

La función de mejor respuesta, o de reacción, $s_i^* = R_i(\mathbf{s}_{-i})$, asigna a cada perfil de los rivales la respuesta óptima del jugador.

Dominancia: s_i está *estrictamente dominada* por s_i' si esta da más pago frente a todo $\mathbf{s}_{-i}$:

$$u_i(s_i', \mathbf{s}_{-i}) > u_i(s_i, \mathbf{s}_{-i}) \quad \forall \mathbf{s}_{-i}$$

Está *débilmente dominada* si s_i' da al menos lo mismo siempre y más en algún caso, y una estrategia dominante domina a todas las demás. Un jugador racional nunca juega una estrategia estrictamente dominada, lo que, si la racionalidad es conocimiento común, justifica su eliminación iterativa, cuyo resultado no depende del orden de supresión. Eliminar las débilmente dominadas es menos seguro, porque el resultado sí puede depender del orden.

1.4. Soluciones a los juegos

1.4.1. Juegos no cooperativos

Los conceptos de solución responden a qué perfiles cabe esperar de jugadores racionales, pero difieren en lo que exigen y en el tipo de juego.

EED: si cada jugador dispone de una estrategia dominante, la juega con independencia de lo que espere de los demás, y ese perfil es el **equilibrio en estrategias dominantes**, la predicción más robusta, pero rara vez existe. La eliminación iterativa de estrategias estrictamente dominadas amplía su alcance. Si tras repetirla sobrevive un único perfil, el juego es *resoluble por dominancia*.

EN: un **equilibrio de Nash** es un perfil en el que la estrategia de cada jugador es mejor respuesta a las de los demás:

$$u_i(s_i^*, \mathbf{s}_{-i}^*) \geq u_i(s_i, \mathbf{s}_{-i}^*) \quad \forall s_i \in S_i, \ \forall i$$

Nadie gana desviándose unilateralmente, como en un acuerdo que se cumple solo. Exige más que el EED, porque cada jugador debe acertar al prever las estrategias rivales, pero se aplica a muchos más juegos, y todo EED es un EN.

- En estrategias puras puede no existir, como en el *matching pennies*, o no ser único, como en la batalla de los sexos.
- En estrategias mixtas, un jugador solo mezcla entre estrategias puras que le dan el mismo pago esperado, de modo que, en equilibrio, sus probabilidades dejan indiferentes a los rivales. El **teorema de Nash** (1950) garantiza que todo juego finito tiene al menos un EN en mixtas. La prueba aplica el teorema de punto fijo de Kakutani a la correspondencia de mejor respuesta.

ENPS: en un juego dinámico, el EN puede apoyarse en *amenazas no creíbles*, acciones previstas para nodos que no se alcanzan y que el jugador no ejecutaría si llegara a ellos. Selten (1965) las descarta exigiendo racionalidad también fuera de la senda. Un *subjuego* empieza en un nodo que forma por sí solo un conjunto de información y comprende todos sus sucesores sin partir ningún conjunto de información. Un **equilibrio de Nash perfecto en subjuegos** induce un EN en cada subjuego, se alcance o no. Con información perfecta se obtiene por *inducción hacia atrás*, desde los últimos nodos hasta el inicial. Cada nodo final se sustituye por el pago de la acción óptima y se retrocede. Todo ENPS es un EN, pero no a la inversa.

EBN: si algún jugador desconoce los pagos de los rivales, la **transformación de Harsanyi** (1967-1968) hace que la naturaleza elija primero el *tipo* de cada jugador según una distribución que es conocimiento común. Cada jugador observa solo su tipo. El **equilibrio de Bayes-Nash** es el EN de ese juego ampliado: cada tipo maximiza su pago esperado, dadas sus creencias. Una estrategia es así un plan que asigna una acción a cada tipo posible, y las creencias se refieren a los tipos de los rivales y a sus estrategias.

EBP: en los juegos dinámicos con información incompleta, las acciones revelan información sobre el tipo. El **equilibrio bayesiano perfecto**

une la racionalidad secuencial con creencias consistentes, obtenidas por la regla de Bayes siempre que sea posible. Fuera de la senda Bayes no impone nada, y esa libertad explica la multiplicidad de equilibrios. Las creencias son una distribución de probabilidad sobre los nodos de cada conjunto de información, y las estrategias han de ser óptimas en cada uno, dadas esas creencias. Refinamientos como el criterio intuitivo acotan la multiplicidad.

1.4.2. Juegos cooperativos

Planteamiento: la **función característica**, $V(S)$, asigna a cada coalición $S \subseteq N$ el valor que puede garantizarse por sí misma, y la pregunta es cómo repartir el de la gran coalición, $V(N)$, mediante un reparto $\mathbf{x} = (x_1, \dots, x_n)$.

Núcleo (Gillies, 1959): un reparto es estable si ninguna coalición puede garantizar a sus miembros más de lo que reciben, porque, si no, lo *bloquearía*. El núcleo es el conjunto de repartos eficientes que ninguna coalición bloquea:

$$\mathcal{C}(V) = \left\{ \mathbf{x} \in \mathbb{R}^n : \sum_{i \in N} x_i = V(N),\ \sum_{i \in S} x_i \geq V(S)\ \ \forall S \subseteq N \right\}$$

Puede contener infinitos repartos, si las demás coaliciones obtienen poco por sí solas, o ser vacío, si obtienen mucho. Las coaliciones de un solo jugador imponen la *racionalidad individual*, $x_i \geq V(\{i\})$. Un ejemplo de núcleo vacío es el de tres jugadores que se reparten una unidad cuando basta una mayoría de dos para quedársela, porque cada pareja debería recibir al menos la unidad y, sumando las tres condiciones, los pagos sumarían al menos 3/2.

Valor de Shapley (1953): busca un único reparto equitativo, caracterizado por cuatro axiomas: eficiencia, simetría, jugador nulo y aditividad. La única regla que los cumple asigna a cada jugador la media de sus contribuciones marginales:

$$\phi_i(V) = \sum_{S \subseteq N \setminus \{i\}} \frac{s!\,(n-s-1)!}{n!}\,[V(S \cup \{i\}) - V(S)]$$

El peso es la probabilidad de que i encuentre formada exactamente la coalición S si los jugadores llegan en un orden aleatorio. Siempre existe y es único, pero puede quedar fuera del núcleo, porque la equidad no garantiza la estabilidad. Sí pertenece a él en los juegos convexos (Shapley, 1971). En ellos, la contribución marginal de cada jugador no disminuye cuando se amplía la coalición a la que se une.

En suma, los conceptos no cooperativos son cada vez más exigentes: el EN pide acertar con las estrategias rivales y existe en mixtas en todo juego finito, y el ENPS y el EBP añaden racionalidad fuera de la senda. En los cooperativos, el núcleo reparte por estabilidad y el valor de Shapley por equidad, y lo equitativo puede no ser estable.

2. Aplicaciones

Cada juego se presenta, en lo posible, con su planteamiento, su equilibrio y su bienestar.

2.1. Juegos estáticos con información completa

2.1.1. Dilema del prisionero

Planteamiento: dos sospechosos, interrogados por separado, pueden confesar (C) o no confesar (NC). Si ambos confiesan, reciben cinco años; si ninguno lo hace, un año, y si solo uno confiesa, queda libre y el otro recibe diez.

Equilibrio: confesar es estrictamente dominante, porque da más sea cual sea la elección del otro ($-5 > -10$ si confiesa y $0 > -1$ si no), de modo que (C, C) es el EED y el único EN.

Bienestar: el resultado es **Pareto-ineficiente**: (NC, NC), con pagos $(-1, -1)$, es mejor para ambos, pero no es sostenible, porque cada uno tiene incentivos a desviarse. Esta tensión entre racionalidad individual y eficiencia colectiva reaparece en la colusión, los bienes públicos o las guerras arancelarias. La cooperación exige compromisos vinculantes o repetición con castigos creíbles.

2.1.2. Batalla del mar de Bismarck (1943)

Planteamiento: es un juego histórico de suma cero. Un convoy japonés elige la ruta norte o sur de Nueva Bretaña y la aviación aliada, dónde concentrar su búsqueda. Los pagos son días de bombardeo.

Equilibrio: para Japón, el norte domina débilmente al sur, porque garantiza como máximo dos días de bombardeo frente a tres posibles. Eliminada la ruta sur, la mejor respuesta aliada es buscar en el norte, y (Norte, Norte) es el único EN en puras, que coincide con la solución *maximin* y con lo que hicieron ambos bandos.

Bienestar: en suma cero, cualquier perfil es óptimo de Pareto.

2.1.3. Cournot: EN único sin dominancia

Dos empresas eligen simultáneamente cantidades de un bien homogéneo. No hay estrategias dominantes, porque la cantidad óptima de cada una depende de la del rival, pero el EN, la intersección de las funciones de reacción, es único bajo los supuestos habituales. Se sitúa entre el monopolio y la competencia perfecta, y muestra que el EN es el concepto general de los juegos estáticos. La unicidad exige que las funciones de reacción tengan pendiente menor que uno en valor absoluto, como con demanda lineal y costes marginales constantes, con pendiente $-1/2$.

2.1.4. *Matching pennies*: sin EN en puras

Dos jugadores muestran a la vez una moneda: si coinciden, gana el 1, y si no, el 2. No hay EN en puras, porque en toda celda alguien gana desviándose. En mixtas, cada uno elige cara con $p = 1/2$, lo que deja al rival indiferente, con pago esperado nulo. Ilustra el teorema de Nash y por qué son necesarias las estrategias mixtas en los juegos de pura oposición.

2.1.5. Batalla de los sexos: equilibrios múltiples

Dos personas quieren pasar la tarde juntas, pero una prefiere el fútbol y la otra la ópera. Hay dos EN en puras, (Fútbol, Fútbol) y (Ópera, Ópera), óptimos de Pareto pero con distinto reparto, y uno en mixtas, con probabilidad 3/5 para la opción favorita, ineficiente porque a veces no se coordinan. La teoría no predice cuál se alcanzará, lo que motiva la selección por puntos focales de Schelling o por comunicación previa. En los juegos de coordinación, la historia y las instituciones importan tanto como los incentivos.

2.2. Juegos con incertidumbre

2.2.1. Información imperfecta

Idea: los jugadores no observan alguna acción de los rivales, propiedad distinta de la completitud, pues el dilema del prisionero tiene información completa pero imperfecta. Se representa con conjuntos de información y se resuelve con los mismos conceptos. En hundir la flota, cada jugador conoce el resultado de sus acciones, pero no la disposición de la flota rival, y el equilibrio se aproxima con reglas heurísticas.

2.2.2. Información incompleta: Bayes-Nash

Planteamiento: con dos estados de la naturaleza, crisis y no crisis, $\Omega = \{\omega_c, \omega_{nc}\}$, y una matriz de pagos distinta en cada uno, el jugador 1 observa el estado y el 2 solo conoce la probabilidad de crisis, P_c. En el EBN, el jugador 1 responde en cada estado con la matriz de ese estado, y el 2 elige la acción de mayor pago esperado. El equilibrio puede ser ineficiente *ex post* aunque sea óptimo *ex ante*. Sin información privada bastaría con hallar el EN de la matriz de pagos esperados, y la predicción es sensible a la distribución supuesta, que ha de ser conocimiento común.

2.2.3. Coches usados (Akerlof, 1970)

Planteamiento: con información asimétrica, los vendedores conocen la calidad de su coche, $\theta \in \{\theta_L, \theta_H\}$, y los compradores solo su distribución, de modo que ofrecen un precio basado en la calidad esperada. Actualizan sus expectativas por la regla de Bayes a partir de la conducta de los vendedores, que deciden si vender según su información privada y el precio.

Equilibrio: si la proporción de coches buenos no es lo bastante alta, al precio medio se retiran los vendedores de alta calidad, lo que degrada la calidad esperada y el precio, en espiral. En el EBN resultante solo se intercambian los coches malos, los *lemons*.

Bienestar: el equilibrio no es óptimo de Pareto, porque quedan sin hacer intercambios beneficiosos. La información privada es un **fallo de mercado** por derecho propio.

2.2.4. Señalización (Spence, 1973; Cho y Kreps, 1987)

Idea: es un juego dinámico con información incompleta, resuelto con el EBP. El emisor informado elige una señal observable, y el receptor responde tras actualizar sus creencias.

Supuestos: un trabajador y varias empresas competitivas que pagan la productividad esperada. La productividad, $\theta \in \{\theta_L, \theta_H\}$, es privada, y el tipo alto tiene probabilidad *a priori* μ_0. La educación, h, no altera la productividad, pero su coste, $c(h, \theta) = h/\theta$, es menor para el tipo productivo (*cruce único*).

Planteamiento: un EBP es una educación por tipo, unas creencias $\mu(h) = \Pr(\theta_H \mid h)$, obtenidas por Bayes en las educaciones de equilibrio, y un salario igual a la productividad esperada, tales que cada tipo resuelve:

$$\max_{h \geq 0}\ w(h) - c(h, \theta), \qquad w(h) = \mu(h)\,\theta_H + (1 - \mu(h))\,\theta_L$$

Equilibrio separador: el tipo bajo no se educa y cobra θ_L, y el alto elige h^* y cobra θ_H, siempre que ningún tipo prefiera imitar al otro:

$$\theta_H - \frac{h^*}{\theta_H} \geq \theta_L \qquad \text{y} \qquad \theta_L \geq \theta_H - \frac{h^*}{\theta_L}$$

Estas condiciones de compatibilidad de incentivos acotan la señal, $\theta_L(\theta_H - \theta_L) \leq h^* \leq \theta_H(\theta_H - \theta_L)$. Cualquier h^* del intervalo es un EBP separador con la creencia $\mu(h) = 1$ si $h \geq h^*$ y nula si no.

Equilibrio agrupador: ambos tipos eligen la misma h_p y cobran $\bar{w} = \mu_0\,\theta_H + (1 - \mu_0)\,\theta_L$, con creencias pesimistas fuera del equilibrio, $\mu(h) = 0$ para toda $h \neq h_p$. Basta con que el tipo bajo no prefiera desviarse a $h = 0$:

$$\bar{w} - \frac{h_p}{\theta_L} \geq \theta_L \implies h_p \leq \mu_0\,\theta_L(\theta_H - \theta_L)$$

Qué equilibrio surge depende del coste de la educación: separa porque es más cara para el tipo bajo, y la separación le cuesta al alto h^*/θ_H frente a una ganancia de $(1-\mu_0)(\theta_H-\theta_L)$.

Criterio intuitivo (Cho y Kreps, 1987): una desviación que no mejoraría al tipo bajo ni cobrando θ_H, pero sí al alto, debe atribuirse a este. Rompe todo agrupador y solo deja el separador de señal mínima, $h^*=\theta_L(\theta_H-\theta_L)$. La misma lógica explica el precio límite como señal de costes bajos (Milgrom y Roberts, 1982a) o la garantía como señal de calidad.

Conclusiones: una actividad improductiva puede ser valiosa como transmisora de información. Pero la libertad de creencias genera una multiplicidad que ningún refinamiento elimina del todo, y la educación es un coste puro.

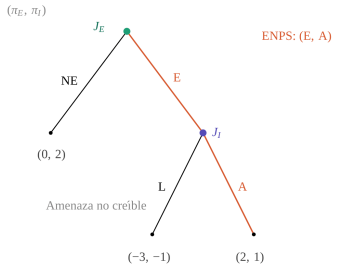
2.3. Juegos dinámicos

2.3.1. Juego de entrada

Planteamiento: una entrante decide entrar (E) o no (NE). Si entra, la incumbente lucha (L) o acomoda (A). Los pagos (entrante, incumbente) son $(0,2)$ sin entrada, $(-3,-1)$ con lucha y $(2,1)$ con acomodación.

EN y ENPS: hay dos EN en puras, (E, A) y (NE, L), pero el segundo descansa en una amenaza no creíble, porque ante una entrada real la incumbente acomodaría ($1>-1$). Por inducción hacia atrás, la entrante anticipa la acomodación y entra, de modo que el ENPS es (E, A), con pagos $(2,1)$. Para comprometerse a luchar, la incumbente puede hundir una inversión C_0, con π^M, π^D y π^L los beneficios de monopolio, duopolio y lucha. La entrada queda disuadida si la inversión hace creíble la lucha, esta causa pérdidas a la entrante y la inversión compensa:

$$\pi^L>\pi^D-C_0,\qquad \pi^L<0,\qquad \pi^M-C_0>\pi^D$$



Juego de entrada: la amenaza de lucha no es creíble y el ENPS es (E, A).

Conclusiones: la incumbente puede hacer creíble la lucha con costes hundidos, como inversiones irreversibles en capacidad, y el poder estratégico presente aumenta al reducir la propia libertad futura. La otra lección es la **ventaja de mover primero**, como la del líder de Stackelberg, aunque no es universal. En la competencia secuencial en precios con productos diferenciados, el seguidor obtiene más que el líder, porque fija un precio algo inferior.

2.3.2. Dilema del prisionero repetido finito

El juego de etapa se repite T veces, con T conocido. Cooperar (c) es no confesar y no cooperar (nc), confesar. Por inducción hacia atrás:

1. *Periodo T*: no hay futuro ni castigo creíble, luego no cooperar es dominante y se juega el EN de etapa.
2. *Periodo $T-1$ y anteriores*: como en T nadie cooperará, desviarse no tiene coste, y el argumento se extiende hasta el primero.

El único ENPS repite T veces el EN de etapa. Como su pago coincide con el *minimax*, ni siquiera otros EN sostienen la cooperación. Solo con varios EN de etapa cabe cooperar en las primeras rondas, usando la selección del equilibrio final como premio o castigo.

2.3.3. Repetición infinita: colusión tácita

Planteamiento: el juego se repite sin fin (o con probabilidad de continuar en cada periodo) y los jugadores descuentan con $\beta\in(0,1)$, el "superjuego" de Friedman (1971). La historia es $h^{t-1}=\{(s_1^\tau,s_2^\tau)\}_{\tau=1}^{t-1}$ y, con la *estrategia del gatillo*, $s_i^t(h^{t-1})=c$ si ambos han cooperado siempre y, ante una sola desviación, se vuelve para siempre al EN de etapa.

Condición de cooperación: cooperar debe valer al menos tanto como desviarse y sufrir el castigo. Con la serie geométrica:

$$\frac{\pi_{c,c}}{1-\beta}\geq\pi_{nc,c}+\frac{\beta}{1-\beta}\pi_{nc,nc}\implies\beta\geq\frac{\pi_{nc,c}-\pi_{c,c}}{\pi_{nc,c}-\pi_{nc,nc}}$$

El numerador es la ganancia inmediata de traicionar, y el denominador le suma la pérdida que el castigo inflige en cada periodo futuro. El castigo es creíble porque volver al EN de etapa es un equilibrio en cada subjuego. En el dilema del prisionero basta $\beta\geq1/5$, y en el duopolio de Bertrand, $\beta\geq1/2$ (A19).

Folk theorem: Friedman (1971) mostró que, con jugadores pacientes, puede sostenerse cualquier pago factible que mejore para todos el del EN de etapa, y Fudenberg y Maskin (1986) lo extendieron a todo pago factible superior al *minimax*, con una condición de dimensión si hay más de dos jugadores. La repetición resuelve la eficiencia a costa de la multiplicidad. Además del gatillo, caben el *tit-for-tat* (Axelrod, 1984), más indulgente, y el *castigo óptimo*, que castiga durante un número finito de periodos y vuelve después a cooperar. La aplicación central es la **colusión tácita**, cuya estabilidad depende del número de empresas, de la transparencia del mercado y del factor de descuento.

2.3.4. Reputación (Selten, 1978; Kreps y Wilson, 1982)

Paradoja de la cadena de tiendas: una cadena se enfrenta sucesivamente a K entrantes. La inducción hacia atrás dice que nunca debe luchar, pero la intuición dice que conviene luchar al principio para ganar reputación de dureza.

Extensiones: Kreps y Wilson (1982), y Milgrom y Roberts (1982b), introducen una probabilidad $\varepsilon>0$ de que la cadena sea un tipo duro que lucha siempre. En el EBP, acomodar revela el tipo racional, de modo que este imita al duro y soporta hoy el coste de luchar. Con los pagos del juego de entrada, un entrante solo entra si $\Pr(\text{luchar})<2/5$. Con K grande, una ε mínima sostiene la no entrada en casi todos los mercados. Explica la predación como inversión en reputación, pero la predicción depende del tipo duro que se postule. Muestra además la fragilidad de la inducción hacia atrás, porque una perturbación mínima de la información cambia por completo la predicción.

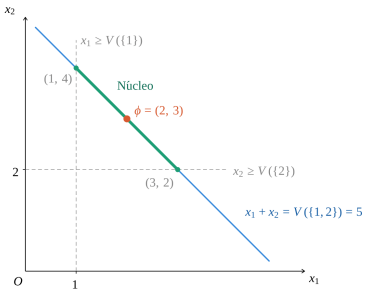
2.4. Juegos cooperativos y negociación

2.4.1. El núcleo: la *joint venture*

Dos empresas obtienen en solitario $V(\{1\})=1$ y $V(\{2\})=2$, y juntas, $V(\{1,2\})=5$. El núcleo son los repartos con $x_1+x_2=5$, $x_1\geq1$ y $x_2\geq2$, es decir, el segmento entre $(1,4)$ y $(3,2)$, estable pero no único.

2.4.2. Valor de Shapley y poder de voto

Promediando las contribuciones marginales en los dos órdenes de formación, el valor de Shapley da $\phi_1=2$ y $\phi_2=3$, un punto del núcleo.



Joint venture: el núcleo es el segmento de repartos eficientes no bloqueados; ϕ, el valor de Shapley.

Poder de voto: en los juegos de votación, el valor de Shapley da el **índice de Shapley-Shubik**, la proporción de ordenaciones en que cada partido es pivotal. El poder no es proporcional a los escaños. Con 45, 35 y 20 escaños y mayoría de 51, cualquier pareja es mayoría y cada partido tiene 1/3.

2.4.3. Negociación: Nash (1950) y Rubinstein (1982)

Solución de Nash: dos agentes se reparten un excedente y, sin acuerdo, obtienen su pago de desacuerdo, $d=(d_1,d_2)$. Nash exige invariancia a transformaciones afines, simetría, independencia de alternativas irrelevantes y eficiencia de Pareto, y la única solución que los cumple es:

$$\max_{(u_1,u_2)\in\mathcal{U}}(u_1-d_1)(u_2-d_2)\qquad\text{s.a.: } u_i\geq d_i,\ i=1,2$$

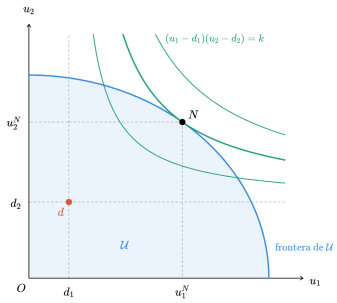
CPO: iguala la pendiente de la frontera de $\mathcal{U}$, en valor absoluto, a la ratio de ganancias netas. Al repartir un excedente Y con utilidades lineales, $u_1=x$ y $u_2=Y-x$, iguala las ganancias netas:

$$-\frac{du_2}{du_1}=\frac{u_2-d_2}{u_1-d_1}$$
$$x^*-d_1=Y-x^*-d_2\implies x^*=d_1+\tfrac{1}{2}(Y-d_1-d_2)$$

Cada parte recibe su alternativa exterior más la mitad del excedente sobre el desacuerdo, y obtiene más quien tiene mejor alternativa exterior.

Rubinstein: con ofertas alternas sin límite de periodos y factor de descuento β, el único ENPS tiene acuerdo inmediato y el proponente se queda $x^*=1/(1+\beta)$, porque ofrece al otro lo que este obtendría, descontado, si rechazase. Cuando $\beta\to1$ converge a la mitad, la solución de Nash simétrica. La ventaja del proponente nace de la impaciencia. El jugador 1 propone en los periodos pares y el 2 en los impares, y la oferta de equilibrio deja al otro indiferente, $1-x^*=\beta x^*$.

Conclusiones: la solución de Nash es la herramienta estándar del reparto bilateral que el mercado no fija, como la negociación salarial. Pero la independencia de alternativas irrelevantes es discutible (Kalai y Smorodinsky, 1975) y ambos modelos suponen información completa.



Solución de negociación de Nash.

En suma, los juegos estáticos recorren los casos del EN. Con incertidumbre, la información privada puede impedir intercambios beneficiosos, y en los dinámicos la disuasión y la cooperación se apoyan en compromisos irreversibles, en la repetición indefinida o en la reputación.

3. Teoría de las subastas

Las subastas aplican el equilibrio de Bayes-Nash a un problema de diseño.

3.1. Concepto (Vickrey, 1961; Myerson, 1981)

Teoría de subastas: iniciada por William Vickrey (1961, Nobel en 1996), analiza los mecanismos de venta con información asimétrica, en la que cada pujador conoce su valoración, pero no la de los demás. Myerson (1981) la extendió al diseño de la subasta óptima para el vendedor. Es la aplicación señera de los juegos con información incompleta.

3.2. Formatos básicos de subasta

Con un único objeto, se distinguen cuatro formatos básicos:

1. *Inglesa, ascendente*: el precio sube abiertamente hasta que queda un solo pujador, que paga el precio al que abandonó el último rival.
2. *Holandesa, descendente*: el precio baja desde un nivel alto hasta que un pujador lo acepta y paga ese precio.
3. *En sobre cerrado de primer precio*: con ofertas selladas, gana la más alta y paga su propia oferta.
4. En sobre cerrado de segundo precio, o de Vickrey: gana la oferta más alta, pero paga la segunda.

3.3. Pujas de equilibrio (Vickrey, 1961)

Idea: en la subasta de primer precio, pujar alto aumenta la probabilidad de ganar pero reduce lo que se gana, y la puja óptima rebaja la valoración (*bid shading*).

Supuestos: $n \geq 2$ pujadores neutrales al riesgo, con valoraciones privadas independientes, uniformes en $[0,1]$; un vendedor que valora el objeto en cero, y un equilibrio simétrico con pujas $b(v)$ crecientes.

Planteamiento: en la de primer precio, el pujador de valoración v maximiza la ganancia si vence por la probabilidad de que las $n-1$ valoraciones rivales queden por debajo:

$$\max_{b} \; (v-b)\cdot\left[b^{-1}(b)\right]^{n-1}$$

Equivale a elegir la valoración que aparenta, $\tilde{v}=b^{-1}(b)$:

$$\max_{\tilde{v}} \; (v-b(\tilde{v}))\cdot\tilde{v}^{n-1}$$

CPO: si en equilibrio no quiere aparentar una valoración distinta de la propia, el beneficio marginal de pujar un euro más, el aumento de la probabilidad de ganar por el excedente, iguala su coste marginal, la probabilidad de ganar:

$$\frac{(n-1)\,v^{n-2}}{b'(v)}\,[v-b(v)] = v^{n-1}$$

CSO: la derivada del objetivo respecto a $\tilde{v}$ tiene el signo de $v-\tilde{v}$, de modo que a quien aparenta menos de lo que valora le compensa subir la puja, y a quien aparenta más, bajarla. Por tanto, $\tilde{v}=v$ es el máximo global.

Solución: la CPO equivale a una ecuación diferencial que, integrada con $b(0)=0$, da la puja de equilibrio:

$$\frac{d}{dv}\left[v^{n-1}b(v)\right] = (n-1)\,v^{n-1} \implies b(v)=\frac{n-1}{n}\,v$$

En este equilibrio de Bayes-Nash simétrico, la puja es una fracción constante de la valoración, la mitad con dos pujadores, que tiende a la valoración cuando $n\to\infty$, porque la competencia elimina la rebaja.

Vickrey: pujar la propia valoración, $b(v)=v$, es **débilmente dominante**, porque pujar más solo añade victorias con pérdida y pujar menos solo hace perder victorias rentables. El ganador paga la segunda valoración, cuya esperanza es $\frac{n-1}{n}\,v$, lo mismo que puja en la de primer precio.

Conclusiones: la rebaja estratégica es una respuesta racional a la información privada, pero la solución cerrada depende de la simetría, la neutralidad al riesgo y la independencia.

3.4. Teorema de equivalencia de ingresos

Origen: Vickrey (1961) observó que, con valores privados independientes y uniformes, las subastas de primer y de segundo precio dan el mismo ingreso esperado, y Riley y Samuelson (1981) y Myerson (1981) lo generalizaron.

Teorema: con pujadores neutrales al riesgo y simétricos, valores privados independientes extraídos de la misma distribución continua y un formato en el que gane el de mayor valoración y el de valoración mínima obtenga excedente nulo, los cuatro formatos dan **el mismo ingreso esperado** al vendedor.

Razonamiento: en la de Vickrey y en la inglesa se puja la verdadera valoración. En la de primer precio y en la holandesa se rebaja, pero el ingreso esperado no cambia, porque depende de quién obtiene el objeto y no de las reglas de pago. Por el teorema de la envolvente, la utilidad de equilibrio del tipo v es $U(v)=\int_0^v P(x)\,dx$. Como la probabilidad de ganar, P, es la misma en todos los formatos, también lo son el pago esperado de cada tipo y el ingreso del vendedor, $\frac{n-1}{n+1}$ en el caso uniforme.

Extensiones: qué cambia al fallar los supuestos.

- *Precio de reserva (Myerson, 1981)*: el vendedor eleva su ingreso esperado fijando un $\underline{r}$ con $\underline{r}-\frac{1-F(\underline{r})}{f(\underline{r})}=0$, como un monopolista que iguala su coste nulo al valor virtual. En el caso uniforme $\underline{r}^*=1/2$, y con dos pujadores el ingreso pasa de $1/3$ a $5/12$. El valor virtual es el ingreso marginal de vender al tipo $\underline{r}$, descontada la renta informacional que cede a los tipos superiores, y $\underline{r}^*$ no depende de n.
- *Aversión al riesgo*: rompe la equivalencia a favor de la de primer precio y la holandesa, en las que los pujadores rebajan menos. Pujar algo más reduce el riesgo de quedarse sin el objeto, mientras que en la de Vickrey y en la inglesa pujar la valoración sigue siendo débilmente dominante y la recaudación no cambia.
- *Valores comunes*: surge la *maldición del ganador*, porque gana quien más ha sobrestimado el valor. Hay que pujar según el valor esperado condicionado a ganar, y más competidores no siempre elevan las pujas. La inglesa la mitiga, porque las retiradas revelan las señales.
- *Valores afiliados (Milgrom y Weber, 1982)*: la inglesa recauda al menos tanto como las de sobre cerrado, y la de segundo precio al menos tanto como la de primero, porque revelan más información.

Conclusiones: es la referencia de todos los formatos, pero sus supuestos rara vez se cumplen a la vez, y no contempla la colusión ni la venta de muchas unidades.

3.5. Aplicaciones

El Nobel de 2020 a Paul Milgrom y Robert Wilson premió nuevos formatos, de múltiples objetos y combinatorios. Cuando se venden muchas unidades la equivalencia falla: en la subasta de *precio uniforme* todos pagan el precio marginal, lo que invita a rebajar la puja de las últimas unidades, y en la *discriminatoria* cada uno paga su puja. Ninguna domina a la otra en general.

- *Deuda pública*: el Tesoro Público usa la subasta española u holandesa modificada, en la que las ofertas competitivas aceptadas pagan su precio si queda entre el mínimo aceptado y el medio ponderado, y el medio ponderado si lo igualan o superan (Orden ECM/2/2026). El Tesoro de EE. UU. aplica precio uniforme desde 1998.
- *Política monetaria*: el Eurosistema adjudicó sus operaciones principales de financiación a tipo variable y precios múltiples entre junio de 2000 y octubre de 2008. Desde entonces aplica tipo fijo con adjudicación plena, confirmado en 2024, y la subasta deja de ser un juego estratégico.
- *Derechos de emisión*: el RCDE UE subasta hasta el 57 % de sus derechos generales de 2021-2030 en la European Energy Exchange de Leipzig, en ronda única, sobre cerrado y precio uniforme.
- *Espectro*: Milgrom y Wilson diseñaron para la Comisión Federal de Comunicaciones de EE. UU. la subasta simultánea de rondas múltiples, que en julio de 1994 vendió diez licencias en 47 rondas por 617 M$. Vende a la vez todas las licencias, con pujas ascendentes en rondas sucesivas, para que cada operador ajuste sus pujas por licencias complementarias a la vista de los precios.
- *Contratación pública*: el riesgo es la colusión entre licitadores (*bid rigging*). En Long Island, el orden de las ofertas de cobertura del cártel no respondía a sus costes (Porter y Zona, 1993).

A15. La empresa: el tamaño eficiente y sus límites. Mención especial de la economía de los costes de transacción. La Teoría de la Organización Industrial: barreras a la entrada y mercados impugnables

1. Teoría de la empresa

La **teoría de la empresa** estudia cuatro ámbitos: su existencia y sus objetivos, por qué nacen las empresas y qué persiguen; sus límites, qué determina su tamaño; su organización interna, y su comportamiento en los mercados.

1.1. Existencia y objetivos

1.1.1. Enfoques tradicionales

Maximización de beneficios: la empresa existe como unidad de decisión que combina factores para maximizar la diferencia entre ingresos y costes. Esta visión es operativa y potente para la teoría de mercados, pero supone una unidad monolítica, sin conflictos internos, y con toda la información necesaria para optimizar.

Escuela austriaca: para Menger, Hayek y Kirzner, el orden económico es un descubrimiento incesante y no un equilibrio en reposo. El empresario no optimiza mecánicamente, sino que está *alerta* a las ocasiones de beneficio, y un beneficio extraordinario indica una oportunidad aún sin aprovechar, no un monopolio. Hayek (1945) subrayó, además, que el sistema de precios es un mecanismo descentralizado para transmitir información dispersa. Competir es, para ellos, un proceso que transcurre en el tiempo, no un estado.

1.1.2. Enfoque directivista o gerencial

Origen: desde Berle y Means (1932), la propiedad, de los accionistas, está separada de la gestión, de los directivos. Si estos tienen discrecionalidad, la empresa puede perseguir objetivos distintos del beneficio.

Baumol (1959): los directivos maximizan los ingresos por ventas, de los que dependen su remuneración y prestigio, sujetos a un beneficio mínimo que satisfaga a los accionistas:

$$\max_x \ \mathrm{IT}(x) = p(x)\,x \quad \text{s.a.:} \quad \pi(x) \geq \pi_{\min}$$

El lagrangiano es $\mathcal{L} = p(x)\,x + \lambda\,[\pi(x) - \pi_{\min}]$, y las condiciones de Kuhn-Tucker, necesarias y suficientes con ingreso total y beneficio cóncavos, son:

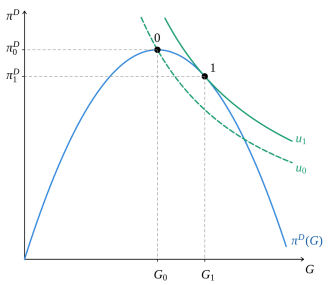
$$\mathrm{Img}(x) + \lambda\,\pi'(x) = 0$$
$$\lambda \geq 0, \qquad \lambda\,[\pi(x) - \pi_{\min}] = 0$$

Con la restricción activa, $\mathrm{Img}(x_B) = -\lambda\,\pi'(x_B) > 0$, de modo que la empresa opera a la derecha del máximo de beneficios, con $x_\pi < x_B < x_{\mathrm{IT}}$ y $p_B < p_\pi$. Produce más y vende más barato que el maximizador de beneficios, con un beneficio igual a $\pi_{\min}$. Si la restricción no es activa, $\lambda = 0$, produce donde $\mathrm{Img}(x_{\mathrm{IT}}) = 0$, con elasticidad unitaria, lo que solo es factible si ese beneficio cubre el mínimo. Cuando $\pi_{\min}$ tiende al beneficio máximo, el modelo converge al tradicional.

Estática comparativa: un impuesto de suma fija, o un aumento de los costes fijos, no altera la producción ni el precio del maximizador de beneficios, pero reduce la producción y sube el precio del maximizador de ventas. Que las empresas trasladen a precios los costes fijos fue el principal apoyo empírico de Baumol. La razón es que el impuesto desplaza hacia abajo la curva de beneficios, cuyo corte derecho con $\pi_{\min}$ se mueve a la izquierda.

Otras hipótesis: sustituyen las ventas por otros objetivos de los directivos.

- *Utilidad gerencial (Williamson, 1964)*: los directivos maximizan una utilidad que depende del gasto en plantilla, G, y de los beneficios discrecionales, π^D, sujetos a un beneficio mínimo. Eligen la tangencia de la frontera de beneficios con su curva de indiferencia más alta, con un gasto G_1 mayor que el que maximiza el beneficio, G_0 (*preferencia por el gasto*). La frontera crece mientras el gasto es productivo y decrece desde su máximo, el punto 0 que preferirían los propietarios, y en el punto 1 el directivo acepta un beneficio π_1^D menor.



Williamson (1964): el directivo elige un gasto en plantilla G_1 superior al que maximiza el beneficio, G_0.

- *Crecimiento (Marris, 1964)*: maximizan la tasa de crecimiento de la empresa sujetos a una restricción de valoración bursátil. Si la acción cae demasiado, la amenaza de una adquisición hostil disciplina a los directivos, y el mercado de control corporativo actúa así como mecanismo disciplinante externo.
- *Ineficiencia-X (Leibenstein, 1966)*: sin presión competitiva y con información asimétrica, nadie se esfuerza al máximo y la empresa opera por debajo de su frontera técnica, aunque se critica que ello puede ser el óptimo, dado el coste del control.

Conclusiones: Baumol deriva por primera vez la conducta de la empresa de un objetivo distinto del beneficio, con una predicción contrastable, y el enfoque anticipa la teoría de la agencia. Pero los objetivos se postulan en lugar de derivarse de los incentivos, el beneficio mínimo es exógeno y la competencia tiende a expulsar a quien no maximiza el beneficio. Por eso la discrecionalidad gerencial solo se sostiene con poder de mercado y barreras a la entrada.

En suma, estos enfoques son complementarios y explican qué persigue la empresa, pero no por qué ciertas transacciones se organizan dentro de ella y no en el mercado.

1.2. Límites

1.2.1. Enfoque neoclásico: la caja negra

El enfoque neoclásico atiende a la tecnología, los costes y el tamaño del mercado. El tamaño de la empresa lo fija el mínimo del coste medio a largo plazo, en forma de U. La dimensión óptima es única, y la empresa es, en lo demás, una caja negra tecnológica.

1.2.2. Enfoque tecnológico de la TOI

Horquilla de capacidad (Stigler, 1939): las empresas mantienen una reserva de capacidad que les da flexibilidad ante la incertidumbre. La evidencia apunta a un coste medio a largo plazo en forma de L, es decir, alcanzada la escala mínima eficiente (EME), es casi constante. No hay dimensión óptima única, lo que explica la coexistencia de tamaños muy distintos.

Subaditividad (Baumol, Panzar y Willig, 1982): una sola empresa produce el *output* total a menor coste que cualquier reparto entre varias, definición moderna de **monopolio natural**. En el caso uniproducto, las economías de escala son condición suficiente pero no necesaria. El límite del tamaño eficiente lo fija la subaditividad, no la pendiente del coste medio. En el caso multiproducto, con vectores de producción $\mathbf{x}_i$, la subaditividad exige:

$$C\left(\sum_i \mathbf{x}_i\right) < \sum_i C(\mathbf{x}_i)$$

Economías de alcance (Panzar y Willig, 1981): producir varios bienes en una misma empresa cuesta menos que hacerlo por separado, por los *inputs* públicos internos, como la marca o el conocimiento, las indivisibilidades y las estructuras comunes. Explican la empresa multiproducto y fijan el límite horizontal del tamaño.

1.2.3. La empresa como estructura de contratos

Pregunta: Coase (*The Nature of the Firm*, 1937; Nobel en 1991) planteó por qué existen las empresas si el mecanismo de precios asigna eficientemente los recursos y, a la inversa, por qué una sola no organiza toda la producción.

Idea: usar el mercado tiene **costes de transacción**: búsqueda de contrapartes, negociación, supervisión y previsión de contingencias. La empresa existe porque, para ciertas actividades, organizar internamente la producción cuesta menos que recurrir al mercado.

Razonamiento: Williamson (1975; Nobel en 2009) distingue costes *ex ante*, de previsión, redacción y negociación de los contratos, y *ex post*, de regateo, ejecución y litigios. El coste de acudir al mercado decrece con la dimensión y el de organización interna crece por las deseconomías gerenciales. La dimensión óptima minimiza su suma, y en ella ambos costes marginales se igualan:

$$\min_x \ \mathrm{CT}(x) = \mathrm{CE}(x) + \mathrm{CI}(x) \quad \Rightarrow \quad \mathrm{CI}'(x^*) = -\mathrm{CE}'(x^*)$$

Desarrollo gráfico: el coste total es la suma vertical de ambas curvas y tiene forma de U, y la dimensión óptima, x^*, corresponde a su mínimo, C^*. Con curvas simétricas, ese punto está sobre el corte de $\mathrm{CE}(x)$ y $\mathrm{CI}(x)$, donde $\mathrm{CE}(x^*) = \mathrm{CI}(x^*)$.

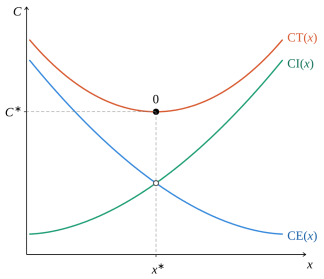
Conclusiones: sitúa los costes de usar el mercado en el centro del análisis. Pero el concepto es amplio y poco preciso, lo que permite explicar *a posteriori* cualquier frontera, y no aclara qué permite a la organización interna evitar esos costes.

En suma, el enfoque tecnológico y el contractual convergen en que la empresa minimiza el coste total de producir e intercambiar, pero dejan abierta su organización interna.

1.3. Organización

1.3.1. Enfoque de derechos de propiedad

Idea: la propiedad no es el bien, sino un conjunto de derechos sobre él, de uso, de apropiación de sus rendimientos y de transferencia, y quién



Coase (1937): la dimensión óptima, x^*, minimiza la suma del coste de mercado y del de organización interna.

los tiene determina los incentivos.

Primera generación (Alchian, 1965; Demsetz, 1967; Alchian y Demsetz, 1972): los derechos aparecen cuando la ganancia de internalizar una externalidad supera el coste de establecerlos. En la producción en equipo la contribución individual no se observa por separado y surge el escaqueo (*shirking*). La solución es un supervisor que sea **perceptor residual** del beneficio. El empresario capitalista emerge así como respuesta eficiente a un problema de medición. Para Alchian, la conducta depende de la estructura de los derechos, y en la empresa pública, cuya propiedad no puede venderse, quien decide no carga con las ganancias ni con las pérdidas de sus decisiones.

Nuevo enfoque (Grossman y Hart, 1986; Hart y Moore, 1990): con contratos incompletos, la propiedad asigna los **derechos residuales de control**, sobre todo lo que el contrato no especifica. Las inversiones específicas se hacen antes de negociar, y cada parte solo se apropia de una parte de su rendimiento, de modo que invierte menos de lo eficiente (*hold-up*). La propiedad óptima va a la parte cuya inversión específica es más importante y reúne los activos complementarios. Los costes de integrar nacen del mismo mecanismo que sus beneficios. La propiedad de un activo refuerza la posición negociadora de su titular y sus incentivos a invertir, pero debilita los de la otra parte, que por eso invierte menos.

Conclusiones: da una razón de eficiencia a la propiedad privada y precisa los costes de la integración vertical. Pero la primera generación es verbal, y los fundamentos de los contratos incompletos los discuten Maskin y Tirole (1999).

1.3.2. Enfoque transaccional

Idea: en el marco de Williamson, la estructura de gobernanza óptima de cada transacción, mercado, formas híbridas o jerarquía, depende de tres atributos.

1. *Frecuencia*: las transacciones recurrentes justifican amortizar estructuras especializadas, y las ocasionales se resuelven mejor en el mercado.
2. *Incertidumbre*: deja más contingencias sin prever y encarece las renegociaciones, lo que, con activos específicos, favorece la jerarquía.
3. *Especificidad de los activos*: las inversiones no reutilizables fuera de la relación generan cuasirrentas apropiables y exponen al *hold-up*, lo que favorece la integración vertical. La cuasirrenta apropiable es el exceso del valor del activo en su uso actual sobre el de su mejor uso alternativo (Klein et al., 1978).

Razonamiento: la elección entre mercado y jerarquía se convierte en un problema de minimización de costes de transacción, que explica la integración vertical, las relaciones a largo plazo y las formas híbridas, como las franquicias o las alianzas.

Conclusiones: hace operativa la intuición de Coase y da predicciones contrastables sobre la integración vertical. Pero es en gran medida verbal, y la especificidad de los activos es difícil de medir. La racionalidad limitada y el oportunismo se introducen como supuestos sin formalizar.

1.3.3. Enfoque de contratos completos

Idea: aplica la teoría de la agencia (Ross, 1973; Jensen y Meckling, 1976) a la relación entre un principal, el propietario, y un agente, el directivo o trabajador, con contratos que contemplan todas las contingencias verificables.

Planteamiento: el principal, neutral al riesgo, elige el esfuerzo e y el contrato salarial $w(x)$ que maximizan su beneficio esperado, sujeto a la restricción de participación del agente, con unas probabilidades $p_i(e)$ que el esfuerzo desplaza hacia los resultados altos:

$$\max_{e,\,w(\cdot)} \mathbb{E}[\pi] = \sum_i p_i(e)\,[x_i - w(x_i)]$$

CPO: el producto marginal esperado del esfuerzo iguala la compensación que el agente exige por su desutilidad marginal.

CSO: con resultado esperado cóncavo en el esfuerzo y utilidad del agente cóncava, el programa es cóncavo y la CPO es necesaria y suficiente.

Solución: si el esfuerzo es observable, se alcanza el *first best*, en el que el principal exige el esfuerzo eficiente y paga al agente averso al riesgo un salario fijo.

Conclusiones: si el esfuerzo no es observable, el pago debe depender del resultado y se sacrifica aseguramiento, lo que da un *second best*. Los **costes de agencia** suman la supervisión, la garantía y la pérdida residual, que el gobierno corporativo reduce sin eliminar.

1.4. Comportamiento

1.4.1. Modelo de variaciones conjeturales

Idea: los modelos de oligopolio difieren en lo que cada empresa espera que haga la rival cuando altera su producción. Recogida esa expectativa en la **variación conjetural** (Bowley, 1924; Frisch, 1933), $v_i \equiv \partial x_j/\partial x_i$, todos son casos de una misma CPO.

Planteamiento: cada empresa elige su cantidad tomando la conjetura como dada:

$$\max_{x_i}\ \pi_i = p(x_i + x_j)\cdot x_i - \mathrm{CT}_i(x_i)$$

CPO: como $dX/dx_i = 1 + v_i$, la empresa iguala su coste marginal a un ingreso marginal percibido o, dividiendo por el precio, con la cuota s_i y la elasticidad de la demanda, fija un margen de Lerner que depende de la conjetura:

$$p(X) + p'(X)\,(1+v_i)\,x_i = \mathrm{CT}_i'(x_i)$$
$$\iff \frac{p - \mathrm{CT}_i'}{p} = \frac{(1+v_i)\,s_i}{|\varepsilon_{Xp}|}$$

La empresa anticipa que el precio caerá por su propia expansión y por la reacción que conjetura en la rival, de modo que, cuanto mayor es la conjetura, más restringe la producción y mayor es el margen. El valor de v_i selecciona el modelo:

- Con $v_i = -1$, la rival compensa la variación (competencia perfecta o Bertrand, con margen nulo).
- Con $v_i = 0$, la rival no reacciona (Cournot).
- Con $v_i = \varphi_j'(x_i)$, la rival sigue su función de reacción (líder de Stackelberg).
- Con $v_i = 1$, la rival replica la variación (cártel).

Con n empresas simétricas, el margen es $[1+(n-1)v]/(n\,|\varepsilon_{Xp}|)$: crece con la conjetura y decrece con la elasticidad y, salvo en el cártel, con el número de empresas. El **parámetro de conducta**, $\theta \equiv 1+(n-1)v \in [0,n]$, es el que estima la NEIO.

Conclusiones: ofrece una lectura común de los modelos estáticos de oligopolio y un índice continuo del poder de mercado. Pero carece de fundamento estratégico, porque es pseudodinámico y sus conjeturas son *ad hoc*. Con conjeturas consistentes (Bresnahan, 1981), demanda lineal y coste marginal constante, solo cabe $v = -1$, y la teoría de juegos lo sustituyó. Con $p = a - X$ y coste marginal c, la rival que conjetura v responde con $x_j = (a - c - x_i)/(2+v)$, de pendiente $-1/(2+v)$, que solo coincide con la conjetura si $v = -1$. Con costes marginales crecientes, la conjetura consistente queda entre -1 y 0.

1.4.2. Diferenciación horizontal (Hotelling, 1929)

Los consumidores se distribuyen uniformemente en un segmento, con costes de transporte, y las empresas eligen localización y precio. Con precios fijos, se aglomeran en el centro (el principio de **mínima diferenciación**). Con precios endógenos y costes cuadráticos, se separan al máximo para relajar la competencia en precios (d'Aspremont et al., 1979). El resultado es muy sensible a los costes de transporte, y con costes lineales puede no existir equilibrio. La localización es así una decisión estratégica, porque acercarse al rival capta demanda pero intensifica la competencia en precios.

1.4.3. Innovación (Schumpeter, 1942; Arrow, 1962)

Schumpeter: la innovación crece con el tamaño y el poder de mercado, porque solo la gran empresa financia la I+D y diversifica su riesgo, y solo el beneficio extraordinario permite recuperar la inversión.

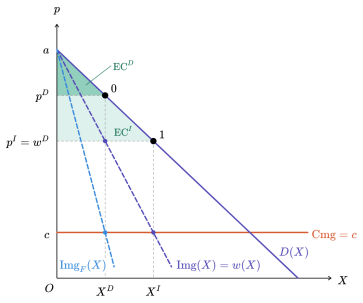
Arrow: el monopolista innova menos por el **efecto reemplazo**, porque la innovación canibaliza el beneficio que ya obtiene. Con una innovación de proceso no drástica, el monopolista gana la diferencia entre dos beneficios positivos y la empresa competitiva, la diferencia entre uno positivo y cero. La competitiva, que parte de un beneficio nulo, cobra algo menos que el coste anterior, se queda con todo el mercado y gana el ahorro de coste sobre toda la demanda.

Como el excedente social supera a ambos, el orden de incentivos es planificador, competencia y monopolio. Hay dos matices: el *efecto eficiencia* (Gilbert y Newbery, 1982), porque el monopolista amenazado puja más por la patente, y la financiación de la I+D, núcleo del argumento schumpeteriano.

Evidencia: el gasto en I+D crece en proporción al tamaño, y Aghion, Bloom, Blundell, Griffith y Howitt (2005) hallan una **U invertida**: con poca competencia domina el efecto de escapar de la competencia, y con mucha, el schumpeteriano. Ni la competencia ni el monopolio alcanzan el incentivo óptimo.

1.4.4. Otros modelos de comportamiento

1. *Doble marginalización* (Spengler, 1950): si un fabricante monopolista vende con precio mayorista lineal a un minorista también monopolista, cada uno añade su margen sin internalizar el del otro. Con demanda lineal, la cadena vende la mitad que la empresa integrada, más cara, con un beneficio conjunto un 25 % inferior y la cuarta parte del excedente del consumidor. La integración, o una restricción vertical que elimine el segundo margen, mejora a empresas y consumidores (regla de la razón de Chicago), aunque puede cerrar el mercado a los rivales.



Doble marginalización con demanda lineal y coste marginal constante.

2. *Diferenciación y variedad* (Spence, 1976; Dixit y Stiglitz, 1977): margen constante, $p = \frac{\sigma}{\sigma-1}$ Cmg, y beneficio nulo con libre entrada (A18).
3. *Diferenciación vertical* (Shaked y Sutton, 1982): calidades distintas para suavizar la competencia en precios y número finito de empresas (A18).

En suma, la estructura de costes delimita cuántas empresas caben, los contratos dónde terminan y la agencia qué persiguen, y la conducta altera la propia estructura del mercado.

2. Teoría del mercado

El recorrido parte de Harvard y Chicago, sigue con los determinantes de la estructura de los mercados y termina con su regulación.

2.1. Antecedentes

2.1.1. Paradigma ECR: Harvard (años cuarenta)

Idea: para Mason y Bain, la estructura del mercado, número de empresas, concentración, barreras y diferenciación, determina la conducta, en precios, inversión o publicidad, y esta el resultado. La causalidad es unidireccional, de modo que toda industria con pocas empresas resulta sospechosa, y la política de competencia mide la concentración con el índice de Herfindahl-Hirschman y los ratios de concentración.

Conclusiones: inaugura el análisis empírico sistemático de los mercados y da instrumentos operativos. Pero es estático, sin conducta estratégica, y su causalidad es discutible, porque las empresas más rentables crecen y concentran el mercado.

2.1.2. Escuela de Chicago (años cincuenta y sesenta)

Idea: para Stigler, Demsetz, Posner y Bork, la concentración no indica necesariamente poder de mercado, por dos razones. Las grandes empresas pueden serlo por ser más eficientes, de modo que la causalidad va de la eficiencia a la concentración, y la forma en L de los costes medios implica que la gran dimensión no comporta desventaja. Además, la amenaza de entrada disciplina los precios incluso en mercados concentrados. Con libre entrada el mercado se autorregula, y la regulación debe limitarse a eliminar las barreras legales.

Conclusiones: subraya la eficiencia y la entrada potencial, que formalizarán los mercados impugnables. Pero ignora las barreras estratégicas y la predación, que puede ser racional. En suma, Harvard era demasiado estático y Chicago, demasiado optimista.

2.2. Estructura

2.2.1. Causas: las barreras a la entrada

Definición: las **barreras a la entrada** deciden si la concentración es inocua o lesiva, y su definición fija el perímetro de la intervención.

1. *Bain (1956)*: condiciones que permiten a las instaladas fijar persistentemente precios supracompetitivos sin inducir la entrada: economías de escala, diferenciación de producto y ventajas absolutas de costes. Con economías de escala, el entrante ha de elegir entre entrar pequeño con desventaja de costes o grande deprimiendo el precio, y la lealtad de marca le obliga a gastos de introducción ya amortizados por las instaladas.
2. *Stigler (1968)*: costes que soporta el entrante y no soportaron las instaladas. Las economías de escala ya no son barrera. Por eso Harvard era intervencionista y Chicago, abstencionista.
3. *Baumol y Willig (1981)*: los costes fijos no son barrera, pero sí los **costes hundidos**, que crean una asimetría de riesgo entre quien ya los comprometió y quien aún puede evitarlos.

Barreras estratégicas: las crean las instaladas, como el precio límite (Bain-Sylos-Modigliani), la inversión en capacidad como compromiso creíble (Spence, 1977; Dixit, 1980) o la proliferación de marcas (Schmalensee, 1978). Con las legales forman las barreras artificiales, frente a las naturales, que nacen de la tecnología y la demanda.

2.2.2. Mercados impugnables (Baumol, 1982)

Idea: Baumol retoma la distinción de Demsetz entre la competencia en el mercado y por el mercado y, con Panzar y Willig, define un mercado **impugnable** como aquel en que los entrantes pueden entrar y salir sin desventaja: igual acceso a la tecnología, costes fijos pero no hundidos, y entrada más rápida que la reacción de la instalada. La reacción se retrasa por causas como los costes de menú, mientras la demanda responde al instante al entrante, y sin costes hundidos la inversión es recuperable al salir.

Razonamiento: incluso un monopolio natural fija un precio igual al coste medio, porque cualquier precio superior atraería una entrada relámpago (*hit-and-run*). La disciplina procede de la competencia potencial, y la estructura deja de ser el indicador relevante. Formalmente, solo es de equilibrio una configuración *sostenible*, que atiende la demanda sin pérdidas y no permite una entrada rentable a un precio igual o menor:

$$\sum_i x_i = X(p), \qquad p\,x_i - C(x_i) \geq 0$$
$$p^e x^e - C(x^e) \leq 0 \quad \forall\, p^e \leq p,\ x^e \leq X(p^e)$$

Ello implica beneficio nulo, eficiencia productiva y, con dos o más empresas, $p = \text{Cmg}$. El monopolio natural queda en $p = \text{Cme} > \text{Cmg}$, el *second best*.

Conclusiones: invierte la carga de la prueba del paradigma ECR y fue muy influyente en la desregulación del transporte aéreo y las telecomunicaciones. Pero la ausencia total de costes hundidos rara vez se da.

2.2.3. Mercados de varias caras (Rochet y Tirole, 2003)

Idea: una plataforma pone en contacto a dos grupos con **externalidades cruzadas de red**: el valor para un lado depende del número de participantes del otro, como en las tarjetas de pago, entre comercios y titulares, los sistemas operativos o los mercados en línea.

Razonamiento: la estructura de precios no es neutral, es decir, importa el reparto de los precios entre los dos lados y no solo su nivel agregado. Suele ser óptimo subvencionar al lado más sensible al precio, o al que genera más externalidad, incluso con precios nulos o negativos. Un precio inferior al coste en un lado no prueba por sí solo la predación, y el mercado relevante debe analizarse con ambos lados.

Conclusiones: explica los servicios gratuitos o subvencionados para uno de los lados. Pero los resultados dependen mucho de las externalidades y de si los usuarios operan en varias plataformas.

2.2.4. Predación (Areeda y Turner, 1975)

Idea: la **predación** sacrifica beneficios presentes con precios bajos para expulsar a un rival y recuperarlos después. La escuela de Chicago (McGee, 1958) la consideraba irracional.

Regla de Areeda-Turner: aproxima el coste marginal por el coste variable medio, y un precio inferior se presume predatorio. Es demasiado permisiva, porque la predación por reputación puede darse por encima de ese coste, y estática, porque ignora la recuperación. Bolton, Brodley y Riordan (2000) proponen dos fases: primero, la plausibilidad de la recuperación, como exige en EE. UU. *Brooke Group* (1993), y después, la prueba de costes. La segunda fase examina los precios por debajo del coste y la ausencia de justificación de eficiencia.

Doctrina europea: *AKZO* (1991) presume abusivos los precios inferiores al coste variable medio, y los situados entre este y el coste total medio, si forman parte de un plan de exclusión, sin exigir la recuperación. El contraste entre la recuperación exigida y la intención suficiente refleja pesos distintos de los errores de tipo I y II. Las reglas de costes dan seguridad jurídica, pero solo capturan la predación más burda, y las de dos fases incorporan la teoría estratégica a costa de mayor complejidad probatoria.

2.3. Regulación

Cuando una sola empresa es la estructura eficiente, el problema es regularla sin la información que ella tiene.

2.3.1. Competencia por el mercado (Demsetz, 1968)

Idea: subastar periódicamente el derecho exclusivo a servir el mercado (*franchise bidding*). Con licitación competitiva, el precio converge al coste medio y se replica una regulación de beneficio nulo sin regulador permanente ni conocer los costes de la empresa. Williamson (1976) objetó que los contratos de concesión son incompletos, que los activos específicos dificultan la reversión y que el incumbente acumula ventajas informativas en las relicitaciones.

2.3.2. Competencia por comparación (Shleifer, 1985)

Idea: el regulador fija el precio de cada empresa con los costes de empresas comparables de otros mercados, p. ej. distribuidoras regionales (*yardstick competition*). Fijarlo con los costes declarados por la propia empresa destruiría sus incentivos a reducirlos. Con la comparación, en

cambio, sus reducciones de costes no alteran su precio, y la empresa se beneficia íntegramente de ellas. Exige comparadores similares, o corregir econométricamente sus diferencias, y ausencia de colusión entre ellos.

2.3.3. Regulación por incentivos (Laffont y Tirole, 1986)

Idea: el regulador, el principal, no conoce la eficiencia de la empresa, el agente, y le ofrece un menú de contratos entre los que se autoselecciona, revelando su eficiencia a cambio de una renta informacional.

Supuestos: un monopolio natural tiene costes $\mathrm{CT}(x) = (\varphi - e)\,x + F$, donde φ es su eficiencia intrínseca y e su esfuerzo, ambos no observables, de modo que se combinan selección adversa y riesgo moral. El esfuerzo cuesta $\psi(e)$, convexa, y cada euro de fondos públicos cuesta $1+\lambda$.

Planteamiento: por el principio de revelación (Myerson, 1979), el regulador elige un mecanismo directo que maximiza el bienestar esperado sujeto a la participación y a la compatibilidad de incentivos de cada tipo. Con la producción normalizada a uno y sin costes fijos, el coste observado es $C = \varphi - e$, que el regulador reembolsa junto con una transferencia neta t, y la renta de la empresa es $U = t - \psi(e)$. Con W el bienestar que genera cada tipo, G y g la distribución y la densidad de los tipos, $\hat\varphi$ el tipo anunciado y $C(\hat\varphi) = \hat\varphi - e(\hat\varphi)$ su coste asignado, el mecanismo resuelve:

$$\max_{\{t(\varphi),\,e(\varphi)\}} \int_{\underline\varphi}^{\bar\varphi} W(t(\varphi), e(\varphi); \varphi)\, g(\varphi)\, d\varphi$$

$$\text{s.a.:} \begin{cases} U(\varphi) = t(\varphi) - \psi(e(\varphi)) \geq 0 & \forall \varphi \\ U(\varphi) \geq t(\hat\varphi) - \psi(\varphi - C(\hat\varphi)) & \forall \varphi, \hat\varphi \end{cases}$$

CPO: el ahorro marginal del esfuerzo iguala su desutilidad marginal más el coste de la renta informacional que hay que ceder a los tipos más eficientes:

$$1 = \psi'(e(\varphi)) + \frac{\lambda}{1+\lambda} \cdot \frac{G(\varphi)}{g(\varphi)} \cdot \psi''(e(\varphi))$$

Un poco más de esfuerzo ahorra una unidad de coste, pero si el regulador lo pide al tipo φ, su contrato atrae a los tipos más eficientes, de masa $G(\varphi)$, que podrían imitarlo con menos esfuerzo, y hay que subirles la renta. Cada euro de renta cuesta λ, frente al valor social $1+\lambda$ del ahorro, de ahí el factor $\lambda/(1+\lambda)$.

El esfuerzo se distorsiona a la baja respecto al *first best*, $\psi'(e) = 1$, tanto más cuanto mayor es λ, salvo para el tipo más eficiente (**no hay distorsión en la cima**).

Solución: el mecanismo se implementa con un **menú de contratos** lineales en el coste realizado, cuya potencia es la fracción del sobrecoste que recae sobre la empresa. Con un pago fijo a, un coste objetivo C^* y una potencia $b \in [0,1]$, que vale 1 en el precio fijo y 0 en el coste más margen, el menú es:

$$t(\hat\varphi, C) = a(\hat\varphi) - b(\hat\varphi) \cdot [C - C^*(\hat\varphi)]$$

- La empresa eficiente elige un contrato de precio fijo, de modo que asume todo el riesgo de costes y retiene todo el ahorro de su esfuerzo, y percibe una renta informacional para que no le compense hacerse pasar por ineficiente.
- La ineficiente elige el contrato de menor potencia, el más próximo al de coste más margen: el regulador absorbe buena parte de sus sobrecostes, con un esfuerzo distorsionado a la baja, pero sin ceder renta.
- Los tipos intermedios combinan un componente fijo y un reparto parcial de los sobrecostes.

El menú resuelve el *trade-off* entre la eficiencia, que pide contratos de alta potencia, y la extracción de rentas. Como la empresa ajusta su esfuerzo hasta $\psi'(e) = b$, el menú induce el esfuerzo óptimo con $b(\hat\varphi) = \psi'(e(\hat\varphi))$, igual a 1 para el tipo más eficiente y decreciente con φ.

Conclusiones: fundamenta la regulación *price-cap* frente a la de tasa de retorno, y valió a Tirole el Nobel de 2014. Pero exige mucha información al regulador, es estático, expuesto al efecto trinquete, e ignora la captura del regulador.

En suma, la teoría del mercado ha pasado de tomar la estructura como dato a explicarla, y la regulación, de fijar precios a resolver el problema informativo del regulador.

3. Nueva teoría de la organización industrial empírica

Origen: Bain (1951) halla, para 42 industrias manufactureras estadounidenses de 1936-1940, beneficios mayores en las que las ocho mayores empresas superan el 70 % de las ventas. Demsetz (1973) objetó que, si fuera colusión, todas las empresas se beneficiarían, pero solo las grandes son más rentables, por lo que concentración y beneficio son endógenos. La lectura estructuralista veía en la correlación la colusión que la concentración facilita, y la de Demsetz, la eficiencia de Chicago, de modo que la disputa no podía zanjarse con regresiones entre industrias.

Idea: la **NEIO** (Bresnahan, 1989) sustituye esas regresiones por el análisis estructural de industrias concretas, especificando la demanda y la conducta para estimar directamente el poder de mercado.

Parámetro de conducta: la NEIO escribe el margen como $(p - \mathrm{Cmg})/p = \theta/(n\,|\varepsilon_{X_p}|)$, con $\theta = 0$ en competencia, 1 en Cournot y n en el cártel. Bresnahan (1982) y Lau (1982) muestran que θ está identificado si la demanda sufre rotaciones exógenas, de modo que el grado de competencia pasa a ser estimable. Así puede distinguirse la concentración que refleja eficiencia de la que refleja colusión.

3.1. Delimitación del mercado relevante

El análisis parte de delimitar el **mercado relevante**, de producto y geográfico, con el test del monopolista hipotético (SSNIP). Solo dentro de él tienen sentido los índices.

3.2. Medición del poder de mercado

3.2.1. Índices de poder de mercado

Índice de Lerner: el margen relativo entre precio y coste marginal, de 0 en competencia perfecta a 1. En Cournot, $I_{L,i} = s_i/|\varepsilon_{X_p}|$, y la media ponderada por las cuotas es la suma de sus cuadrados (el HHI en tanto por uno) dividida por la elasticidad (Cowling y Waterson, 1976):

$$I_{L,i} = \frac{p - \mathrm{Cmg}_i}{p} \in [0,1]$$

$$I_L = \sum_{i=1}^{n} s_i\, I_{L,i} = \frac{\sum_{i=1}^{n} s_i^2}{|\varepsilon_{X_p}|} \in [0,1]$$

En el monopolio, el índice iguala la inversa de la elasticidad de la demanda. Como el coste marginal no se observa, se aproxima con el margen contable, (ingresos − CV)/ingresos, o se estima con la NEIO.

***q* de Tobin**: el valor de mercado de la empresa entre el coste de reposición de sus activos. Con libre entrada tiende a 1, y un valor persistentemente superior indica rentas que la competencia no erosiona. No exige conocer el coste marginal, pero omite los intangibles, lo que lo sesga al alza.

3.2.2. Índices de posición

La **cuota de mercado**, $s_i = x_i/X$, va de 0 a 1. En la Unión Europea, una cuota del 50 % o superior permite presumir posición dominante (*AKZO*, 1991), aunque puede reflejar eficiencia (Demsetz, 1973).

3.2.3. Índices de concentración

Ratio de concentración: $\mathrm{CR}_k = \sum_{i=1}^{k} s_i$, la cuota de las k mayores empresas. Es sencillo, pero ignora el reparto de las cuotas. Va de k/n, con cuotas iguales, a 1, y Bain (1951) utilizó el CR_8.

Índice de Herfindahl-Hirschman: suma los cuadrados de las cuotas en porcentaje (de $1/n$ a 1 en tanto por uno, cuya inversa es el número equivalente de empresas). Una fusión de i y j lo eleva en $2\,(100\,s_i)(100\,s_j)$:

$$\mathrm{HHI} = \sum_{i=1}^{n} (100\,s_i)^2 \in \left[\frac{10\,000}{n},\ 10\,000\right]$$

- *Unión Europea (2004)*: los problemas son improbables con un HHI posterior inferior a 1000, entre 1000 y 2000 con un aumento inferior a 250, o superior a 2000 con un aumento inferior a 150.
- *Estados Unidos (2023)*: más de 1800 indica un mercado muy concentrado, y se presume lesiva la fusión que lo crea o refuerza con un aumento superior a 100.

3.3. Evidencia empírica

Lerner: Porter (1983) estima la demanda y la oferta del cártel ferroviario del Joint Executive Committee (1880-1886): en los periodos colusivos el margen se acerca al de Cournot, y las guerras de precios son castigos temporales que sostienen la colusión. Ellison (1994) halla un margen más próximo al de monopolio. Su modelo, más rico, estima una demanda más elástica e indicios de rebajas secretas de precios, que apoyan débilmente a Green y Porter (1984) frente a Rotemberg y Saloner (1986).

Cuota: Gal (2003) muestra que, en las economías pequeñas, la EME elevada hace inevitable e incluso eficiente una concentración alta, de modo que los umbrales no son trasladables entre jurisdicciones.

Otros: entre el 5,3 y el 7,4 % de las adquisiciones farmacéuticas buscan interrumpir un proyecto rival, las *killer acquisitions* (Cunningham et al., 2021), y Rysman (2004) estima externalidades cruzadas en las guías telefónicas y prefiere una estructura más competitiva al monopolio.

A16. Análisis de mercados (I). El modelo de competencia perfecta. Análisis de equilibrio parcial: corto y largo plazo; dinámicas de ajuste y estabilidad del equilibrio. Análisis de eficiencia y bienestar

1. Fundamentos del modelo marshalliano

La formalización del equilibrio parcial competitivo nace con Marshall (1890) y se completa con Hicks, Allen y Pigou. Sus supuestos se agrupan en cinco bloques.

1.1. Supuestos

Metodología: el análisis es de equilibrio parcial, con dos bienes: el bien final y un bien compuesto que engloba los demás mercados y actúa como numerario. La simplificación permite tratar el mercado como aislado y representar su oferta y su demanda.

Entorno: no hay fallos de mercado: ni externalidades, ni bienes públicos, ni poder de mercado, ni información imperfecta, porque es completa y simétrica, ni costes de transacción.

Agentes: sin sector público, hay consumidores racionales que maximizan su utilidad y empresas dirigidas por sus propietarios, sin conflicto de agencia, que maximizan el beneficio económico. Están en gran número, de modo que el mercado está atomizado y todos son **precio-aceptantes**, $\partial p/\partial x_i^s = \partial p/\partial x_i^d = 0$, es decir, cada uno compra o vende cuanto quiere al precio de mercado, y la curva que afronta es perfectamente elástica.

Tecnología: el modelo parte de cuatro supuestos.

- *Producto homogéneo*: todas las empresas producen un único bien normal, $x = f(\mathbf{z})$, idéntico, con elasticidad de sustitución infinita entre variedades, y perfectamente divisible.
- Divisibilidad perfecta de los factores, de modo que las funciones de producción y de costes son continuas.
- Rendimientos mixtos, a corto plazo en el factor variable y a largo plazo en la escala: la producción tiene forma de S y el coste medio, de U, con una dimensión óptima de planta.
- *Homogeneidad tecnológica*: todas las empresas tienen la misma tecnología y el mismo acceso al mercado de factores.

De los supuestos de entorno y tecnología se deriva la **ley del precio único**: bienes idénticos tienen precios idénticos.

Mercado de factores: es perfecto y grande en relación con el del bien final, y las empresas son también precio-aceptantes en él, sin economías ni deseconomías de escala externas. Contratan o alquilan cuantos factores quieran al precio de mercado, $\partial p_i^z/\partial z_i = 0$.

1.2. Demanda de mercado

Definición: la **demanda de mercado** es la suma horizontal de las demandas individuales de los H consumidores, que proceden de la maximización de la utilidad, con la renta, las preferencias y los demás precios constantes:

$$D_x(p_x) = \sum_{h=1}^{H} d_x^h(p_x)$$

Propiedades: la suma horizontal no conserva todas las de las demandas individuales.

1. Hereda la continuidad, aunque presenta quiebros en los precios a los que un consumidor entra en el mercado o sale de él.
2. Hereda la homogeneidad de grado cero en precios y renta (ausencia de ilusión monetaria).
3. Tiene pendiente negativa si la tienen las individuales, condición suficiente pero no necesaria.
4. Cumple la ley de Walras agregada y es observable, pero no hereda la simetría de los efectos sustitución ni la ley de la demanda compensada, y no puede tratarse como la de un consumidor representativo. Solo depende de la renta agregada, y no de su distribución, si las curvas de Engel son rectas paralelas (Gorman, 1953), como con preferencias cuasilineales.

1.3. Oferta: la decisión de la empresa

Beneficio: el **beneficio económico** es la diferencia entre el ingreso total y el coste total, $\pi(x) = \mathrm{IT}(x) - \mathrm{CT}(x)$.

Ingreso: el ingreso total es $\mathrm{IT}(x) = p(x) \cdot x$. Su derivada, el ingreso marginal, es la contribución de la última unidad vendida, $\mathrm{Img}(x) = p + (\partial p/\partial x)\, x = p\,(1 + 1/\varepsilon_{xp})$, y el ingreso medio, $\mathrm{Ime}(x) = p(x)$, es la curva de ventas de la empresa. Como es precio-aceptante, $\partial p/\partial x = 0$, de modo que $\mathrm{Img} = \mathrm{Ime} = p$ y la curva de ventas es horizontal. Por encima del precio de mercado no vende nada.

Coste: el coste total procede de minimizar el gasto para cada nivel de producción. Los rendimientos mixtos dan un coste medio en U y un coste marginal creciente en el tramo relevante. A corto plazo hay costes fijos y variables, y a largo plazo todos son variables.

1.3.1. Equilibrio: enfoque de mercado

Planteamiento: la empresa elige su nivel de producción:

$$\max_x \ \pi = p \cdot x - \mathrm{CT}(x) \quad \text{s.a.: } x \geq 0$$

CPO: iguala el ingreso marginal al coste marginal, lo que, al ser precio-aceptante, equivale a igualar a este el precio, sin margen alguno:

$$\mathrm{Img}(x) = \mathrm{Cmg}(x) \ \Leftrightarrow \ p = \mathrm{Cmg}(x)$$

Resulta de derivar el beneficio, $\partial\pi/\partial x = p - \partial\mathrm{CT}/\partial x = 0$, y ya la derivó Cournot (1838). Con $p > \mathrm{Cmg}$ conviene ampliar la producción, porque la última unidad deja más ingreso del que cuesta, y con $p < \mathrm{Cmg}$, recortarla.

CSO: el coste marginal debe ser estrictamente creciente en el óptimo, $\partial\mathrm{Cmg}/\partial x > 0$, porque en el tramo decreciente la empresa siempre ganaría produciendo una unidad más. Con la condición de beneficio no negativo a largo plazo, $p \geq \mathrm{Cme}$, hace incompatibles la competencia perfecta y los rendimientos crecientes. En términos de la segunda derivada del beneficio:

$$\frac{\partial^2\pi}{\partial x^2} < 0 \ \rightarrow \ \frac{\partial \mathrm{Cmg}}{\partial x} > 0$$

1.3.2. Equilibrio: enfoque de la producción

Planteamiento: la empresa elige las cantidades de factores:

$$\max_{\mathbf{z}} \ \pi = p \cdot f(\mathbf{z}) - \mathbf{p}^z \cdot \mathbf{z} \quad \text{s.a.: } \mathbf{z} \geq 0$$

CPO: se leen de cuatro formas.

1. *Minimización del coste*: la relación marginal de sustitución técnica iguala el precio relativo de los factores, $\mathrm{RMST}_{21} = p_1^z/p_2^z$, de modo que hay eficiencia económica y no solo técnica.
2. *Igualdad de productividades por euro*: $f_1'/p_1^z = f_2'/p_2^z = 1/\mathrm{Cmg}$.
3. *Maximización del beneficio*: cada factor se contrata hasta que el valor de su productividad marginal iguala su precio, $p \cdot f_i' = p_i^z$.
4. *Regla de precios*: $p = p_i^z/f_i' = \mathrm{Cmg}$, sin margen alguno, consecuencia de que la empresa sea precio-aceptante.

CSO: la producción debe ser estrictamente cóncava en el tramo relevante, $f_{11} < 0$ y $f_{11}f_{22} - f_{12}^2 > 0$, es decir, rendimientos marginales decrecientes con efectos cruzados acotados, no rendimientos decrecientes a escala.

1.3.3. Funciones de oferta y demanda de factores

Demandas incondicionadas de factores: $z_i^d = d_i^z(p, \mathbf{p}^z)$ admiten que varíe la producción.

- Son continuas y homogéneas de grado cero en todos los precios, de modo que la inflación pura no altera las elecciones.
- El efecto precio propio es no positivo, más potente que el de Shephard porque incluye el efecto producción. El cruzado es indeterminado, pero simétrico. Con dos factores, su signo es el opuesto al de f_{12}, de modo que, si son complementarios en la producción, $f_{12} > 0$, la subida del precio de uno reduce la demanda del otro. Por el lema de Hotelling y el teorema de Schwarz, además, el efecto del precio de un factor sobre la oferta es el opuesto al del precio del bien sobre la demanda de ese factor:

$$\frac{\partial d_i^z}{\partial p_j^z} = \frac{\partial d_j^z}{\partial p_i^z} \qquad \frac{\partial s}{\partial p_i^z} = -\frac{\partial d_i^z}{\partial p}$$

- Coinciden con las demandas condicionadas evaluadas en la producción óptima, porque quien maximiza el beneficio minimiza también el coste:

$$d_i^z(p, \mathbf{p}^z) = d_i^c(\mathbf{p}^z, s(p, \mathbf{p}^z))$$

- Son observables y estimables, se obtienen de la función de máximo beneficio por el lema de Hotelling y, con la oferta, permiten recuperarla si forman un sistema de ofertas netas con matriz de derivadas simétrica y semidefinida positiva.

Oferta individual: $x^s = s(p, \mathbf{p}^z)$ es el tramo creciente del coste marginal por encima del precio de cierre. Es continua salvo en ese umbral, donde salta de cero al *output* que minimiza el coste variable medio, diferenciable por tramos, homogénea de grado cero y no decreciente en el precio del bien, por la convexidad de la función de beneficio.

1.3.4. Función de máximo beneficio

Definición: la **función de máximo beneficio**, $\pi^* = \Pi(p, \mathbf{p}^z)$, recoge el beneficio máximo para cada nivel de precios.

Propiedades: es continua, homogénea de grado uno, no decreciente en el precio del producto, no creciente en los de los factores y **convexa** en todos los precios. Al subir el precio del bien, el beneficio crece cada vez más deprisa, porque la empresa reajusta su producción. Al subir el de un factor, cae cada vez más despacio, porque la empresa lo sustituye. Por la homogeneidad, $\Pi(\lambda p, \lambda \mathbf{p}^z) = \lambda\,\Pi(p, \mathbf{p}^z)$, un aumento equiproporcional de todos los precios eleva el beneficio máximo en la misma proporción sin alterar las decisiones de la empresa.

Lema de Hotelling: sus derivadas recuperan la conducta de la empresa, como la identidad de Roy y el lema de Shephard en el consumidor:

$$x^s = \frac{\partial \Pi}{\partial p} \qquad z_i^d = -\frac{\partial \Pi}{\partial p_i^z}$$

Con la convexidad, la oferta no disminuye con el precio del bien ni la demanda de un factor aumenta con su precio. Toda función con estas

propiedades es la de máximo beneficio de alguna tecnología y, como es observable, permite estimar la oferta y las demandas de factores sin estimar la tecnología (el enfoque dual). Es, además, el beneficio del enfoque de mercado evaluado en la oferta óptima, con $\mathrm{CT}(x) = c(\mathbf{p}^z, x)$, de modo que ambos enfoques llevan al mismo beneficio máximo:

$$\Pi(p, \mathbf{p}^z) = p \cdot s(p, \mathbf{p}^z) - c(\mathbf{p}^z, s(p, \mathbf{p}^z))$$

Así ocurre p. ej. con la tecnología $x = L^{1/4}K^{1/4}$ y $\mathbf{p}^z = (w, r)$, cuyo coste es $\mathrm{CT}(x) = 2x^2(wr)^{1/2}$. La CPO, $p = 4x(wr)^{1/2}$, da la oferta y una función de máximo beneficio, homogénea de grado uno y convexa en p, cuya derivada respecto al precio la recupera:

$$x^s = \frac{p}{4(wr)^{1/2}} \qquad \Pi(p, w, r) = \frac{p^2}{8(wr)^{1/2}}$$

En suma, el modelo marshalliano es la referencia normativa de la teoría de mercados, porque da funciones de oferta y de demanda de factores con propiedades análogas a las del consumidor y un equilibrio eficiente. Pero sus supuestos rara vez se cumplen a la vez, y es estático y de equilibrio parcial.

2. Análisis temporal del modelo

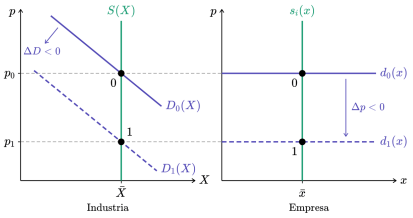
Siguiendo a Marshall, el modelo se estudia en tres horizontes, que se distinguen por la capacidad de ajuste de cada empresa ante los cambios de la demanda. Esa capacidad da forma a la oferta y decide si la empresa puede tener beneficios o pérdidas.

2.1. Muy corto plazo: subastas

Supuestos: en el **periodo de mercado** de Marshall, el número de empresas es fijo, sin libre entrada ni salida, y la cantidad ofrecida está dada, como un *stock* de bienes perecederos ya producido.

Equilibrio: con la producción hecha, cada empresa ofrece lo mismo sea cual sea el precio, y la oferta individual y la de mercado son verticales. El precio lo fija solo la demanda. Como en una subasta, su única función es de **racionamiento**, porque reparte las existencias para que el mercado se vacíe, sin señalar nada a los productores. La empresa, sin poder ajustar su oferta, puede tener beneficios o pérdidas.

Desarrollo gráfico: la oferta de mercado, $S(X)$, es vertical en la cantidad dada, $\bar{X}$, y la de cada empresa, $s_i(x)$, en $\bar{x}$. El corte de la demanda decreciente, $D_0(X)$, con $S(X)$ fija el precio p_0, al que cada empresa, frente a una demanda horizontal, $d_0(x)$, vende $\bar{x}$.



Muy corto plazo: con la oferta rígida en $\bar{X}$, la demanda fija el precio.

Estática comparativa: un desplazamiento de la demanda se traslada íntegramente al precio, en su mismo sentido, sin alterar la cantidad, y el precio varía tanto más cuanto menos elástica es la demanda. La oferta solo cambia si lo hacen las existencias, no los costes, que solo afectan a la producción futura. Una pérdida de parte del *stock* reduce la cantidad en toda su cuantía y sube el precio, tanto más cuanto menos elástica es la demanda.

2.2. Corto plazo: posibilidad de pérdidas

Supuestos: el número de empresas sigue fijo, pero la producción no. Cada empresa tiene al menos un factor fijo y elige su producción con el variable. Los precios de los factores son constantes, porque la industria es pequeña en relación con los mercados de factores.

Oferta individual: la empresa produce en el tramo creciente del coste marginal, y su viabilidad depende de dónde quede el precio.

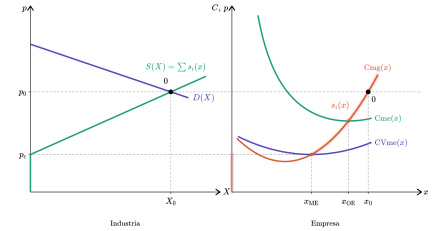
1. Precios bajos, $p < \mathrm{CVme}(x_{\mathrm{ME}})$: el ingreso no cubre siquiera los costes variables, y la empresa cierra, porque produciendo perdería más que los costes fijos.
2. Precios medios, entre el **mínimo de explotación** y el **óptimo de explotación**, $\mathrm{Cme}(x_{\mathrm{OE}})$: la empresa pierde dinero, pero le conviene producir, porque cubre todo el coste variable y parte del fijo.
3. Precios altos, $p \geq \mathrm{Cme}(x_{\mathrm{OE}})$: cubre todos sus costes y, por encima del óptimo de explotación, obtiene beneficios extraordinarios.

La oferta a corto plazo es, por tanto, discontinua:

$$x^s_{\mathrm{CP}} = \begin{cases} s(p, \mathbf{p}^z_0) & \text{si } p \geq \mathrm{CVme}(x_{\mathrm{ME}}) \\ 0 & \text{si } p < \mathrm{CVme}(x_{\mathrm{ME}}) \end{cases}$$

Oferta de mercado: es la suma horizontal de las individuales y hereda sus propiedades: discontinua en cada precio de cierre, aunque la agregación la suaviza, homogénea de grado cero y no decreciente, con una elasticidad que es la media de las individuales ponderada por las cuotas. A diferencia de la demanda de mercado, se comporta como la oferta de un único productor que maximizase el beneficio agregado. Parte del eje de precios en el precio de cierre, p_c, el mínimo de $\mathrm{CVme}(x)$, por debajo del cual ninguna empresa produce.

Equilibrio: lo fija el corte de oferta y demanda, la **cruz marshalliana**. Marshall la comparó con una tijera, de la que no puede decirse qué hoja corta. El precio cumple una doble función: raciona la demanda y señala a las empresas cuánto producir. En el periodo de mercado lo fijaba solo la demanda, y a corto plazo lo fijan ambas.



Corto plazo: al precio de la industria, p_0, la empresa produce x_0 por encima del óptimo de explotación y obtiene beneficios.

Estática comparativa: las perturbaciones de demanda, que nacen de las preferencias, la renta, los precios de los bienes relacionados o los impuestos a los compradores, mueven precio y cantidad en el mismo sentido. Las de oferta, que nacen del cambio técnico, los precios de los factores, el número de productores o los impuestos a los vendedores, los mueven en sentidos opuestos.

- Un aumento de la demanda eleva el precio y la cantidad. Cada empresa amplía su producción a lo largo de su coste marginal, y el precio sube tanto más cuanto menos elástica es la oferta. Cada empresa aumenta así su beneficio o reduce sus pérdidas.
- Un impuesto específico, t, desplaza la oferta hacia arriba si lo paga el vendedor, o la demanda hacia abajo si lo paga el comprador, con los mismos efectos. Abre una **cuña fiscal**, $p_d - p_s = t$, reduce la cantidad y genera un exceso de gravamen, porque la pérdida de excedentes supera la recaudación. La incidencia no depende de quién lo pague legalmente, sino de las elasticidades, según la ley de Dalton, y recae más sobre el lado más inelástico:

$$\frac{\partial p_d}{\partial t} = \frac{\varepsilon_S}{\varepsilon_S + |\varepsilon_D|} \qquad \frac{\partial p_s}{\partial t} = -\frac{|\varepsilon_D|}{\varepsilon_S + |\varepsilon_D|}$$

- Un aumento del coste fijo no altera el equilibrio a corto plazo, porque no cambia el coste variable medio ni el marginal. Uno del coste variable contrae las ofertas: cada empresa produce menos y el precio sube, con la carga repartida según las elasticidades.

2.3. Largo plazo: beneficio nulo

Supuestos: todos los factores son variables y hay libre entrada y salida, de modo que el número de empresas lo determina el mercado. Junto a $p = \mathrm{Cmg}$ rige la **condición de beneficio nulo**, $p = \mathrm{Cme}$, que ninguna empresa persigue, sino que impone la libre entrada.

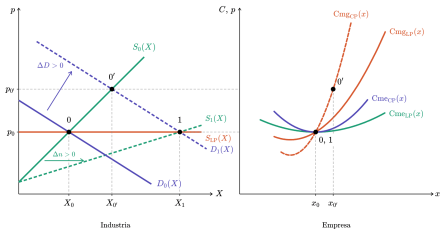
Oferta individual: su umbral es el mínimo del coste medio a largo plazo, porque a largo plazo la empresa puede evitar todos sus costes cerrando y solo produce si $p \geq \mathrm{Cme}_{\mathrm{LP}}$. Por el teorema de Le Châtelier-Samuelson (1947), la oferta de largo plazo responde al precio al menos tanto como la de corto, porque con todos los factores ajustables la empresa se adapta mejor. La diferencia crece con el número de factores que no puede ajustar a corto plazo. Como a corto plazo, la oferta es discontinua, ahora en el mínimo del coste medio a largo plazo:

$$x^s_{\mathrm{LP}} = \begin{cases} s(p, \mathbf{p}^z_0) & \text{si } p \geq \mathrm{Cme}_{\mathrm{LP}}(x_{\mathrm{OE}}) \\ 0 & \text{si } p < \mathrm{Cme}_{\mathrm{LP}}(x_{\mathrm{OE}}) \end{cases}$$

Oferta de mercado: ya no es la simple suma de las individuales, porque la entrada y la salida cambian el número de empresas hasta que se cumple la condición de beneficio nulo. Con empresas idénticas, cada entrada gira hacia la derecha la oferta a corto plazo, y la oferta a largo plazo, que une los sucesivos equilibrios, es horizontal al nivel del mínimo del coste medio, porque la cantidad se ajusta por la variación del número de empresas.

Equilibrio: todas las empresas activas operan en su dimensión óptima, $\mathrm{Cme}_{\mathrm{CP}} = \mathrm{Cmg}_{\mathrm{CP}} = \mathrm{Cmg}_{\mathrm{LP}} = \mathrm{Cme}_{\mathrm{LP}}$, con beneficio nulo. El precio cumple una tercera función, la de **asignar los recursos entre industrias**, como señal que provoca la entrada y la salida.

Estática comparativa: con precios de los factores constantes, la oferta a largo plazo es horizontal, de modo que el precio solo depende de los costes y la demanda determina la cantidad.



Largo plazo: tras un aumento de la demanda, la entrada de empresas devuelve el precio al mínimo del coste medio.

– Un aumento permanente de la demanda eleva primero el precio y crea beneficios extraordinarios, que atraen entrantes hasta que el precio vuelve a p_0. La cantidad sube por la entrada de empresas, cada una de las cuales produce lo mismo que antes. Si la expansión altera los precios de los factores, el precio sí varía.
– Un aumento permanente del coste fijo eleva el coste medio y expulsa empresas. El precio sube, la cantidad cae y cada empresa que se queda produce más. La carga recae íntegramente en los consumidores, porque la oferta es perfectamente elástica.
– Un aumento del coste variable sube el precio y reduce la cantidad, con efecto indeterminado sobre la producción de cada empresa y el número de empresas, porque cada empresa opera en el nuevo mínimo del coste medio.

2.4. Dinámicas de ajuste

Walras (1874): en el tanteo, que la tradición personificó en un subastador ficticio, el precio varía en proporción al exceso de demanda, y solo se intercambia en el equilibrio:

$$\frac{dp}{dt} = \kappa_p \left[D(p) - S(p)\right]$$

En torno a p^*, con $\kappa_p > 0$ la velocidad de ajuste, la desviación del precio evoluciona según:

$$\frac{d(p-p^*)}{dt} \approx \kappa_p \left[D'(p^*) - S'(p^*)\right](p-p^*)$$

Es estable si el exceso de demanda decrece con el precio, $D'(p^*) - S'(p^*) < 0$, lo que se cumple siempre con demanda decreciente y oferta creciente. Entonces, si la demanda aumenta en ΔD, el precio sube en $\Delta D/[S'(p^*) - D'(p^*)]$, positivo precisamente porque el equilibrio es estable (**principio de correspondencia** de Samuelson, 1947). Converge desde cualquier precio inicial si el exceso de demanda solo cambia de signo en p^*. Se adapta a los mercados en que el precio se forma antes de intercambiar, como los de capitales, pero con varios mercados interdependientes deja de estar garantizada, salvo, por ejemplo, con sustituibilidad bruta.

Marshall (1890): la cantidad varía según la diferencia entre el precio de demanda y el de oferta, y se intercambia fuera del equilibrio:

$$\frac{dX}{dt} = \kappa_X \left[p_d(X) - p_s(X)\right]$$

Aquí $p_d(X)$ y $p_s(X)$, las inversas de la demanda y de la oferta, son los precios a los que los consumidores comprarían y los productores ofrecerían exactamente la cantidad X. En torno a X^*, la desviación evoluciona según:

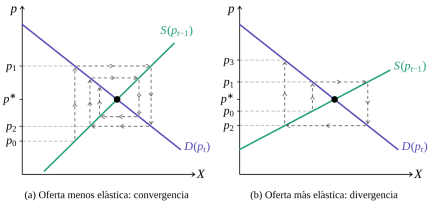
$$\frac{d(X-X^*)}{dt} \approx \kappa_X \left[p_d'(X^*) - p_s'(X^*)\right](X-X^*)$$

Es estable si $p_s'(X^*) > p_d'(X^*)$, es decir, si la oferta corta a la demanda desde abajo. Se adapta a los mercados industriales, en que la producción se decide antes de conocer el precio, pero no explica cómo se llega al precio de demanda.

Telaraña (Kaldor, 1934): formalizado en 1930 por Ricci, Schultz y Tinbergen y bautizado por Kaldor, recoge mercados como los agrarios, en que la producción tarda en adaptarse: la oferta responde al precio anterior, $X_t^s = S(p_{t-1})$, y la demanda al corriente, $X_t^d = D(p_t)$. Así ocurre en las cosechas, la construcción o el ganado, en que la cantidad se decide con el precio conocido entonces. El precio de cada periodo vacía el mercado para la cantidad decidida con el anterior, y la secuencia de precios dibuja una telaraña en torno al equilibrio.

Si la oferta es menos elástica que la demanda, las oscilaciones se amortiguan y el precio converge. Si es más elástica, se amplifican, y con elasticidades iguales, el precio oscila indefinidamente. Supone expectativas ingenuas, que las expectativas racionales superan, porque si los productores anticipan el ciclo, lo eliminan.

Comparación: con demanda decreciente y oferta creciente, ambos criterios se cumplen a la vez. Si las dos pendientes tienen el mismo signo,



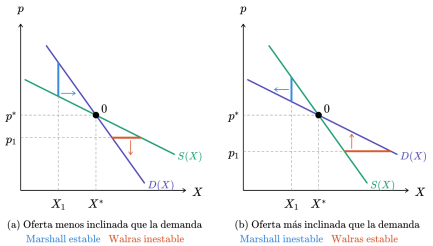
(a) Oferta menos elástica: convergencia (b) Oferta más elástica: divergencia

Telaraña: con oferta rezagada, el precio converge si la oferta es más inclinada que la demanda (a) y diverge en caso contrario (b).

como con la oferta decreciente de una industria con economías externas, la oferta de trabajo que se vuelve hacia atrás o la demanda creciente de un bien Giffen, los criterios se excluyen: lo estable para Marshall es inestable para Walras. No es una paradoja, porque cada uno describe una conducta distinta ante el desequilibrio. Formalmente, como $D'(p) = 1/p_d'(X)$ y $S'(p) = 1/p_s'(X)$, los criterios de Walras y de Marshall se escriben con las pendientes del plano (X, p):

$$\frac{p_s' - p_d'}{p_d'\,p_s'} < 0 \qquad p_s' - p_d' > 0$$

Con ambas pendientes negativas, el equilibrio es estable para Marshall si la oferta es menos inclinada que la demanda y para Walras si lo es más, y con ambas positivas, al revés. Con demanda creciente y oferta decreciente fallan los dos.



(a) Oferta menos inclinada que la demanda
Marshall estable Walras inestable

(b) Oferta más inclinada que la demanda
Marshall inestable Walras estable

Estabilidad de Walras y de Marshall con oferta y demanda decrecientes.

2.5. Propiedades del equilibrio

Definición: un **equilibrio competitivo** es un precio no negativo y una asignación en que consumidores y empresas optimizan al precio dado, el mercado se vacía y, a largo plazo, la libre entrada ha eliminado los beneficios extraordinarios.

Existencia: si la oferta y la demanda son continuas y la cantidad demandada supera a la ofrecida a precios bajos y queda por debajo a precios altos, las curvas se cortan al menos una vez, porque el exceso de demanda, $Z_x(p_x) = D_x(p_x) - S_x(p_x)$, es continuo y cambia de signo (teorema de Bolzano). La oferta, sin embargo, tiene un salto en el precio de cierre, por debajo del cual coincide con el eje de precios, y si la demanda pasa por el hueco, ningún precio vacía el mercado. Tampoco hay equilibrio si el coste medio mínimo supera la disposición máxima a pagar, porque oferta y demanda no se cortan para cantidades positivas y el bien no se produce.

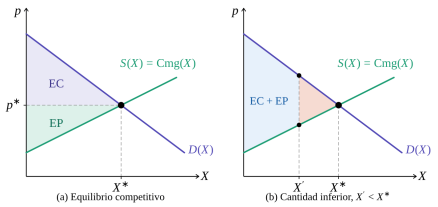
Unicidad: en un mercado basta una demanda decreciente y una oferta creciente, que se cortan como mucho una vez, y en general, la sustituibilidad bruta. Puede fallar, como con la oferta de trabajo que se vuelve hacia atrás, porque el efecto renta domina al de sustitución, o con costes lineales, y los rendimientos crecientes comprometen la propia existencia. Gráficamente, una oferta no monótona puede cortar a la demanda varias veces.

Estabilidad: el equilibrio es **estable** si el ajuste lo devuelve a él tras una perturbación, en un entorno del equilibrio (estabilidad local) o desde cualquier posición inicial (global). La condición depende de la dinámica de ajuste, y en un equilibrio inestable la estática comparativa pierde sentido.

Excedente total: la suma del excedente del consumidor, el área entre la demanda y el precio, y del productor, entre el precio y la oferta, es máxima en la cantidad competitiva, donde $p_d(X^*) = \mathrm{Cmg}(X^*)$. Con menos se dejan de producir unidades que valen más de lo que cuestan, y se pierde el triángulo entre ambas curvas, mientras que con más, las adicionales cuestan más de lo que valen.

Primer teorema del bienestar: con preferencias localmente no saciables y sin fallos de mercado, todo equilibrio competitivo es **eficiente en el sentido de Pareto**. Se ha visto como la formalización de la mano invisible de Smith, referida a la optimalidad paretiana, y tiene cuatro corolarios.

– No desperdicia recursos, aunque no sea necesariamente equitativo.



Excedente total: máximo en el equilibrio competitivo, X^* (a); con $X' < X^*$ se pierde el triángulo entre demanda y oferta (b).

– Sirve de guía de política económica.
– Economiza información, porque a cada agente le basta conocer los precios.
– Sus fallos justifican la intervención del Estado por eficiencia.

Segundo teorema: con preferencias y tecnologías convexas, todo óptimo de Pareto puede alcanzarse como equilibrio competitivo tras redistribuir las dotaciones con transferencias de suma fija. Separa eficiencia y equidad, aunque choca con la falta de transferencias que no distorsionen. Un planificador benevolente podría corregir la distribución y dejar después que el mercado asigne eficientemente.

En suma, la capacidad de ajuste de la oferta decide quién fija el precio: la demanda en el periodo de mercado, ambas a corto plazo y los costes a largo plazo. Si es estable, el equilibrio maximiza el excedente total y, sin fallos de mercado, es eficiente en el sentido de Pareto.

3. Extensiones

Las extensiones relajan los supuestos sobre la tecnología, los precios de los factores y el funcionamiento del mecanismo de precios.

3.1. Tecnología sin rendimientos mixtos

Rendimientos constantes: con $\mathrm{CT}(x) = cx$, el coste medio y el marginal a largo plazo coinciden en una horizontal de altura c, que es la oferta de la empresa, y la oferta a largo plazo es infinitamente elástica. Caben múltiples equilibrios, y quedan indeterminados el tamaño de cada empresa y cuántas operan. No hay economías ni deseconomías de escala, $\varepsilon_{Cx} = 1$, y a corto plazo los costes mantienen su forma de U, con ofertas crecientes.

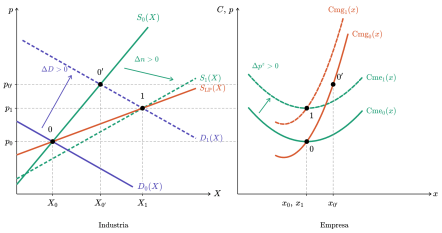
Rendimientos decrecientes: con coste total convexo, el coste marginal está siempre por encima del medio, de modo que $p = \mathrm{Cmg}$ genera beneficios y atrae entrantes sin límite. Operan infinitas empresas de producción infinitesimal, y la asignación es un óptimo de Pareto. Hay deseconomías de escala, $\varepsilon_{Cx} > 1$, como con $\mathrm{CT}(x) = cx^3$, y la oferta de la empresa es toda la curva de coste marginal.

Rendimientos crecientes: con coste total cóncavo, como $\mathrm{CT}(x) = cx^{1/2}$, hay economías de escala, $\varepsilon_{Cx} < 1$. El coste marginal está siempre por debajo del medio, y no se cumple la CSO. Aplicar $p = \mathrm{Cmg}$ da pérdidas, y las empresas salen hasta que queda un **monopolio natural** que sube el precio al menos hasta el coste medio. Hay ineficiencia asignativa, porque se produce menos y más caro, y explotación monopolística de los factores, que se retribuyen por debajo del valor de su productividad marginal. No existe curva de oferta competitiva.

3.2. Externalidades pecuniarias

Si la producción agregada altera los precios de los factores, la oferta a largo plazo deja de ser horizontal.

Deseconomías externas: externas a la empresa pero internas a la industria, elevan el coste medio al crecer la producción agregada, por la escasez de factores, cuyo precio sube, o por la congestión. Un aumento de la demanda eleva el precio y crea beneficios, que atraen entrantes. Estos encarecen los factores y desplazan hacia arriba las curvas de costes, y el nuevo equilibrio, con beneficio nulo en el mínimo del nuevo coste medio, tiene un precio mayor. La nueva dimensión óptima puede ser menor, igual o mayor. La oferta a largo plazo es **creciente**.



Industria de costes crecientes: la entrada encarece los factores y la oferta a largo plazo, S_{LP}, es creciente.

Economías externas: las de aglomeración, como en un clúster, ahorran costes de transporte y suministro y difunden conocimientos. La entrada abarata los factores y los costes, y la oferta a largo plazo es **decreciente**. De ahí el argumento de la industria naciente, que propone proteger a la industria nacional hasta que su expansión reduzca sus costes medios al nivel internacional. Así, la pendiente de la oferta a largo plazo depende de la capacidad de la industria para alterar el precio de sus factores, y las economías externas son mucho más difíciles de medir que las internas.

3.3. Mecanismo de precios

3.3.1. Controles de precios

Idea: un **precio máximo** por debajo del de equilibrio, como en los alquileres, crea un exceso de demanda y reduce la cantidad a la ofrecida, $S(p_{\mathrm{máx}})$. El bien se raciona con colas, asignación administrativa, peor calidad o mercados negros. Un **precio mínimo**, como el salario mínimo o los precios de garantía agrarios, crea un exceso de oferta que se absorbe con compras públicas, cuotas o, en el trabajo, desempleo. El precio deja así de vaciar el mercado, y el exceso de demanda, $D(p_{\mathrm{máx}}) - S(p_{\mathrm{máx}})$, o el de oferta, $S(p_{\mathrm{mín}}) - D(p_{\mathrm{mín}})$, persiste.

Bienestar: en ambos casos se intercambia la cantidad del lado corto del mercado y se pierde, como mínimo, el triángulo entre demanda y oferta, como en el impuesto, pero sin recaudación. El racionamiento por precio se sustituye por otro menos eficiente y no necesariamente más equitativo, porque el bien va a quien llega antes y no a quien más lo valora. La intervención puede justificarse por razones distributivas, pero el modelo fija el coste de eficiencia que ha de compensar. Con todo, con monopsonio, un salario mínimo moderado puede aumentar el empleo, y el triángulo no recoge los recursos que consume el racionamiento, como el tiempo de las colas.

3.3.2. Costes de transacción (Coase, 1937)

Idea: usar el mercado exige buscar a la contraparte, informarse, negociar y asegurar el cumplimiento. Esos costes abren una cuña entre el precio del comprador y el del vendedor, como un impuesto: reducen la cantidad y reparten la carga según las elasticidades. Pero la cuña consume recursos reales o remunera a un intermediario, en lugar de ser una transferencia. Explican los intermediarios, que cobran la cuña a cambio de reducirla, la existencia de la empresa y la ausencia de mercados cuyos costes de transacción superan las ganancias del intercambio. El modelo los trata, sin embargo, como una cuña exógena y proporcional a la cantidad, cuando muchos son fijos por operación o dependen de las instituciones. Además, su medición directa es difícil.

3.3.3. Racionamiento del crédito (Stiglitz y Weiss, 1981)

Idea: el tipo de interés es el precio del crédito, pero también informa sobre la calidad de los prestatarios. Subirlo expulsa los proyectos seguros, por selección adversa, y empuja hacia otros más arriesgados, por riesgo moral. A partir de cierto tipo, el rendimiento esperado del banco cae, y este prefiere racionar, de modo que hay un equilibrio con exceso de demanda de crédito. Explica que un mercado no se vacíe sin rigideces de precios y da fundamento microeconómico al canal del crédito de la política monetaria. Pero depende de que los bancos no puedan discriminar entre prestatarios, cuando las garantías revelan parte de su riesgo, y es un modelo estático de equilibrio parcial.

3.4. Otras extensiones

Barreras a la entrada: si existen a largo plazo, la empresa sigue igualando precio y coste marginal, pero pueden persistir beneficios extraordinarios, y pérdidas si hay además barreras de salida.

Poder de mercado: cuando la empresa no es precio-aceptante, fija un margen sobre el coste marginal, sea cual sea la estructura de mercado, medido por el índice de Lerner, $I_L = (p - \mathrm{Cmg})/p \in [0, 1]$, nulo en competencia perfecta.

Renta ricardiana (Ricardo, 1817): es el excedente sobre su coste de oportunidad de los propietarios de recursos escasos, como la tierra. La oferta crece porque, al subir el precio, se cultivan también las tierras pobres, y la libre entrada no erosiona la renta.

Economía abierta y equilibrio general: los aranceles, las cuotas y las restricciones no cuantitativas, que actúan como un "arancel implícito", reducen los intercambios beneficiosos y redistribuyen renta. El equilibrio general trata los mercados interrelacionados y complementa el parcial, que basta para las medidas antimonopolio.

A17. Análisis de mercados (II). Teoría del monopolio. Discriminación de precios. Monopolio natural. Producción conjunta. Análisis de eficiencia y bienestar. Monopsonio. Monopolio bilateral

1. El monopolio de oferta

El estudio parte de sus causas, sus supuestos y la demanda a la que se enfrenta, y sigue con la decisión del monopolista y el análisis de su equilibrio.

1.1. Concepto, causas y supuestos

Definición: el **monopolio de oferta** es la estructura en que un único vendedor abastece toda la demanda de un bien sin sustitutivos cercanos. Su rasgo definitorio es la ausencia de rivales actuales y potenciales, porque las **barreras a la entrada** impiden que los beneficios extraordinarios atraigan a nuevas empresas. El origen de esas barreras decide la regulación.

1. *Naturales o estructurales*: una sola empresa abastece la demanda a menor coste que varias, por subaditividad, típicamente con una escala mínima eficiente grande respecto al mercado. Solo protegen si la entrada exige costes hundidos. Sin ellos, el mercado es impugnable (Baumol y Willig, 1981). Estas barreras son las del monopolio natural, cuya regulación se centra en el control de precios, en su financiación y, rara vez, en la producción pública.
2. *Estratégicas*: conductas de la instalada para disuadir la entrada, como el precio límite, la inversión en capacidad o la proliferación de marcas, o para expulsar rivales. Solo disuaden si son creíbles, con un compromiso irreversible (Dixit, 1980) o con reputación (Kreps y Wilson, 1982; Milgrom y Roberts, 1982b). La defensa de la competencia es aquí cautelosa, porque un precio bajo puede ser simple competencia y la integración vertical o la exclusividad pueden ganar eficiencia.
3. *Legales*: patentes, licencias, concesiones o derechos exclusivos. En licencias y concesiones, la regulación controla los precios y revisa periódicamente el privilegio. En las patentes, el poder de mercado temporal es el precio de incentivar la innovación, y la política se centra en su duración y su alcance.

Supuestos: información perfecta y sin costes de transacción, como en competencia perfecta, pero con barreras. Hay un único bien normal, con rendimientos mixtos; un monopolista que maximiza el beneficio frente a consumidores precio-aceptantes, y un mercado de factores perfecto. Así, la empresa es el mercado, $x = X$, su demanda no es perfectamente elástica y puede dejar de cumplirse la ley del precio único, lo que hace posible la discriminación de precios.

1.2. Demanda de mercado

Definición: es la suma horizontal de las individuales, como en competencia perfecta, pero el monopolista se enfrenta a toda ella, de modo que cada cantidad que decide vender determina su precio. Con una demanda lineal, $p(X) = a - bX$, hay dos implicaciones.

- El ingreso marginal también es lineal, $\mathrm{Img}(X) = a - 2bX$, con la misma ordenada en el origen y el doble de pendiente, siempre por debajo del precio. Procede de derivar el ingreso total, $\mathrm{IT}(X) = aX - bX^2$.
- La elasticidad no es constante: crece de cero a infinito a lo largo de la recta y vale uno en su punto medio, donde el ingreso marginal se anula. Como el monopolista solo produce donde la demanda es elástica, elige siempre en la mitad superior. La elasticidad es $|\varepsilon_{Xp}| = p/(a-p)$, nula en $X = a/b$ e infinita en $X = 0$.

1.3. La decisión del monopolista

1.3.1. Equilibrio: enfoque de mercado

Planteamiento: el monopolista maximiza el beneficio y, como se enfrenta a toda la demanda, elegir precio o cantidad es equivalente:

$$\max_X \ \pi = p(X)\cdot X - \mathrm{CT}(X)$$

CPO: iguala el ingreso marginal al coste marginal, lo que equivale a fijar un margen sobre este:

$$\mathrm{Img}(X) = \mathrm{Cmg}(X) \ \Leftrightarrow \ p = \underbrace{\frac{1}{1-1/|\varepsilon_{Xp}|}}_{1+\mu}\cdot \mathrm{Cmg}(X)$$

El ingreso marginal es el precio de la unidad adicional menos la pérdida de ingreso en las ya vendidas, que ahora se venden más baratas, $(\partial p/\partial X)\cdot X$. El margen, $\mu > 0$, existe porque la elasticidad es finita, y el monopolista opera en el **tramo elástico**, porque en el inelástico, producir menos subiría el precio en mayor proporción de lo que caen las ventas, y el ingreso total aumentaría. Tampoco produce con elasticidad unitaria, donde el ingreso marginal es nulo, salvo con coste marginal nulo, caso en que maximizar el beneficio equivale a maximizar el ingreso.

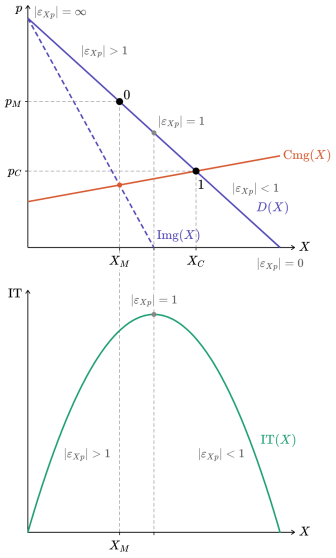
Índice de Lerner (1934): el margen relativo es la inversa de la elasticidad:

$$I_L \equiv \frac{p - \mathrm{Cmg}(X)}{p} = \frac{1}{|\varepsilon_{Xp}|} \in (0,1]$$

Vale cero en competencia perfecta y se acerca a uno cuando la elasticidad tiende a la unidad. El poder de mercado disminuye conforme los consumidores tienen alternativas. En el oligopolio de Cournot con producto homogéneo, el de cada empresa es $I_{L,i} = s_i/|\varepsilon_{Xp}| \in (0,1]$, donde $s_i = x_i/X$ es su cuota de mercado. Con demanda isoelástica, $X = k\cdot p^{-|\varepsilon_{Xp}|}$, el índice es constante, y ante un cambio del coste marginal el monopolista ajusta el precio manteniendo el margen relativo.

CSO: basta con que el ingreso marginal sea más inclinado que el coste marginal, $\partial\mathrm{Img}/\partial X < \partial\mathrm{Cmg}/\partial X$. A diferencia de la competencia perfecta, el monopolista puede maximizar también en el tramo decreciente del coste marginal, de modo que el monopolio admite rendimientos crecientes.

Desarrollo gráfico: el ingreso marginal queda por debajo de la demanda y se anula donde la elasticidad es unitaria, que maximiza el ingreso total. El monopolista produce X_M, donde corta al coste marginal, a la izquierda de ese punto, y fija sobre la demanda el precio p_M, superior al coste marginal y al precio competitivo, p_C, que corresponde a la cantidad X_C, donde la demanda corta al coste marginal.



Equilibrio del monopolio: produce X_M donde Img = Cmg y fija p_M sobre la demanda, por encima de p_C.

1.3.2. Equilibrio: enfoque de la producción

Planteamiento: el monopolista elige las cantidades de factores, con la producción, $f(\mathbf{z}) = X$, valorada según la demanda inversa:

$$\max_{\mathbf{z}} \ \pi = p(X)\cdot f(\mathbf{z}) - \mathbf{p}^z\cdot\mathbf{z} \quad \text{s.a.: } \mathbf{z}\geq 0$$

CPO: como la empresa competitiva, el monopolista minimiza el coste, $\mathrm{RMST}_{21} = p_1^z/p_2^z$, de modo que produce cualquier cantidad al mínimo coste. La diferencia está en que contrata cada factor hasta que el **ingreso de su productividad marginal** iguala su precio. Como $p_i^z/f_i' = \mathrm{Cmg}$, la condición equivale, de nuevo, al margen sobre el coste marginal:

$$\mathrm{IPmg}_i = \mathrm{Img}\cdot f_i' = p_i^z \ \Leftrightarrow \ p = \frac{1}{1-1/|\varepsilon_{Xp}|}\cdot\frac{p_i^z}{f_i'}$$

Como el ingreso marginal es menor que el precio, cada factor cobra menos que el valor de su productividad marginal, $p\cdot f_i'$, lo que cobraría en competencia. La CSO exige un beneficio estrictamente cóncavo en los factores, aunque, como en el enfoque de mercado, basta un ingreso marginal que decrezca con rapidez suficiente. Mientras maximice el beneficio, el monopolista es eficiente en cómo produce, pero no en cuánto, es decir, cumple la eficiencia productiva y viola la asignativa.

1.3.3. Ausencia de función de oferta

Idea: la empresa competitiva asocia a cada precio una cantidad. El monopolista carece de curva de oferta, porque su producción depende de la forma de la demanda. Dos demandas distintas pueden llevarle a vender lo mismo a precios distintos, o cantidades distintas al mismo precio. No hay estática comparativa de validez general, y es posible la discriminación de precios.

1.3.4. Función de máximo beneficio

Definición: $\pi^*(\mathbf{p}^z)$ da el mayor beneficio para cada vector de precios de los factores, dada la demanda. No depende de un precio del producto, porque el monopolista lo elige, y no es homogénea en los precios de los factores, porque la demanda queda fija en términos monetarios. Es continua, no creciente y convexa en esos precios, de modo que la demanda de cada factor no aumenta con su precio. Por el lema de Hotelling, su

derivada, cambiada de signo, es la demanda de cada factor, que coincide con la condicionada evaluada en la producción de monopolio, porque produce al mínimo coste. Es observable, porque se miden los precios de los factores y el beneficio. Solo es homogénea, de grado uno, si la demanda inversa se multiplica en la misma proporción que esos precios. El lema de Hotelling y la dualidad se escriben:

$$\frac{\partial \pi^*}{\partial p_i^z} = -z_i^*(\mathbf{p}^z) \qquad z_i^*(\mathbf{p}^z) = d_i^c(\mathbf{p}^z, X_M(\mathbf{p}^z))$$

1.4. Análisis del equilibrio

1.4.1. Análisis positivo

Propiedades: el equilibrio existe, es único y es estable mientras persistan la necesidad del producto y las barreras. A igualdad de costes, la cantidad es siempre menor y el precio siempre mayor que en competencia perfecta. El beneficio y la escala a la que opera dependen del plazo.

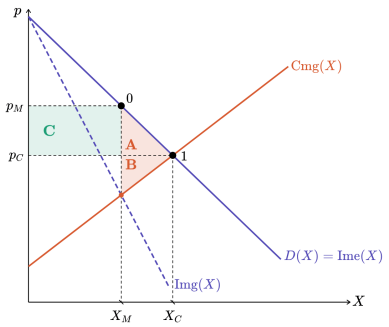
Corto plazo: al menos un factor es fijo. El monopolista produce si $p_M \geq \text{CVme}(X_M)$, evaluado en la cantidad que elige y no en el mínimo de explotación, porque no toma el precio como dado. Puede operar con beneficios, si $p_M > \text{Cme}(X_M)$, o con pérdidas, si el coste medio queda por encima de la demanda para cualquier cantidad.

Largo plazo: solo produce si cubre todos sus costes, $p_M \geq \text{Cme}_{\text{LP}}(X_M)$, de modo que no soporta pérdidas. Hay dos diferencias con la competencia perfecta.

1. *Rentas del monopolio*: puede obtener beneficios positivos, porque las barreras impiden la entrada. Esas rentas pueden deberse a derechos sobre un recurso, como una patente, por los que otros pagarían. Están en la base de la regulación mediante subastas.
2. No opera necesariamente en la dimensión óptima, porque su curva de ventas no es perfectamente elástica ni se le impone el beneficio nulo. Según el tamaño del mercado, puede producir en la escala óptima, en el mínimo del coste medio; en una subóptima, con capacidad excedentaria, como en la competencia monopolística, o en una sobreutilizada, con deseconomías de escala, si el mercado es tan grande que exige una planta mayor que la dimensión óptima.

1.4.2. Análisis normativo: eficiencia y bienestar

Ineficiencia asignativa: hay unidades no producidas por las que los consumidores pagarían más que su coste marginal, pero el monopolista no las produce, porque tendría que bajar el precio de todas las unidades. La demanda recoge la valoración social de cada unidad y el ingreso marginal, la privada del monopolista, y la pérdida nace de esa divergencia. Parte del excedente del consumidor se transfiere al productor, el rectángulo C, y otra se pierde, el triángulo A del consumidor y el B del productor (el **triángulo de Harberger**).



Ineficiencia del monopolio: el rectángulo C se transfiere al productor y los triángulos A y B son la pérdida irrecuperable.

Con demanda $p = a - X$ y coste marginal constante c, el monopolio produce la mitad que la competencia, $X_C = a - c$, su beneficio es la mitad del excedente competitivo, $(a-c)^2/2$, y la pérdida irrecuperable, $(a-c)^2/8$, la cuarta parte:

$$X_M = \frac{a-c}{2} \qquad p_M = \frac{a+c}{2} \qquad \pi_M = \frac{(a-c)^2}{4}$$

Coste social: Harberger (1954) lo estimó en apenas el 0,1 % de la renta nacional en la industria manufacturera estadounidense. Tullock (1967) y Posner (1975) señalaron que el rectángulo no es una mera transferencia, porque la **búsqueda de rentas**, con presión política, litigios o capacidad excesiva, puede disiparlo, y Cowling y Mueller (1978) elevaron la pérdida a entre el 4 % y el 13 % del producto. En sentido contrario, la tradición schumpeteriana recuerda que el poder de mercado puede ser el precio de la innovación.

Explotación monopolística: el monopolista contrata menos factores y los paga según su $\text{IPmg}_i < \text{VPmg}_i$, con un grado de explotación igual a $1/|\varepsilon_{X_P}|$, aunque los emplea con eficiencia técnica:

$$\text{IPmg}_i = \text{Img} \cdot f_i' < p \cdot f_i' = \text{VPmg}_i \qquad \frac{p \cdot f_i' - p_i^z}{p \cdot f_i'} = \frac{1}{|\varepsilon_{X_P}|}$$

Ineficiencia en costes: sin competencia se debilitan los incentivos a reducir costes, la ineficiencia-X de Leibenstein (1966). Sin presión competitiva ni supervisión, los trabajadores se relajan y los costes superan el mínimo.

1.4.3. Estática comparativa

- *Demanda*: un aumento eleva la cantidad solo si eleva el ingreso marginal en la cantidad inicial, lo que depende de cómo cambie la elasticidad. Con un desplazamiento paralelo de una demanda lineal y coste marginal constante o creciente suben cantidad y precio. Si la demanda se vuelve más inelástica, la cantidad puede bajar mientras sube el precio, y si se vuelve más elástica, el precio puede bajar mientras sube la cantidad.
- *Costes*: un aumento del coste marginal se traslada al precio en parte con demanda lineal, la mitad con coste marginal constante, y más que íntegramente con demanda isoelástica, $\Delta p = (1+\mu) \cdot \Delta\text{Cmg}$.
- *Impuestos (Suits y Musgrave, 1953)*: para la misma recaudación, el impuesto *ad valorem* reduce menos la producción que el específico, porque grava el ingreso marginal, menor que el precio, y no cada unidad. Da un precio menor y una cantidad mayor, y es preferible en mercados no competitivos.
- *Precio máximo*: fijado entre el coste marginal y el precio de monopolio, aumenta la producción, porque hasta la cantidad demandada a ese precio el ingreso marginal coincide con él. Si iguala al competitivo, se alcanza la cantidad de competencia perfecta, lo que anticipa la regulación del monopolio natural.

En suma, el monopolio muestra cómo el poder de mercado rompe la igualdad entre precio y coste marginal y da la referencia de la ineficiencia asignativa. Pero supone un precio único, información completa y ausencia de competencia potencial.

2. Principales monopolios

La sección estudia los monopolios que discriminan precios y el monopolio natural, con una mención a su regulación.

2.1. Monopolios discriminadores de precios

Definición: la **discriminación de precios** cobra precios distintos por un mismo bien sin diferencias de coste, para apropiarse del excedente que el precio único deja sin capturar. Exige tres condiciones.

1. Poder distinguir, directa o indirectamente, las valoraciones de los consumidores.
2. Impedir el arbitraje, la reventa de quien compra barato a quien pagaría caro.
3. Tener libertad para cobrar precios distintos.

Según la información disponible, adopta los tres grados de Pigou (1920), más la intertemporal, llamada de cuarto grado por extensión.

2.1.1. Primer grado (perfecta)

Idea: con información completa, el monopolista cobra por cada unidad el precio de reserva de su comprador. Vender una unidad más no obliga a abaratar las anteriores, de modo que la demanda hace de ingreso marginal, y el ingreso medio, que promedia los precios de reserva de las unidades vendidas, queda por encima de ella. Por eso vende hasta que el precio de reserva iguala al coste marginal, sin margen sobre este:

$$\text{Img}(X) = \text{Cmg}(X) \;\Leftrightarrow\; p(X) = \text{Cmg}(X)$$

Equilibrio: alcanza la **cantidad competitiva**, $X_M = X_C$, sin ineficiencia asignativa, pero con todo el excedente en sus manos, y la intervención se apoyaría en la equidad, no en la eficiencia. Su beneficio es todo el excedente que en competencia sería de los consumidores. Muestra que la pérdida del monopolio simple procede del precio único, no del poder de mercado, y sirve de referencia para los demás grados. Pero exige una información que casi nunca existe. Se aproxima en las subastas, en el regateo del zoco o en el médico rural que cobra según la capacidad de pago. Con la demanda $p = a - X$ y coste marginal c, vende $a - c$, como en competencia, y su beneficio es $(a-c)^2/2$.

2.1.2. Segundo grado: menús de precios

Idea: el monopolista no observa las valoraciones, pero sabe que hay tipos distintos y diseña un menú para que cada uno se autoseleccione: tarifas en dos partes, descuentos por cantidad o *bundling* y versiones de calidad. La **tarifa en dos partes** (Oi, 1971), $T(x) = F + p \cdot x$, cobra una cuota fija por el derecho a comprar y un precio por unidad. El monopolista conoce las demandas de cada tipo y su número, pero no puede identificar quién es quién, y la cuota es personal e intransferible, de modo que no hay arbitraje.

Planteamiento: el monopolista elige cuota y precio sujeto a que participen los consumidores servidos:

$$\max_{F,p} \; \pi = F\,n + (p-c)\,X(p) \quad \text{s.a.: } \text{EC}_i(p) \geq F$$

CPO: con consumidores idénticos, $p = c$ y $F = \mathrm{EC}(c)$: vende al coste marginal y se apropia de todo el excedente con la cuota. Con dos tipos, la cuota la limita el tipo bajo, $F = \mathrm{EC}_B(p)$, y subir el precio unitario extrae parte de la mayor valoración del tipo alto:

$$n_A\,[x_A(p) - x_B(p)] = (p - c)\,[-X'(p)]$$

Subir el precio un euro rebaja en x_B la cuota de cada consumidor, pero eleva en un euro el ingreso de cada unidad. Para el tipo bajo ambos efectos se compensan, y el tipo alto compra $x_A - x_B$ unidades más, sobre las que el monopolista gana sin perder cuota. A cambio, se venden menos unidades, cada una con margen $p - c$. Como la ganancia sobre el tipo alto es positiva, en el óptimo $p > c$.

CSO: el beneficio es cóncavo en p si las demandas no son demasiado convexas, p. ej. con demandas lineales de pendientes parecidas, de modo que la CPO caracteriza el máximo.

Estrategias: una tarifa exclusivista, con $F = \mathrm{EC}_A$, que expulsa al tipo bajo y conviene si abundan los de alta valoración; una no exclusivista, con $F = \mathrm{EC}_B$, o un menú discriminatorio sujeto a compatibilidad de incentivos. El *bundling* vende un lote a un precio inferior a la suma de sus partes, $p(x_1 + x_2) < p(x_1) + p(x_2)$, como los paquetes de *software*. La **elección de calidad** (Mussa y Rosen, 1978) ofrece versiones con diferencias de precio muy superiores a las de coste, como la tapa dura y la edición de bolsillo.

Conclusiones: explica por qué las tarifas reales son no lineales, pero exige ausencia de arbitraje y conocer la distribución de tipos, y el precio superior al coste marginal reintroduce una pérdida, menor que la del monopolio simple.

2.1.3. Tercer grado: segmentación de mercados

Idea: el monopolista usa una característica observable, como la edad o la residencia, para separar grupos con distintas elasticidades, y fija un precio por grupo.

Planteamiento: con dos grupos, A y B, de demandas independientes, elige la cantidad que vende en cada uno:

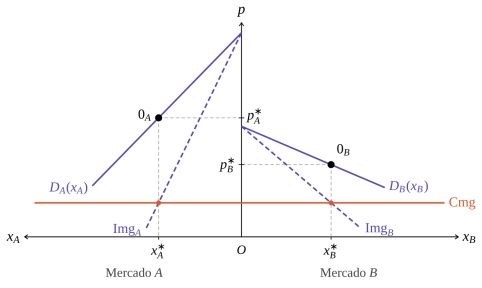
$$\max_{x_A,\,x_B}\ \pi = p_A(x_A)\,x_A + p_B(x_B)\,x_B - \mathrm{CT}(x_A + x_B)$$

CPO: el ingreso marginal de cada grupo se iguala al coste marginal común, lo que equivale a fijar un margen distinto en cada grupo:

$$\mathrm{Img}_A = \mathrm{Img}_B = \mathrm{Cmg}(X)\ \Leftrightarrow\ p_i = \frac{1}{1 - 1/|\varepsilon_i|}\cdot \mathrm{Cmg}(X), \quad i = A, B$$

Si un grupo tuviera un ingreso marginal mayor, trasladar ventas hacia él aumentaría el ingreso sin alterar el coste. El grupo de demanda más inelástica paga el precio mayor (**regla de la elasticidad inversa**), como en los descuentos a estudiantes o jubilados o en las tarifas distintas por países. Basta, como CSO, que cada ingreso marginal sea decreciente y el coste marginal no decreciente, $\partial \mathrm{Img}_i/\partial x_i < \partial \mathrm{Cmg}/\partial X$.

Desarrollo gráfico: con los dos mercados espalda con espalda, en cada uno el monopolista vende la cantidad en que su ingreso marginal corta al coste marginal común, x_A^* y x_B^*, y cobra el precio que marca su demanda, p_A^* y p_B^*.



Tercer grado: el mercado A, de demanda más inelástica, paga un precio mayor que el B, con el mismo coste marginal.

Bienestar: el grupo elástico gana y el inelástico pierde. Por el criterio de Schmalensee (1981) y Varian (1985), solo puede aumentar el bienestar si aumenta la producción total, porque precios distintos reasignan las unidades de forma ineficiente. Con demandas lineales y ambos grupos servidos la producción no cambia y el bienestar cae. Solo mejora si permite servir a un grupo antes excluido. Exige grupos identificables, sin reventa y con demandas independientes, y no da una regla general para la política de competencia.

2.1.4. Cuarto grado: intertemporal (Coase, 1972)

Idea: el monopolista de un bien duradero crea su propia competencia, porque lo que vende hoy reduce la demanda de mañana, y rebaja los precios con el tiempo. Por la **conjetura de Coase**, si los consumidores son pacientes y previsores, retrasan sus compras y el monopolista pierde su poder. Le conviene comprometerse a no rebajar los precios, por ejemplo alquilando el bien. Explica también la reducción de la vida útil del bien o los compromisos de no rebajar precios. La pérdida completa del poder exige consumidores muy pacientes y cambios de precio frecuentes, y con pocos compradores puede fallar, como en la estrategia *Pacman* (Bagnoli et al., 1989), que cobra en cada periodo la mayor valoración de quienes aún no han comprado. Si el bien se deprecia, si entran continuamente nuevos compradores o si ajustar el precio es costoso, el monopolista conserva también parte de su poder.

2.2. Monopolio natural

2.2.1. Concepto

Una sola empresa abastece el mercado a menor coste que varias por la tecnología, no por barreras artificiales: costes fijos elevados y hundidos, típicos de las redes (electricidad, gas, agua, ferrocarril), con economías de escala amplias respecto al mercado y, con varios productos, de alcance. A menudo solo la red es monopolio natural.

2.2.2. Monoproducto

Existencia: hay **monopolio natural** cuando la función de costes es estrictamente subaditiva, $\mathrm{CT}(\sum x_i) < \sum \mathrm{CT}(x_i)$, es decir, una sola empresa produce a menor coste que varias, como en los sectores de red. Las economías de escala son condición suficiente pero no necesaria, y la subaditividad es necesaria y suficiente para la deseabilidad desde la eficiencia productiva, no la asignativa. La subaditividad puede existir con deseconomías de escala si hay costes fijos, y se extiende más allá de la escala mínima eficiente, hasta el corte del coste medio de una empresa con el de dos que se reparten la producción.

Sostenibilidad: con libre entrada y la misma tecnología, caben tres casos según dónde corte la demanda al coste medio.

1. Deseable y sostenible, con la demanda en el tramo de rendimientos crecientes: compartir el mercado haría perder al entrante. Si es impugnable (Baumol, Panzar y Willig, 1982), solo sobrevive con $p = \mathrm{Cme}$, con beneficio nulo, porque un entrante podría fijar un precio algo menor y atender toda la demanda antes de que reaccionara. Si la entrada supone repartir el mercado, es sostenible con beneficios positivos.
2. Deseable pero no necesariamente sostenible, entre el coste medio mínimo y el corte con el coste medio de dos empresas: sin reacción, un entrante cobra menos y surge un duopolio. Según la teoría moderna, el monopolista que reacciona puede disuadirlo bajando el precio al coste medio mínimo, y si la amenaza es creíble, el entrante no entra.
3. No deseable e insostenible, más allá de ese corte, donde dos empresas producen más barato: el mercado evoluciona al duopolio. Aunque reaccione, al bajar el precio el monopolista alcanza un punto en que le conviene más compartir el mercado, y el entrante, que lo anticipa, entra.

2.2.3. Multiproducto

Existencia: las economías de escala de cada bien ya no bastan para la subaditividad. Basta, como condición suficiente, la complementariedad de costes de Sharkey (1982), que exige que el coste marginal de un bien sea menor cuanto más se produce de él o de los demás.

Sostenibilidad: solo resiste la entrada sin subsidios cruzados, es decir, si el ingreso de cada bien cubre su coste incremental y no supera su coste aislado. Se comprueba con las pruebas del coste incremental y de no aislamiento:

$$p_i \cdot x_i \geq \mathrm{CT}(\mathbf{x}) - \mathrm{CT}(\mathbf{x}_{-i}) \qquad p_i \cdot x_i \leq \mathrm{CT}(x_i)$$

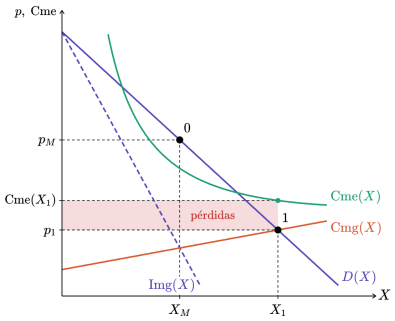
2.3. Una mención a la regulación de los monopolios naturales

Si el monopolio no es natural, se eliminan sus barreras con la liberalización y la defensa de la competencia. Si lo es, no conviene desmantelarlo, porque produce al menor coste, y su regulación depende de que sea sostenible.

2.3.1. Sostenible: regulación de precios

Sin regulación fijaría un precio muy superior al coste marginal. Las reglas se ordenan según la prioridad que se dé a la eficiencia asignativa frente a la viabilidad financiera de la empresa.

First best: se obliga a fijar $p = \mathrm{Cmg}$, que restablece la eficiencia asignativa. En el tramo decreciente del coste medio da pérdidas, que exigen una subvención, con problemas de información, de financiación con impuestos distorsionantes y de riesgo moral; en el creciente, beneficios que cabría extraer con impuestos de suma fija, difíciles de aplicar porque el regulador desconoce su cuantía.



First best: la tarificación al coste marginal (punto 1) es eficiente, pero genera pérdidas (área sombreada).

Second best: se busca el mayor bienestar sin pérdidas. El precio igual al coste medio es sencillo, pero puede no existir, mantiene un precio superior al competitivo y no incentiva a reducir costes. Lo refinan cuatro reglas.

- Tarifa en dos partes, con $p = \text{Cmg}$ y una cuota que cubre los costes fijos, que excluye a los de baja disposición a pagar.
- Tarifas pico-valle, que cargan la capacidad a los consumidores de punta y que, con rendimientos constantes a escala, hacen compatible $p = \text{Cmg}$ con el beneficio nulo.
- Loeb-Magat (1979), que cede al monopolista el excedente del consumidor para que elija $p = \text{Cmg}$, a un coste fiscal que lo hace inviable.
- *Precios de Ramsey-Boiteux*: en el multiproducto con demandas independientes, márgenes relativos inversamente proporcionales a la elasticidad de cada producto, con riesgo de regresividad, porque los productos de demanda más inelástica soportan una carga mayor.

Third best: se acota la rentabilidad o la senda de precios. La regulación de la tasa de beneficio exige poca información, pero provoca el efecto Averch y Johnson (1962), la sobrecapitalización, porque, si el rendimiento autorizado supera el coste del capital, invertir más eleva el beneficio permitido. El límite IPC − X (Littlechild, 1983) incentiva la eficiencia, con el riesgo de deteriorar la calidad, y la competencia por comparación (Shleifer, 1985) retribuye según los costes de empresas comparables, reproduciendo artificialmente la disciplina competitiva.

Información asimétrica: si el regulador no conoce los costes ni observa el esfuerzo, la regulación óptima es un **menú de contratos** (Laffont y Tirole, 1986). La empresa eficiente elige un contrato de precio fijo, con incentivo pleno y una renta informacional; la ineficiente, el de menor potencia, próximo al reembolso de costes, sin renta. El *trade-off* es incentivos frente a extracción de rentas, y su principal problema dinámico es el efecto trinquete. Si el regulador usa la información revelada para endurecer la regulación futura, la empresa deja de revelarla, lo que exige un compromiso creíble del regulador.

2.3.2. No sostenible: regulación de precios y acceso

La entrada duplicaría los costes fijos. Se cierra con una licencia exclusiva y se regula el precio, o se licita periódicamente el derecho a explotar el mercado, la competencia por el mercado (Demsetz, 1968), limitada por la incompletitud de los contratos de concesión y la ventaja del incumbente (Williamson, 1976). Si el monopolio no es deseable, se deja entrar.

En suma, la discriminación permite apropiarse del excedente y su efecto depende de la producción, y el monopolio natural puede ser deseable sin ser sostenible: si lo es, se regula su precio, y si no, también el acceso al mercado.

3. Otros monopolios

El modelo se traslada al lado de la demanda, al mercado con un único vendedor y un único comprador y al monopolista con varios productos o varias plantas.

3.1. Monopsonio

Idea: el **monopsonio** es la estructura simétrica: un único comprador afronta una oferta de elasticidad finita, como un gran empleador en un mercado de trabajo local, un distribuidor dominante frente a los productores agrarios o el Estado como único demandante de armamento pesado.

Planteamiento: el comprador se enfrenta a una oferta creciente del factor, $w(L)$ con $w' > 0$, y elige la cantidad que maximiza su beneficio:

$$\max_L \ \pi = p \cdot f(L) - w(L) \cdot L$$

CPO: el valor del producto marginal iguala el gasto marginal, que supera al precio de oferta porque comprar más encarece todas las unidades:

$$p \cdot f'(L) = w(L) + w'(L) \cdot L \ > \ w(L)$$

El monopsonista compra menos y paga menos que en competencia, con **explotación monopsonística** y una pérdida de bienestar análoga a la del monopolio, porque el coste privado del factor, su gasto marginal, supera al social, su precio de oferta.

CSO: el valor del producto marginal corta al gasto marginal desde arriba, $p\,f'' < 2w' + w''L$, lo que se cumple con $f'' < 0$ y oferta lineal o convexa.

Conclusiones: la ineficiencia procede del poder de mercado, no del lado en que se ejerce, y por eso la regulación invierte su signo. Un precio mínimo, como el salario mínimo, puede aumentar el empleo, y la negociación colectiva sirve de contrapeso. Pero supone un único comprador que conoce la oferta, cuando el poder de compra suele repartirse entre pocos, y el efecto favorable del precio mínimo exige que no se fije muy por encima del competitivo.

3.2. Monopolio bilateral

Idea: un único vendedor se enfrenta a un único comprador, como un sindicato y una patronal. El vendedor querría el resultado del monopolio y el comprador, el del monopsonio, y entre ambos precios la teoría no determina el resultado, que depende del poder de negociación, la paciencia y las alternativas exteriores. Ambos prefieren una cantidad inferior a la competitiva, aunque en general no la misma.

Equilibrio: si negocian cooperativamente, maximizan el excedente conjunto en la cantidad competitiva y se lo reparten con el precio, de modo que el resultado es eficiente, pero su reparto, indeterminado (el terreno de la solución de Nash). Con información incompleta, la negociación puede fracasar o retrasarse. La eficiencia depende de que exista negociación y no de la estructura del mercado, y la regulación se limita a fijar el marco de la negociación y a ofrecer arbitraje.

3.3. Monopolios multidimensionales

Multiproducto (varios mercados): al fijar p_1 y p_2, el monopolista tiene en cuenta el efecto de cada precio sobre las ventas del otro bien:

$$I_{L,i} = \frac{1}{|\varepsilon_{ii}|}\left(1 + I_{L,j}\,\frac{\text{IT}_j}{\text{IT}_i}\,\varepsilon_{ji}\right)$$

Con demandas independientes y costes no separables, las economías de alcance abren la puerta a subsidios cruzados. Con demandas dependientes y costes separables, si los bienes son sustitutivos brutos ($\varepsilon_{ji} > 0$) el poder de mercado aumenta, $I_{L,i} > 1/|\varepsilon_{ii}|$, porque el monopolista internaliza la externalidad de fijar precios altos. Si son complementarios, como las videoconsolas y sus juegos, disminuye, $I_{L,i} < 1/|\varepsilon_{ii}|$, y fija precios más bajos, lo que hace posibles, de nuevo, los subsidios cruzados. El signo de la relación de demanda decide así el poder de mercado, pero la regla exige conocer las elasticidades cruzadas y el coste conjunto.

Multiplanta (cuasicártel, como el de A19): el monopolista iguala el ingreso marginal del mercado al coste marginal de cada planta, con el mismo margen sobre este que en el monopolio simple:

$$\text{Img}(X) = \text{Cmg}_i = \text{Cmg}_j \ \Leftrightarrow \ p = \frac{1}{1 - 1/|\varepsilon_{Xp}|} \cdot \text{Cmg}_i$$

La producción total se fija como en una sola planta, con el coste marginal agregado, suma horizontal de los de cada planta, y se reparte para que la última unidad cueste lo mismo en todas. Si una costara menos, trasladar producción hacia ella reduciría el coste. Con costes marginales constantes y distintos, sin límite de capacidad, toda la producción va a la planta más barata. Si son crecientes, la planta con ventaja en costes produce más, pero no toda la producción. El modelo supone plantas bajo un solo control que conoce sus costes. Aplicado al cártel, el reparto no es estable, porque cada miembro tiene incentivos para desviarse.

1. Fundamentos de la competencia monopolística

Los modelos de primera generación, el espacial de Hotelling (1929) y el no espacial de Chamberlin (1933), parten de un mismo concepto, la diferenciación de producto. La publicidad es el instrumento con el que la empresa la crea y la mantiene.

1.1. Concepto de diferenciación

Definición: un producto está **diferenciado** cuando los consumidores no consideran sus variedades sustitutivos perfectos, de modo que cada empresa conserva parte de su clientela aunque su precio sea algo mayor que el de las rivales.

Diferenciación horizontal: a igualdad de precio, unos consumidores prefieren una variedad y otros la otra. No hay una ordenación de calidad compartida, sino gustos distintos, como el color de un coche o la localización de una tienda. Se modeliza con dos enfoques:

- *Espacial* o de direcciones: las variedades son puntos de un espacio de características y cada consumidor elige la más próxima a su posición, corregida por el precio (Hotelling, Lancaster y Salop).
- *No espacial* o de consumidor representativo: los bienes diferenciados entran como tales en la utilidad, son sustitutivos imperfectos y la diversidad reporta utilidad por sí sola (Chamberlin y Dixit-Stiglitz).

Diferenciación vertical: a igualdad de precio, todos prefieren la misma variedad, porque hay una ordenación de calidad compartida. Los consumidores solo difieren en cuánto pagan por la calidad, normalmente por su renta, y la variedad de menor calidad solo se vende si su precio es suficientemente menor (Shaked y Sutton).

A ellas se añade la **diferenciación informacional**, que procede de lo que los consumidores saben de los productos y no de sus características.

1.2. Enfoque espacial: ciudad lineal (Hotelling, 1929)

Idea: Hotelling (1929) representa los niveles de características como localizaciones en una calle. Elegir dónde situarse es elegir qué grado de característica producir, y la rivalidad es local, entre vecinas, de modo que el mercado se compone de oligopolios solapados.

1.2.1. Supuestos

Entorno: una calle de longitud unitaria, tramo acotado de un espacio de características más amplio.

Agentes: dos empresas en $z_1 = a$ y $z_2 = 1 - b$ venden un bien idéntico con el mismo coste unitario c, de modo que solo difieren en la ubicación. Los consumidores ocupan la calle con densidad uniforme, compran una unidad o ninguna y el mercado queda cubierto, porque su precio de reserva queda holgadamente por encima de los precios.

Restricciones (costes de transporte): cada unidad de distancia resta utilidad y mide lo que cuesta conformarse con una variedad distinta de la deseada, es decir, la heterogeneidad de los gustos. Hotelling los supuso lineales, pero aquí son cuadráticos, td^2, porque con costes lineales el equilibrio puede no existir.

1.2.2. Localizaciones exógenas: precios

Planteamiento: con las localizaciones dadas, cada lado del mercado tiene su problema.

- *Demanda*: parte del **consumidor indiferente**, $\hat{z}$, al que le cuesta lo mismo, precio más transporte, comprar a una u otra empresa:

$$p_1 + t(\hat{z} - z_1)^2 = p_2 + t(z_2 - \hat{z})^2$$

  Despejando, la demanda de la empresa 1 tiene tres componentes, la *cautiva* o del patio de atrás, a, la *compartida*, la mitad del tramo entre ambas, y la *disputada*, que se mueve con la diferencia de precios:

$$x_1^d = a + \frac{1 - a - b}{2} + \frac{p_2 - p_1}{2t(1 - a - b)}$$

- *Oferta*: cada empresa elige su precio tomando el de la rival como dado:

$$\max_{p_i} \ \pi_i = (p_i - c)\, x_i^d$$

CPO: las funciones de reacción son crecientes, con pendiente 1/2, de modo que los precios son **complementarios estratégicos**, y su ordenada crece con el coste de transporte y con la distancia entre empresas:

$$p_1 = \frac{p_2 + c}{2} + \frac{t(1 - a - b)(1 + a - b)}{2}$$
$$p_2 = \frac{p_1 + c}{2} + \frac{t(1 - a - b)(1 - a + b)}{2}$$

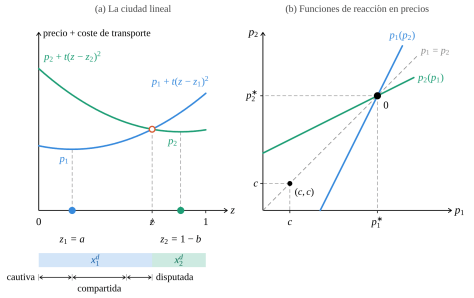
CSO: el beneficio es estrictamente cóncavo en el propio precio, $\partial^2 \pi_i / \partial p_i^2 = -1/[t(1 - a - b)] < 0$, así que la función de reacción es la mejor respuesta.

Equilibrio: en el corte de las funciones de reacción está el equilibrio de Nash en precios, con sus beneficios:

$$p_1^* = c + t(1 - a - b)\left(1 + \frac{a - b}{3}\right) \qquad \pi_1^* = \frac{t}{2}(1 - a - b)\left(1 + \frac{a - b}{3}\right)^2$$

El precio es el coste marginal más un margen proporcional al coste de transporte y a la distancia con la rival. El beneficio se reparte a favor de la empresa con más patio de atrás, que cobra algo más y vende más. Dos casos lo ilustran:

- *Diferenciación máxima*, $a = b = 0$: $p^* = c + t$ y $\pi^* = t/2$. Aunque el bien es idéntico, para quitar un cliente a la rival hay que compensarle todo el trayecto.
- *Diferenciación mínima*, $a + b = 1$: $p^* = c$ y beneficios nulos, el resultado de Bertrand.



Hotelling: consumidor indiferente en la calle y equilibrio de Nash en precios.

Estática comparativa: cuanto mayor es t o la distancia entre empresas, más cautivos son los consumidores y mayores los precios y los beneficios, al desplazarse hacia fuera las funciones de reacción. Con $t = 0$ se recupera la paradoja de Bertrand. La ubicación es a la vez la fuente de la diferenciación y del poder de mercado.

1.2.3. Localizaciones endógenas

Planteamiento: si las empresas eligen también dónde situarse, el juego tiene dos etapas, localización y después precios, y se resuelve hacia atrás.

- *Demanda*: es la del modelo anterior, ahora con localizaciones variables.
- *Oferta*: en la segunda etapa, los precios de equilibrio dependen de las localizaciones, $p_i^*(a, b)$. En la primera, cada empresa elige su localización anticipándolos:

$$\max_{a} \ \pi_1(a, b) = (p_1^*(a, b) - c)\, x_1^d(a, b, p^*(a, b))$$

CPO: el efecto de la localización sobre el beneficio se descompone en dos efectos de signo contrario:

$$\frac{\partial \pi_1}{\partial a} = (p_1^* - c)\left[\underbrace{\frac{\partial x_1^d}{\partial a}}_{\text{efecto demanda}} + \underbrace{\frac{\partial x_1^d}{\partial p_2}\frac{\partial p_2^*}{\partial a}}_{\text{efecto estratégico}}\right]$$

- *Efecto demanda*, positivo: acercarse al centro, y a la rival, aumenta la cuota (**principio de diferenciación mínima** de Hotelling).
- *Efecto estratégico*, negativo: con las empresas juntas, la competencia en precios se endurece, y separarse la suaviza. Llevado al límite, este efecto da el **principio de diferenciación máxima**.

El dilema es parecerse al rival, con más cuota, o alejarse, con más margen (producción en masa o nicho). Lo deciden los costes de transporte y la distribución de los consumidores. La CSO no se cumple en general.

Equilibrio: con costes cuadráticos, el beneficio de la empresa 1, $\pi_1^*(a, b)$ del modelo anterior, decrece en a para cualquier b, de modo que cada empresa se aleja todo lo posible de la rival, sea cual sea su posición. Domina el efecto estratégico:

$$a^* = b^* = 0 \qquad p_1^* = p_2^* = c + t \qquad \pi_1^* = \pi_2^* = \frac{t}{2}$$

El precio es el coste marginal más un margen igual al coste de transporte, y el beneficio se reparte a partes iguales, porque las posiciones son simétricas.

Bienestar: el planificador solo minimiza el coste de transporte y situaría cada empresa en el punto medio de la mitad de calle que abastece, 1/4 y 3/4. El mercado diferencia demasiado para relajar la competencia en precios, y la pérdida es el exceso de coste de transporte.

1.2.4. Conclusiones

Aspectos positivos: muestra con la mayor economía de medios cómo la diferenciación crea poder de mercado, que depende de la distancia entre variedades, de la paradoja de Bertrand a un margen igual al coste de transporte.

Limitaciones: el equilibrio es frágil ante la forma de los costes de transporte y el reparto de los consumidores. El modelo es de duopolio, con demandas unitarias y mercado cubierto, y no determina el número de variedades. Con costes cuadráticos hay equilibrio en precios para cualquier localización, y con lineales solo si las empresas están bastante alejadas.

1.3. Enfoque no espacial: Chamberlin (1933)

Origen: Chamberlin (1933) acuña el término para una estructura que combina el poder de cada empresa sobre su variedad, propio del monopolio, con muchas empresas y libre entrada, propias de la competencia perfecta.

Idea: hay varios bienes sustitutivos imperfectos, $0 < \sigma_{ij} < \infty$. Cada empresa se enfrenta a una demanda decreciente pero muy elástica, y sus beneficios extraordinarios desaparecen a largo plazo con la entrada.

Supuestos:

- *Grupo* de variedades poco diferenciadas, sustituibles en igual grado (**supuesto de simetría**).
- Muchas empresas, cada una con una variedad y costes medios en U. Como pesan poco, no hay interacción estratégica y cada una supone que las rivales no tocarán su precio (comportamiento *miope*).
- *Dos curvas de demanda*: la *planeada*, dd, la que la empresa cree tener si nadie más mueve el precio, muy elástica; y la *efectiva*, DD, la que obtiene cuando todas cambian el precio a la vez, menos elástica y que se desplaza con la entrada y la salida, porque la misma demanda de mercado se reparte entre más o menos variedades.

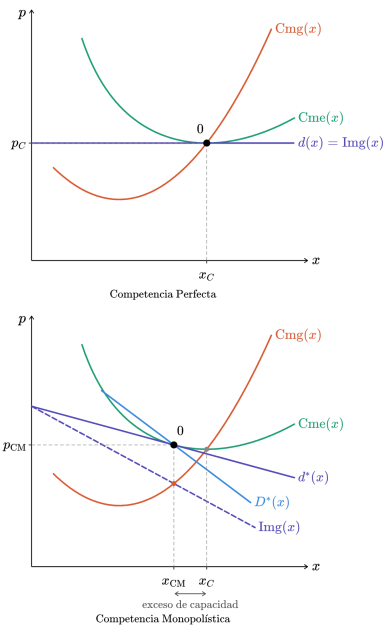
Razonamiento: a corto plazo, cada empresa iguala el coste marginal al ingreso marginal de dd, pero como las rivales hacen lo mismo, sus ventas caen sobre DD. El proceso sigue en zigzag hasta que $\text{Img}_{dd} = \text{Cmg}$ y $dd = DD$. Cada empresa es monopolista de su variedad y puede tener beneficios o pérdidas. Su precio es más alto, y su cantidad más baja, que en competencia.

A largo plazo, los beneficios atraen entrada, que desplaza la demanda de las instaladas hasta agotarlos. En equilibrio, $d^* = D^*$ y la demanda planeada es tangente al coste medio de largo plazo. Cada empresa sigue igualando el ingreso marginal al coste marginal, con un margen pequeño porque su demanda planeada, de elasticidad $|\varepsilon_{xp}|$, es muy elástica, y la entrada anula el beneficio:

$$\text{Img}(x) = \text{Cmg}_{\text{LP}}(x) \iff p = \frac{|\varepsilon_{xp}|}{|\varepsilon_{xp}| - 1}\,\text{Cmg}_{\text{LP}}(x)$$
$$p = \text{Cme}_{\text{LP}}(x)$$

Como la demanda tiene pendiente negativa, la tangencia está en el tramo descendente del coste medio. Por el **teorema del exceso de capacidad**, cada empresa produce por debajo de su escala óptima y deja economías de escala sin explotar. Demsetz (1959) objetó que, con gastos de venta, no hay necesariamente exceso de capacidad.

Desarrollo gráfico: en competencia perfecta, la referencia, el equilibrio está en el mínimo del coste medio, con x_C y p_C. En competencia monopolística, la demanda planeada, $d^*(x)$, es tangente al coste medio en x_{CM}, donde su ingreso marginal corta al coste marginal, y la efectiva, $D^*(x)$, más inclinada, la corta en ese punto. La distancia entre x_{CM} y x_C es el exceso de capacidad.



Chamberlin: tangencia de la demanda con el coste medio y exceso de capacidad.

Bienestar: el equilibrio es ineficiente por dos vías: *asignativa*, como el monopolio, con $p > \text{Cmg}$, y *productiva*, propia de esta estructura, con p por encima del mínimo del coste medio. Cada empresa produce al mínimo coste la cantidad que elige, pero por debajo de su escala eficiente, y la producción del sector cuesta más de lo que costaría con menos empresas. Con demandas muy elásticas ambas pérdidas son pequeñas, y la variedad adicional, que el consumidor valora, las compensa en parte, comparación que formaliza el modelo de Dixit-Stiglitz.

Conclusiones: incorpora a la empresa la decisión de cuánto diferenciar y cuánto gastar en ventas, y su equilibrio de tangencia es el punto de partida de la teoría de la variedad. La crítica le reprocha cuatro incoherencias: ignora la estrategia entre variedades parecidas, atribuye a las empresas un error sistemático sobre las rivales y no precisa ni cuántas empresas hacen falta ni qué elasticidades cruzadas definen el grupo. Las dos primeras las resuelven los modelos neohotellianos, con vecinos y estrategia, y las dos últimas, el de Dixit-Stiglitz, con un número de variedades endógeno.

1.4. Publicidad: Dorfman-Steiner (1954)

Origen: Dorfman y Steiner (1954) formalizan la decisión sobre la publicidad, el principal de los gastos de venta de Chamberlin.

Planteamiento: con una demanda $x(p, A)$ creciente en el gasto en publicidad con rendimientos decrecientes, $x_{AA} < 0$, la empresa elige precio y publicidad, $\max_{p,A}\,(p-c)\,x(p,A) - A$.

CPO: respecto del precio, la regla del margen, que iguala el índice de Lerner a la inversa de la elasticidad-precio, y respecto de la publicidad, que el margen de la última venta que esta provoca iguale su coste:

$$\frac{p-c}{p} = \frac{1}{|\varepsilon_{xp}|} \in (0,1] \qquad (p-c)\,x_A = 1$$

Solución: combinándolas se obtiene la **condición de Dorfman-Steiner**, donde $\varepsilon_{xA} = x_A A/x$ es la elasticidad de la demanda respecto de la publicidad:

$$\frac{A}{p\,x} = \frac{\varepsilon_{xA}}{|\varepsilon_{xp}|} = \varepsilon_{xA}\,\frac{p-c}{p}$$

La empresa anuncia más cuanto más responde su demanda a la publicidad y menos al precio, de modo que cuanto más diferenciada es la variedad, mayor es su intensidad publicitaria. Como esta es proporcional al margen relativo, la condición fundamenta la asociación entre margen y ratio publicidad-ventas entre sectores. A largo plazo, la publicidad forma parte del coste medio y la entrada sigue llevando el beneficio a cero.

Conclusiones: explica con un cociente de elasticidades por qué la intensidad publicitaria difiere entre sectores, pero no dice si la publicidad informa o persuade e ignora la respuesta de las rivales, que en oligopolio convierte la publicidad en un instrumento estratégico.

En suma, la primera generación explica el poder de mercado sin pocas empresas, por la distancia entre variedades o por la sustituibilidad imperfecta, pero ningún enfoque determina a la vez la localización, el número de variedades y su valoración social.

2. Diferenciación horizontal: segunda generación

Los modelos neohotellianos conservan el espacio de características y endogenizan la entrada, y los neochamberlinianos conservan el consumidor representativo y le dan microfundamentos.

2.1. Modelos neohotellianos

2.1.1. Demanda de características (Lancaster, 1966)

Idea: Lancaster (1966) sostiene que el consumidor busca las características de los bienes, no los bienes como tales. Cada bien es una combinación fija de atributos y elegir una variedad es elegir la combinación más cercana a la deseada al menor coste. Generaliza el espacio de Hotelling a cualquier número de atributos.

Planteamiento: con una tecnología de consumo lineal, el consumidor maximiza la utilidad de las características:

$$\max_{\mathbf{x}}\; u(a_1,\ldots,a_m) \quad \text{s.a.:} \begin{cases} a_j = \sum_i \beta_{ji} x_i \;\forall j \\ \sum_i p_i x_i = M_0 \end{cases}$$

Equilibrio: cada bien es un rayo desde el origen en el plano de características, cuya pendiente es la proporción en que combina las características y cuya longitud depende del precio. Gastando toda la renta en cada uno se obtienen puntos cuya unión es la **frontera eficiente de consumo**, cóncava. El óptimo es la tangencia de la curva de indiferencia más alta con la frontera, en un vértice, con una sola variedad, o en un tramo, combinando dos. La cuasiconcavidad de la utilidad y la concavidad de la frontera aseguran que la tangencia es el máximo.

Estática comparativa: si sube el precio de un bien, su rayo se acorta y puede quedar fuera de la frontera y dejar de consumirse. Una nueva variedad dominada no se compra, y si amplía la frontera desplaza a sus vecinas. La demanda de cada variedad es, así, discontinua.

Conclusiones: explica la diferenciación y la aparición y desaparición de variedades como fenómenos de demanda, funda los precios hedónicos y se aplica a carteras y a la estrategia comercial: segmentación, nicho y posicionamiento. Pero supone características objetivas y tecnología lineal, y no determina el número de variedades.

2.1.2. Ciudad circular (Salop, 1979)

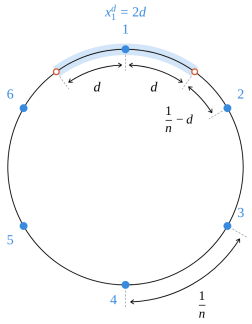
Idea: cerrar la calle de Hotelling en una circunferencia elimina los extremos y la asimetría entre empresas, y permite estudiar cuántas entran con costes fijos.

Supuestos: ciudad circular de perímetro unitario, como los horarios de vuelo o la programación televisiva; n empresas endógeno, con coste marginal c y coste fijo de entrada f, situadas equidistantes; consumidores uniformes, con demandas unitarias y costes de transporte lineales t.

Planteamiento: con n empresas equidistantes, cada lado del mercado tiene su problema.

- *Demanda*: el consumidor indiferente está a una distancia d tal que $p_1 + td = p + t(1/n - d)$, y la demanda de la empresa 1 es $x_1^d = 2d = 1/n + (p - p_1)/t$.
- *Oferta*: cada empresa elige su precio tomando como dados el precio p de las vecinas y el número de empresas:

$$\max_{p_1} \ \pi_1 = (p_1 - c)\left[\frac{1}{n} + \frac{p - p_1}{t}\right] - f$$



Salop: empresas equidistantes en la ciudad circular y demanda de la empresa 1.

CPO: el precio óptimo, que iguala el ingreso marginal al coste marginal, c, crece con el de las vecinas, con el coste de transporte y con la distancia entre empresas, $1/n$:

$$p_1 = \frac{1}{2}\left(p + c + \frac{t}{n}\right)$$

CSO: $\partial^2\pi_1/\partial p_1^2 = -2/t < 0$, luego la CPO define la mejor respuesta.

Equilibrio: con simetría se obtiene el precio a corto plazo, y con la libre entrada, $\pi = 0$, el número de empresas y el precio de largo plazo:

$$p^* = c + \frac{t}{n} \qquad n^* = \sqrt{t/f} \qquad p^{**} = c + \sqrt{tf}$$

Estática comparativa: el margen crece con t, que mide la diferenciación, y cae con n, que mide la competencia. El número de empresas crece con la diferenciación y decrece con el coste fijo. El precio de largo plazo crece con t y con f.

Bienestar: el planificador elegiría $n^{\text{opt}} = \frac{1}{2}\sqrt{t/f}$, la mitad. La libre entrada genera un **exceso de variedades**, porque cada entrante roba demanda ya servida sin internalizar que su coste fijo es un coste social adicional.

Conclusiones: reproduce con microfundamentos el resultado de Chamberlin, beneficio nulo con poder de mercado, y determina el número de variedades. Pero impone la localización equidistante, y con demandas elásticas la entrada puede ser insuficiente.

2.2. Neochamberlinianos (Dixit y Stiglitz, 1977)

Idea: Spence (1976) y Dixit y Stiglitz (1977) reformulan a Chamberlin con microfundamentos, en equilibrio general y con un consumidor representativo para el que la variedad es un bien en sí mismo. El número de variedades depende de cuánto se valora la variedad y de cuánto cuesta producir cada una.

2.2.1. Supuestos

Entorno: competencia monopolística con libre entrada y pleno empleo, de modo que el trabajo total limita el número de variedades y su escala.

Agentes:

- *Empresas*: tecnología idéntica con rendimientos crecientes y un solo factor, el trabajo, con costes $\text{CT} = w(l_0 + lx_i^s)$, donde l_0 son los empleados fijos y l los empleados por unidad. Cada una produce una variedad y es miope.
- *Consumidores*: idénticos, con **preferencia por la variedad**: con el mismo gasto prefieren n variedades a $n-1$. El parámetro $\theta \in (0,1)$ da la elasticidad de sustitución entre variedades, $\sigma = 1/(1-\theta) > 1$. Como preferencias y costes son idénticos, en equilibrio todas las variedades tienen el mismo precio.

2.2.2. Modelo

Consumidores: el consumidor representativo maximiza una utilidad CES:

$$\max_{x_i} \ u = \left(\sum_{i=1}^{n} x_i^{\theta}\right)^{1/\theta} \qquad \text{s.a.:} \ \sum_{i=1}^{n} p_i x_i = M_0$$

En las CPO, la utilidad marginal de cada variedad iguala su precio por el valor marginal de la renta, $x_i^{\theta-1} = \lambda p_i$, la demanda frischiana. Por tanto, la RMS entre dos variedades iguala su precio relativo, y la composición del consumo responde más a los precios cuanto mayor es σ:

$$\frac{x_i}{x_j} = \left(\frac{p_j}{p_i}\right)^{\sigma}$$

La **demanda marshalliana** de cada variedad decrece con su precio, crece con el índice de precios P y con la renta, y es isoelástica, con $|\varepsilon_{xp}| = \sigma$:

$$x_i^d = \left(\frac{p_i}{P}\right)^{-\sigma} \frac{M_0}{P}$$

Con consumidores idénticos, la demanda de mercado de cada variedad es proporcional al tamaño del mercado, L, y conserva la elasticidad σ. El índice de precios de las variedades es:

$$P = \left(\sum_i p_i^{\theta/(\theta-1)}\right)^{(\theta-1)/\theta}$$

Empresas: cada una fija su precio para maximizar el beneficio frente a esa demanda, con la producción, x_i^s, igual a la demanda de mercado de su variedad:

$$\max_{p_i} \ \pi_i = p_i x_i^s - w(l_0 + lx_i^s)$$

La CPO iguala el ingreso marginal al coste marginal, wl, lo que equivale a fijar un margen constante sobre este, independiente de los precios ajenos y mayor cuanto menor es la sustituibilidad:

$$\text{Img} = p_i\left(1 - \frac{1}{\sigma}\right) = wl \iff p_i = \frac{\sigma}{\sigma - 1}\, wl = \frac{1}{\theta}\, wl$$

Equilibrio: la libre entrada, $\pi = 0$, fija la escala de cada empresa, y el pleno empleo, $L = \sum_i (l_0 + lx_i^s)$, el número de variedades:

$$x_i^s = \frac{l_0}{l}(\sigma - 1) \qquad n = \frac{L}{l_0 \sigma}$$

Estática comparativa: la escala crece con el coste fijo y con la sustituibilidad, que reduce el margen. El número de variedades crece con el tamaño del mercado y decrece con el coste fijo y la sustituibilidad. Un mercado mayor no altera la escala de cada empresa. El salario real, $w/p = \theta/l$, no depende del tamaño del mercado, y las ganancias llegan por la variedad, no por el precio. Con economías de escala internas, el número de variedades está, así, limitado por el tamaño del mercado.

Bienestar: con preferencias CES, el mercado produce el número de variedades del óptimo restringido, con beneficio nulo, porque el margen afecta por igual a todas. En el original, con un sector numerario, produce menos que el óptimo sin restricción. Con preferencias más generales puede haber demasiadas o demasiado pocas, según domine el robo de demanda o el excedente que la empresa no se apropia.

2.2.3. Economía abierta (Krugman, 1979b)

Idea: la apertura equivale a un aumento del tamaño del mercado y genera comercio aun sin ventaja comparativa.

Con la CES, la escala y el número de empresas de cada país no cambian, y la ganancia procede de la variedad, porque a las variedades nacionales se suman las importadas. Krugman supone, en cambio, que la elasticidad de la demanda aumenta al caer el consumo individual de cada variedad. Entonces cada país se especializa en menos variedades, las empresas supervivientes ganan escala y bajan su precio (el **efecto escala**). Con empresas heterogéneas, Melitz (2003) añade el efecto selección, que expulsa a las menos eficientes.

Ambos países ganan aunque no difieran en recursos ni tecnología, por la mayor variedad y, con el supuesto de Krugman, por la mayor escala. La teoría tradicional no podía explicar este resultado, que da cuenta del comercio intraindustrial entre países similares.

Conclusiones: obtiene a la vez el comercio y sus ganancias de las economías de escala internas, y es una obra fundacional de la nueva teoría del comercio, ampliada luego con costes de transporte (Krugman, 1980) y empresas heterogéneas. Pero el patrón de comercio queda indeterminado, el efecto escala descansa en un supuesto específico sobre la demanda y, con empresas idénticas, no predice cuáles sobreviven a la apertura.

2.2.4. Conclusiones

Aspectos positivos: la utilidad con preferencia por la variedad y la tecnología con costes fijos son el bloque básico de la nueva economía keynesiana (Blanchard y Kiyotaki, 1987), del crecimiento endógeno (Romer, 1990) y de la nueva teoría del comercio.

Limitaciones: deja implícitas las características, supone que todas las variedades son igual de próximas y, con un margen constante, excluye que la entrada intensifique la competencia en precios.

En suma, la segunda generación endogeniza el número de variedades, con una valoración normativa distinta: en Salop el equilibrio dobla el óptimo, y con preferencias CES el mercado alcanza el óptimo restringido.

3. Otros desarrollos

Dos desarrollos amplían la diferenciación por gustos: la diferenciación vertical, por calidad, y la informacional.

3.1. Competencia en calidad (Shaked y Sutton)

Idea: Shaked y Sutton (1982; 1983) muestran que, si las variedades se ordenan por calidad, la calidad alta atrae a todos si su precio no es mucho mayor, y solo pocas empresas pueden coexistir, con independencia del tamaño del mercado.

3.1.1. Economía cerrada (Shaked y Sutton, 1982)

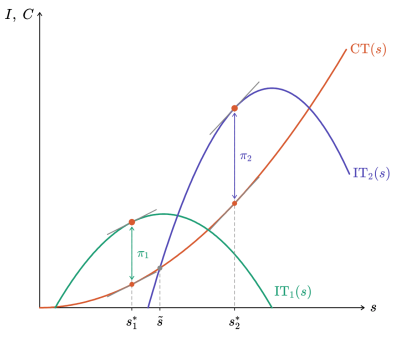
Supuestos: un continuo de consumidores con preferencias idénticas y demanda unitaria, con utilidad neta $u = vs_i - p_i$, donde s_i es la calidad y v la disposición a pagar, que crece con la renta. El coste variable es común e independiente de la calidad, que se obtiene con I+D, un coste fijo y hundido, creciente y convexo en ella, de modo que también lo es el coste total.

Planteamiento: un juego en tres etapas, entrada, calidad y precios, que se resuelve hacia atrás:

1. *Precios*: con calidades próximas, una pequeña rebaja quita al rival casi toda su clientela y los márgenes tienden a cero, como en Bertrand.
2. *Calidad*: acercarse al rival endurece la competencia en precios. Quien no alcanza la calidad alta prefiere situarse deliberadamente por debajo, lo que traslada a la calidad la diferenciación máxima y da origen a la **erosión de la calidad**.
3. *Entrada*: solo sobreviven calidades suficientemente separadas dentro del rango que fija la dispersión de la renta, y un entrante adicional no recuperaría su I+D.

Equilibrio: a corto plazo, la competencia en precios expulsa a las calidades que no encuentran compradores. A largo plazo, si la renta está poco dispersa, solo caben dos empresas, un **oligopolio natural**, con beneficios positivos, mayores para la de calidad alta, y diferenciación máxima en calidad.

Desarrollo gráfico: en el plano calidad-ingreso y coste, el coste total es creciente y convexo en la calidad. Dada la calidad de la rival, cada empresa elige aquella en la que su ingreso marginal iguala al coste marginal. La empresa 1 no sube su calidad, porque acercarse a s_2^* endurece la competencia en precios y hunde su ingreso. La empresa 2 no la baja, porque su beneficio se anularía en $\tilde{s}$.



Shaked y Sutton: calidades de equilibrio del oligopolio natural, con $\pi_2 > \pi_1$.

Propiedad de finitud: el número de empresas con cuota positiva está acotado por grande que sea el mercado, a diferencia de Salop o Krugman. Se cumple si, con todas las calidades a precio igual a su coste marginal, todos eligen la más alta (Shaked y Sutton, 1983), lo que ocurre cuando la calidad recae sobre todo en costes fijos y no en variables. Entonces la calidad alta no encarece la unidad y ninguna empresa de calidad baja puede competir en precio.

3.1.2. Economía abierta (Shaked y Sutton, 1984)

La finitud se mantiene, con un límite menor que la suma de las dos autarquías a corto plazo, porque el mercado integrado sostiene un único oligopolio, no la suma de los nacionales. A largo plazo surge un **duopolio natural internacional** y comercio intraindustrial de calidades distintas entre países de renta similar. El modelo no determina en qué país se sitúa cada empresa.

3.1.3. Conclusiones

Aspectos positivos: da un fundamento no tecnológico a la concentración en las industrias intensivas en I+D y publicidad, y presenta la calidad como instrumento para relajar la competencia en precios.

Limitaciones: Bonanno (1986) muestra que no se mantiene si se compite en cantidades, y supone que los consumidores conocen las calidades.

3.2. Diferenciación informacional

La información de los consumidores es en sí una fuente de diferenciación, y la publicidad su instrumento. Para la visión favorable, informa y reduce los costes de búsqueda. Benham (1972) halló gafas más caras en los estados que restringían la publicidad. Para la adversa, es persuasión y la diferenciación que crea no es real (Galbraith, 1967). Cuál se ajusta mejor depende del tipo de bien.

3.2.1. Búsqueda y experiencia (Nelson, 1970)

Clasificación: según cuándo se conoce la calidad:

- *Bienes de búsqueda*: se comprueba antes de comprar, como la ropa. La publicidad informa de la existencia y el precio, con un contenido verificable.
- *Bienes de experiencia*: solo tras consumirlos, como un restaurante. El propio gasto en publicidad señala la calidad, porque solo compensa a quien espera compras repetidas (Nelson, 1974), y lo mismo hacen la reputación, la marca y las garantías.
- *Bienes de confianza* (Darby y Karni, 1973): la calidad no se conoce ni después, como en una reparación, y la diferenciación descansa en la reputación y la regulación profesional.

3.2.2. Bienes de búsqueda (Butters, 1977)

Supuestos: empresas con un producto homogéneo que solo informan de su precio con anuncios aleatorios costosos; consumidores con demanda unitaria, $u = s_0 - p$, que compran al menor precio recibido. Cada anuncio llega a cada consumidor con probabilidad $1/N$, y con m anuncios la probabilidad de no recibir ninguno es $1-\phi \cong e^{-m/N}$. Cada anuncio cuesta c' y producir cada unidad, c, y $\ln[1/(1-\phi)] \cong m/N$ es el número de anuncios por consumidor.

Equilibrio: es una distribución de precios anunciados. Un precio alto solo vende a quien no ha recibido otro anuncio, y uno bajo, a todos los que lo reciben. Cada empresa elige cuántos anuncios enviar y a qué precio.

Bienestar: el planificador maximiza el excedente de los informados menos el coste de la publicidad, y su objetivo es estrictamente cóncavo en la proporción de informados, ϕ. Su condición coincide con la del equilibrio de mercado, ϕ_c:

$$s_0 - c = \frac{c'}{1-\phi^*} \;\Rightarrow\; \phi^* = \phi_c$$

La publicidad de competencia monopolística es socialmente óptima, porque un anuncio aumenta el bienestar en $s_0 - c$ si el consumidor no recibe otro, y eso es exactamente lo que gana la empresa. Apoya la visión favorable, aunque descansa en demandas unitarias y producto homogéneo.

3.2.3. Experiencia (Bain, 1956; Schmalensee, 1982)

Idea: probar un producto cuesta, de modo que el consumidor no trata igual las marcas probadas y las nuevas. Bain (1956) vio en la lealtad a las marcas una barrera a la entrada, y Schmalensee (1982) formalizó la **ventaja del pionero**.

Razonamiento: con entrada secuencial y consumidores escépticos ante las marcas nuevas, la primera marca satisfactoria se convierte en estándar, aun sin bajar precios. El resultado es una barrera de entrada que no procede de la tecnología ni de la estrategia, sino de la información. La publicidad, la reputación y las garantías sustituyen a la inspección previa. El consumidor que conoce una marca satisfactoria no tiene incentivo a asumir el coste de probar otra.

Conclusiones: conecta la diferenciación con la economía de la información y apoya la visión adversa de la publicidad. Pero los precios de introducción, las muestras o las garantías pueden vencer el escepticismo, y la ventaja puede reflejar una calidad realmente superior, de modo que su valoración normativa es ambigua.

1. Modelos de oligopolio no cooperativos

Sin acuerdos vinculantes, cada empresa elige su estrategia como mejor respuesta a lo que espera de sus rivales. Los modelos se distinguen por la variable estratégica, cantidades o precios, y por el orden de las decisiones, simultáneas o secuenciales.

1.1. Cantidades simultáneas (Cournot, 1838)

Idea: las empresas eligen a la vez sus cantidades y el precio se ajusta para vaciar el mercado. Cada una fija un margen sobre su coste marginal proporcional a su cuota, entre el resultado competitivo y el de monopolio.

1.1.1. Supuestos

Entorno: el de la competencia perfecta salvo por el poder de mercado: bien homogéneo, información completa y consumidores precio-aceptantes. Un *subastador walrasiano* ficticio anuncia el precio que vacía el mercado, $p = p(X)$, con demanda inversa decreciente.

Agentes: las empresas eligen su cantidad con costes $\mathrm{CT}_i(x_i)$. Son miopes: reconocen que su cantidad afecta al precio, pero suponen que no afecta a las demás, $\partial x_j/\partial x_i = 0$, es decir, sus **variaciones conjeturales** son nulas.

Restricciones (condiciones de Hahn): las cantidades son **sustitutos estratégicos**, $\partial^2\pi_i/\partial x_i\partial x_j < 0$, y la pendiente del coste marginal supera a la de la demanda inversa, $p' - \mathrm{CT}_i'' < 0$. Así, las funciones de reacción son continuas, con pendiente entre -1 y 0, y el beneficio es estrictamente cóncavo en la propia cantidad. Estas condiciones garantizan la existencia, la unicidad y la estabilidad del equilibrio.

1.1.2. Modelo

Planteamiento: cada empresa maximiza su beneficio dada la cantidad del resto:

$$\max_{x_i}\ \pi_i = p(X)\cdot x_i - \mathrm{CT}_i(x_i)$$

CPO: el ingreso marginal iguala al coste marginal, lo que equivale a fijar un margen sobre este, y el índice de Lerner de la empresa es su cuota, $s_i = x_i/X$, dividida por la elasticidad de la demanda:

$$\mathrm{Img}_i = \mathrm{Cmg}_i \ \Leftrightarrow\ p = \frac{1}{1 - s_i/|\varepsilon|}\cdot \mathrm{Cmg}_i$$

$$\frac{p - \mathrm{Cmg}_i}{p} = \frac{s_i}{|\varepsilon|} \in (0,1]$$

La condición anida los casos polares:

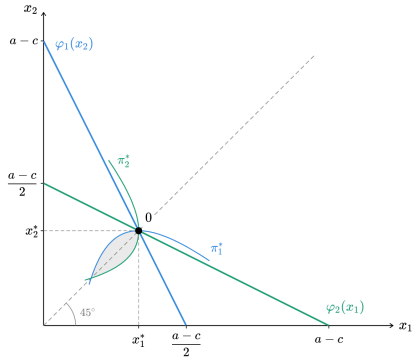
- *Competencia perfecta*: si $s_i \to 0$ o $|\varepsilon| \to \infty$, el margen desaparece y $p = \mathrm{Cmg}_i$.
- *Monopolio*: con $s_i = 1$ se obtiene la regla del monopolio.
- *Oligopolio*: las empresas con menor coste marginal tienen mayor cuota, de modo que la eficiencia relativa determina la estructura del mercado.

CSO: se cumple si la demanda no es demasiado convexa y el coste marginal no decrece, y la CPO define entonces la mejor respuesta:

$$\frac{\partial^2\pi_i}{\partial x_i^2} = 2p' + p''x_i - \mathrm{CT}_i'' < 0$$

Equilibrio: de la CPO sale la **función de reacción**, $x_i^* = \varphi_i(x_j)$. El **equilibrio de Cournot-Nash** es el par de cantidades en el que cada empresa produce su mejor respuesta a lo que produce la otra. Con empresas simétricas, $x_i^* = x^*$ para todas.

Desarrollo gráfico: en el plano (x_1, x_2), la función de reacción une los máximos de las curvas isobeneficio de cada empresa, es decreciente y corta los ejes en las cantidades de monopolio y de competencia. El equilibrio está en el corte de las funciones de reacción, sobre la bisectriz si los costes son iguales, y es estable porque φ_1 es más inclinada que φ_2.



Cournot: las funciones de reacción se cortan en el equilibrio de Nash en cantidades.

Estática comparativa: con n empresas simétricas, coste marginal c y demanda $p = a - X$, cada una produce $x^* = (a-c)/(n+1)$ y el precio es $p^* = (a+nc)/(n+1)$. Al crecer n, el precio tiende al coste marginal, de modo que la competencia perfecta es el límite del oligopolio de Cournot. Si baja el coste de una empresa, su función de reacción se desplaza hacia fuera, gana cuota, la rival produce algo menos, la producción total aumenta y el precio baja. La rival se contrae menos de lo que crece la otra porque la pendiente de su función de reacción es inferior a la unidad en valor absoluto.

Bienestar: el equilibrio no es eficiente ni para la sociedad, porque el precio supera el coste marginal, ni para las empresas, porque sus isobeneficios no son tangentes. El área encerrada entre esas curvas se conoce como **lente de Pareto** y representa las ganancias de cooperar.

Extensiones: ponderando por cuotas, el margen medio de la industria es el índice de Herfindahl-Hirschman dividido por la elasticidad, $\mathrm{HHI}/|\varepsilon|$, lo que fundamenta el uso de los índices de concentración. Con n empresas simétricas, el margen de cada una es $1/(n|\varepsilon|)$, y en general, con cuotas en tanto por uno:

$$\sum_i s_i\,\frac{p - \mathrm{Cmg}_i}{p} = \frac{\sum_i s_i^2}{|\varepsilon|} = \frac{\mathrm{HHI}}{|\varepsilon|}$$

1.1.3. Conclusiones

Aspectos positivos: la predicción es razonable y anidada, $X_M < X_{\mathrm{OC}} < X_C$, el equilibrio existe, es único y estable, y el modelo es la referencia de la teoría del oligopolio.

Limitaciones: es poco realista, con empresas que compiten en cantidades y un subastador que fija el precio, y las empresas son miopes. A lo primero responde Bertrand, con precios, y a lo segundo, Stackelberg, con anticipación.

1.2. Cantidades secuenciales (Stackelberg, 1934)

Idea: una empresa líder elige primero su cantidad anticipando la reacción de la seguidora, lo que le da la **ventaja de mover primero**.

Supuestos: los de Cournot, pero en dos periodos. La líder tiene una ventaja exógena, como una inversión previa en capacidad, que es un compromiso creíble. Su variación conjetural es $\partial x_2/\partial x_1 = \varphi_2'(x_1) < 0$. La seguidora observa la cantidad de la líder y su conjetura es nula.

Planteamiento: por inducción hacia atrás, la seguidora responde con su función de reacción de Cournot, $x_2 = \varphi_2(x_1)$, y la líder maximiza su beneficio sujeta a ella:

$$\max_{x_1}\ \pi_1 = p(x_1 + x_2)\cdot x_1 - \mathrm{CT}_1(x_1) \quad \text{s.a.: } x_2 = \varphi_2(x_1)$$

CPO: es la de Cournot con un término adicional positivo, el **efecto estratégico**, que recoge cómo la líder anticipa su impacto sobre la seguidora:

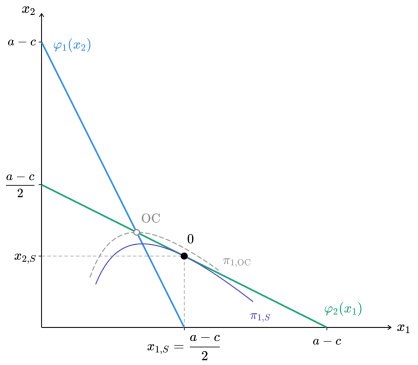
$$\mathrm{Img}_1 + \frac{\partial p}{\partial x_2}\cdot\frac{\partial \varphi_2}{\partial x_1}\cdot x_1 = \mathrm{Cmg}_1$$

Equilibrio: es un ENPS en el que la líder elige, sobre la función de reacción de la seguidora, el punto tangente a su isobeneficio de mayor beneficio. Respecto a Cournot, la líder produce y gana más, y la seguidora produce y gana menos. Como la seguidora reduce su producción menos de lo que la líder la aumenta, la producción total sube y el precio baja:

$$X_C > X_S > X_{\mathrm{OC}} > X_M$$

Con demanda $p = a - X$ y coste marginal común c, la líder produce y gana el doble que la seguidora, frente a $(a-c)/3$ y $(a-c)^2/9$ de cada empresa en Cournot:

$$x_1 = \frac{a-c}{2} > x_2 = \frac{a-c}{4}, \qquad \pi_1 = \frac{(a-c)^2}{8} > \pi_2 = \frac{(a-c)^2}{16}$$



Stackelberg: la líder elige sobre la reacción de la seguidora el punto de mayor beneficio.

Estática comparativa: con demanda lineal y coste marginal común, las cantidades son proporcionales a $a-c$ y la líder conserva dos tercios de la producción total.

Bienestar: el liderazgo es procompetitivo, porque el excedente del consumidor aumenta respecto a Cournot y, con costes marginales constantes e idénticos, la pérdida irrecuperable es menor, aunque el equilibrio sigue sin ser eficiente en el sentido de Pareto.

Conclusiones: introduce el tiempo y la ventaja de mover primero. Pero los papeles de líder y seguidora son exógenos, si ambas intentan liderar no hay equilibrio, la guerra de Stackelberg, y el compromiso solo es creíble si es irreversible.

1.3. Precios simultáneos (Bertrand, 1883)

Idea: Bertrand (1883) critica el realismo de Cournot y hace del precio la variable estratégica. Con producto homogéneo, costes marginales constantes e idénticos y sin restricciones de capacidad, dos empresas reproducen el resultado competitivo.

1.3.1. Supuestos

Los de Cournot, pero las empresas fijan precios a la vez, con rendimientos constantes, costes idénticos y capacidad para abastecer todo el mercado. Cada una supone que la rival mantendrá su precio. Los consumidores compran al más barato y, con precios iguales, se reparten a partes iguales, de modo que la demanda de cada empresa es discontinua en el precio de la rival. La demanda de la empresa i es:

$$d_i(p_i,p_j)=\begin{cases} X(p_i) & \text{si } p_i<p_j \\ X(p_i)/2 & \text{si } p_i=p_j \\ 0 & \text{si } p_i>p_j \end{cases}$$

1.3.2. Modelo

Planteamiento: cada empresa elige su precio para maximizar su beneficio dado el de la rival, con la demanda discontinua de los supuestos:

$$\max_{p_i}\ \pi_i=(p_i-c)\,d_i(p_i,p_j)$$

CPO: no existen como tales. Como la demanda salta en $p_i=p_j$, el beneficio no es derivable en ese punto, y la mejor respuesta no se obtiene igualando una derivada a cero, sino comparando el beneficio de cada tramo.

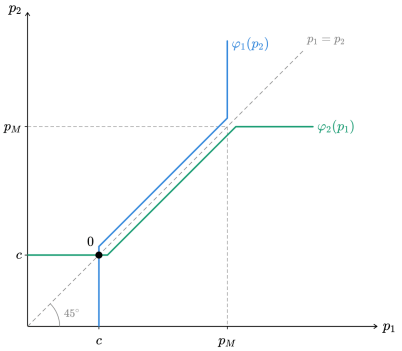
CSO: tampoco existe, pero esa comparación ya garantiza un máximo.

Solución: la función de reacción se define por tramos. Ante cualquier precio de la rival entre el coste marginal y el de monopolio, a la empresa le conviene cobrar apenas un poco menos, $p_j-\eta$, y quedarse con toda la demanda. Si la rival cobra más que el monopolio, fija p_M, y si cobra menos que el coste, fija c. Con $\eta>0$ arbitrariamente pequeño:

$$p_i\equiv\varphi_i(p_j)=\begin{cases} p_M & \text{si } p_j>p_M \\ p_j-\eta & \text{si } c<p_j\le p_M \\ c & \text{si } p_j\le c \end{cases}$$

Equilibrio: el único equilibrio de Nash es $p_1^*=p_2^*=c$, con beneficios nulos. Subir el precio en solitario deja a la empresa sin clientes, y bajarlo la obliga a vender con pérdidas.

Desarrollo gráfico: en el plano (p_1,p_2), la función de reacción de la empresa 1 discurre justo por encima de la bisectriz entre el coste marginal y el precio de monopolio, y fuera de ese tramo es vertical en p_M y en c. La de la empresa 2 es su simétrica respecto a la bisectriz. Las dos solo se cortan en (c,c).



Bertrand con bienes homogéneos: las funciones de reacción solo se cortan en (c,c).

Estática comparativa: una variación de c se traslada íntegra al precio, y el número de empresas, a partir de dos, no altera el equilibrio, a diferencia de Cournot.

Bienestar: es ineficiente para las empresas, que ganarían cooperando, pero eficiente para la sociedad, sin pérdida irrecuperable.

1.3.3. Paradoja de Bertrand y soluciones

Paradoja: competir en precios es más verosímil que en cantidades, pero la conclusión lo es menos, porque bastan dos empresas para el resultado competitivo, cuando en las industrias con pocos competidores los precios son supracompetitivos. Cada solución relaja un supuesto:

1. *Restricciones de capacidad*: con ellas no se puede acaparar toda la demanda. Para Edgeworth (1897), el equilibrio puede no existir en estrategias puras y los precios siguen ciclos, como los del carburante. Para Kreps y Scheinkman (1983), si las empresas eligen primero la capacidad y compiten después en precios, con racionamiento eficiente y demanda cóncava, el resultado coincide con el de Cournot, y la capacidad es, así, la verdadera variable estratégica.
2. *Diferenciación de producto*: las funciones de reacción son continuas y crecientes, los precios son **complementarios estratégicos** y el equilibrio tiene precios por encima del coste marginal.
3. *Costes asimétricos*: la empresa de menor coste fija un precio algo inferior al coste de la rival y obtiene beneficios.
4. *Interacción repetida*: la cooperación tácita puede sostenerse con amenazas de castigo.

Conclusiones: muestra que la variable estratégica importa, y su resultado competitivo es la referencia frente a la que se miden los demás modelos, pero descansa en supuestos extremos.

1.4. Precios secuenciales (Gal-Or, 1985)

Idea: Gal-Or (1985) muestra que mover primero es una ventaja si las funciones de reacción son decrecientes, como con cantidades, y una desventaja si son crecientes, como con precios y producto diferenciado.

Supuestos: los de Bertrand, en dos periodos y con producto diferenciado, con demanda lineal $x_i=a-p_i+\gamma\,p_j$, donde $\gamma\in(0,1)$ es el grado de sustituibilidad. Con producto homogéneo, la seguidora rebajaría cualquier precio de la líder y el resultado sería el de Bertrand.

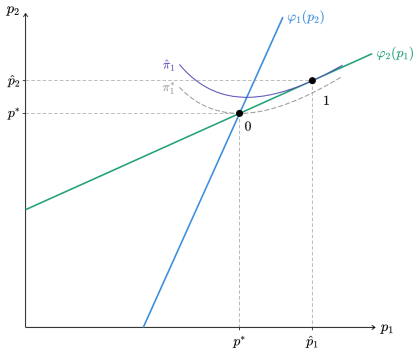
Planteamiento: por inducción hacia atrás, la seguidora responde a p_1 como en Bertrand y la líder maximiza su beneficio sujeta a esa respuesta:

$$\max_{p_1}\ \pi_1=(p_1-c)\,(a-p_1+\gamma\,p_2)\quad \text{s.a.: } p_2=\frac{a+c+\gamma\,p_1}{2}$$

CPO: añade a la de Bertrand un efecto estratégico positivo. Como subir p_1 lleva a la seguidora a subir p_2, la líder percibe una demanda menos elástica que en Bertrand y cobra más.

Equilibrio: en el ENPS, los dos precios superan el de Bertrand simultáneo, p^*, y la seguidora cobra algo menos que la líder: $\hat p_1>\hat p_2>p^*$. Los beneficios cumplen $\hat\pi_2>\hat\pi_1>\pi^*$, es decir, la líder gana más que en Bertrand, porque podría fijar p^*, pero la seguidora gana aún más (**ventaja de mover segundo**).

Desarrollo gráfico: en el plano (p_1,p_2), las funciones de reacción de Bertrand son crecientes y se cortan en el equilibrio simultáneo, (p^*,p^*). La de la seguidora, de pendiente $\gamma/2$, es la restricción de la líder, cuyas isobeneficio tienen forma de U y dan más beneficio cuanto más altas. La líder elige sobre ella el punto de tangencia con su isobeneficio más alta.



Precios secuenciales: la líder elige sobre la reacción creciente de la seguidora el punto de mayor beneficio.

Estática comparativa: las subidas de precio crecen con la sustituibilidad, y con ellas la ventaja de la seguidora. Si $\gamma\to 0$, cada empresa fija su precio de monopolio y el orden deja de importar.

Bienestar: con esta demanda lineal, los consumidores pierden más de lo que ganan las empresas y el bienestar total cae, de modo que, con precios, el liderazgo es anticompetitivo.

Conclusiones: muestra que la pendiente de las funciones de reacción decide quién gana con el orden de juego y da un fundamento no colusivo al liderazgo en precios. Pero los papeles son exógenos y el compromiso es poco creíble con precios, que se revisan con facilidad.

En suma, la variable estratégica determina el grado de competencia: con

cantidades, el equilibrio está entre el monopolio y la competencia, mientras que con precios bastan dos empresas para el resultado competitivo. La capacidad, la diferenciación, la asimetría de costes y la repetición devuelven poder de mercado a las empresas.

2. Colusión: soluciones cooperativas

Los equilibrios no cooperativos son equilibrios de Nash, pero no óptimos de Pareto para las empresas, que ganarían coordinándose. El problema es que esa cooperación no es estable. El marco dinámico añade la repetición, la respuesta a la historia, la credibilidad, con el ENPS como equilibrio relevante, y la inducción hacia atrás.

2.1. Alianzas no sostenibles

Clasificación: los acuerdos pueden ser *vinculantes*, si un tercero obliga a cumplirlos, objeto de los juegos cooperativos, o *no vinculantes*, sostenidos solo por los incentivos de las partes; y *explícitos*, con comunicación y acuerdo formal, como el cártel, o *tácitos*, nacidos de la mera repetición. Los acuerdos colusivos ilegales son siempre no vinculantes.

2.1.1. Cártel: cuasimonopolio multiplanta

Planteamiento: el cártel maximiza el beneficio conjunto como un monopolista multiplanta, y lo que decide es el reparto de la producción:

$$\max_{\{x_1,\dots,x_n\}} \pi = p(X)\cdot X - \sum_{i=1}^{n} \mathrm{CT}_i(x_i)$$

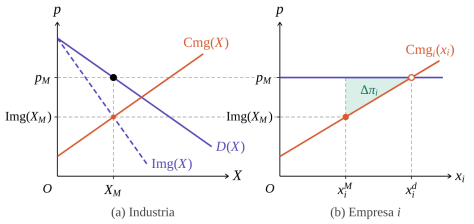
CPO: el ingreso marginal, que solo depende del total, iguala el coste marginal de cada planta:

$$\mathrm{Img}(X) = \mathrm{Cmg}_i(x_i) = \mathrm{Cmg}_j(x_j) \quad \forall\, i,j$$

Con costes marginales constantes y sin límite de capacidad, la planta más barata lo produce todo. Con costes crecientes, se igualan los costes marginales entre plantas. El beneficio conjunto es el máximo posible, con pérdida irrecuperable de eficiencia.

Equilibrio: el cártel no es un equilibrio de Nash. Si una empresa toma el precio del cártel, p_M, como dado, le conviene producir hasta que su coste marginal lo iguale y obtener beneficios extraordinarios. Incumplir es una **estrategia dominante**, como en el dilema del prisionero.

Desarrollo gráfico: el cártel produce X_M, donde el ingreso marginal corta al coste marginal agregado, suma horizontal de los de las empresas, y vende a p_M. Cada empresa produce su cuota, donde su coste marginal iguala $\mathrm{Img}(X_M)$, y el área $\Delta\pi_i$ mide lo que gana desviándose.



Cártel: reparto de la producción y beneficio adicional de desviarse, $\Delta\pi_i$.

Conclusiones: fija el beneficio máximo de la industria, la referencia de toda colusión, y muestra el incentivo a desviarse y la razón de ser de las leyes *antitrust*. Pero en un marco estático el cártel no es sostenible.

2.1.2. Horizonte finito

Razonamiento: con información completa y un último periodo T conocido, el juego se resuelve hacia atrás. En T no hay futuro con el que castigar, y no cooperar es la estrategia dominante. Anticipándolo, el mismo argumento vale para $T-1$ y, retrocediendo, para cada periodo hasta el primero, de modo que la certeza del final deshace la cooperación desde el principio.

Equilibrio: el ENPS consiste en T repeticiones del equilibrio de Nash estático, siempre que este sea único, como en Bertrand o en el dilema del prisionero. Con varios equilibrios en el juego de etapa puede sostenerse cierta cooperación al principio.

Conclusiones: la repetición no basta por sí sola. Pero el resultado depende de que el final se conozca con certeza y de que la racionalidad sea de conocimiento común: basta una pequeña probabilidad de que el rival sea cooperativo para sostener la cooperación durante buena parte del juego.

2.2. Alianzas sostenibles (Friedman, 1971)

Idea: si el juego se repite un número indeterminado de periodos, las ganancias a largo plazo de cumplir pueden superar las ganancias a corto plazo de incumplir.

Supuestos: dos jugadores racionales con información completa y horizonte infinito sostienen un acuerdo tácito con una **estrategia del gatillo**: cada empresa coopera mientras ambas hayan cooperado en todos los periodos anteriores, y una sola desviación rompe la cooperación para siempre. Con historia h^{t-1} y acciones $s_i^\tau \in \{c, nc\}$, cooperar o no cooperar:

$$s_i^t(h^{t-1}) = c \;\Leftrightarrow\; (s_1^\tau, s_2^\tau) = (c,c) \quad \forall \tau < t$$

Condición de cooperación: el beneficio descontado de cooperar debe ser al menos el de desviarse una vez y sufrir después el castigo, $\pi_{c,c}/(1-\beta) \geq \pi_{nc,c} + \beta\,\pi_{nc,nc}/(1-\beta)$. Operando, el factor de descuento ha de superar un umbral:

$$\beta \geq \frac{\pi_{nc,c} - \pi_{c,c}}{\pi_{nc,c} - \pi_{nc,nc}}$$

En un duopolio de Bertrand, quien se desvía captura todo el beneficio de monopolio y el castigo lo anula, de modo que basta $\beta \geq 1/2$.

Equilibrio: la cooperación es un ENPS si se cumple la condición. Requiere estrategias del gatillo creíbles, supervisión adecuada, percepción de juego a largo plazo y agentes que valoren el futuro.

Estática comparativa: con n empresas que compiten en precios y se reparten el beneficio de monopolio, $\pi_{c,c} = \pi_M/n$, $\pi_{nc,c} = \pi_M$ y $\pi_{nc,nc} = 0$, de modo que la condición es $\beta \geq 1 - 1/n$. La colusión es más difícil cuantas más empresas hay, porque la ganancia de desviarse crece y la parte de cada una cae. La dificultan también la impaciencia y una interacción poco frecuente.

Extensiones: para Stigler (1964), la colusión exige detectar las desviaciones, y detectarlas es más difícil con más empresas, más heterogéneas y transacciones menos transparentes. El modelo aplica el **teorema popular**: con un factor de descuento próximo a uno, cualquier asignación factible e individualmente racional puede sostenerse. Dos extensiones explican las guerras de precios:

- *Información imperfecta* (Green y Porter, 1984): si solo se observa el precio, un precio bajo puede deberse a una desviación o a un *shock*, y se castiga durante un periodo finito. Las guerras de precios ocurren en equilibrio en las fases de demanda baja.
- *Ciclos de demanda* (Rotemberg y Saloner, 1986): la tentación de desviarse es mayor en los auges, de modo que el precio colusivo debe bajar en las expansiones y las guerras de precios son procíclicas.

Conclusiones: la colusión descansa en cómo se castiga a quien se desvía, y los programas de clemencia explotan la fragilidad del acuerdo. El modelo no dice cómo se acuerdan las empresas, y la reversión permanente no es, en general, el castigo creíble más severo. Abreu (1986) propone la "zanahoria y el palo", un castigo breve y más duro.

En suma, la colusión solo se sostiene con horizonte indefinido y un factor de descuento suficiente, umbral que crece con el número de empresas. Todos estos modelos toman ese número como dado.

3. Entrada y disuasión

Con un número de empresas endógeno, la pregunta es por qué los beneficios extraordinarios no atraen entrada hasta agotarlos. La respuesta está en las barreras a la entrada, en la conducta de las instaladas y en la amenaza de entrada.

3.1. Barreras a la entrada

Definición: para Bain (1956), son las condiciones que permiten a las instaladas mantener precios supracompetitivos sin inducir la entrada, como las economías de escala, las ventajas absolutas de costes o la diferenciación. Para Stigler (1968), son un coste que soporta el entrante y no soportaron las instaladas. Para Baumol y Willig (1981), la barrera la crean los costes hundidos, no los fijos.

Causas:

1. *Naturales o estructurales*: una escala mínima eficiente grande respecto al mercado. En el caso polar, esa escala da lugar a un monopolio natural, con costes subaditivos.
2. *Estratégicas*: las conductas de las instaladas para impedir la entrada, como el precio límite, la publicidad agresiva o la inversión en capacidad.
3. *Legales*: patentes, licencias, concesiones y derechos exclusivos.

3.2. Disuasión estratégica de la entrada

3.2.1. Precio límite (Bain, Sylos-Labini, Modigliani)

Idea: la instalada sostiene un precio tan bajo que el entrante, anticipando el precio que regiría tras entrar, no cubriría sus costes medios.

Supuestos: el entrante tiene un coste medio $\mathrm{Cme}_E(x_E)$ decreciente hasta la escala mínima eficiente. Por el **postulado de Sylos**, espera que la instalada mantenga su producción x_I tras la entrada, de modo que afronta la demanda residual $p(x_I + x_E)$.

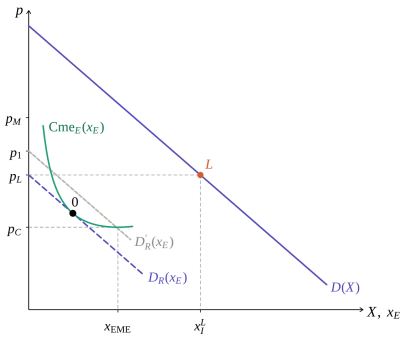
Equilibrio: la instalada elige la menor producción, y el mayor precio, que deja la demanda residual del entrante por debajo de su coste medio:

$$p(x_I^L + x_E) \leq \mathrm{Cme}_E(x_E) \quad \forall\, x_E > 0$$

Gráficamente, en el precio límite la demanda residual es tangente por debajo al coste medio del entrante, y $p_C < p_L < p_M$: la instalada sacri-

fica beneficio presente para conservar el mercado. En la formulación de Modigliani, el entrante solo puede entrar a la escala mínima eficiente o por encima, y la condición se cumple con igualdad en esa escala, de modo que x_I^L lo deja exactamente sin beneficio:

$$p(x_I^L + x_{\text{EME}}) = \text{Cme}_E(x_{\text{EME}})$$



Precio límite: la demanda residual del entrante es tangente a su coste medio.

Estática comparativa: según la aproximación de Modigliani, el margen sobre el precio competitivo crece con la escala mínima eficiente relativa al mercado y decrece con la elasticidad de la demanda:

$$\frac{p_L - p_C}{p_C} \approx \frac{x_{\text{EME}}}{X_C\,|\varepsilon|}$$

Extensiones: el postulado de Sylos no es creíble, porque tras la entrada a la instalada le conviene acomodarla. En la versión de **señalización** de Milgrom y Roberts (1982a), un precio bajo revela costes bajos, aunque en equilibrio no reduce necesariamente la entrada.

Conclusiones: es la primera explicación de precios persistentes entre el competitivo y el de monopolio, ligados a la escala mínima eficiente. Pero la amenaza no es creíble en su versión original y el modelo es estático.

3.2.2. Capacidad (Spence, 1977; Dixit, 1980)

Idea: la credibilidad la aporta una inversión irreversible. Si la instalada hunde antes una capacidad elevada, producir mucho tras la entrada pasa a ser su mejor respuesta y no una amenaza vacía.

Supuestos: en la primera etapa, la instalada elige una capacidad k con coste r por unidad. En la segunda, el entrante decide si entra pagando un coste fijo F, y ambos compiten a la Cournot con demanda $p = a - X$. La instalada produce hasta k con coste marginal c y por encima con $c + r$; el entrante, siempre con $c + r$.

Planteamiento: por inducción hacia atrás, en la segunda etapa, dado k y supuesta la entrada, cada empresa maximiza su beneficio dada la cantidad de la otra, con el coste rk ya hundido:

$$\max_{x_E} \pi_E = (a - x_I - x_E - c - r)\,x_E - F$$

$$\max_{x_I} \pi_I = (a - x_I - x_E - c)\,x_I - r \cdot \max\{x_I - k, 0\}$$

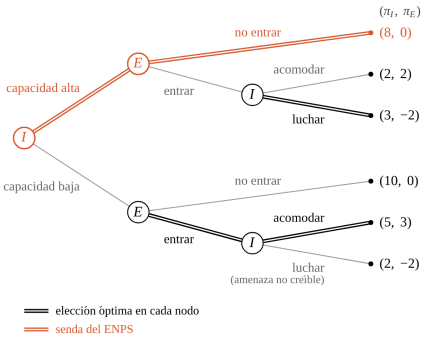
CPO: la función de reacción del entrante es $x_E = (a - c - r - x_I)/2$. La de la instalada tiene un *quiebro* en k, porque su coste marginal salta de c a $c + r$, de modo que, al elegir k, la instalada elige en qué punto de la función de reacción del entrante se sitúa el equilibrio. La función de reacción de la instalada es:

$$x_I = \begin{cases} (a - c - x_E)/2 & \text{si } (a - c - x_E)/2 < k \\ (a - c - r - x_E)/2 & \text{si } (a - c - r - x_E)/2 > k \\ k & \text{en otro caso} \end{cases}$$

Equilibrio: si la instalada usa toda su capacidad, el beneficio del entrante, $\pi_E(k) = [(a - c - r - k)/2]^2 - F$, decrece en k, y la capacidad que lo anula es $k_D = a - c - r - 2\sqrt{F}$. Frente a la capacidad de monopolio, k_M, hay tres regímenes:

- *Entrada bloqueada*, $k_D \le k_M$: la instalada actúa como monopolista sin hacer nada.
- *Entrada disuadida*, $k_D > k_M$ y disuadir es rentable: la instalada instala y utiliza k_D, por encima de su producción de monopolio.
- *Entrada acomodada*: si F es pequeño, disuadir es demasiado caro y la instalada elige la capacidad de líder de Stackelberg.

Desarrollo gráfico: en la forma extensiva, la instalada elige capacidad alta o baja, el entrante decide si entra y la instalada acomoda o lucha. Con capacidad baja, luchar no es creíble y el entrante entra, mientras que con capacidad alta, ya pagada, luchar es la mejor respuesta y el entrante no entra. La senda del ENPS, capacidad alta y no entrada, es la entrada disuadida.



Juego de entrada con compromiso de capacidad: la capacidad alta hace creíble la lucha.

Estática comparativa: cuanto menor es F, mayor es k_D y más caro disuadir. Un aumento de r acerca el mercado a la disuasión o al bloqueo, porque también el entrante paga la capacidad, y un mercado mayor lleva del bloqueo a la disuasión y, después, a la acomodación. Con $A \equiv a - c - r$ y $k_M = A/2$, en la versión simplificada, que ignora las restricciones de credibilidad sobre la capacidad, la entrada es:

$$\begin{cases} \text{bloqueada} & \text{si } A \le 4\sqrt{F} \\ \text{disuadida} & \text{si } 4\sqrt{F} < A \le (8 + 4\sqrt{2})\sqrt{F} \approx 13{,}7\sqrt{F} \\ \text{acomodada} & \text{si } A > (8 + 4\sqrt{2})\sqrt{F} \end{cases}$$

Conclusiones: resuelve la credibilidad con los costes hundidos, sin capacidad ociosa en equilibrio (Dixit), y hace del número de empresas el resultado de decisiones de inversión. Exige una capacidad observable e irreversible y competencia en cantidades, y con varias instaladas la disuasión es un bien público con problema de *free rider*.

3.2.3. Predación (Kreps y Wilson; Milgrom y Roberts)

Idea: la instalada puede expulsar a un rival ya presente fijando temporalmente precios por debajo del coste y recuperar después las pérdidas.

Con información completa, la predación no es creíble, pero con información incompleta es racional, como **reputación** de agresividad que disuade entradas futuras (Kreps y Wilson, 1982; Milgrom y Roberts, 1982b), o como predación financiera, que explota la dependencia del rival de la financiación externa. La jurisprudencia toma como referencia el coste variable medio, aunque la predación puede ser racional por encima de él.

Conclusiones: da fundamento a su persecución como abuso de posición dominante, pero no ofrece un criterio operativo para separarla de la competencia agresiva.

3.2.4. Costes hundidos endógenos (Sutton, 1991)

Idea: la concentración puede no disolverse con la entrada si se compite en calidad, publicidad o I+D, atributos que se compran con costes hundidos. Si la calidad recae en costes fijos, solo un número finito de empresas sobrevive, sea cual sea el tamaño del mercado, resultado conocido como **propiedad de finitud** (Shaked y Sutton, 1983).

Sutton (1991) distingue los costes hundidos *exógenos*, como una planta, con los que la concentración cae al crecer el mercado, y los *endógenos*, como la publicidad o la I+D, con los que tiene una cota inferior que no tiende a cero, porque la escalada de gasto expulsa a quien no la sigue.

3.3. Mercados impugnables

Idea: Baumol, Panzar y Willig (1982) invierten el argumento. Sin costes hundidos ni barreras de salida, y si el entrante actúa antes de que la instalada reaccione, la amenaza de entrada relámpago, *hit-and-run*, obliga a fijar precios iguales al coste medio, con cualquier número de empresas.

El número de empresas no es, por sí solo, un indicador del poder de mercado, porque importan los costes hundidos, fundamento a la vez de la disuasión y de los oligopolios naturales. Sus supuestos son extremos, pero la teoría ha influido mucho en la desregulación.

En suma, el análisis de la entrada explica por qué persisten los beneficios y hace endógeno el número de empresas, pero la frontera entre barreras naturales y estratégicas es difusa y las predicciones dependen de la variable estratégica.

A20. Poder de mercado y regulación óptima. Definición del mercado relevante. Desarrollos en presencia de información asimétrica. Aplicaciones prácticas

1. Poder de mercado

El poder de mercado se mide con el índice de Lerner, aparece en varias estructuras y nace de las barreras a la entrada, que ordenan también las causas de la regulación.

1.1. Concepto y estructuras

Definición: hay **poder de mercado** cuando una o varias empresas pueden alterar el precio, lanzando cantidades inferiores a las competitivas a un precio superior. Se pierde eficiencia paretiana, porque quedan sin producir unidades cuyo coste es inferior a lo que los consumidores darían por ellas.

Índice de Lerner: la medida directa del poder de mercado es el margen relativo sobre el coste marginal, que en el monopolio iguala la inversa de la elasticidad y en Cournot es proporcional a la cuota:

$$I_L = \frac{p - \text{Cmg}}{p} = \frac{1}{|\varepsilon_{Xp}|} \in [0,1], \qquad I_{L,i} = \frac{s_i}{|\varepsilon_{Xp}|} \in [0,1]$$

El índice vale cero en competencia perfecta y crece con la capacidad de fijar el precio por encima del coste marginal. Pero el coste marginal suele ser inobservable, lo que motiva los índices indirectos y el enfoque estructural de la nueva organización industrial empírica. En los mercados de varias caras, el índice convencional puede ser negativo en el lado subvencionado, y el margen de cada lado se mide frente al coste neto de la externalidad que genera.

Estructuras: además del monopolio y del oligopolio, el poder de mercado aparece en la **competencia monopolística**, con margen positivo pero beneficio nulo a largo plazo.

Tipos de monopolio: el diseño de la regulación depende del tipo de monopolio:

- *Discriminador*: la discriminación, de primer, segundo o tercer grado, puede aumentar la cantidad y la eficiencia, pero transfiere excedente al productor. La regulación se centra en las que restringen la competencia, como los precios predatorios o los descuentos de fidelidad.
- *Natural*: si es sostenible, es el objeto central de la regulación de precios y, si no, se subasta el derecho a explotarlo (competencia por el mercado).
- *Monopsonio*: un único comprador, como un gran distribuidor frente a los productores agrarios o un gran empleador en un mercado de trabajo local, que compra menos y a menor precio. Se regula con precios o salarios mínimos.
- *Bilateral*: un único vendedor frente a un único comprador, como un sindicato y una patronal o un fabricante y un distribuidor exclusivo. La teoría no determina el precio, que depende del poder de negociación, pero con negociación eficiente sí la cantidad, la que maximiza el excedente conjunto. La regulación fija el marco de la negociación y arbitra.

1.2. Causas del poder de mercado

Origen: las **barreras a la entrada**, que impiden que los beneficios extraordinarios atraigan a nuevas empresas:

- *Naturales o estructurales*: la estrechez del mercado frente a la escala mínima eficiente, que da lugar al monopolio natural. Solo protege a la instalada si la entrada exige costes hundidos (Baumol y Willig, 1981).
- *Estratégicas*: las conductas de las instaladas para disuadir la entrada o expulsar a rivales, como el precio límite, la inversión en capacidad, la proliferación de marcas o la predación. Conductas que parecen restringir la competencia pueden responder, sin embargo, a razones de eficiencia, como las sinergias o el ahorro de costes de transacción.
- *Legales*: patentes, licencias, concesiones y derechos exclusivos.

Definición de barrera: Bain (1956) la define por sus resultados, como las condiciones que permiten a las instaladas fijar precios supracompetitivos sin inducir la entrada. Stigler (1968) la restringe a los costes que soporta el entrante y no soportaron las instaladas, y Baumol y Willig (1981) la sitúan en los costes hundidos, idea que anticipa la teoría de los mercados impugnables. La distinción entre barreras naturales y artificiales ordena, además, las causas de la regulación.

1.3. Causas de la regulación

Equidad: garantizar el acceso a servicios como la energía, el agua o las telecomunicaciones con independencia de la *disparidad geográfica* y de la *disparidad de renta*, mediante obligaciones de servicio universal y tarifas sociales, que se financian con subsidios cruzados o con el presupuesto. Sin ellas, un operador no serviría las zonas rurales o insulares, o lo haría a precios prohibitivos, y excluiría a quienes no pueden pagar.

Eficiencia: corregir la ineficiencia asignativa. El *monopolio no natural*, sostenido por barreras artificiales, se combate eliminándolas con la liberalización y la defensa de la competencia, mientras que el *monopolio natural*, sostenido por la tecnología, no conviene desmantelarlo, sino regularlo.

Monopolio natural: existe cuando la función de costes es estrictamente **subaditiva**, de modo que ninguna combinación de empresas produce un vector dado a menor coste que una sola: $\text{CT}(\mathbf{x}) < \sum_j \text{CT}(\mathbf{x}^j)$ para toda descomposición $\sum_j \mathbf{x}^j = \mathbf{x}$ en al menos dos vectores no nulos. La escala mínima eficiente es grande en relación con la demanda. Desmantelarlo sacrificaría eficiencia productiva, y no regularlo, eficiencia asignativa. La empresa suele operar con costes medios decrecientes, Cmg < Cme, en el tramo relevante de la demanda, aunque la subaditividad puede mantenerse algo por encima de la escala mínima eficiente, donde Cme = Cmg.

Disyuntiva: cuanto más se obliga al monopolista a fijar $p = \text{Cmg}$, mayores son sus pérdidas, porque el coste marginal suele ser inferior al medio en el tramo relevante. Hay que elegir entre eficiencia asignativa y viabilidad financiera. Las modalidades de regulación se ordenan según esa disyuntiva y según la información de que disponga el regulador.

En suma, el poder de mercado se mide por el margen sobre el coste marginal y nace de las barreras a la entrada. Las artificiales se eliminan y el monopolio natural se regula, con la disyuntiva entre eficiencia asignativa y viabilidad financiera.

2. Regulación óptima

La regulación óptima se estudia con información perfecta, con información asimétrica, con varias empresas y en los mercados de varias caras.

2.1. Regulación con información perfecta

Si el regulador conoce los costes y la demanda, la regulación se ordena en tres niveles según la prioridad entre eficiencia asignativa y viabilidad financiera: el *first best* financia las pérdidas con transferencias, el *second best* busca el precio eficiente con beneficio nulo y el *third best* solo acota los beneficios. Este último lo hace con un tope a la rentabilidad o a la senda de precios, sin fijar directamente el precio.

2.1.1. *First best*: precio igual al coste marginal

Planteamiento: el regulador maximiza la suma del excedente del consumidor y el beneficio, sin restricción financiera:

$$\max_p \; W = \int_0^{x(p)} p(q)\,dq - \text{CT}(x)$$

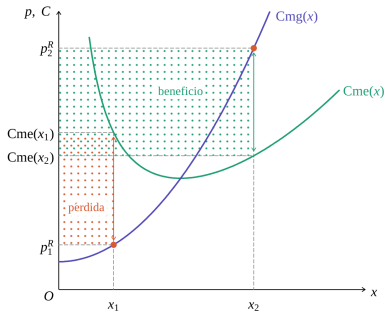
CPO: $p = \text{Cmg}$, la condición de eficiencia asignativa.

CSO: el bienestar es cóncavo si la demanda inversa decrece más deprisa que el coste marginal, $p'(x) < \text{Cmg}'(x)$, lo que se cumple también con el coste marginal decreciente habitual del monopolio natural si decrece más despacio que la demanda.

Solución: si el coste marginal es inferior al medio, como suele ocurrir en el monopolio natural, la empresa pierde $x \cdot (\text{Cme} - \text{Cmg})$, que se cubre con una subvención de suma fija. Si lo supera, un impuesto de suma fija retira el beneficio.

Desarrollo gráfico: con un coste medio en U, el coste marginal marca el precio regulado. En el tramo decreciente del coste medio, $p_1^R = \text{Cmg}(x_1) < \text{Cme}(x_1)$ y la pérdida la cubre la subvención. En el creciente, $p_2^R > \text{Cme}(x_2)$ y el beneficio se detrae con un impuesto. La pérdida y el beneficio son las áreas:

$$x_1 \cdot [\text{Cme}(x_1) - p_1^R] \qquad \text{y} \qquad x_2 \cdot [p_2^R - \text{Cme}(x_2)]$$



Tarificación al coste marginal: pérdida en el tramo decreciente del coste medio y beneficio en el creciente.

Conclusiones: es la única solución plenamente eficiente, pero exige información perfecta, la subvención exige impuestos distorsionantes y la empresa que sabe cubiertas sus pérdidas no tiene incentivos a reducir costes, un problema de riesgo moral.

2.1.2. *Second best*: beneficio nulo

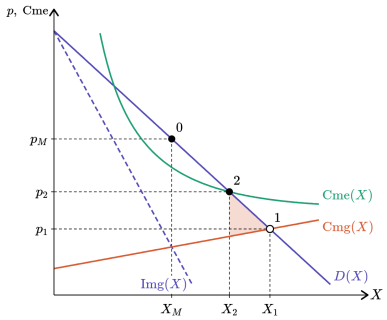
Planteamiento: en el monopolio monoproducto se añade la restricción de beneficio no negativo:

$$\max_p W \quad \text{s.a.:} \quad p \cdot x(p) - \text{CT}(x) \geq 0$$

CPO: la restricción se satura y $p = \text{Cme}$.

CSO: el bienestar decrece con el precio por encima del coste marginal, luego el máximo es el menor precio compatible con beneficio nulo.

Solución: el precio se fija donde la demanda corta al coste medio, con una cantidad inferior a la eficiente y un precio superior al coste marginal. La pérdida de eficiencia asignativa, el triángulo entre la demanda y el coste marginal, es el precio de la viabilidad financiera.



Regulación *second best* con beneficio nulo.

Extensiones: dos mecanismos alcanzan la asignación del *first best* sin transferencias, por lo que son también soluciones de primer óptimo:

- *Tarifa en dos partes*: $T(x) = F + p \cdot x$, con $p = \text{Cmg}$ y una cuota fija que reparte la pérdida entre los consumidores. El consumo marginal es eficiente y la empresa cubre costes, pero la cuota excluye a quien tiene poca disposición a pagar. Con n consumidores, la cuota es $F = x \cdot (\text{Cme} - \text{Cmg})/n$, que con coste marginal constante se reduce a los costes fijos, $F = \text{CF}/n$.
- *Tarifas pico-valle*: con demanda cíclica, bien no almacenable y capacidad ociosa en los valles, $p_{\text{valle}} = \text{Cmg}$ y $p_{\text{pico}} = \text{Cmg} + \text{coste de capacidad}$. Los consumidores de punta pagan la capacidad y los ingresos cubren los costes sin transferencias. Con costes unitarios de operación y de capacidad constantes, cada periodo paga su coste marginal, de corto plazo en el valle y de largo plazo en la punta.

Conclusiones: es la solución básica y la más sencilla, pero puede no existir si la demanda no corta al coste medio, mantiene un precio superior al competitivo y, como cubre siempre el coste, no incentiva su reducción (la regulación *cost-plus*).

2.1.3. *Second best* multiproducto: Ramsey-Boiteux

Planteamiento: con varios productos de demandas independientes, el regulador maximiza la suma de excedentes y beneficio sujeto a que la empresa cubra sus costes comunes:

$$\max_{p_1,\dots,p_n} \sum_i \text{EC}_i(p_i) + \pi \quad \text{s.a.:} \quad \sum_i p_i x_i - \text{CT}(x_1,\dots,x_n) \geq 0$$

CPO: cada producto soporta un margen inversamente proporcional a la elasticidad de su demanda, la **regla de la elasticidad inversa**:

$$\frac{p_i - \text{Cmg}_i}{p_i} = \frac{k}{|\varepsilon_i|}$$

El **número de Ramsey**, $k = \lambda/(1+\lambda) \in (0,1)$, depende del multiplicador de la restricción financiera, λ, que mide el coste en bienestar de exigir un euro más de beneficio y equivale a la ganancia de bienestar de relajar esa restricción en un euro. Ramsey (1927) formuló la regla para la imposición óptima, y Boiteux (1956), para la tarificación de los monopolios públicos. Con demandas interdependientes, las elasticidades propias se sustituyen por superelasticidades que recogen los efectos cruzados.

CSO: con demandas independientes y costes marginales constantes, el excedente total de cada mercado es cóncavo en su precio y el beneficio, cuasicóncavo en el tramo relevante, de modo que se cumplen las condiciones de suficiencia de Arrow-Enthoven.

Solución: los servicios de demanda más inelástica cubren una parte mayor de los costes comunes, lo que minimiza la pérdida de eficiencia. No hay subsidio cruzado mientras cada servicio cubra al menos su coste incremental. Si el monopolio no es sostenible frente a la entrada, ninguna regla de precios lo sostiene, y la solución es licitar el derecho a explotarlo (Demsetz, 1968). Que sea sostenible depende de la tecnología y de la demanda, porque la subaditividad lo justifica por la eficiencia productiva, pero no le asegura sobrevivir sin protección.

Conclusiones: lleva a la tarificación la lógica de la imposición óptima y es la referencia de las tarifas de las industrias de red. Pero exige conocer costes y elasticidades de todos los productos y puede ser regresiva, porque los bienes inelásticos pesan más en las cestas de renta baja.

2.1.4. *Third best*: tasa de beneficio

Planteamiento: el regulador fija una tasa máxima de rentabilidad sobre el capital, s, superior a su coste, r:

$$\max_{L,K} p(x) \cdot x - wL - rK \quad \text{s.a.:} \quad \frac{p(x) \cdot x - wL}{K} \leq s$$

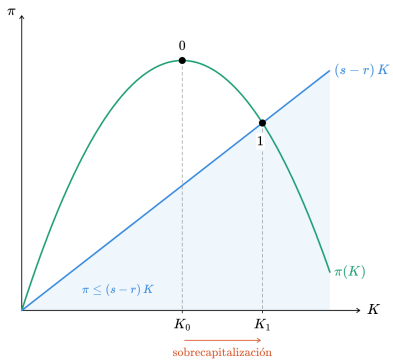
CPO: la restricción se satura y la relación marginal de sustitución técnica no iguala w/r, porque el coste efectivo del capital para la empresa es inferior a r.

CSO: el conjunto factible no es convexo, y Averch y Johnson suponen que la solución saturada es un máximo.

Solución: como cada unidad de capital amplía el beneficio permitido en $s - r > 0$, la empresa se sobrecapitaliza, con una ratio capital-trabajo superior a la de coste mínimo (el **efecto Averch-Johnson**, 1962).

Desarrollo gráfico: en el plano capital-beneficio, sin regulación la empresa iría al máximo de $\pi(K)$, con K_0. Regulada, su beneficio no puede superar la semirrecta $(s-r)\,K$, y se queda en el corte de ambas, con más capital, K_1. La distancia entre K_1 y K_0 es la sobrecapitalización. El beneficio regulado ha de cumplir:

$$\pi(K) \leq (s-r)\,K$$



Efecto Averch-Johnson: la regulación de la tasa de beneficio lleva a sobrecapitalizar.

Conclusiones: exige poca información, porque basta fijar la tasa y comprobar el capital y los costes declarados. Pero induce el efecto Averch-Johnson, la tasa se elige sin criterio objetivo y los costes se trasladan al consumidor, sin estímulo a reducirlos.

2.1.5. *Third best*: límites de precios (Littlechild, 1983)

Idea: Littlechild (1983), en su informe para la privatización de British Telecom, propone el *price cap*: una senda máxima de precios para el periodo regulatorio, normalmente de cinco años, desvinculada de los costes propios de la empresa, indexada a la inflación y descontando las ganancias de productividad esperadas:

$$p_t \leq p_{t-1} \cdot (1 + \text{IPC} - \text{X})$$

En la senda, p_t es el precio regulado o el índice de precios de la cesta regulada, y X, el factor de productividad esperada de la industria.

Razonamiento: como el precio no depende de los costes propios, cada ahorro eleva el beneficio hasta la siguiente revisión, y la eficiencia productiva pasa a ser la principal vía de aumentar el beneficio. La competencia por comparación de Shleifer (1985) extiende esta lógica al caso de varias empresas.

Conclusiones: da potentes incentivos a la eficiencia y evita la sobrecapitalización. Pero es difícil acertar con X, el estímulo decae cerca de la revisión y la empresa puede ahorrar empeorando la calidad.

2.2. Información asimétrica

Cuando la empresa sabe más que el regulador sobre sus costes, la regulación ha de inducirla a revelarlos y a reducirlos.

2.2.1. Loeb y Magat (1979)

Idea: el regulador paga una subvención igual al excedente del consumidor, $z = \text{EC}(p)$, y deja elegir el precio, de modo que la empresa resuelve $\max_p p \cdot x(p) - \text{CT}(x) + \text{EC}(p)$. Al internalizar el excedente total, elige $p = \text{Cmg}$ sin que el regulador conozca los costes.

Conclusiones: alcanza el *first best* aun con información asimétrica. Pero, a diferencia de la tarifa en dos partes y de las tarifas pico-valle, no

renuncia a las transferencias, y su coste fiscal y la redistribución hacia el monopolista lo hacen políticamente inviable. Sus autores proponen combinarlo con una subasta previa de la franquicia que devuelva esas rentas al sector público.

2.2.2. Laffont y Tirole (1986)

Idea: la nueva teoría de la regulación aplica la teoría de los incentivos al diseño regulatorio, con Laffont y Tirole (1986) como referencia y su tratado (Laffont y Tirole, 1993) como síntesis. El regulador afronta una disyuntiva entre **eficiencia**, incentivar el esfuerzo de reducción de costes, y **extracción de rentas**, minimizar las transferencias a la empresa, y la resuelve con un menú de contratos entre los que la empresa se autoselecciona.

Razonamiento: un monopolio natural conoce su eficiencia intrínseca, φ, el "tipo" de la empresa, y elige un esfuerzo costoso. El regulador solo observa el coste realizado, y el problema combina selección adversa y riesgo moral. Una empresa eficiente puede hacerse pasar por ineficiente, imitar su coste con menos esfuerzo y quedarse la diferencia, la **renta informacional**. Como cada euro transferido cuesta a la sociedad $1+\lambda$, el regulador quiere incentivar el esfuerzo cediendo la menor renta posible. El mecanismo óptimo es un **menú de contratos** lineales en el coste, $t(\hat{\varphi}, C) = a(\hat{\varphi}) - b(\hat{\varphi}) \cdot [C - C^*(\hat{\varphi})]$, con $\hat{\varphi}$ el tipo anunciado, cuya potencia $b(\hat{\varphi}) \in [0,1]$, la fracción del sobrecoste o del ahorro que recae sobre la empresa, vale uno en el de precio fijo y cero en el de coste más margen, y decrece con φ:

1. *Empresa eficiente*: elige el contrato de precio fijo, asume todo el riesgo de costes y retiene todo el ahorro, con el esfuerzo del *first best*, sin distorsión "en la cima", pero con la mayor renta informacional.
2. *Empresa ineficiente*: elige el contrato de menor potencia, el más próximo al *cost-plus*, con el esfuerzo distorsionado a la baja y renta nula. El esfuerzo no se anula, y el contrato de coste más margen puro, $b = 0$, es solo el caso límite al que se llegaría si la distorsión lo llevase a cero.
3. *Tipos intermedios*: contratos con reparto parcial del riesgo, rentas decrecientes con φ y esfuerzo también distorsionado a la baja.

Para reducir la renta de los tipos eficientes, el regulador acepta menos esfuerzo de los ineficientes. La regulación por incentivos sustituye así el control de costes por la elección entre contratos, y el menú materializa la disyuntiva entre eficiencia y extracción de rentas.

Conclusiones: la regulación óptima con información asimétrica es un menú de contratos que equilibra incentivos y rentas, y la asimetría causa necesariamente una pérdida respecto al *first best*, porque las rentas son el precio de los incentivos. Los esquemas reales, del *cost-plus* al *price cap*, son puntos del menú, y el modelo ha sido muy influyente en la regulación de las telecomunicaciones, la energía y el transporte. Pero en entornos dinámicos aparece el efecto trinquete (apartado 2.2.3), y el propio Tirole, premio Nobel en 2014 en buena parte por esta obra, insiste en que no hay recetas universales. Cada mercado exige un análisis específico, y ninguno se regula por sí solo.

2.2.3. Efecto trinquete

Si la empresa anticipa que sus ahorros serán confiscados en la próxima revisión, no tiene incentivos a reducir costes. El remedio es un compromiso creíble del regulador, con periodos regulatorios largos y reglas de revisión fijadas de antemano. El efecto trinquete es un límite dinámico del modelo de Laffont y Tirole, porque el regulador aprende el tipo de la empresa y le retira la renta.

2.3. Regulación con varias empresas

Con más de una empresa, real o potencial, el regulador puede apoyarse en la competencia para obtener la información que le falta o para disciplinar al monopolista. Puede hacerlo mediante tres mecanismos, la competencia por el mercado, la competencia por comparación y la competencia potencial.

2.3.1. Competencia por el mercado (Demsetz, 1968)

Idea: cuando la subaditividad impide la competencia en el mercado, el derecho exclusivo a servirlo se subasta periódicamente y se adjudica a quien se comprometa a cobrar menos. Con licitación competitiva, las ofertas convergen al coste medio y se replica la regulación de beneficio nulo sin regulador permanente, es decir, la competencia *ex ante* sustituye a la imposible *ex post*.

Razonamiento: cada licitador sabe que cualquier oferta por encima del coste medio será batida por otro, de modo que las rentas del monopolio se disipan en la puja. El regulador pasa así de fijar precios a organizar subastas.

Conclusiones: ahorra información y funda las concesiones de servicios públicos. Pero, como señaló Williamson (1976), los contratos son incompletos y exigen renegociación, los activos específicos dificultan la reversión y el incumbente acumula ventajas en las relicitaciones.

2.3.2. Competencia por comparación (Shleifer, 1985)

Idea: el regulador fija el precio de cada empresa según los costes de empresas comparables de otros mercados, no según los propios, de modo que cada una retiene íntegramente sus ahorros y el conjunto converge a la eficiencia.

Razonamiento: fijar el precio según los costes que declara la propia empresa destruye sus incentivos a reducirlos. La comparación con empresas de otros mercados, como las distribuidoras regionales, reproduce la disciplina de la competencia sin que compitan en el mismo mercado.

Conclusiones: es la técnica habitual con las distribuidoras de electricidad, gas y agua. Exige comparadores genuinamente similares y que no se coludan para inflar sus costes. Las diferencias entre comparadores pueden corregirse econométricamente, y la técnica es una de las aplicaciones más directas de la teoría de la agencia con varios agentes.

2.3.3. Mercados impugnables (Baumol, Panzar y Willig)

Idea: Baumol, Panzar y Willig (1982) generalizan la competencia perfecta a mercados con pocas empresas y retoman la idea de Demsetz de que la amenaza de competencia disciplina al monopolio, aunque por la competencia potencial y no por la licitación. Llaman **impugnable** a un mercado en el que los entrantes compiten sin desventajas. Exige tres condiciones: el mismo acceso a la tecnología, costes fijos no hundidos y un tiempo de entrada inferior al de reacción del incumbente.

Razonamiento: así son posibles las estrategias *hit and run*, y la mera amenaza de entrada obliga al monopolista a fijar $p = \text{Cme}$. Un entrante puede aprovechar cualquier precio por encima del coste medio y retirarse sin pérdida antes de que el incumbente reaccione, de modo que el beneficio nulo se impone aunque el mercado lo ocupe una sola empresa.

Conclusiones: para la liberalización, lo relevante no es el número de empresas, sino las barreras de entrada y salida. Pero el *hit and run* puede no ser creíble y los costes hundidos son frecuentes y difíciles de medir.

2.4. Mercados de varias caras (Rochet y Tirole)

Idea: en los mercados de dos caras, formalizados por Rochet y Tirole (2003) y por Armstrong (2006), una plataforma conecta a dos grupos entre los que hay **externalidades cruzadas de red**, como las tarjetas de pago o los *marketplaces*. La estructura de precios no es neutral, es decir, importa su reparto entre los lados, no solo su nivel. Rochet y Tirole parten de las tarjetas de pago, con precios por transacción, y Armstrong modeliza cuotas de participación, el marco que se sigue aquí.

Razonamiento: una plataforma monopolista cobra a cada grupo una cuota p_i y soporta un coste f_i por miembro. Cada miembro obtiene $u_i = \alpha_i n_j - p_i$, donde α_i es la externalidad cruzada, y la participación crece con la utilidad, $n_i = N_i(u_i)$. Rebajar el precio de un grupo atrae miembros que elevan lo que se puede cobrar al otro, y el precio de cada lado es su **coste neto de la externalidad** más un margen de monopolio:

$$p_i = f_i - \alpha_j n_j + \frac{n_i}{N_i'(u_i)}$$

El último término es el margen de monopolio, menor cuanto más responde la participación del grupo a su utilidad neta. Con η_i, la elasticidad de la participación respecto del precio, el índice de Lerner gana un **término de dos caras**:

$$\frac{p_i - f_i}{p_i} = \frac{1}{\eta_i} - \frac{\alpha_j n_j}{p_i}$$

La plataforma aplica a cada lado la regla de la elasticidad inversa, pero frente al coste neto de la externalidad y no frente a f_i.

Un grupo paga menos cuanto más elástica es su participación y cuanto más valora su presencia el otro grupo. Si $\alpha_j n_j$ supera el margen de monopolio, su precio queda por debajo del coste, y puede ser nulo o negativo si $\alpha_j n_j$ iguala o supera f_i más ese margen. El índice de Lerner convencional, frente a f_i, es bajo o negativo en el lado subvencionado y alto en el que paga, y ninguno mide por separado el poder de la plataforma, que se aprecia en el índice ajustado, frente al coste neto, y en la estructura de precios.

Extensiones: un precio por debajo del coste en un lado no prueba la predación, y el mercado relevante debe considerar ambos lados. Si un grupo se afilia a una sola plataforma y el otro a varias, *multihoming*, cada plataforma compite por sus usuarios exclusivos y cobra caro al otro grupo (**cuellos de botella competitivos**, Armstrong, 2006).

Conclusiones: explica precios nulos sin predación y da el índice de Lerner ajustado de cada lado. Pero supone un monopolio con cuotas fijas, y con varias plataformas o tarifas por transacción ofrece pocas reglas generales. Su aplicación exige un análisis caso por caso, y adaptar el mercado relevante y las pruebas de abuso a las plataformas digitales es hoy el terreno central de la regulación de la competencia.

En suma, con información perfecta los instrumentos se ordenan por la prioridad entre eficiencia asignativa y viabilidad financiera, y sus reglas siguen vigentes en las industrias de red. Pero la empresa sabe más que el regulador, y la teoría moderna busca el mecanismo que la induce a revelar sus costes y reducirlos, con el menú de contratos y la competencia como fuente de información. Exige compromiso creíble, comparadores independientes y ausencia de costes hundidos.

3. Mercado relevante y aplicaciones prácticas

Esta sección trata de la puesta en práctica por las autoridades de competencia y los reguladores sectoriales, cuya herramienta de partida es el mercado relevante. Sigue tres pasos: medir el poder de mercado, intervenir, en las industrias de red con la liberalización y la reestructuración, y evaluar la intervención y sus límites.

3.1. Mercado relevante

Definición: el **mercado relevante** es el conjunto de productos y áreas geográficas que ejercen una presión competitiva efectiva sobre la empresa. Se delimita en tres pasos:

1. *Dimensión de producto*: los sustitutivos razonables, aproximados por las elasticidades cruzadas.
2. *Dimensión geográfica*: el área con condiciones de competencia suficientemente homogéneas, aquella a cuyos proveedores acudirían los consumidores ante una subida del precio.
3. *Test del monopolista hipotético* o SSNIP: si un monopolista hipotético del conjunto candidato no podría subir el precio de forma rentable un 5-10 %, se añade el sustitutivo más próximo. El mercado relevante es el conjunto más pequeño en el que la subida sería rentable. Su delimitación es el punto de partida de todo expediente de competencia.

El test debe aplicarse a precios competitivos, porque a precios ya monopolísticos todo parece sustitutivo y el mercado se sobrestima, la **falacia del celofán** (caso DuPont, 1956).

3.2. Medición empírica

Del índice de Lerner, que exige conocer el coste marginal, se pasa a los indicadores estructurales. La cuota mide la posición de cada empresa, y los índices de concentración y de barreras, la estructura del conjunto y su sostenibilidad.

3.2.1. Índices de posición y concentración

Cuota de mercado: es el umbral habitual de las presunciones de dominio, pero una cuota alta puede reflejar eficiencia superior, como señala la crítica de Chicago, y depende de la delimitación del mercado.

Ratios de concentración: CR_k, la cuota conjunta de las k mayores empresas, son el indicador más simple, pero ignoran la distribución de las cuotas, tanto entre las k mayores como entre las restantes:

$$\mathrm{CR}_k = \sum_{i=1}^{k} s_i$$

Índice de Herfindahl-Hirschman: suma los cuadrados de las cuotas, ponderando cada empresa por su propia cuota, y va de 0 a 10.000:

$$\mathrm{HHI} = \sum_{i=1}^{n} (100\, s_i)^2 \in \left[\frac{10\,000}{n},\ 10\,000\right]$$

Las *Merger Guidelines* de 2010 distinguían concentración baja, por debajo de 1500, moderada, hasta 2500, y alta. Las de 2023 califican de muy concentrado un mercado con $\mathrm{HHI} > 1800$ y presumen que reduce sustancialmente la competencia la fusión que lo crea o refuerza con $\Delta\mathrm{HHI} > 100$. La variación del índice que causa una fusión es, junto con su nivel, uno de los criterios principales de la Comisión Europea y de las autoridades estadounidenses para decidir si la operación requiere un examen en profundidad.

3.2.2. Índices de barreras a la entrada

Una concentración alta preocupa si la entrada es difícil. Las barreras se aproximan con la **escala mínima eficiente relativa**, EME/X, la medida tradicional de Bain; la intensidad de los costes hundidos, como el gasto en publicidad, I+D y capital específico sobre las ventas, y las tasas de entrada y salida. Junto con la concentración, permiten calificar un mercado de impugnable o no. Una escala mínima eficiente relativa alta obliga al entrante a capturar una parte sustancial del mercado, y los costes hundidos miden lo que perdería si fracasa. Tasas de entrada y salida altas indican un mercado impugnable aunque esté concentrado.

3.3. Liberalización de las industrias de red

Diagnóstico: en el ferrocarril, la electricidad, el gas, el agua o las telecomunicaciones, una empresa integrada reúne la red, con rasgos de monopolio natural, y el servicio, potencialmente competitivo, y al dueño de la red le interesa negar o encarecer el acceso a sus rivales. La estrategia es regular la red, para conservar la eficiencia productiva, y liberalizar el servicio, para lograr la asignativa. En la Unión Europea se ha instrumentado con sucesivos paquetes de directivas. Estos imponen, sobre todo en energía y ferrocarril, la separación entre la gestión de la red y la prestación del servicio, y garantizan el acceso de terceros en condiciones transparentes y no discriminatorias.

Cuestiones transversales: los *costes varados*, inversiones rentables bajo protección que dejan de serlo en competencia; los *costes de transición a la competencia*, compensaciones que no deben encubrir ayudas públicas; el *reequilibrio tarifario*, que elimina los subsidios cruzados, y la garantía del *servicio universal*. Los subsidios cruzados podrían financiar precios predatorios, y el coste del servicio universal debe repartirse entre los operadores sin distorsionar la competencia.

Fórmulas de reestructuración:

- *Licitación de la titularidad*, como las concesiones de agua o autopistas: disipa las rentas en la puja, con riesgo de colusión y los límites de Williamson. Las rentas van a los consumidores si se licita por el menor precio, o al sector público si se licita por el mayor canon.
- *Regulación del acceso*, como la desagregación del bucle local: preserva la eficiencia productiva, con riesgo de discriminación a favor de la filial propia. Un canon de acceso alto refuerza el poder del dueño de la red, y uno bajo genera una entrada excesiva y desincentiva la inversión en ella.
- *Separación de la propiedad*, como en el ferrocarril, con gestor de infraestructura independiente, y en el transporte eléctrico, con la separación patrimonial que prevé el tercer paquete energético (2009): elimina la discriminación, pero pierde economías de alcance. La red puede querer reintegrarse, con lo que el problema de la discriminación reaparecería.
- *Propiedad compartida*, como los sistemas de reservas de las aerolíneas: sin incentivo a discriminar, pero con riesgo de cierre a entrantes y colusión. La cercanía entre red y servicio adapta enseguida la primera a lo que necesita el segundo, pero con muchos socios la gestión de la red se complica.
- *Separación operativa*, como Openreach, la red de acceso de British Telecom, separada funcionalmente en 2006 y jurídicamente desde 2018: evita expropiar, pero la entidad puede ser capturada e invertir poco. Si la gobierna el regulador, funciona casi como una regulación del acceso, con las empresas en su gobierno se parece a la propiedad compartida y, con plena independencia, a la separación de la propiedad.
- *Partes recíprocas*, como la telefonía móvil con interconexión obligatoria: adecuada cuando el poder procede de las economías de red y no de escala, de modo que solo hay ineficiencia asignativa y conviene introducir competencia también en la red. Exige que el acceso funcione en los dos sentidos. Cada operadora abre su red a los clientes de las demás porque, si no, estas le negarían el acceso a las suyas.
- *Partes no recíprocas*, como la distribución eléctrica y de gas: separa distribución y comercialización, pero deja intacto el poder de la distribución. Exige, además, que el consumidor elija comercializador, con los costes de información y de cambio que ello supone.

3.4. Evaluación y límites

Análisis coste-beneficio: asigna valores monetarios a los costes y beneficios sociales de un proyecto para juzgar su deseabilidad, elegir entre alternativas y evaluar *ex post*. Nace con Dupuit (1844), que introdujo el excedente del consumidor para valorar obras públicas, y su primera adopción formal es la Flood Control Act (1936). Plantea la valoración con precios sombra, la elección de la tasa social de descuento y los efectos redistributivos. Pese a la dificultad de monetizar intangibles y al riesgo de manipulación, sigue siendo la referencia de la Comisión Europea y del BEI, y fue de aplicación obligada en los grandes proyectos cofinanciados por los fondos estructurales hasta el periodo 2014-2020.

Captura del regulador: Olson (1965) muestra que los grupos pequeños y concentrados se organizan mejor que los grandes y difusos. Para Stigler (1971), la regulación "es adquirida por la industria y diseñada y aplicada fundamentalmente en su beneficio". De ahí la **búsqueda de rentas**, que despilfarra recursos en la presión, y la **captura**, por presión directa, dependencia informativa, puertas giratorias o captura cultural. Peltzman (1976) añade que el regulador reparte las rentas entre grupos para maximizar su apoyo, y Becker (1983), que la competencia entre grupos limita las ineficiencias. La teoría funda la desregulación y la independencia de los reguladores, pero puede explicar *a posteriori* casi cualquier regulación, le cuesta explicar las desregulaciones que perjudicaron a industrias organizadas y no modeliza la asimetría de información.

Los demás fallos del sector público (Niskanen, Nordhaus, la aversión al riesgo burocrática y la ilusión fiscal) y sus mecanismos correctores se estudian en el tema A23, apartados 3.2 a 3.4.

1. Modelo walrasiano de competencia perfecta

Walras (*Elementos de economía política pura*, 1874) funda la teoría del equilibrio general sobre la competencia perfecta, que buscaba justificar como mecanismo de asignación óptimo. Los mercados de bienes y de factores se analizan conjuntamente, y un vector de precios que cada agente toma como dado coordina las decisiones hasta vaciar todos los mercados a la vez.

1.1. Supuestos

Metodología: equilibrio general, con n mercados de bienes y factores interrelacionados que se analizan simultáneamente.

Entorno: no hay fallos de mercado: ni externalidades, ni bienes públicos, ni poder de mercado, ni costes de transacción ni rigideces. Los precios, salarios y tipos de interés son flexibles y la información es perfecta. Los consumidores poseen las dotaciones de bienes y factores y las empresas.

Agentes: consumidores que maximizan su utilidad y empresas que maximizan su beneficio, todos numerosos y precio-aceptantes. No hay sector público.

Tecnología: cada empresa produce un único bien homogéneo y divisible, $x = f(\mathbf{z})$, con rendimientos estrictamente decrecientes, y las empresas de un mercado comparten tecnología, lo que garantiza la ley del precio único. Con rendimientos decrecientes, los pagos de los factores no agotan el producto y los beneficios son positivos, pero la frontera de posibilidades de producción sigue siendo cóncava.

Mercado de factores: es también perfecto, la oferta de factores procede de las dotaciones de los consumidores y sus precios se determinan en el equilibrio. Formalmente, $\partial p/\partial x = 0$ y $\partial p_l^z/\partial z_l = 0$.

1.2. Modelo descentralizado con producción

1.2.1. Demanda: consumidores

Planteamiento: con preferencias racionales, monótonas, continuas y estrictamente convexas, y utilidades independientes, el consumidor h maximiza su utilidad sujeto a una renta **endógena**, que suma el valor de sus dotaciones de bienes y de factores y sus participaciones en los beneficios:

$$\max_{\mathbf{x}^h} \; u^h(\mathbf{x}^h) \quad \text{s.a.:} \; \sum_i p_i x_i^h \le \sum_i p_i x_{i0}^h + \sum_l p_l^z z_{l0}^h + \sum_f \alpha_f^h \pi^{f*}$$

La renta tiene tres fuentes: la *renta dotacional de bienes*, valor de mercado de sus dotaciones iniciales; la *renta dotacional de factores*, y la *renta de las participaciones* en los beneficios de las empresas, con $\sum_h \alpha_f^h = 1$.

CPO: la relación marginal de sustitución, u_i'/u_j', valoración subjetiva de cada consumidor que varía de uno a otro y de una cesta a otra, iguala el precio relativo, tasa objetiva y común a la que el mercado permite intercambiar los bienes. Si fuesen distintas, reasignar el gasto mejoraría la utilidad. Derivando el lagrangiano y eliminando el multiplicador:

$$\frac{u_i'}{u_j'} = \frac{p_i}{p_j}, \qquad \forall\, i \neq j, \quad \forall\, h = 1, \dots, H$$

CSO: la cuasiconcavidad estricta de la utilidad hace suficiente la CPO y única la demanda.

Solución: cada consumidor obtiene una demanda homogénea de grado cero en precios y renta, sin ilusión monetaria. Como la renta es el valor de las dotaciones a los precios vigentes, la demanda de mercado, suma horizontal de las individuales, depende solo del vector de precios y es homogénea de grado cero en él:

$$D_i(p) = \sum_h x_i^h(p, M_0^h)$$

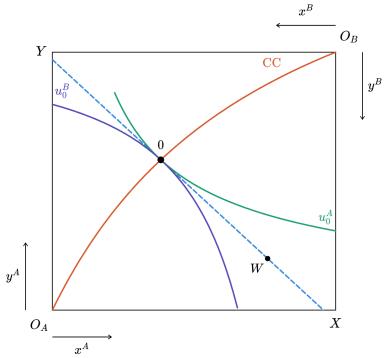
Equilibrio: como todos afrontan el mismo precio relativo, todos igualan su RMS a él:

$$\text{RMS}_{ji}^A = \frac{p_i}{p_j} = \text{RMS}_{ji}^B$$

La condición se generaliza a H consumidores. El vector de precios es así un **mecanismo de coordinación**, pues obliga a todos a ajustar sus valoraciones a una misma tasa objetiva, y sin precios comunes no habría razón para que las RMS se igualaran.

Desarrollo gráfico: en la caja de Edgeworth-Bowley, cuyas dimensiones son las dotaciones totales, las cantidades de A se miden desde O_A y las de B desde el vértice opuesto, O_B, de modo que cada punto es un reparto factible. La recta presupuestaria común pasa por la dotación inicial W con pendiente $-p_x/p_y$. En el equilibrio es tangente a la vez a las curvas de indiferencia de ambos consumidores, en un punto de la curva de contratos. La curva de contratos une las tangencias entre las curvas de indiferencia, los repartos en los que las RMS coinciden, y el vector de precios selecciona ese punto entre las infinitas asignaciones factibles.

Bienestar: la igualdad de las RMS es la condición de eficiencia en el consumo. Llevando la curva de contratos al plano de utilidades se obtiene la **frontera de posibilidades de utilidad**, que da la máxima utilidad de A para cada nivel de utilidad de B. Tiene pendiente negativa, porque sobre la curva de contratos mejorar a uno exige empeorar al otro, y el punto que alcanza el equilibrio depende de la dotación inicial, que fija la renta de cada consumidor. En cada punto de la curva de contratos, la RMS común indica el precio relativo que sostendría ese reparto como equilibrio competitivo.



Equilibrio del consumo en la caja de Edgeworth-Bowley, dadas la dotación inicial y los precios.

1.2.2. Oferta: empresas

Planteamiento: con rendimientos estrictamente decrecientes y sin externalidades, cada empresa elige los factores que maximizan su beneficio:

$$\max_{\mathbf{z}^f} \; \pi^f = p_i\, f_i(\mathbf{z}^f) - \sum_l p_l^z z_l^f$$

CPO: tres condiciones equivalentes:

$$p_i\, \text{Pmg}_l^f = p_l^z$$

$$\text{RMST}_{kl}^f = \frac{\text{Pmg}_l^f}{\text{Pmg}_k^f} = \frac{p_l^z}{p_k^z}$$

$$\text{Img}_i = \text{Cmg}_i \iff p_i = \frac{p_l^z}{\text{Pmg}_l^f} = \text{Cmg}_i$$

1. *Maximización del beneficio*: cada factor se contrata hasta que el valor de su productividad marginal iguala su precio. Mientras la última unidad aporte más de lo que cuesta, conviene contratar más.
2. *Minimización del coste*: la relación marginal de sustitución técnica iguala el precio relativo de los factores. Si no, reasignar factores abarataría la misma producción.
3. *Regla de precios*: el ingreso marginal, que para una empresa precio-aceptante es el precio, iguala el coste marginal, la condición estándar de oferta competitiva.

CSO: la concavidad estricta de la producción hace suficientes las CPO.

Solución: la oferta y las demandas de factores son únicas, continuas y homogéneas de grado cero en precios, y sumadas dan la oferta de mercado y la demanda de cada factor. Las de cada empresa se reúnen en el vector de producción neta, positivo en el bien y negativo en los factores. El beneficio se escribe entonces $\pi^f = \mathbf{p} \cdot \mathbf{y}^f$, y el exceso de demanda suma en una sola expresión la oferta de bienes y la demanda de factores de las empresas.

Equilibrio: con un precio relativo de los factores común, todas las empresas igualan su RMST:

$$\text{RMST}_{KL}^1 = \frac{w}{r} = \text{RMST}_{KL}^2$$

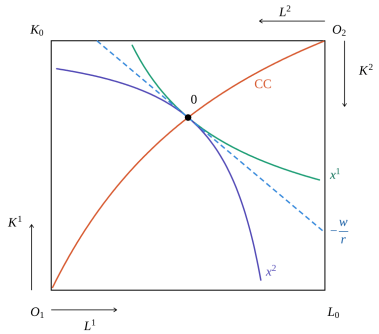
El precio relativo w/r coordina así a todas las empresas.

Desarrollo gráfico: en la caja de factores, cuyas dimensiones son las dotaciones totales de trabajo y capital, la recta isocoste común es tangente a la vez a las isocuantas de ambas empresas, en un punto de la curva de contratos de la producción.

Bienestar: es la condición de eficiencia en la producción. Llevada la curva de contratos al plano de bienes, se obtiene la **frontera de posibilidades de producción**, cóncava, cuya pendiente es la relación marginal de transformación, $\text{RMT}_{yx} = \text{Cmg}_x/\text{Cmg}_y$, el coste de oportunidad de x en unidades de y. Como el precio iguala el coste marginal, en equilibrio $\text{RMT}_{yx} = p_x/p_y$, y la producción es, así, el punto de la frontera tangente a la recta isoingreso más alejada del origen.

1.2.3. Ajuste del sistema

Exceso de demanda: a un vector de precios cualquiera, los planes de unos y otros no tienen por qué ser compatibles. La discrepancia, $Z_i(p)$, es la demanda de mercado menos la producción neta de las empresas y las dotaciones. Es una función, y no una correspondencia, continua para precios positivos y homogénea de grado cero, de modo que solo se determinan precios relativos. La teoría no le impone una pendiente general,



Equilibrio en la producción: la isocoste común es tangente a las isocuantas de ambas empresas.

pero si todos los bienes son **sustitutivos brutos**, $\partial Z_i/\partial p_j > 0$ para $j \neq i$, el equilibrio es único y estable. Formalmente, con $Z_i(kp) = Z_i(p)$ para todo $k > 0$:

$$Z_i(p) = \sum_{h=1}^{H} x_i^h(p) - \left(\sum_{f=1}^{F} y_i^f(p) + \sum_{h=1}^{H} x_{i0}^h \right)$$

Integrabilidad: no hereda la simetría ni la ley de la demanda compensada. Por el teorema de **Sonnenschein, Mantel y Debreu** (1972-1974), cualquier función continua, homogénea de grado cero y que cumpla la ley de Walras es, lejos de los precios nulos, la de alguna economía de intercambio con al menos tantos consumidores como bienes, de modo que no puede tratarse como la de un consumidor representativo. Además, solo es observable en el equilibrio, porque fuera de él recoge demandas y ofertas nocionales.

Bienes escasos y libres: si el bien es escaso, el exceso de demanda corta el eje en un precio positivo y, si es decreciente, hay exceso de demanda por debajo de ese precio y de oferta por encima. Si es libre, hay exceso de oferta incluso a precio nulo, y su precio de equilibrio debe ser cero. Si los productores pueden eliminar sin coste el exceso de producción, estos bienes se incorporan al modelo.

Ley de Walras (1874): el valor de los excesos de demanda es nulo para cualquier vector de precios, no solo en equilibrio, porque cada consumidor gasta exactamente su renta:

$$\sum_{i=1}^{n} p_i Z_i(p) = 0$$

Los mercados están interconectados, pues lo que se gasta en uno deja de gastarse en otro. De ahí cuatro corolarios:

1. *Excesos compensadores*: un exceso en un mercado implica otro de signo contrario en otro.
2. *Precios relativos*: con $n-1$ ecuaciones independientes solo se determinan precios relativos, y un bien se toma como **numerario**.
3. *El dinero es un velo*: en esta economía de trueque, el numerario es solo unidad de cuenta.
4. *Economía abierta*: $\sum_i p_i Z_i(p) = -\text{BC}$: un exceso de demanda interna se cubre con importaciones netas, y en el modelo de dos países basta vaciar un mercado para obtener la relación real de intercambio. Esa relación es la pendiente de la recta por el origen sobre la que son compatibles las ofertas de exportación y las demandas de importación de ambos países. Si $\text{BC} = 0$, hay equilibrio externo.

Tâtonnement: en la reconstrucción habitual, un subastador ficticio guía la economía antes de cualquier transacción:

1. Anuncia un vector de precios.
2. Los agentes calculan sus demandas y ofertas nocionales a esos precios.
3. El subastador observa los excesos en cada mercado.
4. Si algún mercado no se vacía, no hay transacciones: sube los precios con exceso de demanda y baja los de exceso de oferta.
5. Repite hasta vaciar todos los mercados, y solo entonces se ejecutan las transacciones.

A diferencia del ajuste marshalliano vía cantidades, el ajuste es simultáneo, vía precios, y las demandas nocionales coinciden con las efectivas. Esta propiedad se pierde con rigideces de precios, punto de partida de los neokeynesianos del desequilibrio. Formalmente, el subastador transforma precios normalizados en precios normalizados:

$$T_i(p') = \frac{\max\{p_i' + \lambda_i Z_i(p'),\ 0\}}{\sum_{j=1}^{n} \max\{p_j' + \lambda_j Z_j(p'),\ 0\}}$$

El equilibrio es un **punto fijo**, $T(p^*) = p^*$, lo que conecta el subastador con la prueba de existencia. En la regla, $\lambda_i > 0$ es la velocidad de ajuste y el operador máx impide precios negativos. Si $Z_i > 0$, el precio sube, y si $Z_i < 0$, baja.

1.2.4. Equilibrio general competitivo

Definición: el **equilibrio general competitivo** es el vector de precios p^* en el que se detiene el ajuste, con el que consumidores y empresas optimizan y todos los mercados se vacían. En cada mercado hay una demanda, una oferta y una condición de equilibrio, con precios y cantidades como incógnitas, de modo que el equilibrio es la solución de un sistema de ecuaciones simultáneas. Que haya tantas ecuaciones como incógnitas no basta, porque las ecuaciones deben ser independientes, y, por la ley de Walras, una es redundante. Se expresa de dos formas:

– *Igualación de valoraciones*: la valoración subjetiva de los consumidores iguala el precio relativo, y este, el coste de oportunidad de producir:

$$\text{RMS}_{ji}(p^*) = \frac{p_i^*}{p_j^*} = \frac{\text{CT}_i'}{\text{CT}_j'} = \text{RMT}_{ji}(p^*)$$

– *Ausencia de excesos*: los mercados de bienes escasos se vacían y los bienes libres, con precio nulo, admiten excesos de oferta: $Z_i^* \leq 0$, $p_i^* \geq 0$, $p_i^* Z_i^* = 0$. El ajuste ante excesos se supone instantáneo.

El equilibrio se caracteriza, así, por cuatro condiciones: los consumidores maximizan su utilidad y las empresas su beneficio dados los precios, todos los mercados se vacían a la vez y solo se determinan precios relativos, con un bien como numerario. La triple igualdad sintetiza la eficiencia del sistema competitivo. En el modelo $2 \times 2 \times 2 \times 2$, con dos bienes, dos factores, dos empresas y dos consumidores, hay 14 incógnitas y 14 ecuaciones.

En suma, el modelo integra las decisiones de consumidores y empresas y muestra que un vector de precios puede coordinarlas a la vez. Las funciones de exceso de demanda conservan la continuidad y la homogeneidad y cumplen la ley de Walras, aunque no heredan el resto de la estructura de las demandas (SMD), lo que permite plantear la existencia, la unicidad y la estabilidad. Pero el *tâtonnement* es una ficción, el modelo es estático y con mercados completos, y descansa en la convexidad y en la ausencia de fallos.

2. Análisis del equilibrio

Definido el equilibrio, se pregunta si existe, si es único y si es estable, y después si es eficiente.

2.1. Análisis positivo

2.1.1. Existencia: punto fijo de Brouwer

Origen: Walras se limitó a contar ecuaciones e incógnitas, lo que no garantiza la existencia. La prueba definitiva se debe a Arrow y Debreu (1954), que se apoyaron en la teoría de juegos, con preferencias y conjuntos de producción convexos, y Debreu (1959) la extendió a la incertidumbre.

Idea: por el **teorema del punto fijo de Brouwer** (1912), toda función continua de un conjunto no vacío, cerrado, acotado y convexo de $\mathbb{R}^n$ en sí mismo tiene un punto fijo, $f(x^*) = x^*$. Basta aplicarlo a la función del subastador. El teorema es condición suficiente para la existencia, porque el punto fijo $T(p^*) = p^*$ equivale al vaciado de todos los mercados de bienes escasos.

Razonamiento: la existencia se demuestra en tres pasos:

1. *Precios*: como solo importan los relativos, se normalizan para que sumen uno, $p_i' = p_i / \sum_j p_j$, con al menos un bien escaso. Mantienen sus valores relativos y cada uno está en $[0,1]$, de modo que el conjunto está acotado, y es cerrado y convexo porque lo definen una igualdad y desigualdades débiles. En símbolos, $p_i'/p_j' = p_i/p_j$ y $\sum_i p_i' = 1$.
2. *Continuidad*: para los bienes libres la demanda podría ser infinita. La **saciabilidad condicional**, según la cual los consumidores se sacian de todo salvo de algún bien escaso, la hace finita y la función del subastador, continua.
3. *Punto fijo*: para un bien libre, como $p_i^* = 0$, el punto fijo exige $\lambda_i Z_i(p^*) \leq 0$, es decir, exceso de oferta o equilibrio. Para uno escaso, la ley de Walras obliga a que el denominador de T valga uno, y el punto fijo se reduce a $Z_i(p^*) = 0$, el vaciado del mercado.

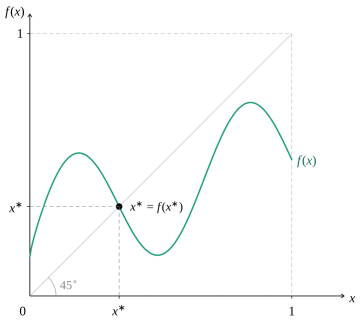
Desarrollo gráfico: una función continua de $[0,1]$ en sí mismo que parte de la bisectriz o de encima y llega a ella o a debajo corta necesariamente la bisectriz en un punto fijo. La misma lógica vale para el exceso de demanda. Si es continuo, para pasar de exceso de oferta a exceso de demanda tiene que cruzar el eje.

Conclusiones: prueba la coherencia lógica del sistema walrasiano con el mismo argumento de punto fijo que el equilibrio de Nash. Pero solo prueba la existencia, sin decir cuántos equilibrios hay ni si se alcanzan, y excluye los rendimientos crecientes y las indivisibilidades.

2.1.2. Unicidad

La unicidad importa porque, con equilibrios múltiples, la estática comparativa pierde su punto de referencia. Dos teoremas dan condiciones suficientes:

Teorema de la preferencia revelada: si el exceso de demanda agregado cumple el ADPR, según el cual si $Z(p) \neq Z(p')$ y $pZ(p') \leq 0$, entonces $p'Z(p) > 0$, el conjunto de equilibrios es convexo, porque toda



Existencia: toda función continua del simplex en sí mismo tiene un punto fijo (Brouwer).

combinación convexa de dos equilibrios es otro equilibrio. Si además están aislados, el equilibrio es único. En efecto, si p^1 y p^2 son equilibrios y su combinación convexa no lo fuese, el ADPR haría positivos los dos sumandos de su ley de Walras, que debe valer cero.

Teorema de los sustitutivos brutos: si todos los bienes son sustitutivos brutos en sentido estricto, el equilibrio es único. Con dos bienes, si p_2 baja respecto al equilibrio, hay exceso de oferta del bien 1 y, por la ley de Walras, exceso de demanda del 2. Con sustitutivos brutos, la homogeneidad de grado cero implica $\partial Z_i/\partial p_i < 0$.

Equilibrios múltiples: con efectos renta intensos, rendimientos crecientes u ofertas de trabajo de pendiente negativa, la función de exceso de demanda corta el eje varias veces: con sustitutivos brutos lo corta una sola vez, en p^*, y sin ellos puede cortarlo en p^1, p^2 y p^3. En la oferta de trabajo, la pendiente negativa aparece cuando el efecto renta domina al efecto sustitución a salarios altos.

Conclusiones: dan la unicidad y, con ella, la referencia de la estática comparativa. Pero son muy restrictivos: el ADPR agregado exige casos como preferencias homotéticas idénticas, y la sustituibilidad bruta excluye la complementariedad bruta y que dominen los efectos renta.

2.1.3. Estabilidad

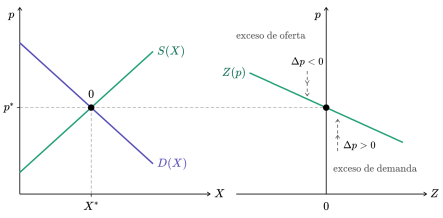
Definición: el equilibrio es estable si, tras una perturbación, el ajuste lo devuelve a él, $\lim_{t\to\infty} p(t) = p^*$; globalmente, desde cualquier vector inicial, o localmente, en su entorno. En un mercado aislado exige $\partial Z_i/\partial p_i < 0$ cerca de p^*.

Función de distancia: con dos bienes, sustitutivos brutos y la regla $\dot{p}_i = \lambda_i Z_i(p)$, la distancia al equilibrio cae monótonamente:

$$\Phi = \frac{(p_1(t)-p_1^*)^2}{2\lambda_1} + \frac{(p_2(t)-p_2^*)^2}{2\lambda_2}, \qquad \frac{d\Phi}{dt} < 0$$

El equilibrio es estable en sentido walrasiano (Arrow et al., 1959).

Desarrollo gráfico: si el exceso de demanda decrece con el precio, por encima de p^* hay exceso de oferta y el subastador baja el precio, y por debajo lo sube, de modo que el precio vuelve a p^*. Si crece, por encima de p^* hay exceso de demanda y el precio sigue subiendo, y por debajo sigue bajando, con lo que el tanteo aleja el precio del equilibrio, que es inestable. En el plano cantidad-precio, la oferta y la demanda se cortan en p^*, y $Z(p)$ recoge su diferencia a cada precio.



Estabilidad walrasiana: el subastador sube el precio con exceso de demanda y lo baja con exceso de oferta.

Conclusiones: con sustitutivos brutos, el *tâtonnement* converge de forma monótona y permite la estática comparativa. Pero sin ellos la estabilidad no está garantizada (Scarf, 1960), y descansa en un subastador ficticio.

2.2. Optimalidad de Pareto

Definición: una asignación es **eficiente en el sentido de Pareto** si no hay otra factible que mejore a alguien sin empeorar a otro. Tiene tres dimensiones, que no involucran precios:

- *Consumo*: en la curva de contratos, $\text{RMS}^A_{yx} = \text{RMS}^B_{yx}$. Fuera de ella, las curvas de indiferencia se cortan y encierran una "lente" de mejoras para ambos. Sobre ella, mejorar a uno empeora al otro.
- *Producción*: $\text{RMST}^1_{KL} = \text{RMST}^2_{KL}$, de modo que no se puede producir más de un bien sin reducir el otro.
- *Global*: además, la combinación de producto es eficiente, $\text{RMS}_{ji} = \text{RMT}_{ji}$. La RMS, tasa a la que los consumidores intercambian bienes, iguala así la RMT, tasa a la que la economía puede transformar un bien en otro.

2.2.1. Primer teorema del bienestar

Teorema: con preferencias localmente no saciables y sin fallos de mercado, todo equilibrio competitivo es eficiente en el sentido de Pareto. El teorema formaliza la *mano invisible* de Smith, aunque se refiere a la optimalidad paretiana y no a la social. Exige mercados completos y agentes precio-aceptantes, pero no convexidad.

Demostración: como todos los agentes afrontan los mismos precios, cada condición de equilibrio implica la de eficiencia, en el consumo, en la producción y en la combinación de producto:

$$\text{RMS}^A_{yx} = \frac{p_x}{p_y} = \text{RMS}^B_{yx} \;\Rightarrow\; \text{RMS}^A_{yx} = \text{RMS}^B_{yx}$$

$$\text{RMST}^1_{KL} = \frac{w}{r} = \text{RMST}^2_{KL} \;\Rightarrow\; \text{RMST}^1_{KL} = \text{RMST}^2_{KL}$$

$$\text{RMS}_{ji} = \frac{p_i}{p_j} = \text{RMT}_{ji} \;\Rightarrow\; \text{RMS}_{ji} = \text{RMT}_{ji}$$

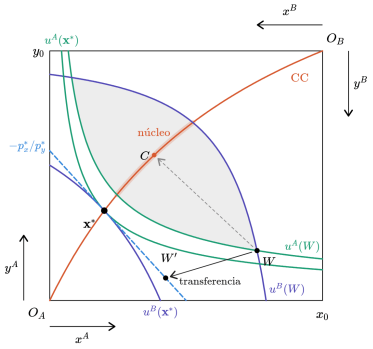
El equilibrio selecciona así necesariamente una asignación Pareto-eficiente, sobre la curva de contratos. Cuando cada agente persigue lo que más le conviene orientado solo por los precios, nadie puede ganar sin que otro pierda, y los precios, mecanismo de coordinación, resultan ser también el vehículo de la eficiencia.

Corolarios: el equilibrio es eficiente, aunque puede no ser equitativo; el mercado alcanza por sí solo un óptimo de Pareto; a cada agente le basta conocer los precios, y los fallos de mercado pueden justificar la intervención.

2.2.2. Segundo teorema del bienestar

Teorema: con preferencias y conjuntos de producción convexos, cualquier óptimo de Pareto puede alcanzarse como equilibrio competitivo redistribuyendo las dotaciones con transferencias de suma fija.

Demostración: el primer teorema sitúa el equilibrio en la curva de contratos, pero el punto depende de la dotación inicial. Desplazarla mueve la recta presupuestaria y selecciona otro punto de la curva. El Gobierno podría elegir la distribución de bienestar deseada, redistribuir las dotaciones *ex ante* y dejar que la competencia alcance esa asignación. La convexidad garantiza que la tangente común a las curvas de nivel sea una recta de precios que sostiene el óptimo como equilibrio.



Segundo teorema fundamental del bienestar.

Corolarios: la *separación entre eficiencia y equidad*, porque basta redistribuir *ex ante* y dejar actuar al mercado, y la *planificación de Lange-Lerner*, que con precios sombra alcanzaría los mismos resultados.

Objeciones: las transferencias de suma fija exigen una información que ningún gobierno tiene, los impuestos reales distorsionan y, según Hayek, la planificación exigiría costes de información muy elevados. La tensión entre ambos teoremas, eficiencia sin equidad en el primero y redistribución previa en el segundo, define el marco normativo del equilibrio general, pero juntos son un **ideal regulativo** más que una prescripción.

En suma, el equilibrio existe, pero su unicidad y su estabilidad exigen condiciones restrictivas, como la sustituibilidad bruta. Con precios comunes es eficiente, y con convexidad cualquier óptimo es alcanzable, aunque con transferencias inviables. El análisis supone, además, un subastador ficticio y una economía sin dinero ni incertidumbre.

3. Extensiones y aplicaciones

El modelo admite otros procesos de ajuste, la introducción del dinero y de la incertidumbre, y aplicaciones computables.

3.1. Procesos de ajuste sin subastador

3.1.1. Ajuste vía cantidades (Marshall, 1890)

Idea: Marshall (*Principles of Economics*, 1890) analiza el ajuste de un mercado aislado desde la producción. Los productores, y no un subastador, amplían la producción si el precio de demanda, lo que los

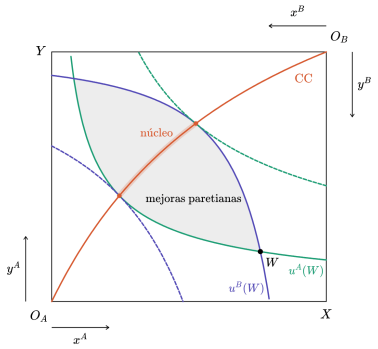
consumidores están dispuestos a pagar por la cantidad corriente, supera al de oferta, el coste marginal, y la reducen en caso contrario. El ajuste es gradual, mercado a mercado y con transacciones a precios de desequilibrio.

Conclusiones: se adapta a los mercados industriales, pero es de equilibrio parcial y, con oferta y demanda de pendiente del mismo signo, su criterio de estabilidad es el contrario del walrasiano.

3.1.2. Caja de Edgeworth-Bowley (1881; 1924)

Idea: Edgeworth (*Mathematical Psychics*, 1881) introduce las curvas de indiferencia y la curva de contratos en el intercambio entre dos agentes, representadas en la caja que dibujó Pareto en 1906 y popularizó Bowley en 1924. Sin precios ni subastador, dos agentes intercambian hasta agotar las ganancias mutuas.

Razonamiento: desde una dotación W fuera de la curva de contratos, las curvas de indiferencia que pasan por ella encierran una lente de intercambios beneficiosos. La negociación termina en el tramo de la curva de contratos dentro de la lente, pero el punto depende del poder de negociación y el resultado es indeterminado. La caja recoge las asignaciones factibles, con $x^A + x^B = X$ e $y^A + y^B = Y$.



Negociación en la caja de Edgeworth-Bowley: el núcleo es el tramo de la curva de contratos dentro de la lente.

Conclusiones: el intercambio voluntario alcanza la eficiencia sin precios, y la caja es la herramienta gráfica básica del equilibrio general. Pero el resultado es indeterminado y solo admite dos agentes y dos bienes.

3.1.3. Núcleo (Debreu y Scarf, 1963)

Idea: Edgeworth (1881) conjeturó que el tramo alcanzable se reduce al crecer el número de agentes. Con coaliciones y acuerdos vinculantes, el **núcleo** es el conjunto de asignaciones que ninguna coalición puede bloquear, de modo que todas son individualmente racionales y Pareto-eficientes. Es individualmente racional si no la bloquea ningún agente aislado, y Pareto-eficiente si no la bloquea la coalición de todos.

Razonamiento: Debreu y Scarf (1963) demuestran que, si se replica la economía indefinidamente, el núcleo converge a las asignaciones de equilibrio competitivo, y la asignación walrasiana siempre pertenece a él.

Conclusiones: valida la competencia perfecta en un entorno más general y natural, porque, aunque Walras es menos realista que Edgeworth, en el límite los resultados coinciden, y ningún grupo desearía retirarse a comerciar entre sus miembros. Los precios competitivos surgen de la negociación entre muchos agentes, sin subastador. Pero el resultado es asintótico y supone coaliciones sin coste. Con pocos agentes, además, el núcleo puede ser amplio.

3.2. Equilibrio general con dinero

Modelos: los *no monetarios* prescinden del dinero o lo tratan como un bien más; los *monetarios* le dan naturaleza propia, como el dinero fiduciario de las economías actuales. La tradición neoclásica mantiene el dinero fuera de la economía real, mientras que las posteriores, generalmente keynesianas o poskeynesianas, lo integran en ella.

Dicotomía clásica: la tradición neoclásica añade el dinero con los saldos de caja de Cambridge, como una fracción $k \in (0,1)$ del valor de las dotaciones, el inverso de la velocidad de circulación:

$$m^{d,h} = k\left(\sum_i p_i x^h_{i0} + \sum_l p^z_l z^h_{l0}\right)$$

Agregando, $M^d = k \cdot PY$, y el vaciado del mercado monetario fija el nivel absoluto de precios sin alterar los relativos. Pero, si los excesos de demanda de bienes dependen solo de los precios relativos, la ley de Walras, $Z_M = -\sum_i p_i Z_i(\mathbf{p})$, hace que, vaciados los bienes, el mercado de dinero se vacíe para cualquier nivel de precios, mientras que $kPY = M^s$ solo se cumple en uno: la dicotomía es inconsistente.

Patinkin (1956): incorpora el **efecto saldos reales**. Los saldos M/P entran en la utilidad, de modo que un cambio de precios altera la riqueza real y las demandas de bienes. Define el nivel de precios como media ponderada y expresa todos los excesos de demanda, de bienes y de dinero, en precios deflactados:

$$P = \sum_i \omega_i p_i, \qquad \omega_i \in (0,1), \quad \sum_i \omega_i = 1$$

$$Z_i = f_i\left(\frac{p_1}{P}, \frac{p_2}{P}, \dots, \frac{M}{P}\right), \qquad i = 1, \dots, n+1$$

Con $n+2$ ecuaciones, una redundante por la ley de Walras, y $n+1$ incógnitas, se determinan los precios absolutos. Todas las funciones son homogéneas de grado cero en precios y dinero, con lo que la neutralidad del dinero es compatible con el abandono de la dicotomía. Multiplicar cada precio y M por un mismo factor no altera los precios relativos ni los saldos reales.

Extensiones: Patinkin observó que, si el dinero no reporta utilidad, nadie mantendría saldos, y que con información perfecta y sin costes de transacción los agentes intercambiarían sin dinero. El dinero solo tiene un papel sustantivo con fricciones, como la incertidumbre, los costes de transacción o las asimetrías de información, que se modelizan con el dinero en la utilidad, la restricción *cash-in-advance*, las generaciones solapadas o la búsqueda. Estas ideas inspiraron a los neokeynesianos del desequilibrio y a la nueva economía keynesiana, que modeló explícitamente las rigideces nominales y las fricciones de mercado.

Conclusiones: integra el dinero en el sistema walrasiano y pasa a la síntesis neoclásica. Pero el dinero en la utilidad es un atajo que postula su demanda sin derivarla de las fricciones, y el efecto saldos reales es débil en la práctica, porque exige caídas de precios muy intensas.

3.3. Incertidumbre (Arrow-Debreu)

Idea: Arrow (1953) y Debreu (*Theory of Value*, 1959) llevan el modelo a un mundo en el que los bienes de mañana dependen de un estado de la naturaleza que nadie controla, y muestran que basta redefinir los bienes: un mismo bien en fechas o estados de la naturaleza distintos son **bienes contingentes** distintos, definidos por su especificación física, su fecha y su estado.

Razonamiento: Debreu prueba la existencia con mercados completos, uno por bien contingente, y preferencias convexas, que con utilidad esperada equivalen a aversión al riesgo, además de los axiomas de preferencias del caso con certidumbre. Arrow, en una economía de intercambio, muestra que no hace falta un mercado por bien contingente: basta un título por estado de la naturaleza, pagadero en dinero, más mercados al contado tras conocerse el estado. Hacen falta así $S + C$ mercados y no $S \times C$, con S estados y C bienes. Ampliar la lista de bienes no altera los conjuntos de producción, porque la empresa maximiza el beneficio valorado a los precios de los bienes contingentes, sin considerar probabilidades.

Conclusiones: extiende la existencia y la eficiencia a la incertidumbre sin nuevos instrumentos y es el marco de referencia de la asignación del riesgo. Pero ninguna economía tiene mercados completos, y el modelo supone información común, sin riesgo moral ni selección adversa. Si la estructura informacional cambia entre periodos, los mercados se reabrirían, y el modelo exige que se abran y cierren definitivamente en el periodo inicial.

3.4. Aplicaciones

Desde los años setenta, tras el algoritmo de Scarf (1967), los **modelos de equilibrio general computable** llevan la teoría a la práctica:

- *Crecimiento*: modelos de horizonte infinito y de generaciones solapadas, que aplican el marco a las decisiones de consumo, ahorro e inversión a lo largo del tiempo.
- *Comercio internacional*: el impacto de las barreras comerciales sobre los salarios reales y el bienestar, como en los estudios del TLCAN y del TTIP, y la evaluación *ex ante* de los acuerdos de integración.
- *Política tributaria*: el impacto de los cambios impositivos y de las transferencias sobre el conjunto de la economía.
- *Política medioambiental*: los efectos de las políticas de reducción de emisiones sobre todos los sectores a la vez, como en el modelo GREEN de la OCDE (1992), con fundamento en la corrección de externalidades.
- *Planificación regional*: costes de transporte y movilidad laboral, como en la nueva economía geográfica.
- *Modelos institucionales*: DSGE como QUEST, de la Comisión Europea, o el NAWM, del BCE, frente a modelos macroeconométricos como el NiGEM, del NIESR, que no es un modelo de equilibrio general en sentido estricto.

1. El óptimo económico y la eficiencia de Pareto

La economía del bienestar parte del criterio de Pareto, que ordena asignaciones sin comparar utilidades entre personas, y caracteriza el **óptimo económico**: las condiciones de eficiencia en la producción, en el consumo y en la relación entre ambos.

1.1. Los criterios de Pareto (1906)

Origen: Vilfredo Pareto formula la eficiencia en su *Manual de economía política* (1906). Hasta entonces, la economía del bienestar, heredera de Bentham y de Marshall, juzgaba una política por el cambio en la suma de utilidades individuales. Eso exigía una utilidad cardinal y comparable entre personas, dos supuestos que Pareto consideraba insostenibles.

Idea: Pareto sustituye la utilidad cardinal por la ordinal, la *ofelimidad*, y construye un criterio sin comparaciones interpersonales. Si nadie empeora y alguien mejora, el juicio es incontrovertible, se midan como se midan las utilidades.

Formula dos criterios encadenados: la *superioridad de Pareto*, una relación entre dos asignaciones, y la *optimalidad de Pareto*, una propiedad de una asignación respecto de todas las factibles. Ambos recaen sobre una *asignación económica*, una lista de magnitudes que describe el estado de una economía, como una cesta de bienes, un plan de producción o un reparto de utilidades.

1.1.1. Supuestos

Toda la economía del bienestar paretiana descansa en cuatro supuestos, y cada uno ha recibido una crítica.

Independencia del proceso: solo importa el resultado, no el mecanismo que asigna los recursos. Pero la sociedad puede valorar el procedimiento en sí, como el intercambio libremente consentido o la democracia.

Individualismo: el bienestar social depende solo del bienestar de los individuos. No permite juzgar indeseable en sí misma una estructura de mercado como el monopolio.

No paternalismo: nadie sabe mejor que el propio individuo qué le conviene. No se sostiene con los bienes preferentes de Musgrave (1959) ni con las adicciones.

Benevolencia: un aumento de la utilidad de un individuo, *ceteris paribus*, mejora el bienestar social. No emite juicio alguno sobre la desigualdad resultante.

1.1.2. Dominancia de Pareto

Definición: la **dominancia de Pareto** es una relación de orden parcial entre vectores de magnitudes deseables: $\mathbf{v}$ domina a $\mathbf{v}'$ si es al menos igual en todas sus componentes y estrictamente mayor en alguna:

$$v_i \geq v'_i \;\; \forall i = 1, \dots, n \qquad \text{y} \qquad \exists\, j : v_j > v'_j$$

El criterio no exige dos individuos: lo que cambia de un ámbito a otro es el contenido del vector.

- En la eficiencia técnica, el plan de producción, con la isocuanta como frontera.
- En la eficiencia productiva, las cantidades producidas, con la frontera de posibilidades de producción (FPP).
- En la eficiencia en el intercambio, las utilidades dadas las dotaciones, con la curva de contratos.
- En la eficiencia asignativa, las utilidades con producción y consumo, con la gran frontera de posibilidades de utilidad (GFPU).

De ahí dos precisiones. La eficiencia económica o en costes no es un criterio de Pareto, sino un problema de optimización que exige precios. La eficiencia técnica es necesaria, pero no suficiente, para la económica. Y el óptimo de Pareto sin calificativo es la eficiencia asignativa u óptimo global, la más exigente, que el equilibrio competitivo verifica.

1.1.3. Superioridad de Pareto

Definición: aplicada al vector de utilidades, $\mathbf{x}$ es **superior en el sentido de Pareto** a $\mathbf{x}'$ si ningún individuo prefiere $\mathbf{x}'$ y al menos uno prefiere estrictamente $\mathbf{x}$. El paso de $\mathbf{x}'$ a $\mathbf{x}$ es una **mejora paretiana**:

$$u^h(\mathbf{x}) \geq u^h(\mathbf{x}') \;\; \forall h = 1, \dots, H \qquad \text{y} \qquad \exists\, j : u^j(\mathbf{x}) > u^j(\mathbf{x}')$$

El criterio es puramente ordinal, es decir, basta la relación de preferencia de cada individuo. Es exigente, porque requiere unanimidad en sentido débil, e incompleto, porque no compara asignaciones en las que uno gana y otro pierde.

La limitación afecta a casi toda la política económica: un arancel, una reforma fiscal o una liberalización generan siempre ganadores y perdedores.

1.1.4. Optimalidad o eficiencia de Pareto

Definición: una asignación factible $\mathbf{x}^*$ es **eficiente u óptima en el sentido de Pareto** si no existe otra asignación factible que le sea Pareto-superior, es decir, solo cabe mejorar a uno empeorando a otro. Los óptimos se obtienen maximizando la utilidad de un individuo con la de los demás dada y sujeta a los recursos y la tecnología.

Propiedades:

1. Superioridad y optimalidad son distintas: una asignación puede ser Pareto-superior a otra sin ser óptima.
2. Es distributivamente neutral: dar todos los recursos a una sola persona es un óptimo de Pareto.
3. Los óptimos son múltiples, toda la gran frontera de posibilidades de utilidad (GFPU), y cuál alcanza el mercado depende de la distribución inicial.
4. Tiene un sesgo hacia el *statu quo*: privilegia la distribución de partida, que no somete a examen.

1.1.5. Conclusiones

Aspectos positivos: sienta un criterio de eficiencia sin comparaciones interpersonales, que solo requiere preferencias ordinales y, como dominancia entre vectores, sirve para la producción, el consumo y la asignación global.

Limitaciones: además de las críticas a los supuestos, la indeterminación, porque la ordenación es incompleta y elegir entre los puntos de la GFPU el *optimum optimorum* exige un criterio adicional. La exigencia de unanimidad lleva a la *tiranía de la minoría*.

1.2. Eficiencia en una economía 2x2x2x2

Pareto estudia una economía con dos bienes, x e y; dos factores, K y L; dos empresas y dos consumidores.

Supuestos: sin fallos de mercado, economía de trueque y factores homogéneos con oferta perfectamente inelástica; productividades marginales positivas y decrecientes, rendimientos no crecientes y distinta intensidad factorial; utilidades ordinales, crecientes y cuasicóncavas.

1.2.1. Óptimo en la producción

Planteamiento: se maximiza $x = f_x(K_x, L_x)$ dados el nivel de producción de y, $f_y(K_y, L_y) = y_0$, y las dotaciones de factores, K_0 y L_0.

CPO: la relación marginal de sustitución técnica debe ser la misma en los dos bienes. Si no, podría producirse más de uno sin reducir el otro:

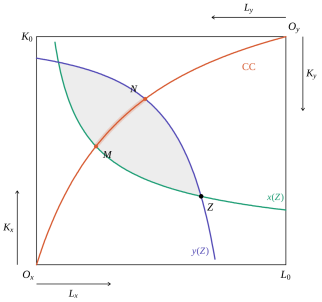
$$\text{RMST}^x_{KL} = \frac{f'_{L_x}}{f'_{K_x}} = \frac{f'_{L_y}}{f'_{K_y}} = \text{RMST}^y_{KL}$$

CSO: la convexidad de las isocuantas garantiza que la tangencia es un máximo de x dado y_0.

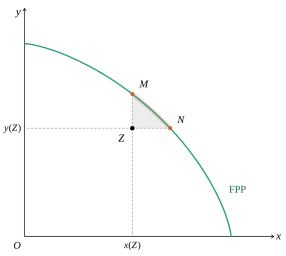
Desarrollo gráfico: en la *caja de Edgeworth de la producción*, de lados L_0 y K_0, las isocuantas de x se miden desde O_x y las de y desde O_y. Donde se cruzan, como en Z, forman una lente de reasignaciones Pareto-superiores.

Las tangencias, como M y N, forman la **curva de contratos de la producción**, cuyos puntos no son comparables según Pareto y cuya forma depende de la intensidad factorial.

Su proyección en el plano x-y es la **FPP**, cuya pendiente en valor absoluto es la relación marginal de transformación, RMT_{yx}, el coste de oportunidad de x en términos de y. Es cóncava por los rendimientos decrecientes a escala, la especialización de los factores y la distinta intensidad factorial. Una asignación ineficiente como Z se proyecta en su interior.



(a) Caja de Edgeworth de la producción



(b) Frontera de posibilidades de producción

Caja de Edgeworth de la producción y FPP: la curva de contratos se proyecta sobre la frontera y Z, en su interior.

1.2.2. Óptimo en el consumo (intercambio)

Planteamiento: se maximiza $u^A(x^A, y^A)$ dados el nivel de utilidad de B, u^B_0, y las cantidades disponibles, x_0 e y_0.

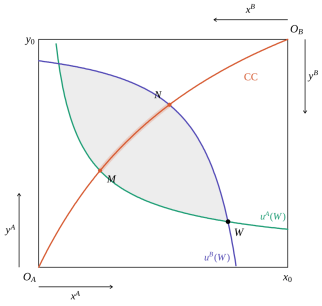
CPO: los dos consumidores tienen la misma relación marginal de sustitución. Si no, quedarían ganancias del intercambio sin explotar:

$$\text{RMS}^A_{yx} = \frac{u'^A_x}{u'^A_y} = \frac{u'^B_x}{u'^B_y} = \text{RMS}^B_{yx}$$

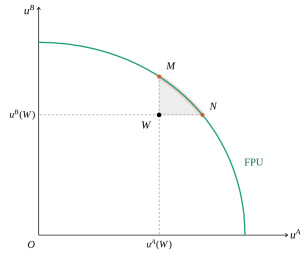
CSO: la convexidad de las curvas de indiferencia garantiza que la tangencia es un máximo de u^A dado u^B_0.

Desarrollo gráfico: en la *caja de Edgeworth del consumo*, las tangencias entre curvas de indiferencia forman la curva de contratos del consumo. Desde la dotación W, solo son alcanzables por intercambio voluntario los puntos de la curva de contratos dentro de la lente, el tramo entre M y N. Esos puntos forman el **núcleo**, que según Debreu y Scarf (1963) converge a los equilibrios competitivos al crecer el número de agentes.

En el plano u^A-u^B, la curva de contratos da la **FPU**, de pendiente negativa. Cada óptimo en la producción tiene su FPU, así que hay infinitas.



(a) Caja de Edgeworth del consumo



(b) Frontera de posibilidades de utilidad

Caja de Edgeworth del consumo y FPU: el núcleo es el tramo de la frontera entre M y N.

1.2.3. Óptimo global (reparto del *output*)

Planteamiento: se maximiza u^A dados u^B_0, la tecnología y las dotaciones de factores.

CPO: la tasa subjetiva de intercambio coincide con la tasa objetiva de transformación:

$$\text{RMS}^A_{yx} = \text{RMS}^B_{yx} = \text{RMT}_{yx}$$

Puede haber, por tanto, asignaciones eficientes en la producción o en el consumo por separado que no sean óptimos globales.

CSO: preferencias convexas y un conjunto de producción convexo, con una FPP cóncava.

Desarrollo gráfico: la caja del consumo se inscribe en la FPP, con O_B en el punto producido. Sobre la curva de contratos, solo son óptimos globales los puntos donde la RMS iguala la RMT en O_B, y puede haber infinitos.

La **GFPU** es la envolvente de las FPU en sus óptimos globales.

En suma, el óptimo de Pareto exige igualar las RMST en la producción, las RMS en el consumo y, en el óptimo global, las RMS con la RMT. Con estas condiciones, Pareto anticipa los dos teoremas, demostrados por Arrow (1951a) y Debreu (1951).

2. Los teoremas fundamentales del bienestar

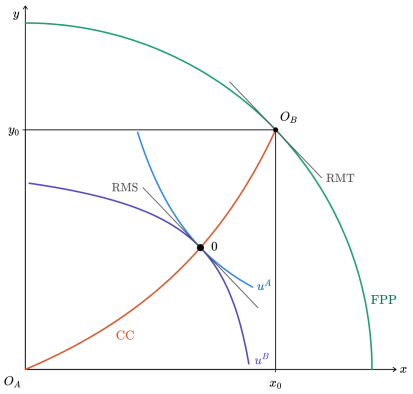
Los dos teoremas conectan el óptimo económico con el equilibrio competitivo: el mercado produce asignaciones eficientes y puede producir cualquiera de ellas si antes se redistribuyen las dotaciones.

2.1. Primer teorema (1TFEB)

Teorema: todo equilibrio competitivo $(\mathbf{p}^*, \mathbf{x}^*)$ es una asignación **eficiente en el sentido de Pareto**. Se ha visto como la formalización de la mano invisible de Adam Smith, aunque se refiere a la optimalidad paretiana y no a la social.

Corolarios:

1. *Eficiencia sin equidad*: el equilibrio no desperdicia recursos, pero puede no ser equitativo.
2. *Referencia normativa*: si se cumplen sus supuestos, la competencia perfecta lleva sin excepción a un óptimo de Pareto.



Óptimo global: la caja de Edgeworth se inscribe en la FPP y la RMT iguala a las RMS.

3. *Economía de información*: cada agente solo necesita conocer los precios.
4. *Justificación de la intervención*: si falla alguna condición suficiente, como ante fallos de mercado, el Estado tiene un motivo de eficiencia para intervenir.

Demostración: es por reducción al absurdo y usa un solo supuesto sobre las preferencias, la **no saciedad local**. Si existiera una asignación factible $\hat{\mathbf{x}}$ Pareto-superior, quien la prefiere estrictamente no pudo pagarla a los precios de equilibrio, y quien es indiferente no puede obtenerla más barata. Sumando:

$$\mathbf{p}^* \sum_h \hat{\mathbf{x}}^h > \mathbf{p}^* \sum_h \mathbf{x}^{*h}$$

Pero las empresas maximizan el beneficio y el equilibrio agota la renta, de modo que ninguna asignación factible puede valer más, por lo que $\hat{\mathbf{x}}$ no es factible. No usa convexidad ni continuidad, lo que lo hace muy general. Sí usa que todos los bienes tienen precio y mercado, que es lo que fallan los bienes públicos y las externalidades.

Conclusiones: formaliza la mano invisible con un solo supuesto, la no saciedad local, y hace de los fallos de mercado el criterio de intervención. Pero nada dice de la equidad y exige mercados completos y agentes precio-aceptantes.

2.2. Segundo teorema (2TFEB)

Queda por saber si el mercado puede alcanzar también la **optimalidad social**, el óptimo de Pareto que la sociedad prefiera por razones distributivas.

Teorema: con preferencias y conjuntos de producción convexos y una redistribución adecuada de las dotaciones iniciales, el mercado competitivo puede alcanzar cualquier óptimo de Pareto.

Corolarios:

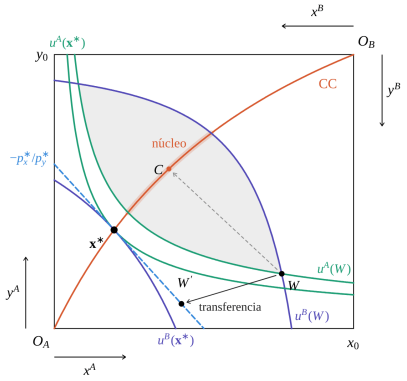
1. *Separación entre eficiencia y equidad*: se reparte con transferencias de suma fija y el mercado se encarga de la eficiencia.
2. *Planificación con precios sombra*: una planificación à la Lange-Lerner podría lograr los mismos resultados que la competencia perfecta, argumento del debate sobre el cálculo económico socialista.

Desarrollo gráfico: en la caja de Edgeworth del consumo, el intercambio voluntario lleva de la dotación W a un punto del núcleo, como C, dentro de la lente que encierran las curvas de indiferencia que pasan por W. Si el planificador prefiere un óptimo $\mathbf{x}^*$ fuera de la lente, una transferencia de suma fija lleva la dotación a W' sin distorsionar decisiones. La recta de precios que pasa por W' y $\mathbf{x}^*$, con pendiente igual a la RMS común, es la recta presupuestaria de ambos, y $\mathbf{x}^*$ es el equilibrio competitivo desde W'.

Demostración: con preferencias convexas, el conjunto de cestas agregadas preferidas a un óptimo $\hat{\mathbf{x}}$ es convexo, y también el de producción si lo son las tecnologías. Como no se cortan en su interior, el **teorema del hiperplano separador** da unos precios que sostienen $\hat{\mathbf{x}}$ como equilibrio. Con dos consumidores y dos bienes, ese hiperplano es la recta de precios tangente en el óptimo a ambas curvas de indiferencia.

Sin convexidad falla, con preferencias no convexas o con rendimientos crecientes. Es menos general que el primero, y por eso el monopolio natural y las indivisibilidades son fallos de mercado.

Muestra que la equidad no exige renunciar al mercado. Pero casi toda transferencia distorsiona, de modo que redistribuir cuesta eficiencia, y, para Hayek, los precios sombra no pueden estimarse sin el mercado que los revela.



Segundo teorema: desde W el intercambio lleva al núcleo (C); tras la transferencia a W', la recta de precios sostiene $\mathbf{x}^*$ como equilibrio.

2.3. Medición de las pérdidas de bienestar

Si fallan los supuestos, por poder de mercado, impuestos distorsionantes, externalidades o bienes públicos, el equilibrio deja de ser óptimo. En el monopolio, el precio supera al coste marginal y la RMS deja de igualar la RMT. La economía queda en un punto interior de la frontera, y Pareto dice que es ineficiente, pero no cuánto.

Excedente: es la diferencia entre lo que los agentes están dispuestos a pagar o a recibir y lo que pagan o reciben. El del consumidor es el área bajo la demanda y sobre el precio, y el del productor, el área sobre la oferta y bajo el precio. El excedente total se maximiza en el equilibrio competitivo, y cualquier otra cantidad genera una **pérdida irrecuperable de eficiencia**. Con elasticidad ε, una distorsión proporcional τ y coste marginal constante, es aproximadamente $\frac{1}{2}\,\varepsilon\,\tau^2\,p\,X$. Crece con el cuadrado de la distorsión, y por eso muchas distorsiones pequeñas cuestan menos que una grande de igual magnitud total. La misma área mide el coste social del monopolio, el exceso de gravamen de un impuesto o el coste de un control de precios.

El excedente es el instrumento del análisis coste-beneficio, pero solo es exacto sin efectos renta, con preferencias cuasilineales, y agregarlo supone dar el mismo valor a un euro para todos.

2.4. Compensación y bienestar social

Los teoremas no eligen entre óptimos. Hay dos vías: los criterios de compensación de la London School of Economics, sin comparaciones interpersonales, y la función de bienestar social, con juicios de valor explícitos.

2.4.1. Criterios de compensación

Se formularon entre 1939 y 1950 en un debate encadenado, en el que cada propuesta respondía a los defectos de la anterior.

Kaldor (1939): una asignación es preferible si los ganadores pueden compensar potencialmente a los perdedores y seguir ganando. La compensación es hipotética, de modo que ya no se exige unanimidad, sino ganancias netas positivas. Amplía las comparaciones, pero no es inequívoco, porque desde la otra asignación podría recomendar el cambio inverso.

Hicks (1939a y 1940): es el criterio simétrico. Una asignación es preferible si los perdedores no pueden compensar a los ganadores para evitar el cambio. Tampoco es inequívoco y puede producir ordenaciones múltiples y cíclicas.

Scitovsky (1941): el criterio de Kaldor puede recomendar pasar de $\mathbf{x}$ a $\mathbf{x}'$ y después volver (la **paradoja de Scitovsky**, que solo se da si las FPU se cruzan). Propone un *criterio doble*, Kaldor y Hicks a la vez, que no evita la intransitividad con tres o más estados, como demostró Samuelson (1950).

Samuelson y Little (1950): para Samuelson, $\mathbf{x}'$ es preferible solo si su FPU es completamente exterior a la de $\mathbf{x}$, lo que es riguroso pero muy exigente, pues rara vez ocurre. Little añade al criterio de Hicks un juicio distributivo: el cambio es deseable si mejora la distribución de la renta y los perjudicados no pueden sobornar a los beneficiados. Concluye que los juicios de valor no pueden eludirse.

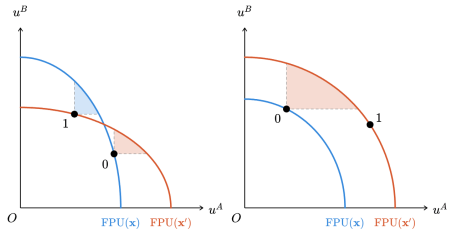
En suma, los criterios amplían la economía del bienestar sin comparaciones interpersonales explícitas y fundamentan el análisis coste-beneficio. Pero son innecesarios si Pareto ya ordena y, si no, su veredicto puede contradecirse, y llevan un juicio de valor implícito: un euro vale igual sea quien sea su dueño.

2.4.2. Función de bienestar social

Origen: Bergson (1938) y Samuelson (1947) resuelven la indeterminación paretiana haciendo explícitos los juicios de valor.

Definición: la **función de bienestar social** (FBS) asigna a cada estado social un nivel de bienestar según las utilidades individuales:

$$W = W(u^1, \dots, u^H), \qquad \partial W/\partial u^h > 0$$



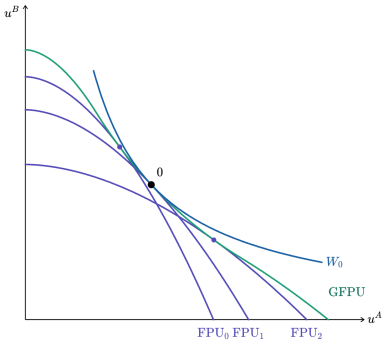
(a) Paradoja de Scitovsky (b) Criterio de Samuelson

Criterios de compensación en el plano de utilidades.

Sus curvas de indiferencia social expresan con su pendiente la tasa a la que la sociedad sacrifica el bienestar de unos por el de otros y, con su curvatura, su aversión a la desigualdad.

Óptimo social: el **optimum optimorum** es la tangencia entre la GFPU y la curva de indiferencia social más alta, y por el segundo teorema puede alcanzarse como equilibrio competitivo tras redistribuir. La forma de la FBS decide cuál se elige:

- La utilitarista, con curvas lineales, es indiferente a la distribución.
- La rawlsiana, con curvas en ángulo recto, solo atiende al peor situado.
- Las intermedias valoran el nivel y la igualdad de las utilidades.



GFPU, envolvente de las FPU, y óptimo social en su tangencia con la curva de indiferencia social W_0.

La FBS completa la teoría, pero al precio de admitir juicios de valor y comparaciones interpersonales, lo que Pareto quería evitar. De dónde proceden esos juicios es la pregunta del teorema de Arrow.

En suma, el primer teorema convierte los fallos de mercado en la justificación de la intervención por eficiencia, y el segundo separa la eficiencia de la equidad. Pero no eligen entre óptimos, y esa separación depende de transferencias de suma fija que no existen.

3. La teoría del *second best*

Los teoremas dependen de que se cumplan sus supuestos. La teoría del *second best* estudia qué ocurre cuando alguno no puede cumplirse.

3.1. El teorema del *second best*

Origen: Viner (1950) y Meade (1955b) advirtieron, al estudiar las uniones aduaneras, que eliminar una distorsión cuando persisten otras puede empeorar el bienestar. Lipsey y Lancaster (1956) lo generalizaron a un teorema.

Idea: la economía real se aleja de la competencia perfecta por bienes públicos, información imperfecta, poder de mercado o impuestos distorsionantes, y la teoría busca los criterios de eficiencia cuando alguna no puede eliminarse.

Teorema: si una de las condiciones de optimalidad paretiana, las del *first best*, no puede cumplirse, satisfacer las demás deja, en general, de ser deseable. El óptimo alcanzable, el **second best**, suele exigir violar también las demás.

3.1.1. Supuestos

Se parte del problema del planificador generalizado a n bienes, con objetivo $u(X)$ y función de transformación G, que da la igualdad entre RMS y RMT. Se añade una restricción que impide cumplirla en el bien k, cuyo precio se aparta del coste marginal en una proporción δ que no puede corregirse. Normalizando G para que su multiplicador sea $\lambda = 1$ en el *first best*, la restricción solo distorsiona si $\delta \neq 1$.

3.1.2. Modelo

Planteamiento: con los multiplicadores λ y μ, el planificador resuelve:

$$\max_X \; u(X) \quad \text{s.a.:} \; \begin{cases} G(X) = 0 \\ u'_k(X) - \delta\, G'_k(X) = 0 \end{cases}$$

CPO: sin la restricción, $\text{RMS}_{ji} = \text{RMT}_{ji}$ para todo par de bienes. Con

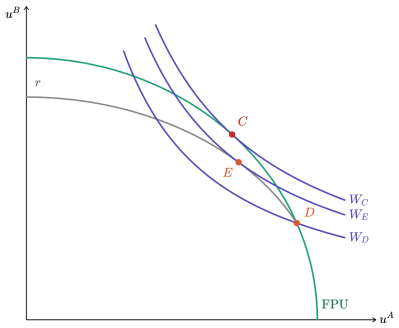
ella, para los bienes no afectados:

$$\frac{u'_i}{u'_j} = \frac{G'_i + \mu\,(u''_{ik} - \delta G''_{ik})/\lambda}{G'_j + \mu\,(u''_{jk} - \delta G''_{jk})/\lambda}$$

La RMS difiere de la RMT salvo que las **derivadas cruzadas** se anulen. La distorsión en k se transmite a los demás sectores por las relaciones de sustitución y complementariedad, y el óptimo restringido exige apartar también sus precios del coste marginal, en una cuantía sin signo general.

CSO: se supone un óptimo restringido interior y un conjunto factible convexo. Lipsey y Lancaster se limitan a las condiciones necesarias.

Desarrollo gráfico: en el plano u^A-u^B, sin restricción el *first best* es C, la tangencia de W_C con la FPU. La restricción r hace inalcanzable C. Dentro de la parte accesible, D, sobre la FPU, cumple las demás condiciones de Pareto, pero E, ineficiente, alcanza una curva de indiferencia social más alta, W_E. El *second best* es E.



Teorema del *second best* (Lipsey y Lancaster, 1956): con la restricción r, la asignación ineficiente E supera a la eficiente D.

Caso: con dos bienes de coste marginal constante y un monopolio no regulable en x que fija $p_x = \delta \cdot \text{Cmg}_x$, con $\delta > 1$, fijar $p_y = \text{Cmg}_y$ hace que los consumidores sustituyan x por y más de lo eficiente. El *second best* es $p_y = \delta \cdot \text{Cmg}_y$, que restablece el precio relativo eficiente, porque una distorsión en y compensa la de x, como en la regulación de empresas públicas que compiten con sectores no competitivos.

3.1.3. Corolarios

1. *Distorsiones compensatorias*: una distorsión inevitable puede hacer deseable mantener otras o introducirlas. Si puede eliminarse y no persisten otras, eliminarla es siempre mejor.
2. *Ruptura con la teoría tradicional*: cumplir todas las condiciones de óptimo al alcance puede ser contraproducente.
3. *Irrelevancia del recuento*: cumplir más condiciones no garantiza más bienestar.
4. *Ambigüedad de las reformas parciales*: suprimir una sola distorsión puede subir, bajar o dejar igual el bienestar.

3.2. Implicaciones para la política económica

Idea: no hay resultados generales. Sin todas las condiciones de optimalidad, ninguna política parcial, *piecemeal policy*, garantiza una mejora, y cada intervención ha de evaluarse en su contexto.

Regulación y empresa pública: si una empresa pública convive con sectores distorsionados, fijar $p = \text{Cmg}$ deja de ser óptimo para el conjunto. La empresa de rendimientos crecientes obligada a cubrir costes tiene como *second best* los precios de Ramsey-Boiteux. Las tarifas en dos partes y los precios pico-valle mantienen $p = \text{Cmg}$ en el margen.

Imposición distorsionante: un segundo impuesto crea un segundo exceso de gravamen, pero puede contrarrestar el sesgo del primero, de modo que la imposición óptima no minimiza cada distorsión por separado. Los impuestos de suma fija, los únicos no distorsionantes, son difíciles de aplicar y, si son uniformes por persona, regresivos, como mostró el fracaso del *poll tax* británico de 1990.

Comercio internacional: es el origen de la teoría. Una unión aduanera suprime los aranceles entre sus miembros, pero los mantiene frente a terceros, y su efecto neto es ambiguo:

- la *creación de comercio* sustituye producción nacional cara por importaciones baratas del socio;
- la *desviación de comercio* sustituye importaciones baratas de terceros por importaciones caras del socio.

Evaluación de proyectos: sin teoremas generales, cada caso se evalúa por separado, y el análisis coste-beneficio, con precios sombra que incorporan las distorsiones, pasa a ocupar el centro de la política económica.

3.3. Desarrollos posteriores

Argumento por bloques: una primera respuesta divide las condiciones en tres bloques, intercambio, producción y global, y sostiene que una distorsión en uno no obliga a revisar los otros. Producir sobre la FPP, por ejemplo, amplía las asignaciones alcanzables. No es general, porque los bloques están ligados por precios y cantidades. Tres desarrollos lo sustituyen por condiciones precisas o reglas prácticas.

3.3.1. Separabilidad (Davis y Whinston, 1965)

Idea: si el sector distorsionado es **separable** del resto, es decir, si las derivadas cruzadas de la utilidad y de la producción con los demás son nulas, la distorsión no se transmite. Los términos que apartan la RMS de la RMT desaparecen y las condiciones de Pareto se recuperan para los demás pares de bienes.

La política parcial recupera así su validez cuando las interdependencias son débiles, como entre sectores lejanos en la cadena de producción y en las preferencias, y se fundamentan el equilibrio parcial y las reglas sectoriales. Pero las derivadas cruzadas nulas son un supuesto fuerte, rara vez exacto y difícil de contrastar.

3.3.2. Eficiencia productiva (Diamond y Mirrlees, 1971)

Idea: Diamond y Mirrlees (1971) obtienen el resultado más importante en sentido contrario al de Lipsey y Lancaster. Aunque haya impuestos distorsionantes sobre los consumidores, la **eficiencia productiva** sigue siendo deseable: la economía debe producir sobre su FPP.

Razonamiento: si el Gobierno puede gravar todas las transacciones con los consumidores y controlar los beneficios de las empresas, cualquier distorsión en la producción puede sustituirse por otra en el consumo que deje igual a los consumidores y recaude más. Las distorsiones deben concentrarse en el consumo.

Por eso la imposición óptima grava los bienes finales y no los intermedios, las empresas públicas usan los mismos precios sombra que las privadas y no se ponen aranceles a los bienes intermedios. Este resultado es la razón teórica del IVA frente a los impuestos en cascada. Si los beneficios puros no pueden gravarse plenamente, puede ser óptimo distorsionar también la producción.

3.3.3. *Third best* (Ng, 1977)

Idea: calcular el *second best* exige conocer las derivadas cruzadas de todos los sectores con el distorsionado. Si el regulador carece de esa información o de capacidad administrativa, ha de buscar un *third best*.

Razonamiento: con mucha incertidumbre sobre las interdependencias, aplicar las reglas del *first best* en los sectores controlables tiene un valor esperado mayor que aproximar un *second best* mal calibrado. Rehabilita así las reglas del *first best* como reglas prácticas y anticipa la teoría de la regulación con información asimétrica. Pero no en general: con buena información, el *second best* sigue siendo preferible.

En suma, el *second best* muestra que las condiciones de optimalidad solo son deseables si se cumplen todas y que la política económica no tiene atajos. Fundamenta los precios de Ramsey-Boiteux, la imposición óptima y la teoría de las uniones aduaneras. Su límite es que no dice por sí solo qué hacer y puede justificar cualquier distorsión, aunque los desarrollos posteriores acotan cuándo valen las reglas del *first best*.

1. Condiciones de eficiencia en una economía competitiva

Arrow (1951a) y Debreu (1954) probaron que, con competencia perfecta, el mecanismo de precios conduce a un equilibrio general óptimo en el sentido de Pareto. Los precios son entonces señales completas: miden la escasez relativa, informan sobre tecnología y preferencias y no dejan fuera ningún coste ni ninguna utilidad.

1.1. Supuestos y condiciones en una economía 2x2x2x2

1.1.1. Supuestos

Entorno: competencia perfecta en todos los mercados de bienes y factores, es decir, mercados completos, agentes precio-aceptantes e información completa.

Agentes: dos empresas, cada una especializada en un bien, y dos consumidores cuya utilidad depende solo de su propio consumo y es creciente y cuasicóncava.

Tecnología: dos bienes privados, x e y, rivales y excluibles, producidos con dos factores, K y L, de oferta perfectamente inelástica. Cada función de producción depende solo de los factores de su empresa, sin externalidades, y tiene productividades marginales positivas y decrecientes y rendimientos no crecientes, de modo que la FPP es cóncava.

1.1.2. Modelo

Producción: los productores igualan la RMST al precio relativo de los factores, que es el mismo para todos:

$$\text{RMST}^x_{KL} = \frac{f'_{L_x}}{f'_{K_x}} = \frac{f'_{L_y}}{f'_{K_y}} = \text{RMST}^y_{KL} = \frac{w}{r}$$

La condición queda recogida en la FPP, cuya pendiente en valor absoluto es la RMT:

$$\text{RMT}_{yx} = \frac{\text{Cmg}_x}{\text{Cmg}_y} = \frac{f'_{L_y}}{f'_{L_x}} = \frac{f'_{K_y}}{f'_{K_x}}$$

Consumo: los consumidores igualan su RMS al precio relativo de los bienes:

$$\text{RMS}^A_{yx} = \frac{u'^A_x}{u'^A_y} = \frac{u'^B_x}{u'^B_y} = \text{RMS}^B_{yx} = \frac{p_x}{p_y}$$

Global: como $w = p_x f'_{L_x} = p_y f'_{L_y}$, el precio relativo iguala la RMT, y la tasa subjetiva de intercambio coincide con la objetiva de transformación:

$$\text{RMS}^h_{yx} = \frac{p_x}{p_y} = \text{RMT}_{yx}$$

Guiado por una *mano invisible*, cada individuo, persiguiendo su interés, asegura una asignación globalmente eficiente (la "tarta" alcanza su tamaño máximo, aunque no necesariamente se maximiza el bienestar social).

1.2. Teoremas fundamentales del bienestar

Primer teorema: todo equilibrio competitivo es una asignación **eficiente en el sentido de Pareto**.

Segundo teorema: cualquier óptimo de Pareto puede alcanzarse como equilibrio competitivo tras redistribuir las dotaciones con transferencias de suma fija, siempre que preferencias y tecnologías sean convexas.

Lo que importa aquí son sus supuestos: bienes privados, rivales y excluibles; ausencia de efectos externos; agentes precio-aceptantes, e información completa. **Cada fallo de mercado es la violación de uno de ellos**, y el sector público interviene para restablecer las condiciones de eficiencia que el fallo impide cumplir.

En suma, las condiciones y los teoremas dan el criterio para identificar los fallos de mercado. Pero nada dicen sobre la equidad, descansan en mercados completos y el segundo exige convexidad y transferencias de suma fija que en la práctica no existen.

2. Teoría de los fallos de mercado

Los **fallos de mercado** hacen que las decisiones descentralizadas no conduzcan a la eficiencia. La teoría nace con Pigou (1920) y se desarrolla con Meade (1952) y Samuelson (1954a). Un fallo es necesario para intervenir por eficiencia, pero no suficiente, porque solo da un argumento *prima facie*. Hay cuatro: bienes públicos, externalidades, poder de mercado e información imperfecta.

2.1. Bienes públicos

2.1.1. Concepto y tipos

La naturaleza de un bien depende de dos propiedades independientes.

Rivalidad: un bien es rival si la unidad que consume uno ya no está disponible para otro, y *no rival* si todos pueden disfrutarlo a la vez. Atender a un consumidor más tiene entonces coste marginal nulo.

Exclusión: un bien es excluible si dejar fuera a quien no paga es viable y no demasiado caro. Si no lo es, nadie tiene incentivos a pagar por lo que puede consumir gratis, y este **problema del free rider**, o gorrón, puede hacer desaparecer la oferta privada.

Clasificación: cruzando ambas propiedades salen cuatro categorías.

1. *Bienes privados*: rivales y excluibles, que el mercado asigna eficientemente.
2. *Bienes públicos puros* (Samuelson, 1954a): no rivales y no excluibles, como la defensa nacional, el alumbrado público o el conocimiento básico.
3. *Bienes de club* (Buchanan, 1965): excluibles, pero no rivales hasta la congestión. Se financian con cuotas, y el problema es el tamaño óptimo del club, donde el beneficio marginal de un nuevo socio iguala el coste de congestión que impone.
4. *Bienes comunales*: rivales pero no excluibles, como los caladeros o los acuíferos. Sufren la *tragedia de los comunes* (Hardin, 1968): cada usuario obtiene todo el beneficio de una unidad más de uso y soporta solo una fracción de su coste, y el recurso se sobreexplota.

De las dos propiedades, la relevante para la naturaleza del bien es la no rivalidad, porque altera las condiciones de eficiencia, mientras que la excluibilidad determina si la provisión puede ser privada. Por su ámbito geográfico, los bienes públicos pueden ser, además, locales, regionales o globales.

2.1.2. Provisión eficiente (Samuelson, 1954a)

Idea: Samuelson (1954a) muestra que, como todos consumen la misma cantidad del bien público, su valoración social es la suma de las de todos, y esa suma ha de igualar el coste marginal.

Supuestos: dos consumidores, A y B, un bien público g y uno privado x, con información perfecta y una FPP, $T(x, g) = 0$. Las preferencias son cuasilineales, $u^h = v^h(g) + x^h$, de modo que $\text{RMS}^h_{xg} = v^{h\prime}(g)$. Por la no rivalidad, $g^A = g^B = g$, mientras que el bien privado se reparte, $x^A + x^B = x$.

Planteamiento: el planificador maximiza $u^A(x^A, g)$ manteniendo $u^B = u^B_0$, sujeto al reparto del bien privado, a la no rivalidad y a la FPP:

$$\max_{x^A, x^B, g} u^A(x^A, g) \quad \text{s.a.:} \begin{cases} u^B(x^B, g) = u^B_0 \\ x^A + x^B = x \\ T(x, g) = 0 \end{cases}$$

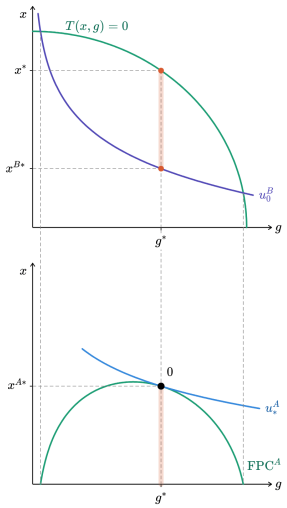
CPO: resulta la **condición de Samuelson**: el bien público se produce hasta que la suma de lo que una unidad más vale para cada usuario coincide con lo que cuesta en bien privado sacrificado:

$$\sum_h \text{RMS}^h_{xg} = \text{RMT}_{xg}$$

Con un bien privado, cada consumidor ajusta su cantidad hasta igualar su valoración marginal al coste, mientras que con el público la cantidad es común y se suman las valoraciones, de modo que en el óptimo, en general, $\text{RMS}^A_{xg} \neq \text{RMS}^B_{xg}$.

CSO: valoraciones marginales decrecientes y una FPP cóncava.

Desarrollo gráfico: restando en vertical la curva de indiferencia u^B_0 a la FPP se obtiene la frontera de posibilidades de consumo de A. Su tangencia con la curva de indiferencia más alta de A, en el punto 0, da $\text{RMS}^A = \text{RMT} - \text{RMS}^B$, que es la condición de Samuelson.



Condición de Samuelson: la frontera de consumo de A es tangente a su curva de indiferencia más alta en el punto 0.

La suma vertical de las demandas, frente a la horizontal de los bienes privados, es el fundamento del análisis coste-beneficio de los proyectos públicos. Pero la condición dice cuánto producir, no cómo. Exige conocer unas valoraciones que nadie tiene incentivos a revelar, no dice cómo repartir el coste y, si la financiación es distorsionante, deja de ser la condición pura (Atkinson y Stern, 1974).

2.1.3. Provisión privada (Bergstrom, Blume y Varian, 1986)

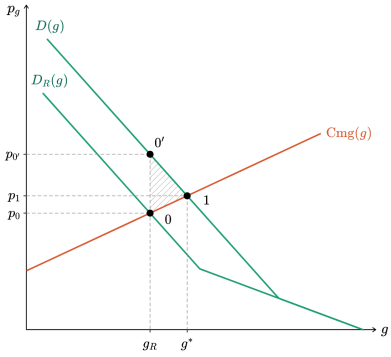
Idea: si la provisión se deja a las contribuciones voluntarias, la cantidad queda por debajo de la eficiente. Prolonga la lógica de la acción colectiva de Olson (1965), y el *free rider* es el reverso de la tragedia de los comunes.

Razonamiento: cada contribuyente aporta hasta que su valoración marginal iguala el coste, $\text{RMS}^h_{xg} = \text{RMT}_{xg}$, como con un bien privado. Como cada RMS^h es menor que la suma, la cantidad provista es inferior a la eficiente, y en el equilibrio de Nash de las contribuciones voluntarias el bien lo financia quien más lo valora y los demás se benefician sin pagar.

- Por la no rivalidad, cada contribuyente ignora el beneficio que genera a los demás.
- Por la no exclusión, nadie puede ser obligado a pagar y los consumidores tienen incentivos a ocultar su valoración.

Si el bien no es excluible, hay infraproducción. Si es de club, el fallo es el infraconsumo, porque se cobra un precio positivo por un bien de coste marginal nulo. El mercado falla en la provisión, no necesariamente en la producción, y la respuesta es que el sector público decida y financie la cantidad, aunque la produzcan empresas privadas.

Desarrollo gráfico: la provisión eficiente, g^*, es el cruce de la demanda social, suma vertical de las individuales, con el coste marginal. Con contribuciones voluntarias se revela una demanda menor, $D_R(g)$, y la provisión, g_R, es insuficiente. En g_R, la valoración social supera al coste marginal, lo que genera una pérdida de eficiencia.



Provisión del bien público: la demanda revelada, $D_R(g)$, deja la provisión g_R por debajo de la eficiente, g^*.

Neutralidad (Warr, 1983): una redistribución pequeña de la renta entre contribuyentes no altera la cantidad total, porque cada uno ajusta su aportación en lo que cambia su renta, y la provisión pública financiada con impuestos de suma fija sobre ellos, si ninguno paga más de lo que aportaba, desplaza sus aportaciones uno por uno. La teoría predice, sin embargo, menos contribuciones de las observadas, que sostienen el altruismo, las normas sociales y la interacción repetida.

2.1.4. Soluciones

Solo el sector público puede obligar a pagar, y le corresponde la provisión, aunque no necesariamente la producción. Debe resolver cómo repartir el coste y cómo conocer las valoraciones.

Financiación: con el *principio de capacidad de pago*, paga más quien más renta tiene, como en el IRPF. Redistribuye, pero distorsiona y no informa de la cantidad. Con el *principio de beneficio*, cada uno paga según lo que el bien le reporta, lo que puede resolver los dos problemas a la vez.

Principio de revelación (Gibbard, 1973; Myerson, 1979): para cualquier mecanismo existe un mecanismo directo equivalente en el que los agentes declaran sus preferencias y les conviene hacerlo verazmente cuando los demás también lo hacen. Basta, por tanto, buscar entre los compatibles con los incentivos.

Precios personalizados (Lindahl, 1919): cada consumidor paga por unidad un precio igual a su valoración marginal, $v^{h\prime}(g) = t^h$, y se provee la cantidad en que la suma de precios iguala el coste marginal:

$$\sum_{h=1}^{H} t^h = \sum_{h=1}^{H} v^{h\prime}(g^*) = \text{Cmg}_g = \text{RMT}_{xg}$$

donde, con coste marginal constante, lo recaudado, $\sum_h t^h g^*$, cubre exactamente el coste de provisión.

Esta condición es la de Samuelson, y paga más quien más valora el bien. Pero los consumidores tienen incentivos a infradeclarar, porque su precio depende de lo que anuncian.

Clarke-Groves-Vickrey (Vickrey, 1961; Clarke, 1971; Groves, 1973): cada agente declara su valoración neta, $\tilde{v}^h = v^h - c^h$, y el bien se provee si su suma es positiva. Solo paga, además de su parte, el **agente pivote**, cuya declaración cambia la decisión, la *tasa de Clarke*, igual al coste que impone a los demás:

$$T^h = \left| \sum_{j \neq h} \tilde{v}^j \right|$$

donde la regla de provisión, $\sum_h \tilde{v}^h > 0$, es el análogo discreto de la condición de Samuelson.

Como la tasa depende solo de los demás, decir la verdad es la estrategia dominante. Pero exige preferencias cuasilineales, es vulnerable a la colusión y no equilibra el presupuesto. Según Green y Laffont (1977), ningún mecanismo con decisión eficiente y sinceridad dominante lo hace en general.

Otras soluciones:

- La votación por mayoría provee la cantidad del votante mediano, que, con un reparto igualitario del coste, es eficiente solo si su valoración marginal iguala la media (Bowen, 1943).
- En los bienes públicos locales, los individuos "votan con los pies" (Tiebout, 1956). Con movilidad perfecta, muchas jurisdicciones y sin efectos desbordamiento, la provisión es eficiente.
- Los usuarios de bienes comunales pueden alcanzar acuerdos voluntarios eficientes (Ostrom, 1990). Este resultado enlaza con los juegos repetidos, en los que la cooperación se sostiene con castigos si los jugadores son suficientemente pacientes.
- El diseño de mecanismos (Hurwicz, 1972) muestra que, en general, ningún mecanismo es a la vez compatible con los incentivos, eficiente y de participación voluntaria.
- La iniciativa privada se apoya en el altruismo, como el altruismo impuro de Andreoni (1990).

2.1.5. Conclusiones

La condición de Samuelson fija un criterio preciso y la contribución voluntaria explica la infraprovisión. Pero la propia intervención será probablemente ineficiente, y como apenas hay bienes públicos puros, cuánto proveer suele estudiarse como un problema de externalidades.

2.2. Externalidades

2.2.1. Concepto y tipos

Definición: Viner (1931) distingue la *externalidad tecnológica*, en la que la actividad de un agente afecta directamente al bienestar o la producción de otro sin transacción de mercado, de la *pecuniaria*, que opera a través de los precios y, en mercados competitivos y completos, no genera ineficiencia. Meade (1952) analizó las primeras, las **externalidades reales**, con los apicultores y los fruticultores, y Arrow (1969) las reinterpretó como **mercados ausentes**.

Clasificación:

- Por el signo, negativas, como la contaminación, o positivas, como la vacunación. Las de red, en las que el valor crece con el número de usuarios, explican las plataformas digitales.
- Por los agentes, de producción o de consumo, y sobre productores o consumidores.
- Por la variable, de flujo, como las emisiones de un año, o de *stock*, como la concentración de CO_2.

Fallo: quien genera el efecto no lo internaliza, y el precio deja de orientar bien las decisiones. Las actividades que perjudican a terceros se llevan más allá de lo eficiente y las que los benefician se quedan cortas. Como el efecto no tiene mercado, la externalidad viola el supuesto de mercados completos del primer teorema.

2.2.2. Provisión eficiente en la producción

Planteamiento: a diferencia del productor, el planificador resta el daño externo, $e(x)$, que la producción impone a terceros:

$$\max_x \; p \cdot x - \text{CT}(x) - e(x)$$

CPO: el precio debe igualar el **coste marginal social**, suma del privado y del daño marginal externo, de modo que la cantidad eficiente es menor que la competitiva:

$$p = \text{Cmg}_P + \text{Cmg}_E = \text{Cmg}_S$$

CSO: $\text{CT}''(x) + e''(x) > 0$, es decir, coste marginal social creciente.

2.2.3. Provisión eficiente en el consumo

Planteamiento: la utilidad de B depende también del consumo de x por A, $u^B(x^B, y^B, x^A)$, y el planificador maximiza la de A manteniendo la de B, sujeto al reparto de los bienes y a la FPP.

CPO: para B la condición es la habitual, pero para A la valoración relevante de x es la suya más el efecto sobre B, medido en unidades de y:

$$\text{RMS}^B_{yx} = \text{RMS}^A_{yx} + \frac{\partial u^B/\partial x^A}{\partial u^B/\partial y^B} = \text{RMT}_{yx}$$

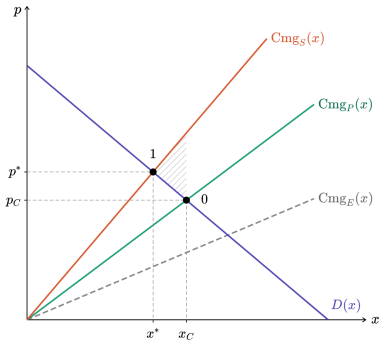
Con externalidad negativa hay sobreconsumo, y con positiva, infraconsumo.

CSO: preferencias convexas de ambos consumidores y conjunto de producción convexo.

2.2.4. Provisión privada: ineficiencia

Desarrollo gráfico: el coste marginal social es la suma vertical del privado, que es la oferta, y del daño marginal. El equilibrio competitivo, en x_C, supera al óptimo, x^*, donde el coste marginal social corta la demanda, y la pérdida es el triángulo entre ambos. Con externalidad positiva hay infraproducción. En el óptimo, el precio suma el coste marginal privado y el daño marginal, ambos evaluados en la cantidad eficiente:

$$p^* = \text{Cmg}_P(x^*) + e'(x^*)$$



Externalidad negativa en la producción: el óptimo exige igualar el coste marginal social, no el privado, al beneficio marginal.

2.2.5. Soluciones

El sector público puede actuar sobre los precios, sobre los derechos de propiedad o sobre las cantidades.

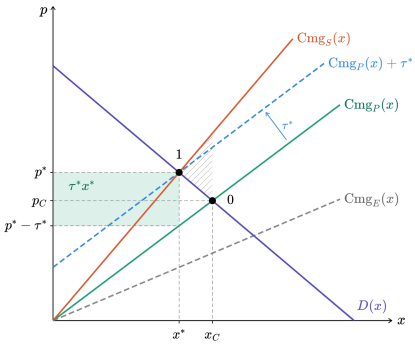
Instrumentos pigouvianos (Pigou, 1920): Pigou identifica la divergencia entre el producto neto marginal social y el privado y propone la **internalización**: crear el precio que falta, con un impuesto por unidad igual al daño marginal externo o una subvención igual al beneficio externo. Con un impuesto unitario τ, la empresa gravada maximiza:

$$\max_x \ \pi = p \cdot x - \text{CT}(x) - \tau \cdot x$$

- *CPO*: el precio iguala el coste marginal privado más el impuesto, y el impuesto que restablece el óptimo es el daño marginal en la cantidad eficiente:

  $$p = \text{Cmg}_P(x) + \tau, \qquad \tau^* = e'(x^*)$$

  Sus efectos son los mismos si recae sobre la producción o sobre el consumo.
- *Estática comparativa*: el impuesto desplaza hacia arriba el coste marginal privado y el equilibrio pasa de x_C a x^*. Desaparece la pérdida y se recauda $\tau^* x^*$.

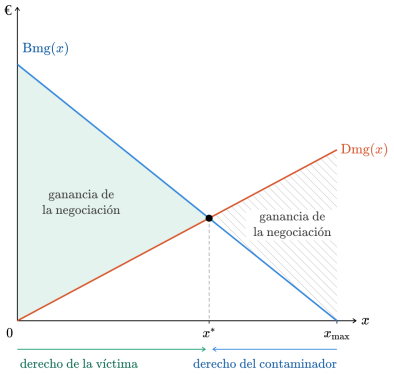


Impuesto pigouviano: $\text{Cmg}_P + \tau^*$ corta la demanda donde lo hace el coste marginal social.

- *Doble dividendo*: si la recaudación reduce otros impuestos distorsionantes, mejora también el sistema fiscal, aunque la interacción con los impuestos existentes tiende a reducir esa ganancia.
- *Conclusiones*: actúa a través del mercado, genera ingresos y deja a cada agente elegir cómo reducir, lo que minimiza el coste. Se aplica en los impuestos sobre carburantes y emisiones y en las tasas sobre plásticos y envases. Pero exige conocer el daño marginal, su efecto es reducido con demanda rígida y, con otras distorsiones, es un problema de *second best*. Con un contaminador monopolista, el impuesto óptimo es inferior al daño marginal (Buchanan, 1969).

Teorema de Coase: en *The Problem of Social Cost* (1960), Coase sostiene que el problema no es la falta de un impuesto, sino de derechos de propiedad sobre el efecto externo. El sector público solo los define y garantiza, y la solución es descentralizada, pero no privada.

- *Teorema* (nombre de Stigler, 1966): con costes de negociación nulos, información perfecta y sin efectos renta, la cantidad pactada es la misma, y eficiente, sea cual sea el reparto inicial de los derechos, que solo cambia la distribución de la renta.
- *Razonamiento*: si el derecho es de la piscifactoría, la acería paga por contaminar mientras su beneficio marginal neto supere el daño marginal. Si es de la acería, la piscifactoría paga por reducir mientras el daño supere ese beneficio. En ambos casos se llega a x^*, donde Bmg iguala Dmg. La compensación por la última unidad iguala el daño marginal en ese punto. Con el derecho de la víctima, la negociación parte de $x = 0$ y avanza hasta x^*, y con el del contaminador, parte de $x_{\max}$ y retrocede. La ganancia de negociar es el área entre ambas curvas a uno u otro lado de x^*.



Teorema de Coase: el acuerdo se alcanza en x^* cualquiera que sea la asignación del derecho.

- *Conclusiones*: para Pigou el fallo está en el precio, y el regulador debe conocer el daño, mientras que para Coase está en los derechos, y las partes revelan sus valoraciones al negociar. La elección es una comparación de costes de transacción. Pero el teorema exige pocas partes identificables y preferencias cuasilineales, e ignora los costes de crear el mercado y la negociación estratégica.

Cantidades y otras soluciones:

- *Mercados de derechos de emisión* (*cap and trade*): el sector público fija un techo de contaminación y reparte derechos que se negocian. Si el techo es la cantidad eficiente, el precio equivale al impuesto pigouviano y reducen quienes pueden hacerlo a menor coste. En la UE, el RCDE UE funciona desde 2005, con techo único y subasta como método general desde 2013, y la reforma de 2023 elimina progresivamente la asignación gratuita en la aviación y, desde 2026, en los sectores del mecanismo de ajuste en frontera por carbono.
- *Precio o cantidad* (Weitzman, 1974): si el regulador desconoce los costes de reducción, el impuesto es preferible cuando el beneficio marginal de reducir varía poco en relación con el coste marginal; con umbrales críticos, el techo.
- *Otras soluciones*:
  - La regulación directa es la única practicable con daños catastróficos o irreversibles, pero ignora las diferencias de coste y no incentiva reducir por debajo del límite. Fijar el límite exige, además, conocer con precisión el efecto externo.
  - La integración de las empresas internaliza el efecto sin negociar.
  - La reasignación de competencias a un ámbito mayor y la responsabilidad civil, que equivale a un impuesto pigouviano fijado por los tribunales. Tiene la ventaja de calcularse sobre el daño efectivo y el inconveniente de los costes del litigio.
  - Los mecanismos de compensación de Varian (1994) y las normas sociales.

2.2.6. Conclusiones

Todas las soluciones buscan que quien genera la externalidad internalice su coste, con la misma disyuntiva que los bienes públicos: los instrumentos centralizados exigen información al regulador; los descentralizados, negociación barata.

2.3. Poder de mercado e información imperfecta

Poder de mercado: la empresa iguala el ingreso marginal, y no el precio, al coste marginal:

$$\text{Img}(x) = \text{Cmg}(x) \ \Leftrightarrow \ p = \frac{1}{1 - 1/|\varepsilon_{xp}|} \cdot \text{Cmg}(x)$$

Así, la RMS deja de igualar la RMT y se produce menos de lo eficiente, con la pérdida del triángulo de Harberger. Las soluciones *first best* buscan restablecer $p = \text{Cmg}$, con regulación de precios o defensa de la competencia, y las *second best* acotan el margen cuando eso no es posible.

Información imperfecta: surgen la selección adversa y el riesgo moral, que violan los mercados completos. Según Greenwald y Stiglitz (1986), el equilibrio ni siquiera es eficiente en sentido restringido, porque hay intervenciones que mejoran a todos aun con la información limitada del sector público, como las obligaciones de información, la regulación de la calidad o los seguros obligatorios.

Bienes públicos globales: el clima, las pandemias o la estabilidad financiera combinan los fallos a escala internacional. Sin un sector público mundial, dependen de la cooperación entre países.

En suma, cada fallo viola un supuesto del primer teorema, y los instrumentos se distinguen por la información que exigen. Todos suponen, sin embargo, un sector público benevolente.

3. Elección pública: los fallos del sector público

Que el mercado falle no basta para justificar la intervención, porque también el sector público puede fallar.

3.1. Origen y concepto

Origen: la **teoría de la elección pública**, con precursores como Wicksell (1896), Black (1948) y Downs (1957), se consolida con Buchanan y Tullock (1962) como análisis económico de las decisiones políticas.

Idea: políticos y burócratas responden a incentivos, como cualquier agente, y las instituciones públicas son el marco que fija esos incentivos. Frente al planificador benevolente, el sector público es una unidad de decisión imperfecta, y los fallos de mercado tienen su réplica en los **fallos del sector público**, que se analizan igual: causas, efectos y soluciones.

3.2. Limitaciones al diseño y la implementación

3.2.1. Grupos de interés (Olson, 1965; Stigler, 1971)

Origen: Olson (1965) explica por qué los grupos pequeños y concentrados se organizan mejor que los grandes y difusos, y Stigler (1971) aplica esa lógica a la regulación.

Idea: los grupos de interés interfieren en las decisiones en favor de intereses privados. Como el beneficio de una decisión favorable se concentra en pocos y su coste se reparte entre muchos, los primeros se organizan y presionan, y los segundos no, por lo que la acción colectiva la ejercen las minorías organizadas.

- *Búsqueda de rentas* (Tullock, 1967; Krueger, 1974): los grupos gastan recursos en obtener privilegios, como aranceles, licencias o subvenciones, y esos recursos se desvían de la producción y se suman al coste social de la intervención.
- *Captura del regulador*: la agencia acaba defendiendo a los regulados, que tienen la información y el interés concentrado. La regulación pasa así de proteger a los consumidores a proteger a las empresas frente a la competencia.
- *Triángulo de hierro*: la alianza entre el grupo, la burocracia del programa y los legisladores que lo supervisan. Los tres ganan con su continuidad y nadie tiene incentivos a evaluarlo.

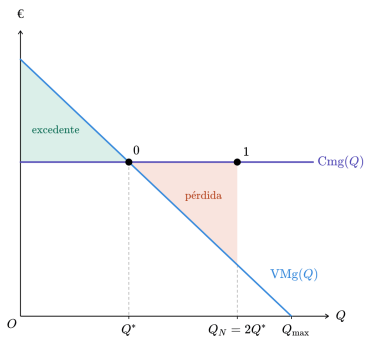
La intervención frente a un fallo de mercado puede acabar sirviendo a quienes debía controlar, y la respuesta es el diseño institucional: independencia, transparencia y evaluación externa.

Explica por qué la intervención favorece a minorías organizadas y funda la teoría económica de la regulación, pero le cuesta explicar la desregulación y la influencia es difícil de medir.

3.2.2. Burocracia (Niskanen, 1971)

Idea: el burócrata maximiza el presupuesto de su departamento, del que derivan su poder y su retribución, y tiene una ventaja informativa sobre los costes.

Razonamiento: departamento y patrocinador forman un monopolio bilateral, y el departamento presenta una oferta de todo o nada. La producción supera la eficiente: si la restricción activa es el presupuesto, llega hasta donde el beneficio total iguala el coste total; con demanda lineal y coste marginal constante, el doble de la cantidad eficiente. Si no, llega hasta donde la valoración marginal del patrocinador se anula. En ambos casos, la intervención tiende a ser mayor de lo necesario.



Burocracia maximizadora del presupuesto (Niskanen, 1971).

Migué y Bélanger (1974) sustituyen el presupuesto por el *presupuesto discrecional*, y la ineficiencia toma la forma de costes excesivos. Explica el crecimiento del sector público desde la oferta, pero supone un patrocinador pasivo y es difícil de contrastar.

3.2.3. Ciclo político (Nordhaus, 1975)

Idea: el Gobierno maximiza votos y explota la curva de Phillips a corto plazo. Con votantes miopes y expectativas adaptativas, una expansión antes de las elecciones reduce el paro cuando se vota, y la inflación se corrige con un ajuste al inicio de la legislatura. Las políticas se deciden así por su rendimiento electoral y no por su eficiencia, lo que introduce una inestabilidad ajena a los fallos de mercado que pretenden corregir.

Justifica bancos centrales independientes y reglas fiscales, pero es incompatible con las expectativas racionales e ignora la ideología que recogen los modelos partidistas. Además, la evidencia muestra ciclos más en el presupuesto que en el paro.

3.2.4. Aversión al riesgo (Stiglitz, 1986)

Idea: el burócrata soporta el coste de los fracasos y no se apropia de los éxitos, de modo que es más averso al riesgo de lo que conviene, lo que frena la innovación. Explica el apego de la Administración a los procedimientos como respuesta racional a sus incentivos, no como un rasgo de las personas, y orienta las reformas hacia la evaluación por resultados. Pero es difícil de contrastar y de distinguir de la prudencia ante decisiones irreversibles.

3.3. Efectos de las imperfecciones

Efectos:

1. *Ineficiencia dinámica*: la regulación puede eliminar los incentivos a innovar, sobre todo en monopolios naturales regulados por coste del servicio.
2. *Distorsión de incentivos*: la intervención altera los precios relativos, y las ganancias en un sector pueden perderse en otros, que es la lógica del *second best*.
3. *Sobredimensionamiento*: por la **ilusión fiscal** (Puviani, 1903), los votantes infravaloran la carga que soportan, por las retenciones, la imposición indirecta o la ilusión de la deuda, y demandan un sector público mayor. Contribuyen también la opacidad del sistema tributario y la rémora fiscal.

3.4. Soluciones a los fallos del sector público

Sobre los incentivos: se trata de que burócratas y políticos soporten las consecuencias de sus decisiones.

- Retribución ligada a la productividad, que rompe la relación entre presupuesto y retribución de Niskanen.
- Supervisión, transparencia y rendición de cuentas, que reducen la ventaja informativa.
- Competencia por comparación entre unidades y libertad de elección de los usuarios.

Sobre las reglas: se trata de limitar de antemano la discrecionalidad, en la línea de la economía constitucional de Buchanan.

- Agencias independientes con mandato definido, aisladas del ciclo político y de la captura.
- Evaluación obligatoria, con análisis coste-beneficio *ex ante* y evaluación *ex post*.
- Presupuesto de base cero, contra la inercia del triángulo de hierro.
- Separación entre provisión y producción, con licitaciones, concesiones o cheques.
- Reglas constitucionales sobre déficit y deuda, que sustraen esas magnitudes a cada legislatura.

Su lógica es la de las soluciones a los fallos de mercado: alinear los incentivos de quien decide con el interés colectivo, bien haciéndole soportar los costes de sus decisiones, bien impidiendo que las tome.

En suma, la elección pública somete la intervención a una prueba adicional: la **comparación de instituciones** (Demsetz, 1969), entre el fallo de mercado y el fallo probable de la intervención, frente al enfoque que compara el mercado real con un Estado ideal. Completa así el argumento, pues los fallos de mercado justifican la intervención y los del sector público la acotan. Su límite es que el interés propio explica mal la participación electoral, la ideología y el servicio público, y puede explicar *a posteriori* casi cualquier decisión.

A24. Economía del bienestar (III). Las funciones de bienestar social. Teoría de la elección colectiva. El teorema de imposibilidad de Arrow y desarrollos posteriores

1. La función de bienestar social

Ordenar los estados sociales de forma completa y transitiva obliga a sopesar lo que ganan unos frente a lo que pierden otros, de modo que la economía del bienestar no puede eludir las comparaciones interpersonales de utilidad.

1.1. Criterio original (Bergson, 1938; Samuelson, 1947)

Origen: Bergson (1938) introduce la **función de bienestar social** (FBS) para hacer explícitas las disyuntivas entre individuos, y Samuelson (1947) la incorpora al núcleo de la economía del bienestar, sin comprometerse con una forma concreta.

Idea: la FBS representa un código de valores colectivos y permite elegir, entre los infinitos óptimos de Pareto, el que maximiza el bienestar social.

1.1.1. Definición y propiedades

Definición: es una función definida sobre las utilidades de los H individuos que refleja las preferencias sociales:

$$W = W(\mathbf{u}), \qquad \mathbf{u} = (u^1, \dots, u^H)$$

Propiedades:

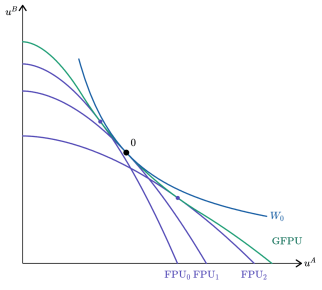
1. *Continuidad y diferenciabilidad*: garantizan un máximo caracterizable por CPO, lo que no ocurre con la *maximin*, no diferenciable en el vértice.
2. *Transformaciones*: exige utilidades comparables entre personas; la utilitarista, cardinales con unidades comparables, y la *maximin*, solo niveles comparables. La FBS resultante es ordinal. Cualquier transformación monótona creciente de W representa, por tanto, los mismos valores colectivos.
3. *Inclinación*: es creciente en cada utilidad, $\partial W/\partial u^h > 0$ (benevolencia), y las mejoras paretianas aumentan el bienestar social. La *maximin* solo es no decreciente, porque una mejora paretiana que no llegue al peor situado deja igual el bienestar social.
4. *Forma*: si la aversión a la desigualdad es deseable, es cuasicóncava, con curvas de indiferencia social convexas. A diferencia de la concavidad, la cuasiconcavidad se conserva ante transformaciones monótonas.
5. *Simetría*: si se desea el anonimato, intercambiar las utilidades de dos individuos no altera el bienestar. Con la cuasiconcavidad, una distribución más igualitaria con la misma suma nunca es peor.

Hereda, además, las propiedades paretianas: independencia del proceso, individualismo, no paternalismo y benevolencia.

1.1.2. Razonamiento

Curvas de indiferencia social (CIS): son las curvas de nivel de la FBS en el plano de utilidades, decrecientes, y su curvatura refleja la aversión a la desigualdad. En la función original son débilmente convexas y sin forma uniforme. La tasa a la que la sociedad sacrifica el bienestar de unos por el de otros decrece conforme la distribución se hace más desigual a favor de estos, pero no a un ritmo fijo.

Punto de felicidad: el *optimum optimorum* es la tangencia entre la CIS más alejada y la gran frontera de posibilidades de utilidad (GFPU). La eficiencia paretiana es necesaria, pero no suficiente, porque un punto ineficiente, dentro de la frontera, puede dar más bienestar que uno eficiente. El punto de felicidad determina la asignación Pareto-óptima que lo genera.



Gran frontera de posibilidades de utilidad y punto de felicidad.

1.1.3. Conclusiones

La FBS resuelve la indeterminación con un juicio de valor explícito y transparente y fundamenta la redistribución. Pero admite cualquier juicio sin decir cuál adoptar, porque no hay método para construirla y su existencia se postula axiomáticamente. El óptimo refleja así los valores de quien la especifica.

1.2. Formas de la FBS: criterios ético-lógicos

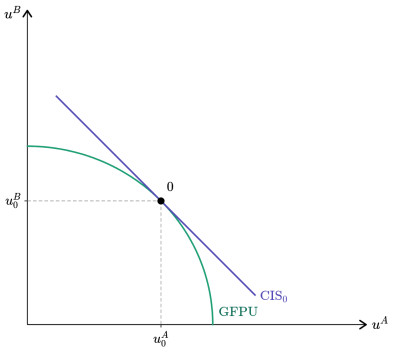
Las formas se ordenan por su aversión a la desigualdad: nula en el utilitarismo, intermedia en Bernoulli-Nash e infinita en el *maximin*.

1.2.1. Utilitarista (Bentham; Harsanyi, 1955)

Idea: el bienestar social es la suma de las utilidades, y solo importa el total, no su reparto.

Velo de la ignorancia: Harsanyi lo justifica porque un individuo que no sabe qué lugar ocupará y maximiza la utilidad esperada de von Neumann-Morgenstern, con posiciones equiprobables, elige la mayor utilidad media.

Planteamiento: con $W = \sum_h u^h(x^h)$, al repartir un bien el óptimo iguala las utilidades marginales, $u'^A = u'^B$. Si difirieran, trasladar una unidad a quien la tiene mayor elevaría la suma. Es un máximo si las utilidades son cóncavas. Las CIS son rectas de pendiente -1, porque las utilidades no se ponderan, y el óptimo es su tangencia con la GFPU.



Criterio utilitarista: CIS rectas de pendiente -1 tangentes a la GFPU en el punto 0.

Respeta el individualismo y el anonimato, pero es indiferente a la distribución: aumenta la utilidad de quien tiene mayor utilidad marginal a costa de quien la tiene menor y es compatible con grandes desigualdades.

1.2.2. Intermedio (Bernoulli; Nash, 1950)

Idea: el bienestar social es el producto de las utilidades y valora su nivel y su igualdad. Sin parámetros, su aversión a la desigualdad es unitaria en la familia de Atkinson.

Planteamiento: con $W = \prod_h u^h(x^h)$, equivalente a $\sum_h \ln u^h$, el óptimo iguala las utilidades marginales divididas por la utilidad, $u'^A/u^A = u'^B/u^B$: el peso social de cada uno es inverso a su utilidad, y la utilidad marginal de quien está peor pesa más. Con utilidades positivas y cóncavas es un máximo. Las CIS son convexas y el óptimo es su tangencia con la GFPU.

Enlaza con la solución de negociación de Nash, pero no permite graduar la aversión y exige un cero bien definido en cada utilidad. Si se desplaza el origen de alguna, el orden social cambia.

1.2.3. *Maximin* (Rawls, 1971)

Idea: en *A Theory of Justice*, el bienestar social es el del más desfavorecido. Tras el velo de la ignorancia, sin atribuir probabilidades, los individuos razonan según el *maximin* y acuerdan dos principios:

- el *principio de libertad*: el conjunto más amplio de libertades básicas compatible con las de los demás;
- el *segundo principio*: las desigualdades solo son legítimas si benefician al más desfavorecido, el *principio de diferencia*, con igualdad equitativa de oportunidades;
- y una *prioridad lexicográfica* del primero sobre el segundo.

Planteamiento: la FBS es de tipo Leontief, $W = \min_h\{u^h(x^h)\}$. No es diferenciable en el vértice y el óptimo exige $u^A = u^B$, la igualdad absoluta de utilidades. No procede de una tangencia, y su suficiencia se sigue de la cuasiconcavidad de W. Las CIS tienen forma de L, y el óptimo está donde la GFPU corta la bisectriz.

Inspira el igualitarismo liberal, pero equivale a una aversión infinita al riesgo, pues ninguna mejora de la mayoría compensa un empeoramiento mínimo del peor situado.

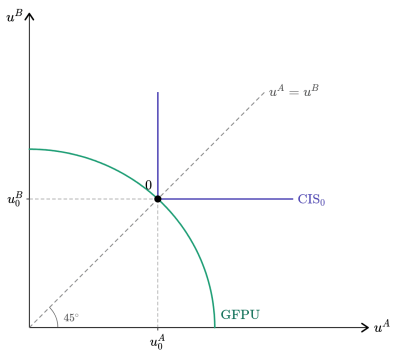
1.2.4. Familia de Atkinson (1970)

Idea: toda medida de desigualdad lleva implícita una FBS. La **renta equivalente igualmente distribuida**, y_e, es la que, repartida por igual, da el mismo bienestar que la distribución real. No supera la media, y el índice mide la pérdida relativa:

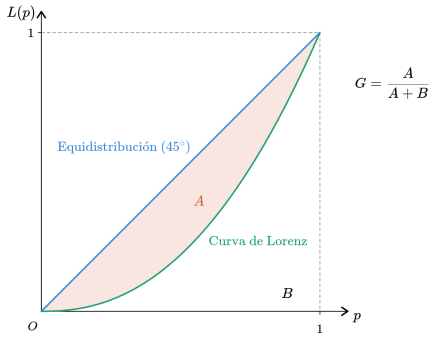
$$A_\varepsilon = 1 - \frac{y_e}{\bar{y}}, \qquad y_e = \left[\frac{1}{H}\sum_h (y^h)^{1-\varepsilon}\right]^{1/(1-\varepsilon)}$$

donde $\bar{y}$ es la renta media.

El parámetro $\varepsilon \geq 0$ es la aversión a la desigualdad: con $\varepsilon = 0$ se obtiene el utilitarismo; con $\varepsilon = 1$, la media geométrica de Bernoulli-Nash, y con $\varepsilon \to \infty$, el *maximin*. Cuanto mayor es ε, mayor es la pérdida que



Criterio *maximin* (Rawls): CIS en ángulo recto y óptimo donde la GFPU corta la bisectriz.



Curva de Lorenz e índice de Gini.

se atribuye a una misma desigualdad. A diferencia del Gini, permite elegir cuánto pesa la desigualdad y es muy usado en la evaluación de las políticas redistributivas. Pero hace explícito el juicio sin resolverlo, porque ε se fija fuera del modelo y, si las curvas de Lorenz se cruzan, la ordenación puede invertirse según el parámetro. Supone, además, una FBS aditiva y simétrica en la que la utilidad se identifica con la renta.

1.3. Más allá del bienestarismo

1.3.1. Capacidades básicas (Sen, 1980)

Idea: Sen sostiene que no importan los bienes primarios, como en Rawls, ni la satisfacción, como en el utilitarismo, sino las **capacidades**: lo que las personas pueden *hacer y ser*, como estar bien nutrido, sano o educado. Dos personas con los mismos bienes pueden tener capacidades muy distintas, por diferencias físicas, sociales o de entorno, y dos con la misma satisfacción pueden vivir en condiciones muy distintas, porque las preferencias se adaptan a las circunstancias. La FBS debe definirse, por tanto, sobre las capacidades. Fundamenta el índice de desarrollo humano, pero las capacidades no se observan directamente, lo que obliga a aproximarlas con indicadores de funcionamientos.

1.3.2. Sin envidia (Foley, 1967; Varian, 1974)

Idea: una asignación es **sin envidia** si nadie prefiere la cesta de otro, y *justa* si además es eficiente. No exige comparar utilidades. Con dotaciones iguales, el equilibrio competitivo es justo. Nadie envidia a nadie, porque todos eligieron con la misma renta, y es eficiente por el primer teorema. Muestra así que equidad y eficiencia no son incompatibles y da contenido concreto al segundo teorema. Con producción y habilidades distintas, puede no alcanzarse, porque el más productivo puede envidiar el ocio del menos productivo.

1.4. Medición de la desigualdad y la pobreza

1.4.1. Desigualdad unidimensional

Curva de Lorenz (1905): representa la proporción acumulada de renta frente a la de población, ordenada de menor a mayor renta. La igualdad perfecta es la bisectriz. Si dos curvas no se cruzan y la renta media es igual, la más cercana a la bisectriz es preferida por cualquier FBS simétrica y cóncava. Si se cruzan, depende de la FBS. Con las rentas ordenadas de menor a mayor, cada punto de la curva es:

$$L\left(\frac{k}{H}\right) = \frac{\sum_{h=1}^{k} y^h}{\sum_{h=1}^{H} y^h}, \qquad k = 0, 1, \dots, H$$

Índice de Gini (1912): es el doble del área entre la curva y la bisectriz, o la mitad de la diferencia media entre rentas en proporción a la media. Vale 0 con igualdad perfecta:

$$G = \frac{1}{2H^2\bar{y}} \sum_{h=1}^{H} \sum_{h'=1}^{H} \left| y^h - y^{h'} \right|, \qquad G \in [0, 1]$$

Vale $1 - 1/H$ si una persona lo recibe todo. Es el más usado, pero es más sensible al centro que a las colas, y su FBS implícita pondera cada renta según su posición, sin parámetro que la sociedad pueda elegir.

Índice de Theil (1967): basado en la entropía, compara la renta de cada individuo con la media y vale entre 0 y $\ln H$:

$$T = \frac{1}{H} \sum_{h=1}^{H} \frac{y^h}{\bar{y}} \ln \frac{y^h}{\bar{y}}, \qquad T \in [0, \ln H]$$

Es **descomponible** (la desigualdad total se separa en la de dentro y la de entre grupos) y permite atribuirla a regiones, sexos o niveles educativos:

$$T = \sum_{j=1}^{J} s_j T_j + \sum_{j=1}^{J} s_j \ln \frac{\bar{y}_j}{\bar{y}}$$

donde s_j es la parte de la renta total que recibe el grupo j, $\bar{y}_j$ su renta media y T_j su índice.

1.4.2. Pobreza (Sen, 1976)

Idea: medir la pobreza exige fijar una línea z, absoluta o relativa, como el 60 % de la mediana en la UE; contar a los pobres, la tasa de incidencia H_p, y medir cuánto les falta. La tasa de incidencia ignora la intensidad y la desigualdad entre los pobres. Así, una transferencia de un pobre a otro menos pobre que lo saque de la pobreza la reduce, aunque empeore al más pobre. La *monotonicidad* exige que la pobreza aumente si baja la renta de un pobre, y la *transferencia*, que aumente si se transfiere renta de un pobre a alguien más rico.

Índice de Sen: exige *monotonicidad* y *transferencia*, y pondera la brecha de cada pobre según su puesto:

$$S = H_p \left[I + (1 - I)\, G_p \right]$$

donde I es la intensidad de la pobreza y G_p el Gini de los pobres. Convierte la medición de la pobreza en un problema de bienestar social. No es descomponible por grupos, lo que motivó la familia FGT (Foster, Greer y Thorbecke, 1984), y toma como dada la línea, que es un juicio de valor previo.

1.4.3. Medición multidimensional

Si importan las capacidades, la pobreza se mide en varias dimensiones: el índice de desarrollo humano (PNUD, 1990), el índice de pobreza multidimensional (Alkire y Foster, 2011) y la tasa AROPE de la UE, que combina la pobreza relativa de renta, la privación material y social severa y la muy baja intensidad laboral del hogar. Qué dimensiones incluir y cómo ponderarlas son juicios de valor.

En suma, la FBS da el óptimo social con valores explícitos, y todo índice lleva implícita una valoración. Pero el óptimo es tecnocrático: depende de una forma funcional que ningún método permite construir.

2. Elección colectiva: el teorema de Arrow

La **teoría de la elección colectiva** se pregunta si los valores colectivos pueden obtenerse de las preferencias individuales mediante un procedimiento que las respete.

2.1. La regla de elección social

Distinción: la FBS de Bergson-Samuelson es una valoración, que asigna un número a cada estado dado un perfil de utilidades cardinales y comparables. La **regla de elección social** (RES), o *constitución*, es un procedimiento. Asocia a cada perfil posible de ordenaciones individuales ordinales y no comparables, $(R^1, \dots, R^H)$, una ordenación social completa y transitiva:

$$R = f(R^1, \dots, R^H)$$

La mayoría, la unanimidad o la regla de Borda son ejemplos. La RES es más amplia que la FBS, porque ha de funcionar para cualquier perfil, y no solo para uno dado. El conjunto de estados sociales tiene al menos tres alternativas, porque con dos el problema es trivial. Cada ordenación individual es completa y transitiva, pero dice qué se prefiere y no cuánto.

Arrow busca una RES que respete unas exigencias mínimas de racionalidad y democracia, y concluye que no existe. Las salidas a la imposibilidad consisten en relajarlas.

2.2. Axiomas de Arrow

Para Arrow, una RES es racional y democrática si cumple cinco axiomas.

1. *Universalidad*: se aplica a cualquier perfil de preferencias, de forma análoga a la completitud.
2. *Unanimidad* (Pareto débil): si todos prefieren A a B, la sociedad también.
3. *Ausencia de dictadores*: nadie determina la elección al margen del resto.
4. *Transitividad*: si $A \succ B$ y $B \succ C$, entonces $A \succ C$, lo que evita los ciclos.

5. *Independencia de alternativas irrelevantes* (IAI): para ordenar A y B basta saber cómo ordena cada uno A y B. Excluye la cardinalización y la manipulación añadiendo o quitando alternativas.

2.3. Teorema de la imposibilidad (Arrow, 1951b)

Teorema: en *Social Choice and Individual Values* (1951b), Arrow prueba que, con un número finito de individuos y tres o más alternativas, **ninguna RES satisface a la vez los cinco axiomas**.

Demostración (Sen, 1970a): un grupo es *decisivo* sobre un par si, cuando todos sus miembros prefieren una alternativa, la sociedad la prefiere.

1. *Expansión*: por Pareto, la sociedad entera es decisiva. Por la IAI, la universalidad, Pareto y la transitividad, un conjunto decisivo sobre un par lo es sobre todos.
2. *Contracción*: todo conjunto decisivo con más de un miembro se divide en dos, y con un perfil sobre tres alternativas, por transitividad, una de las partes resulta decisiva.

Repitiendo, se llega a un conjunto de un solo individuo (un dictador). La universalidad permite construir los perfiles necesarios, la IAI impide usar la intensidad para bloquear la expansión y la transitividad obliga a contraer. Con solo ordenaciones, sin intensidades ni comparaciones, ninguna regla es a la vez coherente y respetuosa con cada individuo, y elegir una RES obliga a abandonar al menos un axioma. Las salidas de la sección 3 son otras tantas formas de pagar ese precio.

2.3.1. Paradoja liberal (Sen, 1970b)

Idea: Sen muestra que basta con Pareto débil y un **liberalismo mínimo**, según el cual cada uno de dos individuos es decisivo sobre algún par de su esfera privada. Con dominio universal, las preferencias *entrometidas* sobre la esfera ajena generan ciclos. El ejemplo canónico es el de un mojigato, un lascivo y *El amante de Lady Chatterley*.

El conflicto entre eficiencia y derechos es lógico: se resuelve excluyendo las preferencias entrometidas o tratando los derechos como restricciones sobre el proceso (Nozick, 1974). Y el criterio de Pareto, aparentemente neutral, no lo es, porque su aplicación irrestricta puede violar la esfera privada. La paradoja introduce los derechos en la economía del bienestar. Su relevancia práctica depende, sin embargo, de cuánto pesen las preferencias entrometidas.

2.3.2. Conclusiones

Aspectos positivos: generaliza a cualquier regla la incoherencia colectiva que Condorcet mostró con un ejemplo, y su método axiomático se extendió a la manipulabilidad y a la implementación.

Limitaciones: la IAI excluye la intensidad de las preferencias, que la regla de Borda recoge, y el dominio universal obliga a ordenar perfiles extremos, cuya exclusión permite a la mayoría simple una ordenación transitiva.

2.4. Rechazo de Arrow (Little, 1952; Bergson, 1954)

Idea: Little y Bergson niegan que la economía del bienestar busque una regla mecánica que sume valores. Su objeto son los juicios de bienestar social. Para Little, esa máquina no garantiza expresar lo que la sociedad valora, y para Bergson la economía del bienestar asesora al individuo, y la regla de Arrow, al funcionario que decide en nombre de la colectividad.

La crítica distingue la fase *deliberativa*, donde se forman los juicios y opera la FBS, de la *decisiva*, donde se agregan y opera la RES. Arrow aceptó la distinción, pero replicó que la cuestión central es qué políticas debe adoptar la sociedad. La crítica fija el alcance del teorema, pero no lo desactiva, pues toda sociedad que decide necesita una regla.

2.5. Conclusiones

En suma, los valores que la FBS daba por supuestos no pueden obtenerse de ordenaciones ordinales con una regla que cumpla los cinco axiomas. El teorema funda la teoría de la elección social, y la sección 3 estudia sus salidas.

3. Desarrollos posteriores: las salidas a la imposibilidad

Las reglas de votación, las que usan las sociedades, renuncian a algún axioma. Las de Borda y Condorcet son anteriores a Arrow, pero se leen a su luz. Las demás salidas amplían la información, estudian la manipulación y la implementación o pasan a la negociación.

3.1. Decisividad y universalidad

3.1.1. Mayoría óptima (Buchanan y Tullock, 1962)

Origen: Wicksell (1896) propone la unanimidad como regla fiscal, la única que garantiza que nadie resulte perjudicado. Buchanan y Tullock (1962) se preguntan qué mayoría exigir, y fundan la teoría económica de las constituciones.

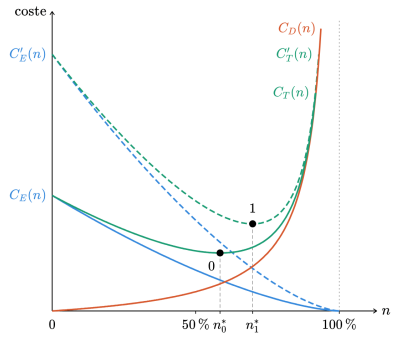
Idea: la **mayoría óptima** minimiza la suma de dos costes opuestos:

$$C_T(n) = C_D(n) + C_E(n)$$

- Los *costes de decisión*, de llegar a un acuerdo, crecen con la mayoría exigida y tienden a infinito con la unanimidad.
- Los *costes externos*, que soportan los perjudicados, decrecen y se anulan con la unanimidad.

La suma tiene forma de U y su mínimo es la mayoría óptima. Las reglas observadas van de la unanimidad, como en la fiscalidad del Consejo de la UE, a la mayoría cualificada, como la doble mayoría del 55 % de los Estados y el 65 % de la población en el Consejo de la UE o tres quintos de los miembros con el 85 % del poder de voto para reformar el convenio del FMI, que da a EE. UU. un veto de hecho, y a las mayorías absoluta y simple, como en el Consejo de Gobierno del BCE. La unanimidad es paretiana y exige pocos juicios de valor, pero favorece el *statu quo* y la *tiranía de la minoría*.

Estática comparativa: la mayoría óptima crece con los costes externos potenciales, como en las reformas constitucionales, y disminuye con los costes de decisión, en decisiones frecuentes y de bajo alcance. En la figura, los costes externos pasan a $C'_E(n)$, los totales a $C'_T(n)$ y la mayoría óptima, de n_0^* a n_1^*.



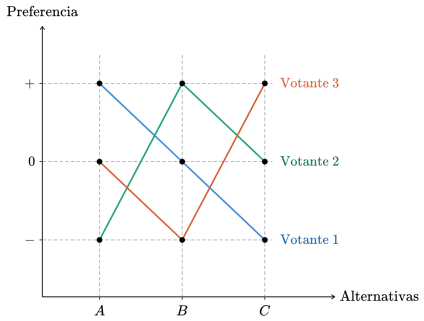
Mayoría óptima (Buchanan y Tullock, 1962): mínimo de los costes totales; crece con los costes externos (trazo discontinuo).

Teorema de May (1952): con dos alternativas, la mayoría simple es la única regla decisiva, anónima, neutral y con respuesta positiva (si una alternativa gana o empata y un votante cambia a su favor, gana). Explica la posición privilegiada de la mayoría simple, y sus problemas aparecen con más alternativas.

La mayoría óptima explica que convivan mayorías distintas en una misma constitución, aunque los costes son difíciles de medir y es más una guía cualitativa que una regla.

3.1.2. Mayoría simple: paradoja de Condorcet (1785)

Idea: la mayoría simple por pares puede dar una ordenación social intransitiva aunque las individuales sean transitivas. Con tres votantes, $A \succ B \succ C$, $B \succ C \succ A$ y $C \succ A \succ B$, A vence a B, B a C y C a A: no hay equilibrio. El problema nace del tercer votante, con dos modas.



Paradoja de Condorcet: el votante 3, con dos máximos, genera el ciclo.

Extensiones: las salidas a la paradoja son preferencias idénticas, que hacen innecesaria la regla; excluir lo ya votado, lo que permite la **manipulación de la agenda**, porque gana lo que entra en la última votación, o exigir preferencias unimodales, renunciando a la universalidad.

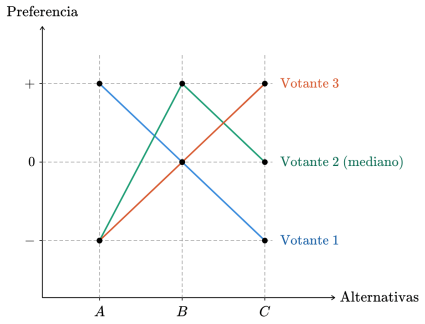
Separa la racionalidad individual de la colectiva y anticipa a Arrow, aunque es un resultado de posibilidad y no de frecuencia: los ciclos solo aparecen con ciertos perfiles y desaparecen cuando las preferencias son unimodales.

3.1.3. Unimodalidad y votante mediano (Black, 1948)

Teorema del votante mediano: si las alternativas se ordenan en una dimensión y las preferencias son **unimodales**, la mayoría elige el óptimo del votante mediano, que deja a cada lado el mismo número de votantes y es un ganador de Condorcet. Se pierde universalidad, pero solo se descartan perfiles extremistas.

Razonamiento: cualquier alternativa situada a la izquierda del óptimo del mediano pierde frente a él con los votos del mediano y de todos los situados a su derecha, y simétricamente las de la derecha. En la figura, B vence a A con los votos de 2 y 3 y a C con los de 1 y 2.

Se considera preferible perder universalidad a perder transitividad. El teorema explica la convergencia de los partidos al centro y fundamenta



Preferencias unimodales: gana B, la alternativa preferida por el votante mediano.

el modelo de Downs. Pero, con bienes públicos y reparto igualitario del coste, la mayoría iguala la valoración del mediano a RMT/H, y solo es eficiente si coincide con la media. En otro caso, hay infraprovisión o sobreprovisión.

3.1.4. Downs-Hotelling (1957)

Idea: Downs (1957) aplica el modelo de localización de Hotelling a la competencia electoral. Con información perfecta, dos partidos que maximizan votos y una sola dimensión ideológica, ambos se desplazan al votante mediano, y este equilibrio de Nash, el *principio de diferenciación mínima*, no tiene por qué ser eficiente. Los votantes, con preferencias unimodales sobre la ideología, votan al partido más cercano. Si un partido propone un programa, el otro puede acercarse más al mediano y acaparar votantes, y el primero reacciona acercándose también, hasta que ambos coinciden con él.

Si las preferencias se desplazan, ambos partidos siguen al mediano, y un cambio en los extremos que no lo altere no modifica el equilibrio. Explica la convergencia al centro en los países bipartidistas. Pero con más partidos aparecen el voto útil, las coaliciones y el extremismo. Con costes de información, los partidos necesitan recursos para sondeos y campañas, lo que da poder a los grupos de presión, y los votantes caen en la abstención racional, porque su voto apenas altera el resultado, la persuasión por carisma y las ilusiones fiscal y de deuda.

3.2. Independencia de alternativas irrelevantes

La mayoría ignora la **intensidad de las preferencias**. Renunciar a la IAI le da entrada.

3.2.1. *Log-rolling* (Buchanan y Tullock, 1962)

Idea: con el intercambio de votos, cada uno apoya la propuesta ajena a cambio de apoyo para la suya, y quien tiene mucho interés en algo lo consigue cediendo en lo que le importa poco. Exige poder pactar antes de votar y que aprobar una propuesta no impida aprobar otra.

Razonamiento: entre tres votantes, el 1 y el 2 se benefician cada uno de un proyecto pagado por los tres y con beneficios totales mayores que su coste. Votados por separado, ambos fracasan. El pacto entre 1 y 2 los aprueba si se votan en un único paquete. Pero falla con muchos votantes, puede explotar a la minoría y llevar a la sobreprovisión. Negociar sale entonces demasiado caro, y la coalición ganadora no está obligada a compensar a los perdedores.

3.2.2. Regla de Borda (1770)

Idea: asigna n puntos a la alternativa preferida, $n-1$ a la segunda, y así sucesivamente. Usa la ordenación completa, evita los ciclos y favorece las alternativas de consenso, pero viola la IAI, porque la puntuación depende de cuántas otras alternativas haya y dónde se sitúen. Genera, además, más empates y permite el voto estratégico (relegar al último puesto a la rival de la opción favorita).

3.3. Base informacional (Sen, 1970a)

Idea: Sen reinterpreta el teorema como un resultado sobre la información, ordinal y no comparable, que usa la regla. Si se admiten utilidades cardinales y comparables, la agregación deja de ser imposible.

Teorema de agregación de Harsanyi (1955): si las preferencias individuales y sociales cumplen la utilidad esperada y la regla respeta Pareto, la FBS es una suma ponderada de utilidades, con pesos iguales si hay anonimato. El utilitarismo se deriva así de la racionalidad bajo incertidumbre, sin postularlo.

D'Aspremont y Gevers (1977): con comparabilidad de niveles, los axiomas de Arrow más anonimato y equidad de Hammond conducen al *leximin*; con comparabilidad de unidades, al utilitarismo; con plena comparabilidad, elegir es un juicio de valor. La comparabilidad de niveles permite decir quién está peor, pero no cuánto, y la de unidades, cuánto gana uno y pierde otro, pero no quién está mejor.

Con información ordinal, restringir el dominio o relajar la transitividad social a la cuasitransitividad permite reglas no dictatoriales, a costa de oligarquías o de indecisión.

Los criterios ético-lógicos aparecen así como consecuencias de la información que se acepta usar. Pero el problema se traslada a justificar las comparaciones interpersonales, sin base empírica aceptada, y los axiomas adicionales, como la equidad de Hammond, son en sí mismos un juicio ético.

3.4. Manipulación e implementación

3.4.1. Manipulación: Gibbard-Satterthwaite (1973)

Teorema: todo mecanismo no dictatorial, con dominio universal y al menos tres alternativas elegibles, es **manipulable**: en algún perfil a alguien le conviene mentir. Equivale al de Arrow, y la salida es restringir el dominio o debilitar el equilibrio. Con preferencias cuasilineales y transferencias, la lógica de Clarke-Groves-Vickrey hace de la verdad la estrategia dominante. El teorema es el origen del diseño de mecanismos y convierte el comportamiento estratégico en un rasgo general que el diseño institucional ha de tener en cuenta, aunque es un resultado de existencia, porque manipular exige conocer las preferencias ajenas.

3.4.2. Implementación (Maskin, 1977)

Idea: Maskin, Nobel en 2007 con Hurwicz y Myerson, plantea la pregunta inversa a la de Arrow. Una regla es implementable en equilibrio de Nash si existe un juego cuyos equilibrios coinciden con lo que prescribe. La condición necesaria es la **monotonicidad de Maskin**. Con tres o más agentes, junto con la ausencia de veto, es suficiente. Si una alternativa elegida no empeora para nadie al cambiar el perfil, ha de seguir eligiéndose: el mecanismo solo reacciona a cambios que los agentes tengan interés en revelar. Fundamenta los mecanismos de revelación y la regulación con información asimétrica, aunque supone que los agentes conocen las preferencias de los demás. Además, sus mecanismos son complejos y poco realistas, y muchas reglas atractivas que eligen una sola alternativa no cumplen la monotonicidad.

3.5. Negociación y juegos cooperativos

3.5.1. Negociación axiomática (Nash, 1950)

Idea: Nash demuestra que cuatro axiomas, eficiencia, simetría, invariancia ante transformaciones afines e IAI, determinan una solución única: maximizar el producto de las ganancias sobre el punto de desacuerdo d:

$$\max_{(u^A,u^B)\in S} (u^A - d^A)(u^B - d^B)$$

donde S es el conjunto de pares de utilidades alcanzables mediante acuerdo y d lo que cada parte obtiene si no lo hay. La compacidad y la convexidad de S y la estricta cuasiconcavidad del producto garantizan que el máximo existe y es único.

Equivale a la FBS de Bernoulli-Nash con origen en el desacuerdo, y, frente a Arrow, obtiene una solución única porque su problema tiene más estructura y renuncia al dominio universal. Es la referencia de la negociación salarial y recibió un fundamento no cooperativo en el modelo de ofertas alternas. Pero no describe el proceso, su axioma de IAI es el más discutido y depende del punto de desacuerdo.

3.5.2. Juegos cooperativos

Núcleo (Gillies, 1959): los repartos que ninguna coalición puede mejorar. Es un concepto de estabilidad, puede ser vacío o contener muchos repartos y, en una economía de intercambio replicada, se reduce a los equilibrios competitivos. **Valor de Shapley** (1953): la media de la contribución marginal de cada jugador en todos los órdenes posibles. Es un concepto de equidad, el único que cumple eficiencia, simetría, tratamiento nulo y aditividad. Existe siempre y es único, pero puede quedar fuera del núcleo. Se aplica al reparto de costes de un bien público o al poder de voto en los organismos internacionales, con el índice de Shapley-Shubik. Ambos conceptos suponen, sin embargo, acuerdos vinculantes dentro de cada coalición y no describen la negociación que conduce a ellos.

3.6. Conclusiones

En suma, cada regla de votación renuncia a algún axioma, y ninguna no dictatorial con dominio universal escapa a la manipulación. La información comparable recupera reglas no dictatoriales, y la negociación obtiene soluciones únicas porque su problema tiene más estructura que el de Arrow. El análisis coste-beneficio y los fallos del sector público se estudian en los temas A20 y A23.

A25. La teoría neoclásica del mercado de trabajo. Análisis intertemporal de la oferta de trabajo. Teoría del capital humano. Función de ingresos de capital humano y evidencia empírica

1. Análisis estático: oferta y demanda de trabajo

Para la teoría neoclásica, el trabajo se intercambia en un mercado parecido al de cualquier otro bien: el salario y el empleo se fijan donde se cruzan la oferta y la demanda.

1.1. Supuestos

Metodología: equilibrio parcial.

Entorno: sin fallos de mercado, es decir, sin externalidades, bienes públicos ni poder de mercado, con información perfecta y sin costes de transacción ni fricciones de emparejamiento. Precios y salarios son plenamente flexibles, de modo que el mercado se vacía.

Agentes: consumidores-trabajadores que maximizan su utilidad y empresas que maximizan el beneficio, numerosos y **salario-aceptantes**, con competencia perfecta en el mercado de bienes.

Tecnología: supuestos neoclásicos habituales, con producción monoproducto y rendimientos mixtos.

1.2. Oferta de trabajo

1.2.1. La decisión de oferta de trabajo

Idea: la oferta de trabajo aplica la teoría del consumidor con dotaciones. El individuo consume un **bien compuesto** c y **tiempo de ocio** o, la única alternativa al trabajo remunerado, que, a diferencia del ocio, es un mal.

Planteamiento: maximiza una utilidad creciente y estrictamente cuasicóncava sujeto a la restricción de gasto, con renta no salarial M_0, y a la de tiempo, $l_0 = l + o$:

$$\max_{o,c} \; u(o,c) \quad \text{s.a.: } p_c c = w(l_0 - o) + M_0$$

donde la renta total, $M = w(l_0 - o) + M_0$, suma la renta salarial, que depende de las horas trabajadas, y la no salarial, exógena.

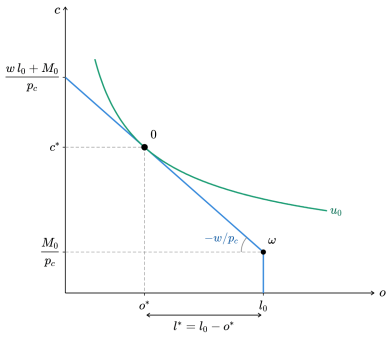
CPO: el individuo gasta toda su renta, y la RMS entre ocio y consumo, la tasa subjetiva de intercambio, iguala el salario real, la tasa objetiva o coste de oportunidad del ocio:

$$\text{RMS}_{co} \equiv \frac{u'_o}{u'_c} = \frac{w}{p_c}$$

El "precio" del ocio es el salario, la renta a la que se renuncia por cada hora no trabajada. La cuasiconcavidad estricta garantiza que la tangencia es un máximo único.

Solución de esquina: si $\text{RMS}_{co}(o = l_0) > w/p_c$, el individuo no trabaja. El **salario de reserva**, $\underline{w}$, es la RMS evaluada en la dotación, y se participa si $w/p_c > \underline{w}$. Depende de la renta no salarial, directamente si el ocio es normal, porque con más M_0 el ocio es más valioso en el margen; del consumo de subsistencia, y de las preferencias.

Desarrollo gráfico: la recta de balance tiene pendiente $-w/p_c$, ordenada en el origen $(wl_0 + M_0)/p_c$, la *renta potencial*, y pasa siempre por la dotación. El equilibrio es la tangencia con la curva de indiferencia más alta, y el trabajo ofrecido es la distancia horizontal entre el ocio elegido y la dotación.



Elección ocio-consumo: tangencia entre la curva de indiferencia y la recta de balance, de pendiente el salario real.

1.2.2. La función de oferta individual

Definición: de la demanda marshalliana de ocio y la restricción de tiempo se obtiene la *oferta de trabajo ordinaria*, y del problema dual, la *compensada*, $l^c(w, p_c, u_0)$, que solo recoge el efecto sustitución:

$$l^s = l_0 - o^*(w, p_c, M_0) = l^s(w^?, p_c^-, M_0^-)$$

Propiedades:

- Son continuas y diferenciables en las soluciones interiores. La ordinaria se anula por debajo del salario de reserva, donde tiene un quiebro.
- Son homogéneas de grado cero: la ordinaria depende solo del salario real y de la renta no salarial real, sin ilusión monetaria.
- El efecto del salario en la ordinaria es ambiguo y el de la renta no salarial, negativo si el ocio es normal. La compensada es no decreciente en el salario.
- La ordinaria, la única observable, coincide con la compensada en la utilidad máxima, $l^s(w, p_c, M_0) = l^c(w, p_c, u^*)$, de donde sale la ecuación de Slutsky.

1.2.3. Estática comparativa

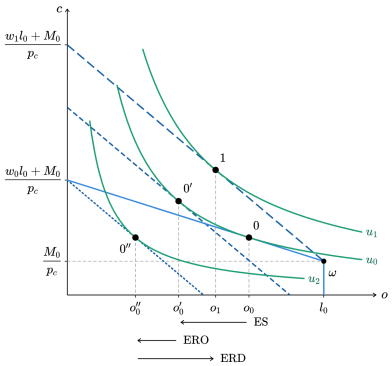
Efectos propios: una subida del salario tiene efecto ambiguo, porque el trabajador es a la vez comprador de ocio y vendedor de tiempo. La **ecuación de Slutsky con dotaciones** lo descompone:

$$\underbrace{\frac{\partial o}{\partial w}}_{\text{ET}} = \underbrace{\left.\frac{\partial o}{\partial w}\right|_{u_0}}_{\text{ES} \le 0} \underbrace{- \frac{\partial o}{\partial M} \cdot o}_{\text{ERO}} \underbrace{+ \frac{\partial o}{\partial M} \cdot l_0}_{\text{ERD}}$$

- El *efecto sustitución* es siempre no positivo, porque el salario es el precio del ocio y, a utilidad constante, el ocio se sustituye por consumo, lo que aumenta la oferta.
- El *efecto renta ordinario* es el de cualquier demanda: el ocio más caro reduce el poder adquisitivo.
- El *efecto renta-dotación* recoge que el salario revaloriza las l_0 horas de la dotación.

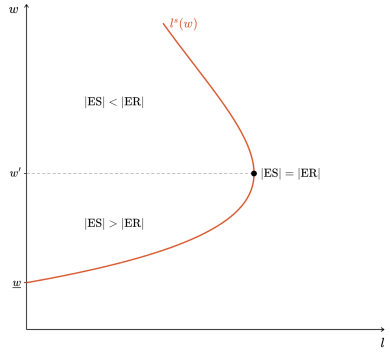
Los dos últimos forman el **efecto renta total**, $(\partial o/\partial M)\, l$, positivo si el ocio es normal. Como quien trabaja vende parte de su dotación, el ERD domina al ERO, el trabajador es más rico y demanda más ocio. La oferta solo es creciente si domina el efecto sustitución y, si domina el efecto renta total, **se vuelve hacia atrás**. Si el ocio fuera inferior, ambos efectos irían en la misma dirección y la oferta sería siempre creciente. Un bien normal puede comportarse así como un bien Giffen para quien lo vende.

Gráficamente, una subida de w hace girar la recta de balance alrededor de la dotación. El ES lleva de 0 a 0 sobre la curva inicial, el ERO de 0 a 0 y el ERD de 0 a 1.



Subida del salario: efecto sustitución (0 → 0), renta ordinario (0 → 0) y renta-dotación (0 → 1).

En el plano trabajo-salario, la oferta alcanza su máximo en w', donde $|\text{ES}| = |\text{ER}|$, y a partir de ahí se vuelve hacia atrás.



Oferta de trabajo individual: el tramo regresivo empieza en w', donde $|\text{ES}| = |\text{ER}|$.

Efectos cruzados: si el consumo es normal, una subida del salario lo eleva, porque el ES y el efecto renta total son positivos.

Otras variables:

1. *Renta no salarial*: si el ocio es normal, reduce la oferta y eleva el salario de reserva.

2. *Horas extraordinarias*: una prima solo sobre las horas adicionales, para quien no superaba la jornada, no revaloriza la dotación y actúa como efecto sustitución puro, de modo que la oferta aumenta.
3. *Imposición*: un impuesto de suma fija, con ocio normal, aumenta la oferta por puro efecto renta. Uno proporcional sobre el salario tiene los tres efectos con signo contrario y es ambiguo. Para igual recaudación, el de suma fija genera más oferta y no produce exceso de gravamen.

1.2.4. Oferta de trabajo agregada

Oferta agregada: es la suma horizontal de las individuales. Se acepta que el tramo relevante es el creciente, porque las subidas salariales atraen a nuevos participantes, el **margen extensivo**, aunque los empleados reduzcan sus horas, el *margen intensivo*.

1.2.5. Evidencia empírica

Elasticidades: las de los trabajadores principales son bajas, la no compensada próxima a cero, y mayores en el margen extensivo. Los efectos renta y sustitución se compensan en buena medida. La curva individual puede tener forma de S invertida, las transferencias no condicionadas reducen poco la oferta y los trabajadores secundarios del hogar son más elásticos.

1.2.6. Conclusiones

La oferta aplica la teoría del consumidor con dotación y Slutsky explica su pendiente ambigua. Pero el modelo es estático, ignora las horas fijas por contrato y trata la participación como una decisión de esquina.

1.3. Demanda de trabajo

1.3.1. La decisión de demanda de trabajo

Idea: procede de la teoría de la empresa competitiva y de Hicks (1932), y es una demanda *derivada* porque el trabajo interesa como medio para obtener la producción.

Planteamiento: la empresa maximiza el beneficio, $\max_{\mathbf{z}} \ \pi = p \cdot f(\mathbf{z}) - \mathbf{p}^z \cdot \mathbf{z}$.

CPO: la RMST iguala el precio relativo de los factores, que es la minimización del coste, y cada factor se contrata hasta que el valor de su productividad marginal iguala su precio:

$$\text{VPmg}_l \equiv p \cdot f_l' = w$$

CSO: el beneficio ha de ser estrictamente cóncavo. Con un solo factor variable basta $f_{ll}'' < 0$.

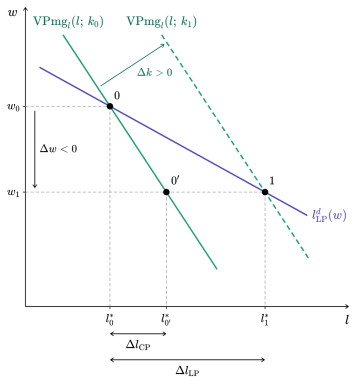
1.3.2. La función de demanda de trabajo

Definición: la *demanda condicionada*, $l_c^d(w, r_K, x_0)$, procede de la minimización del coste con producción dada. La *incondicionada*, $l_i^d(w, r_K, p)$, procede de la maximización del beneficio y es la relevante para el mercado.

Propiedades: son homogéneas de grado cero en los precios y decrecientes en el salario. La incondicionada es la derivada, cambiada de signo, de la función de máximo beneficio (lema de Hotelling), y la condicionada, la de la función de costes (lema de Shephard). La incondicionada es, además, la condicionada evaluada en la producción que maximiza el beneficio, $l_i^d(w, r_K, p) = l_c^d(w, r_K, x^*)$, de donde sale la ecuación de Slutsky de la empresa.

Corto plazo: con el capital fijo, la demanda es el tramo decreciente del VPmg por debajo del máximo del VPme, $l^d(w^-, p^+, f_l'^+, \bar{k}^+)$.

Largo plazo: con todos los factores variables, al efecto sustitución se suma el *efecto escala*, y la demanda es más elástica, porque a corto plazo el capital está fijo y no cabe sustituir factores. Ante una bajada del salario de w_0 a w_1, la empresa se mueve primero sobre su VPmg y después, al contratar más capital, la curva se desplaza hacia fuera.



Demanda de trabajo a largo plazo: con capital variable, más elástica que a corto plazo.

1.3.3. Estática comparativa

Efectos propios: la ecuación de Slutsky de la empresa muestra que ambos efectos van en la misma dirección, de modo que la demanda es **inequívocamente decreciente**:

$$\frac{\partial l_i^d}{\partial w} = \underbrace{\frac{\partial l_c^d}{\partial w}\bigg|_{x_0}}_{\text{ES} \le 0} + \underbrace{\frac{\partial l_c^d}{\partial x} \cdot \frac{\partial x}{\partial w}}_{\text{EP} < 0} < 0$$

donde el ES es no positivo por la concavidad de la función de costes en los precios de los factores, y el efecto producción, negativo, porque un factor más caro eleva el coste marginal y reduce la producción.

Efectos cruzados: sobre la demanda de capital, el ES es no negativo y el EP negativo. Suele predominar el efecto escala, y los factores son complementarios brutos.

1.3.4. Demanda de trabajo agregada

Demanda agregada: al bajar el salario, toda la industria produce más y el precio del producto cae. Por eso no basta la suma horizontal a precio constante y se agregan las **cuasidemandas**, con una curva de mercado menos elástica.

Elasticidad de la demanda derivada: a largo plazo, con rendimientos constantes y oferta de capital perfectamente elástica, $\varepsilon_{L^d w} = -[(1-s)\sigma + s|\varepsilon_{Xp}|]$. Por las **reglas de Marshall y Hicks**, la demanda de trabajo es más elástica cuanto mayores son σ y $|\varepsilon_{Xp}|$, cuanto mayor es la participación del trabajo, s (si $|\varepsilon_{Xp}| > \sigma$), y cuanto más elástica es la oferta de los demás factores.

1.3.5. Evidencia empírica

Elasticidades: Hamermesh (1993) sitúa la de largo plazo, a producción constante, entre $-0{,}15$ y $-0{,}75$, con $-0{,}3$ como mejor estimación. La sustituibilidad es mayor para el trabajo poco cualificado.

Inmigración: en un experimento natural, Card (1990) estudia el éxodo del Mariel (1980), que aumentó un 7 % la fuerza de trabajo de Miami, y no encuentra efectos negativos sobre los trabajadores poco cualificados.

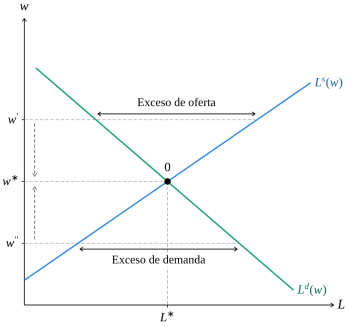
1.3.6. Conclusiones

La demanda hereda las propiedades de las demandas de factores y es decreciente, más elástica a largo plazo. Pero supone trabajo homogéneo, ajuste instantáneo y competencia en el producto.

1.4. Equilibrio y poder de mercado

1.4.1. Equilibrio competitivo

Equilibrio: oferta creciente y demanda decreciente se cruzan en un único punto, (L^*, w^*), estable porque el salario flexible absorbe cualquier exceso. Un aumento de la demanda eleva el salario y el empleo. Un aumento de la renta no salarial, con ocio normal, desplaza la oferta a la izquierda, sube el salario y cae el empleo. Cuanto más rígida es la curva que no se desplaza, más se traduce la perturbación en el salario y menos en el empleo.



Equilibrio competitivo: a w' hay exceso de oferta y a w'', de demanda; en ambos casos el salario vuelve a w^*.

Bienestar: el equilibrio es eficiente en el sentido de Pareto por el primer teorema. El VPmg del último trabajador iguala su salario, y este, la valoración marginal del ocio.

Sin embargo, el equilibrio no implica que todos trabajen, y se distinguen tres tipos de desempleo:

- *Voluntario*: el de quienes tienen un salario de reserva superior al de mercado.
- *Friccional*: temporal, por la imposibilidad de emparejar al instante vacantes y buscadores.
- *Involuntario o clásico*: solo si una rigidez mantiene el salario por encima del de equilibrio.

Extensiones: sin salario-aceptación, por uno u otro lado del mercado, se obtiene la *condición generalizada de contratación*:

$$\left(1 - \frac{1}{|\varepsilon_{Xp}|}\right) p \cdot f_l' = \left(1 + \frac{1}{\varepsilon_{lw}}\right) w$$

donde ε_{Xp} es la elasticidad de la demanda del producto y ε_{lw}, la de la oferta de trabajo a la que se enfrenta la empresa, ambas infinitas en competencia perfecta.

Con competencia en ambos mercados se recupera $\text{VPmg}_l = w$. Con poder en bienes aparece un *mark-up*, y con poder en trabajo, un *mark-down*. En cualquier caso, el empleo es inferior al competitivo.

1.4.2. Sindicatos

Idea: surgen por el **poder compensatorio** frente a las grandes empresas (Galbraith, *American Capitalism*, 1952).

Planteamiento: el sindicato elige sobre la demanda de trabajo el empleo que maximiza su objetivo, con $w(L)$ la demanda inversa y $C(L)$ el coste de oportunidad de los empleados:

$$\max_L \; U = w(L) \cdot L - C(L)$$

CPO: el sindicato monopolista que maximiza la renta económica de sus afiliados iguala el ingreso marginal de la venta de trabajo al salario de reserva, con un *mark-up* inverso a la elasticidad de la demanda de trabajo:

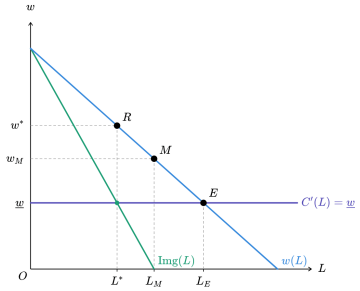
$$w(L^*)\left(1 - \frac{1}{|\varepsilon^d_{Lw}|}\right) = C'(L^*) = \underline{w}$$

CSO: el ingreso marginal debe ser decreciente y el coste marginal de oportunidad no decreciente, como en cualquier monopolio.

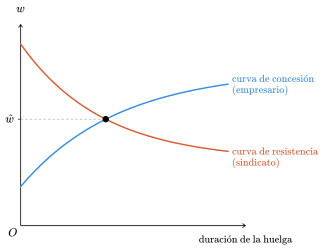
Estrategias:

- maximizar la renta económica, Img = Cmg, con pérdida de eficiencia y desempleo involuntario;
- maximizar la masa salarial, Img = 0, en el punto de elasticidad unitaria;
- maximizar el empleo, con el salario competitivo.

Negociación: en el esquema de Hicks (1932), cuanto más larga se espera la huelga, menos exige el sindicato, su curva de resistencia, y más concede el empresario, su curva de concesión. Con certidumbre, se pacta el salario del cruce y la huelga queda como amenaza.



(a) Estrategias del sindicato



(b) Negociación (Hicks, 1932)

Objetivos y negociación del sindicato.

1.4.3. Monopsonio

Idea: el único demandante se enfrenta a una oferta creciente y, para atraer una unidad más, paga más a todas, de modo que el **gasto marginal** supera al salario:

$$\text{Gmg}_l = w\left(1 + \frac{1}{\varepsilon_{lw}}\right) > w$$

Contrata donde $\text{VPmg}_l = \text{Gmg}_l$, con menos empleo y menor salario que en competencia. La brecha entre el VPmg y el salario es la **explotación monopsonística**. Con monopolio en bienes se suma la explotación monopolística, y salario y empleo son los más bajos.

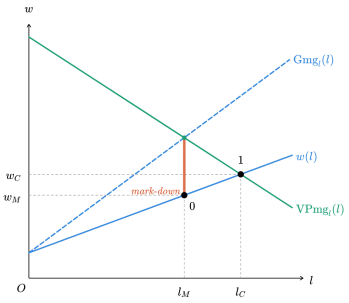
1.4.4. Monopolio bilateral

Con un sindicato frente a una empresa única, el salario queda indeterminado entre el del monopsonista y el del sindicato, según el poder de negociación de cada parte. Si solo se negocia el salario, el empleo no supera el competitivo, y si el salario se fija por encima de este, hay desempleo involuntario.

1.4.5. Evidencia empírica

Sindicatos: la prima salarial sindical en EE. UU. ronda el 15 % en 1967-1979 y el 17-18 % en 1973-2001.

Monopsonio: la elasticidad de la oferta a la que se enfrenta cada empresa es finita, con una mediana de 1,7 en las estimaciones, y pasar del



Monopsonio en el mercado de trabajo.

percentil 25 al 75 de concentración se asocia a salarios entre un 5 % y un 17 % más bajos.

Salario mínimo: Card y Krueger (1994) comparan la restauración de Nueva Jersey, que subió el salario mínimo, con Pensilvania y no encuentran efectos negativos sobre el empleo, lo que es compatible con el monopsonio.

En suma, el equilibrio competitivo es estable y eficiente y, sin rigideces, no genera desempleo involuntario. Sus supuestos, entorno estático, trabajo homogéneo y salario-aceptación, son su programa de investigación.

2. Análisis intertemporal

El análisis intertemporal extiende el modelo estático a varios periodos.

2.1. Supuestos

Entorno: sin fallos de mercado, tiempo discreto, horizonte finito T y tipo de interés real exógeno r.

Agentes: consumidores-trabajadores sin riqueza inicial, con utilidad aditivamente separable en el tiempo, tasa de descuento ρ, con $\beta = 1/(1+\rho)$, y dotación de tiempo normalizada, $o_t + l_t = 1$.

2.2. Oferta intertemporal (Lucas y Rapping, 1969)

Idea: Lucas y Rapping trasladan la elección ocio-consumo a varios periodos y tratan la oferta de trabajo como asignación del tiempo entre periodos. La oferta responde al salario relativo, no a su nivel, y el individuo concentra su trabajo cuando está mejor pagado (la **sustitución intertemporal del ocio**).

Planteamiento: maximiza la utilidad descontada sujeta a que el valor actual del consumo iguale el de la renta salarial:

$$\max_{\{c_t, o_t\}} \sum_{t=0}^{T} \beta^t u(c_t, o_t) \quad \text{s.a.:} \sum_{t=0}^{T} \frac{c_t}{(1+r)^t} = \sum_{t=0}^{T} \frac{w_t(1-o_t)}{(1+r)^t}, \qquad \beta \in (0,1)$$

CPO: la condición intratemporal, $u'_o(t)/u'_c(t) = w_t$, reproduce la estática. La ecuación de Euler iguala la RMS intertemporal del consumo a $1+r$, y la novedad es la condición del ocio, cuyo precio relativo entre periodos es el salario relativo corregido por el tipo de interés:

$$u'_c(t) = \frac{1+r}{1+\rho}\, u'_c(t+1)$$

$$\frac{u'_o(t)}{u'_o(t+1)} = \frac{1+r}{1+\rho} \cdot \frac{w_t}{w_{t+1}}$$

Solución: la oferta relativa crece con el salario relativo y con el tipo de interés y decrece con la tasa de descuento:

$$\frac{l^s_t}{l^s_{t+1}} = f\left[\left(\frac{w_t}{w_{t+1}}\right)^+,\; r^+,\; \rho^-\right]$$

La **elasticidad de sustitución intertemporal** determina cuánto responde la oferta a los cambios transitorios del salario.

Estática comparativa:

1. Un aumento *permanente* del salario no altera el salario relativo. Quedan los efectos estáticos, que tienden a compensarse, y el empleo apenas varía.
2. Un aumento *transitorio* eleva el salario relativo: la oferta aumenta durante la perturbación y cae después.
3. Un aumento del tipo de interés hace más atractivo trabajar y ahorrar hoy, mientras que uno de ρ reduce la oferta presente (efecto *carpe diem*).

Supuso una revolución metodológica y es la base del ciclo real de Kydland y Prescott, con desempleo siempre voluntario. Pero exige cambios salariales transitorios grandes y una elasticidad elevada, cuando las estimaciones la sitúan por debajo de la unidad, y sin costes de ajuste no explica la persistencia del paro.

2.3. Demanda intertemporal

Idea: del lado de la demanda, la dinámica descansa en los **costes de ajuste**, que encarecen contratar y despedir: búsqueda, selección, formación y despido.

Planteamiento: la empresa maximiza el valor actual de sus beneficios reales, con el capital fijo y un coste cuadrático de variar la plantilla, de parámetro $b \geq 0$:

$$\max_{\{l_t\}} V_0 = \sum_{t=0}^{T} \frac{1}{(1+r)^t} \left[f(l_t) - w_t l_t - \frac{b}{2}(l_t - l_{t-1})^2 \right]$$

CPO: la productividad marginal iguala el salario más el coste de ajustar hoy, neto del ahorro descontado en el ajuste de mañana:

$$f_l'(l_t) = w_t + b\,(l_t - l_{t-1}) - \frac{b}{1+r}\,(l_{t+1} - l_t)$$

Solución: con $b = 0$ se recupera la condición estática y el empleo salta a su nivel óptimo. Con $b > 0$, el empleo depende del heredado y de los salarios esperados y se ajusta gradualmente, tanto más despacio cuanto mayor es b. Con costes lineales o fijos, la empresa no ajusta ante perturbaciones pequeñas. El empleo oscila menos que el producto, aunque el efecto sobre su nivel medio es en principio ambiguo, porque los costes de ajuste frenan tanto las contrataciones como los despidos.

2.4. Evidencia empírica

Elasticidad: las estimaciones de la elasticidad de sustitución intertemporal son reducidas, por debajo de la unidad.

Ciclo y tendencia: a corto plazo, el empleo es muy procíclico y el salario real solo moderadamente, por las rigideces, y el ajuste se produce en cantidades. A largo plazo, la tendencia la marcan la productividad, que eleva los salarios reales, y la participación, y los salarios siguen a la productividad con retraso.

En suma, las perturbaciones transitorias mueven el empleo y las permanentes apenas lo hacen, base del ciclo real y de sus críticas. Pero se mantiene el supuesto de trabajo homogéneo.

3. Teoría del capital humano

La idea tiene precedentes en Smith, pero su formulación moderna se debe a la escuela de Chicago. Mincer (1958) trata la formación como una inversión y deriva de ella la distribución de la renta. Schultz (1961) la convierte en una categoría de la contabilidad del crecimiento, y Becker (1964) le da teoría microeconómica. Nelson y Phelps (1966) añaden que el capital humano acelera la adopción de nuevas tecnologías, lo que enlaza con el crecimiento endógeno.

3.1. Capital humano como inversión (Becker, 1964)

Idea: en *Human Capital* (1964), Becker explica por qué difieren los salarios: educarse es invertir en **capital humano**, con un coste presente y un rendimiento futuro, el mayor salario que trae una productividad más alta.

3.1.1. Supuestos

Dos periodos e información perfecta, con mercados de capitales completos y perfectos, de modo que se cumple el **teorema de separación**: la decisión de inversión es independiente de la de consumo. La educación cuesta C_e, suma de un coste explícito, uno de oportunidad y uno psicológico, y quien se educa cobra más, $w_e > w_n$.

3.1.2. Modelo

Planteamiento: el agente elige el consumo de ambos periodos y si educarse, $e \in \{0, 1\}$. La educación solo suma a la riqueza su valor actual neto:

$$c_1 + \frac{c_2}{1+r} \leq W_n + e\ \mathrm{VAN}_e, \qquad \mathrm{VAN}_e = \frac{w_e - w_n}{1+r} - C_e$$

donde $W_n = y_0 + w_n/(1+r)$ es la riqueza sin educación, con y_0 la renta del primer periodo.

CPO: se resuelve en dos etapas. Primero, la inversión, con independencia de las preferencias: se educa si el diferencial salarial descontado cubre el coste, que es la regla del VAN:

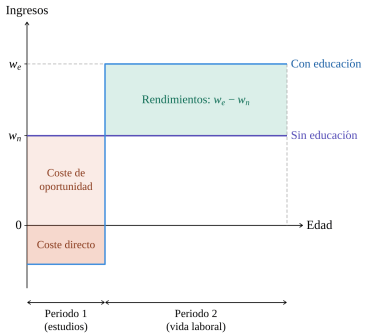
$$e^* = 1 \iff C_e \leq \frac{w_e - w_n}{1+r}$$

Después, reparte el consumo entre periodos con $\mathrm{RMS}_{c_2 c_1} = 1 + r$. La decisión depende directamente del diferencial salarial e inversamente del tipo de interés y del coste.

CSO: la inversión es una elección discreta, en la que basta comparar las dos riquezas.

Equilibrio: los trabajadores son heterogéneos, con un mercado para cada tipo: se sigue contratando por productividad marginal, pero ya no hay un único salario.

Desarrollo gráfico: con los perfiles de ingresos por edad, se educa si el valor actual del área de rendimientos, el diferencial $w_e - w_n$ durante la vida laboral, cubre el área de costes, el coste directo más el salario w_n al que se renuncia mientras se estudia.



Perfiles de ingresos por edad con y sin educación: costes durante los estudios y rendimiento $w_e - w_n$ después.

Formación general y específica: la *general*, útil en cualquier empresa, la paga el trabajador, porque a ninguna empresa le compensa costear lo que otras aprovechan. En la *específica*, la empresa comparte costes y rendimientos para frenar la rotación.

3.1.3. Conclusiones

Reduce la decisión de formarse a la regla del VAN, con predicciones contrastables, y explica quién financia cada formación. Pero no justifica la financiación pública, porque ignora las externalidades del capital humano y las restricciones de crédito, y la señalización discute que la educación eleve la productividad.

3.2. Otras aportaciones

Señal (Spence, 1973): la educación sirve no porque aumente la productividad, sino porque, con información asimétrica, la señala. Como educarse cuesta menos a los más productivos, solo ellos alcanzan el nivel exigido para el salario alto. El mercado se separa por tipos, el rendimiento privado supera al social y la financiación pública pierde su justificación productiva.

3.3. La función de ingresos de Mincer (1958, 1974)

Idea: si la educación se elige como inversión y el mercado es competitivo, el salario de cada nivel ha de compensar exactamente el coste de alcanzarlo.

Razonamiento: si el único coste de un año más es el salario perdido y la vida laboral es larga, en el margen $w_{S+1} = w_S(1+r)$, y en logaritmos $\ln w_S \approx \ln w_0 + rS$, de modo que los ingresos crecen exponencialmente con la educación, y la pendiente es la **tasa de rendimiento de la educación**. Con la formación en el puesto, que se concentra al principio, los ingresos crecen con la experiencia X a ritmo decreciente:

$$\ln w = \beta_0 + \beta_1 S + \beta_2 X + \beta_3 X^2 + \nu, \qquad \beta_1 > 0,\ \beta_2 > 0,\ \beta_3 < 0$$

Explica la distribución de la renta, permite comparar el rendimiento de la escuela con el del capital físico y es estimable con datos individuales. Pero β_1 solo es la tasa de rendimiento bajo los supuestos de la derivación. Sin ellos, es la diferencia porcentual media del salario asociada a un año más, y la forma lineal en S y cuadrática en X es una aproximación que la evidencia no siempre respalda.

3.4. Evidencia empírica

3.4.1. Rendimientos de la educación

Estimaciones: por MCO, entre el 5 y el 10 % por año en la mayoría de los países. Los rendimientos sociales son mayores en los países de renta baja y en la primaria. En los desarrollados, el de la educación superior ha crecido con el cambio técnico sesgado hacia la cualificación. En España se sitúan entre el 5 y el 8 %, con la **sobreeducación** como fenómeno muy estudiado.

Sesgo de habilidad: la habilidad no observada afecta a la educación y al salario. Los estudios de gemelos dan rendimientos similares o algo superiores a los de MCO, y Card (1999) concluye que el rendimiento causal es del orden del 7 al 10 %.

Señal: el año del título rinde más, el *efecto piel de oveja*, pero los rendimientos no caen con la experiencia y los autónomos obtienen rendimientos similares a los de los asalariados. La educación es mayoritariamente productiva, con un componente de señal.

3.4.2. Experimentos naturales

Card, Angrist e Imbens (Nobel de 2021), con Krueger, sustituyen las correlaciones de las regresiones de Mincer por **experimentos naturales**, en los que una ley, una frontera o el azar asignan el tratamiento. Angrist y Krueger (1991) instrumentan con el trimestre de nacimiento y Card (1995) con la proximidad a una universidad, y obtienen rendimientos iguales o superiores a los de MCO, en torno al 13 % en Card frente al 7 % por MCO.

En suma, la función de Mincer convirtió el capital humano en una hipótesis contrastable, confirmada en lo esencial: la educación tiene un rendimiento causal comparable al del capital físico, que explica buena parte de la desigualdad salarial y fundamenta la inversión pública cuando hay restricciones de crédito o externalidades. Pero solo mide rendimientos privados, exige resolver la habilidad no observada, trata los años como homogéneos y la evidencia de los experimentos es local.

A26. Desempleo friccional. La curva de Beveridge, el modelo de búsqueda y emparejamiento de Diamond, Mortensen y Pissarides. Costes de ajuste y dinámica de la demanda de trabajo

1. Búsqueda de empleo y salario de reserva

La teoría de la búsqueda estudia el desempleo friccional desde el lado de la oferta: cómo deciden los trabajadores, con información imperfecta, qué ofertas aceptar y cuánto tiempo seguir buscando.

1.1. Contexto y antecedentes

1.1.1. La descomposición del desempleo

Descomposición: la tasa de desempleo suma el **desempleo de equilibrio**, $\bar{u}_t$, friccional y estructural, no observable, que persiste en el estado estacionario, y el **desempleo cíclico**, u_t^c, la desviación respecto a aquel:

$$u_t = \bar{u}_t + u_t^c$$

La teoría neoclásica no puede explicar $\bar{u}$, porque en su equilibrio nadie dispuesto a trabajar al salario de mercado queda sin empleo. Las teorías de la búsqueda tratan de explicar por qué surge y si es voluntario. Con información imperfecta, encontrar empleo lleva tiempo, y el desempleo es, en parte, el tiempo dedicado a buscar.

1.1.2. Búsqueda simultánea (Stigler, 1962)

Origen: la teoría nace de la economía de la información. Stigler (1962) aplicó al mercado de trabajo su análisis de la dispersión de precios (1961).

Supuestos: el trabajador conoce la distribución de los salarios, pero no el de cada empresa. Cada consulta cuesta c, y la búsqueda es *simultánea*: decide de antemano cuántas empresas, n, contactar y se queda con la mejor oferta.

Razonamiento: el salario esperado de la mejor de n ofertas, $w^e(n)$, crece con n a ritmo decreciente. El trabajador consulta empresas hasta que la ganancia esperada de una consulta más iguala su coste, y no hasta encontrar el salario más alto:

$$\frac{\partial w^e}{\partial n} = c$$

Conclusiones: explica por qué el mercado laboral no cumple necesariamente la ley del precio único. Pero, como señaló McCall, la estrategia simultánea no es óptima, porque si la quinta empresa ofrece un salario muy alto, no tiene sentido visitar las restantes.

1.1.3. Búsqueda secuencial (McCall, 1970)

Idea: la estrategia óptima es *secuencial*: contactar empresas de una en una y decidir, tras cada oferta, si aceptarla o seguir buscando.

Razonamiento: el criterio es una *regla de parada*. Se acepta la oferta w si y solo si alcanza el **salario de reserva**, $w \geq \underline{w}$, fijado donde el coste de una búsqueda más iguala la ganancia esperada de una oferta mejor.

Conclusiones: el salario de reserva pasa a ser el concepto central, del que dependen la duración del desempleo y la tasa de salida. Pero la distribución de salarios se toma como dada, sin preguntarse si las empresas la mantendrían.

1.2. Salario de reserva (Mortensen, 1970)

Idea: Mortensen dio a la búsqueda secuencial forma dinámica: el desempleado compara en cada instante el valor de seguir buscando con el de aceptar la oferta recibida.

1.2.1. Supuestos

Entorno: tiempo continuo, información incompleta y mercados financieros perfectos, con un tipo de interés real r exógeno. El trabajador conoce la distribución de ofertas $H(w)$, pero no el salario de cada empresa.

Agentes: el desempleado es salario-aceptante, busca con intensidad constante y es neutral al riesgo.

Mercado de factores: las ofertas fijan un salario real constante w y llegan de una en una a una tasa de Poisson $\phi > 0$ exógena. El empleo se destruye a una tasa exógena $\lambda > 0$, y la búsqueda es secuencial y costosa: la disyuntiva es seguir buscando o aceptar.

1.2.2. Modelo

Planteamiento: se resuelve con dos ecuaciones de valor, la de buscar y la de aceptar y trabajar, como las de un activo cuyo rendimiento es un flujo más una ganancia esperada de capital:

$$rU_B^e = z + \phi\left(U_\phi^e - U_B^e\right) \quad \rightarrow \quad U_B^e = \frac{1}{r+\phi}\left[z + \phi U_\phi^e\right]$$

$$rU_T^e = w + \lambda\left(U_B^e - U_T^e\right) \quad \rightarrow \quad U_T^e = \frac{1}{r+\lambda}\left[w + \lambda U_B^e\right]$$

- z son los ingresos netos durante la búsqueda (subsidios, producción doméstica) y U_ϕ^e el valor esperado de recibir una oferta, $U_\phi^e = \mathbb{E}[\max(U_T^e(w), U_B^e)]$, porque la oferta se rechaza si vale menos que seguir buscando.
- Despejado, cada valor descuenta su flujo y el valor de la situación siguiente con el factor $1/(r+\phi)$ o $1/(r+\lambda)$. De la segunda, $U_T^e - U_B^e = (w - rU_B^e)/(r+\lambda)$: las ofertas valen menos cuanto más precario es el empleo.

Equilibrio: el salario de reserva iguala ambos valores y es exactamente el flujo de ingresos de la búsqueda:

$$\underline{w} = rU_B^e \quad \rightarrow \quad \underline{w}^* = f(z^+, \phi^+, \lambda^-, r^-)$$

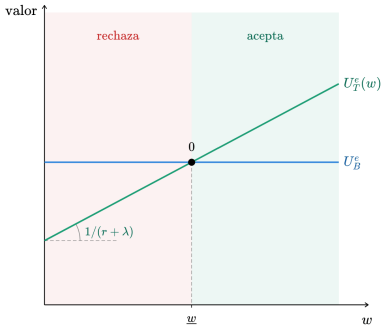
Llevando la ganancia de aceptar a la ecuación de la búsqueda se obtiene su forma implícita, de la que salen esos signos:

$$\underline{w} = z + \frac{\phi}{r+\lambda}\int_{\underline{w}}^{\infty}(w - \underline{w})\,dH(w)$$

Tasa de salida: la probabilidad por unidad de tiempo de recibir una oferta aceptable es $\chi = \phi[1 - H(\underline{w})]$, y la duración media del desempleo es su inversa, $T_u = 1/\chi$. Sus determinantes tienen los signos del salario de reserva, salvo la tasa de llegada, de efecto ambiguo:

$$T_u = \frac{1}{\chi} = f(z^+, \phi^\pm, \lambda^-, r^-)$$

Desarrollo gráfico: en el plano salario-valor, el valor de aceptar, $U_T^e(w)$, es una recta creciente con pendiente $1/(r+\lambda)$, y el de buscar, U_B^e, una horizontal. Su corte, el punto 0, define $\underline{w}$. A su izquierda aceptar vale menos que seguir buscando y se rechaza la oferta, y a su derecha se acepta.



Salario de reserva: corte del valor de aceptar, creciente en w, con el de buscar.

Estática comparativa: lo que eleva U_B^e sube la horizontal y el salario de reserva. Lo que reduce el valor de un empleo aplana la recta y baja también U_B^e.

- *Ingresos de búsqueda*, $\uparrow z$: sube $\underline{w}$, cae χ y sube T_u. La principal implicación de política es que **los subsidios pueden alargar el desempleo friccional**.
- *Tasa de llegada*, $\uparrow \phi$: sube $\underline{w}$, porque esperar es más barato, y el efecto sobre χ y T_u es ambiguo (más oportunidades, pero más exigencia).
- *Tasa de destrucción*, $\uparrow \lambda$: los empleos valen menos, cae $\underline{w}$, sube χ y cae T_u.
- *Tipo de interés*, $\uparrow r$: más impaciencia, cae $\underline{w}$, sube χ y se acorta T_u.

1.2.3. Conclusiones

Aspectos positivos: explica el desempleo friccional por decisiones racionales con información imperfecta y es el marco para analizar las prestaciones y la intermediación. Si el trabajador conoce bien $H(w)$, estar parado es para él la mejor opción, porque rechaza voluntariamente las ofertas inferiores a $\underline{w}$.

Limitaciones: solo considera la oferta, sin vacantes, y no puede decir si todo el desempleo friccional es voluntario. Toma $H(w)$ como dada, lo que cuestionan la paradoja de Diamond (apartado 1.3.1) y Rothschild (1973), y no explica el efecto desánimo.

1.3. Extensiones y evidencia empírica

1.3.1. La paradoja de Diamond (1971)

Idea: Diamond la planteó para el mercado de bienes. Si los trabajadores son idénticos y siguen la estrategia de McCall, ninguna empresa tiene incentivo a pagar más que el salario de reserva común, porque cualquier oferta igual a $\underline{w}$ es aceptada.

Razonamiento: basta un coste de búsqueda positivo, por pequeño que sea, para que la distribución colapse en el **salario de monopsonio** y desaparezca el motivo para buscar.

Extensiones: la teoría respondió por dos vías.

- *Agentes heterogéneos*: distintos salarios de reserva o distinta productividad de las empresas sostienen la dispersión.
- *Competencia entre empresas*: con varias ofertas a la vez o búsqueda desde el empleo (Burdett y Mortensen, 1998), la dispersión aparece incluso con agentes idénticos.

Conclusiones: la dispersión ha de obtenerse en equilibrio, modelizando a la vez trabajadores y empresas. Pero el resultado descansa en supuestos extremos y contradice la dispersión observada.

1.3.2. Utilidad de reserva (Blau, 1991)

Idea: las ofertas difieren también en horas y otras condiciones, y con más de una característica el buscador no fija un salario de reserva sino una **utilidad de reserva**, $\underline{u}$.

Planteamiento: con $u = w^a h^b$, donde el salario es un bien ($a > 0$) y las horas un mal ($b < 0$), la curva de indiferencia de reserva es una recta creciente en el plano $(\ln h, \ln w)$, de pendiente $-b/a > 0$:

$$\ln w = (\ln \underline{u} - b \ln h)/a$$

Decisión: el individuo admite un salario inferior al de reserva si la jornada es corta y rechaza uno superior si es larga. Los dos criterios se separan más cuanto mayor es la aversión a las horas, $|b|$.

Conclusiones: enlaza la búsqueda con la elección entre ocio y consumo, aunque sigue sin demanda de trabajo y con la distribución de ofertas dada.

1.3.3. Otras extensiones

- *Economía conductual* (DellaVigna y Paserman, 2005): con descuento hiperbólico, los impacientes procrastinan, buscan menos y alargan el desempleo. Aunque fijan un salario de reserva menor, domina la menor intensidad de búsqueda.
- *Esfuerzo de búsqueda*: la tasa de llegada depende del esfuerzo, $\phi = \phi(e)$, que se elige igualando su coste marginal a su beneficio marginal y que las prestaciones reducen además de elevar $\underline{w}$.
- *Búsqueda desde el empleo* (Burdett, 1978): explica la movilidad voluntaria y la relación positiva entre salario y antigüedad. Los empleados buscan con un salario de reserva igual al que perciben.
- *Servicios públicos de empleo*: la información sobre vacantes es un bien público y la búsqueda genera externalidades (Pissarides, 1984), de modo que el desempleo de equilibrio puede no ser el eficiente.

1.3.4. Evidencia empírica

Prestaciones y duración: está bien establecido que las prestaciones alargan el desempleo, y las salidas muestran un pico justo antes de agotarse el derecho (Katz y Meyer, 1990). En España, Bover, Arellano y Bentolila (2002) encuentran que, a los tres meses, la tasa de salida de quienes no cobran prestación duplica la de quienes la cobran. El pico de salidas refleja tanto la aceptación de ofertas antes rechazadas como el abandono de la población activa.

Dependencia negativa de la duración: la tasa de salida cae con el tiempo en desempleo, por heterogeneidad no observada, pérdida de capital humano y estigma, y estos dos últimos son la base microeconómica de la histéresis.

Salarios de reserva declarados: se sitúan cerca del último salario y apenas bajan con la duración (Krueger y Mueller, 2016), de modo que la caída de la tasa de salida procede menos de la exigencia que de la llegada de ofertas y del esfuerzo.

En suma, el desempleo friccional se entiende hoy como resultado de una doble búsqueda, de empleo y de trabajadores, y es en general transitorio. Para saber si es voluntario hay que incorporar las vacantes, como hace la teoría del emparejamiento.

2. Emparejamiento y curva de Beveridge

Origen: Diamond (1982), Mortensen (1982) y Pissarides (1985), Nobel en 2010, incorporaron las vacantes y el proceso de emparejamiento. Mortensen y Pissarides (1994) endogenizaron la destrucción de empleo, y Pissarides (2000) dio al modelo su forma canónica.

Idea: con información imperfecta coexisten vacantes y desempleados, algo imposible en un mercado walrasiano: el mercado no se vacía y hay desempleo aun con salarios flexibles.

2.1. Supuestos

Metodología: tiempo continuo y población activa normalizada a uno.

Entorno: información incompleta, porque los agentes conocen el proceso de emparejamiento pero no pueden coordinarse; costes de búsqueda y de transacción, y un tipo de interés real r exógeno.

Agentes: muchas empresas de un único puesto, vacante o cubierto, y muchos trabajadores neutrales al riesgo, una fracción u desempleada. Son idénticos, y su heterogeneidad real se resume en la función de emparejamiento.

Mercado de factores: cada puesto cubierto produce un flujo y; mantener una vacante cuesta c por unidad de tiempo; los puestos se destruyen a la tasa exógena λ, y el salario se negocia bilateralmente tras el emparejamiento.

2.2. Modelo

2.2.1. Funcionamiento del mercado

Ley de movimiento del desempleo: en tiempo continuo, la variación instantánea de la tasa de desempleo es la diferencia entre el flujo de empleados que pierden su empleo y el de desempleados que lo encuentran, a la tasa de salida p:

$$\dot{u} = \lambda(1-u) - p(\theta)\, u$$

Función de emparejamiento: es el elemento central. Como una urna en la que encesta una bola, da los emparejamientos por unidad de tiempo según las vacantes V y los desempleados U, y engloba las fricciones sin hacerlas explícitas. Suele ser Cobb-Douglas:

$$M = A\, V^{\alpha} U^{1-\alpha}$$

- $\alpha \in (0,1)$ es la elasticidad respecto a las vacantes y A la **eficiencia del emparejamiento**, que depende de factores estructurales como la movilidad o la eficacia de los servicios de empleo.
- Con rendimientos crecientes habría efectos de mercado denso y equilibrios múltiples (Diamond, 1982), y con decrecientes, congestión. Se suponen constantes, porque las elasticidades estimadas suman aproximadamente uno (Petrongolo y Pissarides, 2001).

Estrechez: la ratio $\theta \equiv V/U$ mide lo fácil que es encontrar trabajo. Si $\theta > 1$ el mercado es estrecho: es fácil encontrar trabajo y difícil cubrir vacantes. Si $\theta < 1$ es holgado, como España tras 2008. Con rendimientos constantes, las tasas de transición dependen solo de θ:

$$q(\theta) \equiv \frac{M}{V} = A\,\theta^{\alpha-1}, \;\; q' < 0 \qquad p(\theta) \equiv \frac{M}{U} = \theta\, q(\theta), \;\; p' > 0$$

- $q(\theta)$ es la tasa de cobertura de vacantes y $p(\theta)$ la tasa de salida del desempleo.

Un mercado más estrecho favorece a los trabajadores y perjudica a las empresas. Hay **externalidades de búsqueda o congestión**: cada buscador reduce la probabilidad de éxito de los de su lado y aumenta la del contrario. El salario no es el único mecanismo de ajuste, pues también lo son el número de buscadores y el de puestos.

2.2.2. Comportamiento de los agentes

Trabajadores: condición de participación: como en el modelo de Mortensen, pero con una tasa de llegada endógena, $p(\theta)$, el trabajador acepta si la ganancia de emparejarse es no negativa:

$$w \;\geq\; \underline{w} = rU_B^e = z + \frac{p(\theta)}{r+\lambda}\,(w - \underline{w})$$

- El salario de reserva es el ingreso durante la búsqueda más el valor esperado de emparejarse, y crece con la estrechez porque $p(\theta)$ acorta la espera, lo que fundamenta la curva de salarios. Procede de las ecuaciones de valor de Mortensen con la tasa de llegada endógena:

$$rU_B^e = z + p(\theta)\,(U_T^e - U_B^e) \qquad rU_T^e = w + \lambda\,(U_B^e - U_T^e)$$

Empresas: condición de creación de empleo: una vacante cuesta c y se cubre a la tasa $q(\theta)$. Un puesto cubierto rinde $y - w$ y se destruye a la tasa λ. Con libre entrada, el valor de la vacante es nulo y el de un puesto cubierto iguala el coste esperado de cubrirlo, $c/q(\theta)$, lo que da la **condición de creación de empleo** (CCE):

$$y = w + (r+\lambda)\,\frac{c}{q(\theta)}$$

- Esta condición es la CPO neoclásica con fricciones. La empresa crea empleo hasta que la producción iguala el salario más el coste capitalizado de la vacante. A mayor estrechez, más tarda en cubrirse la vacante y menor es el salario que puede pagar: la CCE es decreciente en el plano (θ, w). Procede de los valores de la vacante y del puesto cubierto, con $\Pi_V^e = 0$ por la libre entrada:

$$r\Pi_V^e = -c + q(\theta)\,(\Pi_C^e - \Pi_V^e) \qquad r\Pi_C^e = y - w + \lambda\,(\Pi_V^e - \Pi_C^e)$$

2.2.3. Equilibrio y curva de Beveridge

El equilibrio es un **estado estacionario**: las tasas de desempleo y de vacantes son constantes porque los flujos de entrada y salida se compensan.

Planteamiento: el emparejamiento genera una renta para ambas partes, y el salario la reparte mediante una **negociación à la Nash**, con $\gamma \in [0,1]$ el poder de negociación del trabajador:

$$w^* = \arg\max_{w} \left\{ (U_T^e - U_B^e)^{\gamma}\, (\Pi_C^e - \Pi_V^e)^{1-\gamma} \right\}$$

CPO: reparte el excedente total en proporción $\gamma : (1-\gamma)$, y la solución es el límite del juego de ofertas alternas de Rubinstein (1982) cuando el intervalo entre ofertas tiende a cero. La condición de reparto es:

$$U_T^e - U_B^e = \frac{\gamma}{1-\gamma}\,(\Pi_C^e - \Pi_V^e)$$

CSO: el producto de Nash es log-cóncavo, porque ambos excedentes son lineales en el salario, y el salario negociado es único.

Solución: con las ecuaciones de valor y la libre entrada se obtiene la **curva de salarios** en forma cerrada:

$$w = (1-\gamma)\, z + \gamma\,(y + c\,\theta)$$

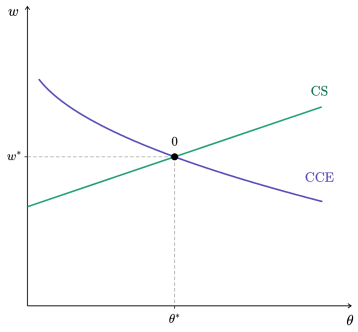
– $c\theta$ es el coste medio de contratación por desempleado, que la empresa ahorra y el trabajador captura según su poder de negociación.

Es creciente en (θ, w) y hace de oferta de trabajo, aunque incorpora el poder de negociación, la opción externa y los costes de búsqueda. Si $\gamma = 0$, el trabajador recibe solo su opción externa, y si $\gamma = 1$, toda la producción más el ahorro de búsqueda. La tasa de destrucción y el tipo de interés se cancelan en la forma cerrada.

Equilibrio: la curva de salarios (CS), creciente, y la CCE, decreciente, determinan θ^* y w^*:

$$\text{CS:}\quad w = (1-\gamma)\, z + \gamma\,(y + c\,\theta)$$
$$\text{CCE:}\quad y = w + (r+\lambda)\,\frac{c}{q(\theta)}$$

Ambos dependen solo de los parámetros y no de u, que se obtiene después.



Equilibrio del emparejamiento: la curva de salarios y la de creación de empleo fijan θ^* y w^*.

– *Desempleo de equilibrio*: en estado estacionario, $\dot{u} = 0$, los flujos de entrada y salida se igualan:

$$u^* = \frac{\lambda}{\lambda + p(\theta)} \quad \rightarrow \quad u^* = f\,[\lambda^+, \theta^-, A^-]$$

– *Curva de Beveridge* (CB): la misma condición en el plano desempleo-vacantes, $\lambda(1-u) = Av^{\alpha}u^{1-\alpha}$, da las combinaciones compatibles con un desempleo constante. Beveridge (1944) observó la relación sin modelarla, y el emparejamiento la deriva del flujo estacionario. Despejando las vacantes, la curva es:

$$v = \left[\frac{\lambda(1-u)}{A\,u^{1-\alpha}}\right]^{1/\alpha}$$

  · *Pendiente negativa*: en una recesión el equilibrio se desplaza sobre ella hacia la derecha.
  · *Convexa*: cada reducción del desempleo exige cada vez más vacantes.
  · *Posición*: más cerca del origen cuanto mayor es A o menor λ.

  Mide la eficiencia del mercado y la posición cíclica. Su corte con la recta $v = \theta^* u$ determina u^* y v^*. Ese desempleo es la **tasa de paro friccional**, igual a la tasa natural cuando el salario es el de equilibrio.

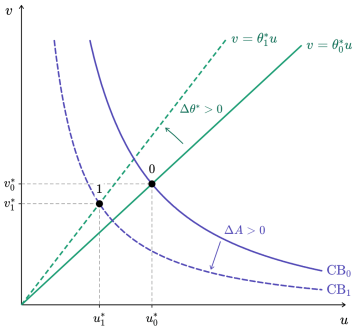
Bienestar: por las externalidades de búsqueda, el equilibrio solo es eficiente si $\gamma = 1 - \alpha$, la **condición de Hosios** (1990). Si γ es mayor, se crean pocas vacantes y hay demasiado desempleo; si es menor, demasiadas. El poder de negociación ha de igualar la elasticidad del emparejamiento respecto al desempleo, y fuera de ese caso hay margen para la intervención.

2.2.4. Estática comparativa y dinámica

Perturbaciones coyunturales (Blanchard y Diamond, 1989): alteran la rentabilidad de crear empleo sin cambiar A ni λ. Giran la recta $v = \theta^* u$ y mueven el equilibrio sobre la curva de Beveridge. Lo que hace menos provechoso crear un puesto reduce θ y las vacantes y eleva el desempleo, con efectos distintos sobre el salario:

– *Tipo de interés o cotizaciones*: bajan la CCE y el salario cae.
– *Subsidio z*: sube la CS al mejorar la opción externa, y el salario sube.
– *Coste de la vacante c*: baja la CCE y sube la CS, pero la forma cerrada muestra que el salario cae, porque la estrechez disminuye más que proporcionalmente y con ella $c\theta$.
– *Productividad y*: produce lo contrario. Suben la estrechez, las vacantes y el salario, y cae el desempleo.

Perturbaciones estructurales: cambian A o λ y desplazan la propia curva de Beveridge. Una mejora del emparejamiento (tecnologías de la información, movilidad, políticas activas) la acerca al origen y lleva el equilibrio de 0 a 1. Un empeoramiento (desánimo, desajuste de cualificaciones, paro de larga duración) la aleja. Con la curva más cerca del origen, cada nivel de desempleo requiere menos vacantes para mantener el empleo.



Mejora de la eficiencia del emparejamiento: la CB se acerca al origen y el equilibrio pasa de 0 a 1.

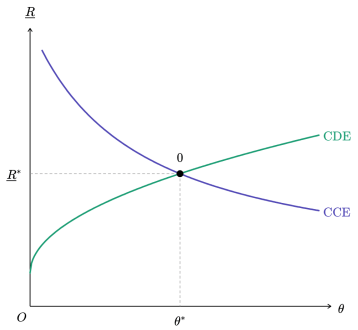
Dinámica: la estrechez es una **variable de salto**, porque abrir o cerrar vacantes no tiene coste de ajuste, y el desempleo una **variable de estado**, que solo cambia con los flujos. La economía salta verticalmente a la nueva recta de creación de empleo y se desplaza después por ella hasta la curva de Beveridge. Esta dinámica genera los bucles en sentido antihorario de los datos. En una recesión las vacantes caen antes de que suba el desempleo, y en la recuperación suben antes de que baje. El desempleo converge a un ritmo $\lambda + p(\theta)$, que tarda meses en completar el ajuste.

2.3. Extensiones y aplicaciones

2.3.1. Destrucción endógena (Mortensen y Pissarides, 1994)

Idea: cada puesto recibe perturbaciones idiosincrásicas de productividad y se destruye cuando esta cae por debajo de la **productividad de reserva**, $\underline{R}$, en la que se anula el excedente conjunto.

Razonamiento: $\underline{R}$ crece con la estrechez, porque mejora la opción externa del trabajador. La curva de destrucción, creciente en $(\theta, \underline{R})$, y la de creación, decreciente, determinan θ^* y $\underline{R}^*$, y con ellos la tasa de destrucción y el desempleo.



Destrucción endógena de empleo (Mortensen y Pissarides, 1994).

Extensiones: los costes de despido reducen $\underline{R}$ y la destrucción, pero también la creación, con efecto ambiguo sobre el desempleo, y las prestaciones elevan $\underline{R}$. El modelo es el marco habitual para evaluar la protección del empleo. La destrucción responde, además, al ciclo y se concentra en las recesiones (Davis y Haltiwanger, 1992).

Conclusiones: hace endógenos los dos flujos brutos, en línea con la evidencia, pero no distingue despidos y abandonos y deja ambiguo el efecto de las instituciones.

2.3.2. Histéresis

Definición: hay **histéresis** cuando el desempleo de equilibrio depende de la trayectoria del efectivo, de modo que una recesión eleva de forma persistente la tasa natural. Hay tres mecanismos:

1. *Salarial o de insiders-outsiders* (Blanchard y Summers, 1986; Lindbeck y Snower, 1988): los empleados negocian salarios que solo atienden a su propio empleo, y tras una recesión el salario no baja lo suficiente para recuperar a los despedidos.
2. *De emparejamiento* (Pissarides, 1992): el paro de larga duración provoca pérdida de capital humano y desánimo, que reducen A. Cuando las vacantes se recuperan, el desempleo es mayor porque la curva de Beveridge se ha desplazado hacia fuera.
3. *De capital* (Bean, 1989): la recesión reduce la inversión y el capital, y con él la demanda de trabajo a cada salario.

Las tres coinciden en que la estabilización tiene efectos permanentes, pero piden remedios distintos: moderación salarial y reforma de la negociación, políticas activas de empleo y estímulo a la inversión.

2.3.3. Los casos español y estadounidense

España: tras 2008 la curva de Beveridge se desplazó hacia fuera, una pérdida de eficiencia y no solo de demanda, por el desajuste de cualificaciones, la menor movilidad por la vivienda en propiedad (Oswald, 1996), el paro de larga duración y las diferencias regionales.

EE. UU. (2021-2022): tras la pandemia, las vacantes alcanzaron máximos con un paro similar al de 2019. Blanchard, Domash y Summers (2022) vieron una pérdida persistente de eficiencia, que exigiría más desempleo para reducir la inflación; Figura y Waller (2022), el tramo casi vertical de una curva estable, lectura que la normalización de 2023 favoreció.

Conclusiones: el emparejamiento es el marco de referencia del desempleo de equilibrio. Explica la coexistencia de vacantes y parados, microfundamenta las curvas de Beveridge y de salarios, y distingue perturbaciones coyunturales y estructurales. Su límite es el **puzzle de Shimer** (2005): calibrado con datos de EE. UU., genera mucha menos volatilidad de la observada, porque el salario negociado absorbe las perturbaciones. Se responde con rigidez salarial (Hall, 2005), más valor del no mercado (Hagedorn y Manovskii, 2008), costes fijos de emparejamiento (Pissarides, 2009) o heterogeneidad de puestos.

3. Costes de ajuste y dinámica de la demanda de trabajo

Si cambiar la plantilla tiene costes, la empresa tarda en llevarla a su nivel óptimo, y eso explica cómo evoluciona el empleo demandado. Del lado de la oferta, la dinámica procede de la sustitución intertemporal de Lucas-Rapping, $l_t^s/l_{t+1}^s = f[(w_t/w_{t+1})^+, r^+, \rho^-]$, sin costes de ajuste, pues el trabajador reasigna sus horas entre periodos sin fricciones. Los costes relevantes son, por ello, los de la demanda.

3.1. Concepto de costes de ajuste

Definición: los **costes de ajuste** son los que, además del precio del factor, paga la empresa por cambiar el ritmo al que incorpora factores: $C = C(\dot{L})$, con $C(0) = 0$ y crecientes con la magnitud del ajuste, $|\dot{L}|$. Formalmente, $C'(\dot{L}) > 0$ si $\dot{L} > 0$ y $C'(\dot{L}) < 0$ si $\dot{L} < 0$.

- *Internos* (Eisner y Strotz, 1963; Lucas, 1967): pérdidas temporales de eficiencia, implícitas, como mermas de producción o formación de los nuevos.
- *Externos*: explícitos, nacen de la relación con terceros, como costes de contratación o indemnizaciones.

Formas funcionales: determinan la dinámica del empleo.

1. *Convexos*: ajustar es tanto más caro cuanto más rápido ($C'' > 0$); simétricos o asimétricos si despedir cuesta más.
2. *Lineales*: proporcionales al número de trabajadores ajustados ($C'' = 0$).
3. *Fijos*: independientes del tamaño del ajuste, como un expediente de regulación de empleo.

En todos los casos **estabilizan el empleo a costa de la eficiencia**: la empresa suaviza la variación o no ajusta ante perturbaciones pequeñas, y el empleo varía menos que el producto.

3.2. Demanda de trabajo

3.2.1. Sin costes de ajuste

Planteamiento: la empresa elige la senda de empleo que maximiza el valor actual de sus beneficios:

$$\max_{\{L_t\}} V(0) = \int_0^\infty e^{-rt}\,[p_t f(L_t) - w_t L_t]\,dt \quad \rightarrow \quad p_t f'_L(L_t^*) = w_t$$

Solución: el problema es separable en el tiempo y la CPO de cada instante es la estática, suficiente porque $f'' < 0$. La demanda es la neoclásica y el empleo salta de inmediato a su óptimo.

3.2.2. Costes convexos (Holt et al., 1960; Eisner y Strotz, 1963)

Planteamiento: Holt, Modigliani, Muth y Simon (1960) y Eisner y Strotz (1963) introdujeron costes cuadráticos, $C = (b/2)\dot{L}^2$, con coste marginal creciente y simétrico. La empresa resuelve:

$$\max_{\{L_t\}} V(0) = \int_0^\infty e^{-rt}\left[p_t f(L_t) - w_t L_t - \frac{b}{2}\dot{L}_t^2\right] dt$$
$$\rightarrow\ p_t f'_L(L_t) = w_t + rb\dot{L}_t - b\ddot{L}_t$$

CPO: la empresa contrata hasta que el valor de la productividad marginal cubre el salario más el coste marginal de ajustar. En el estado estacionario se recupera $pf'_L(L^*) = w$.

CSO: el integrando es estrictamente cóncavo en $(L_t, \dot{L}_t)$, pues $f'' < 0$ y $b > 0$, y la condición de Euler, con la de transversalidad, es suficiente.

Dinámica: ante una perturbación permanente el ajuste es gradual y completo; ante una transitoria, gradual e incompleto, porque no compensa pagar por un cambio que se revertirá. Como el empleo va por detrás de la producción, **la productividad es procíclica**. Con costes asimétricos (Pfann y Palm, 1993) el ajuste es más lento en la dirección más cara, al despedir si las indemnizaciones lo encarecen.

Conclusiones: explica el ajuste gradual y la productividad procíclica, y da ecuaciones de ajuste parcial fáciles de estimar. Pero convexidad y simetría son supuestos de conveniencia que la evidencia no confirma, y no genera inacción ni saltos.

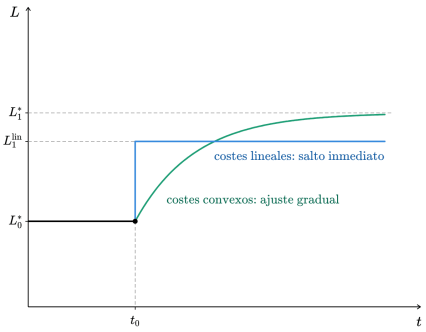
3.2.3. Costes lineales (Nickell, 1986)

Idea: Nickell (1986) mostró que, con un coste marginal de ajuste constante, $C = c|\dot{L}|$, no hay ganancia en la gradualidad y el ajuste, si se produce, es inmediato:

$$p_t f'_L(L_t^*) = w_t \pm rc$$

CPO: la productividad marginal cubre el salario más el flujo equivalente rc del coste de contratar, o menos el de despedir. Entre ambos valores la empresa no ajusta. La no diferenciabilidad en $\dot{L}_t = 0$ genera la **banda de inacción**.

Desarrollo gráfico: ante una perturbación permanente, la senda lineal es un escalón hasta L_1^{lin}, y la convexa se acerca gradualmente a L_1^*, más alto, porque el coste marginal cuadrático se anula en el estado estacionario, mientras que rc queda como cuña permanente sobre el salario.



Sendas de ajuste: escalón inmediato con costes lineales y aproximación gradual con convexos.

Dinámica: ante una perturbación permanente el ajuste es inmediato y de una sola vez, aunque se queda algo por debajo del nivel sin costes. Ante una transitoria pequeña la empresa no ajusta, y ante una grande ajusta de inmediato pero de forma incompleta. Con costes lineales asimétricos, el ajuste sigue siendo inmediato y la banda de inacción es asimétrica.

Conclusiones: explica la inacción y los saltos de los datos de empresa, y su versión asimétrica es el marco de los costes de despido. Pero no recoge el componente fijo de los costes.

3.2.4. Costes fijos (Hamermesh, 1989)

Idea: Hamermesh (1989) mostró, con datos mensuales de establecimientos, que el empleo se ajusta a saltos. Con un coste fijo k, la empresa solo ajusta cuando la ganancia de la plantilla óptima supera k, y entonces lo hace de golpe (una **regla** (S, s)).

Conclusiones: reproduce la inacción y los ajustes grandes e infrecuentes, que agregados dan una dinámica suave. Pero no conduce a una ecuación de ajuste sencilla y es difícil de estimar. La dinámica agregada depende, además, de la distribución de las desviaciones entre empresas.

3.3. Evidencia empírica y política

Balance empírico (Hamermesh y Pfann, 1996): la especificación cuadrática simétrica fue la más usada, pero los datos de empresa muestran costes asimétricos, no necesariamente convexos y con un componente fijo relevante. Importan además el margen intensivo, ajustar horas en lugar de trabajadores, y la heterogeneidad de trabajadores con distintos costes de ajuste.

Costes de despido (Bentolila y Bertola, 1990): con costes lineales y asimétricos, un mayor coste de despido reduce más los despidos en las recesiones que las contrataciones en las expansiones, porque la empresa descuenta el coste futuro al contratar pero lo paga con certeza al despedir. El efecto sobre el empleo medio es **pequeño y positivo**, y sobre los flujos, inequívoco (menos rotación y paro más largo). Hopenhayn y Rogerson (1993) añaden que frena la reasignación hacia las empresas más productivas.

Mercado dual español (Bentolila y Dolado, 1994): con indefinidos de alto coste de despido y temporales sin él, el ajuste recae sobre los temporales con independencia de su productividad, y los indefinidos presionan al alza los salarios como *insiders*. Ello explica la extraordinaria sensibilidad cíclica del empleo español desde 1984.

Política económica: la protección del empleo lo estabiliza a costa de la eficiencia, con sobreempleo en las recesiones y empleo inferior al óptimo en las expansiones. La política afecta al desempleo no solo a través de la búsqueda, sino también de la estructura de los costes de rotación.

1. Contratos implícitos (Baily, 1974; Gordon, 1974; Azariadis, 1975)

Esta línea explica la rigidez del salario real por el reparto del riesgo. Los trabajadores no pueden asegurarse en los mercados financieros frente a las fluctuaciones de su renta, y la empresa, con mejor acceso a ellos, les ofrece ese seguro a través del propio contrato de trabajo.

Origen: muchas relaciones laborales duran años, un rasgo que el mercado walrasiano no recoge. En vez de recurrir al mercado cada periodo, las empresas retienen a su plantilla. Estas teorías explican por qué, con ciertos fallos de mercado, los salarios pueden diferir de los de competencia perfecta.

Idea: las relaciones a largo plazo hacen que los contratos fijen rentas salariales estables, no contingentes al estado de la naturaleza, y actúen como seguro ante una demanda de trabajo fluctuante. La renta se desvincula de las condiciones de cada periodo, aunque el empleo puede seguir fijándose de forma eficiente. La mejor distribución del riesgo a largo plazo supone una renta salarial fija, menor en los estados favorables que el valor de lo producido. Se habla de **contrato implícito** porque esas cláusulas no se escriben: nada fija de forma expresa qué horas y qué salario corresponden a cada realización de la productividad.

1.1. Supuestos

Entorno: información perfecta y precios normalizados, de modo que el consumo real coincide con la renta real.

Agentes:

- *Trabajadores*: son aversos al riesgo, porque no pueden diversificar sus fuentes de renta (toda procede del trabajo, $c_i = w_i L_i$). El consumo da utilidad, con $u' > 0$ y $u'' < 0$, y el trabajo desutilidad, con $v' > 0$ y $v'' > 0$, en una utilidad esperada de von Neumann y Morgenstern:

$$u^e = \sum_{i=1}^{n} p_i[u(c_i) - v(L_i)]$$

  - Tienen una utilidad de reserva $\underline{u}$, y su participación, $u^e \geq \underline{u}$, es irrevocable (no hay movilidad laboral).
- *Empresa*: es neutral al riesgo, porque los mercados financieros le permiten diversificarlo. Su tecnología sufre *shocks* de productividad, A_i, que toman n valores con probabilidad p_i, y su beneficio esperado es:

$$\pi^e = \sum_{i=1}^{n} p_i[A_i f(L_i) - w_i L_i]$$

1.2. Modelo

Planteamiento: la empresa elige las horas y el consumo de cada estado para maximizar su beneficio esperado, sujeta a la participación del trabajador:

$$\max_{L_i, c_i} \pi^e = \sum p_i[A_i f(L_i) - c_i] \quad \text{s.a.: } \sum p_i[u(c_i) - v(L_i)] \geq \underline{u}$$

CPO: con λ el multiplicador de la condición de participación:

$$u'(c_i) = \frac{1}{\lambda} \ \ \forall i, \qquad A_i f'(L_i) = \frac{v'(L_i)}{u'(c_i)}$$

- La primera dice que la utilidad marginal del consumo, y por tanto el consumo y la renta salarial, son constantes en todos los estados (**aseguramiento completo**).
- La segunda, la condición de eficiencia del contrato, fija el empleo de cada estado donde la productividad marginal iguala la relación marginal de sustitución del trabajador, la desutilidad marginal de las horas en unidades de consumo.

CSO: la concavidad de u y de f y la convexidad de v hacen cóncavo el lagrangiano, y las CPO son necesarias y suficientes.

Solución: el contrato **estabiliza la renta laboral, no el salario por hora**. Con consumo constante, el salario por hora, $w_i = c^*/L_i$, se mueve en sentido contrario a las horas, y estas siguen respondiendo a los *shocks* según la condición de eficiencia. El empleo es eficiente en cada estado y no hay desempleo involuntario. Además, el trabajador está mejor en las recesiones, porque trabaja menos con el mismo consumo.

Extensiones: si solo la empresa observa el *shock*, puede declarar el estado que le convenga, y el contrato eficiente deja de ser compatible con los incentivos. El contrato óptimo genera entonces subempleo en los estados desfavorables (Grossman y Hart, 1981) o sobreempleo (Green y Kahn, 1983).

1.3. Conclusiones

Aspectos positivos: los contratos implícitos explican la estabilidad de la renta salarial. El salario cumple una doble función, porque paga el trabajo y asegura al trabajador frente al riesgo tecnológico. Responden a la incertidumbre y al capital humano específico entre empresas y trabajadores aversos al riesgo, y suavizan el consumo, aunque hacen más volátil el empleo.

Limitaciones: hay cuatro objeciones.

- *Ajuste*: con información simétrica el empleo es eficiente, de modo que el modelo no explica el desempleo involuntario (Akerlof y Miyazaki, 1980), y el ajuste recae en las horas más que en los despidos.
- *Nuevas contrataciones*: no explica por qué la empresa no ofrece salarios menores a quienes contrata.
- *Rigidez nominal*: explica la rigidez de la renta real, pero no la de los salarios nominales ante perturbaciones de demanda.
- *Vigencia*: populares en los años ochenta, desde los noventa los han superado los modelos de búsqueda y emparejamiento.

2. Modelos de negociación colectiva

Los modelos de negociación conciben el salario como un reparto de rentas entre la empresa y los trabajadores o su sindicato. El salario no resulta del cruce de oferta y demanda, sino del poder de negociación de cada parte, y el empleo depende de qué variables se negocien.

2.1. Fundamentos

2.1.1. Teoría y proceso de negociación

Definición: la **teoría de la negociación** estudia el reparto pactado, entre agentes racionales, de un excedente que solo surge si cooperan. Empresa y trabajadores se reparten la renta de la relación laboral, lo producido por encima de lo que cada uno obtendría sin acuerdo, fijando el salario y, en su caso, el empleo.

El resultado depende del **poder de negociación**, la capacidad de apropiarse de una fracción mayor del excedente, que aumenta con las alternativas fuera del acuerdo y con la capacidad de resistir un conflicto. La negociación se da en tres niveles: empleador y trabajador individual, empleador y sindicato, o patronal y sindicatos de un sector o de toda la economía.

2.1.2. Centralización (Calmfors y Driffill, 1988)

Idea: los casos extremos, centralizados o descentralizados, dan mejores resultados de desempleo que los intermedios.

Razonamiento: con negociación muy centralizada, los negociadores internalizan los efectos macroeconómicos de sus acuerdos y moderan sus demandas. Con negociación muy descentralizada, la competencia en el mercado de productos disciplina los salarios. La negociación sectorial combina poder suficiente para elevar los salarios con poca internalización. En su muestra, Austria o los países escandinavos y Japón, Suiza o EE. UU. tenían menos paro que Bélgica o los Países Bajos.

Conclusiones: explica una relación **en forma de joroba** entre centralización y desempleo y ha orientado el diseño institucional. Pero omite la coordinación, que puede lograrse sin centralizar, y la evidencia posterior apenas confirma la joroba.

2.1.3. Fundamentos microeconómicos

Edgeworth (1881): el acuerdo debe caer sobre la curva de contratos y ser eficiente en el sentido de Pareto, pero el reparto queda indeterminado.

Hicks (1932): el salario refleja un pulso entre la presión del sindicato, como la amenaza de huelga, y la resistencia de la empresa. Con información cierta, ambas calculan cuánto duraría el conflicto y firman desde el principio el salario que lo evita. Es sencillo, pero trata el poder de negociación como un dato.

Nash (1950): con un enfoque axiomático, muestra que la negociación tiene una única solución, la que maximiza el producto de las ganancias netas, si se cumplen cuatro axiomas.

1. *Optimalidad paretiana*: no queda ninguna ganancia mutua sin aprovechar.
2. *Invarianza ante transformaciones afines positivas* de la utilidad.
3. *Independencia de alternativas irrelevantes*: eliminar opciones no elegidas no altera la solución.
4. *Simetría*: jugadores en posiciones idénticas reciben lo mismo.

Sin el axioma de simetría se obtiene la **solución de Nash generalizada**, con u_i^0 las utilidades de desacuerdo y $\gamma \in [0,1]$ el poder negociador del jugador 1:

$$\max_{(u_1,u_2)\in\mathcal{U}} \Omega = (u_1 - u_1^0)^{\gamma}(u_2 - u_2^0)^{1-\gamma}$$

donde $\mathcal{U}$ es el conjunto de pares de utilidades factibles, y con $\gamma = 1/2$ se recupera la solución simétrica.

Stahl (1972) y **Rubinstein** (1982): la negociación es un juego de ofertas y contraofertas, con horizonte finito o infinito, resuelto como equilibrio de Nash perfecto en subjuegos. El poder procede de la paciencia relativa, y cuando el intervalo entre ofertas tiende a cero el reparto converge a la solución de Nash generalizada, lo que fundamenta γ (Binmore et al., 1986).

2.2. *Right to manage* (Nickell y Andrews, 1983)

Origen: los primeros modelos de negociación colectiva suponen un sindicato monopolista de la oferta de trabajo (Dunlop, 1944). Nickell y Andrews (1983) los generalizan con distintos grados de monopolización: ante el monopolio salarial del sindicato, la empresa retiene su "derecho a gestionar" su demanda de trabajo.

2.2.1. Supuestos

Entorno: información perfecta y negociación secuencial: el sindicato y la empresa negocian solo el salario, y después la empresa decide el empleo. Si la negociación fracasa, la empresa obtiene beneficio nulo y el sindicato, el salario de reserva, $\underline{w}$.

Agentes:

- *Sindicato*: agrupa a N^s afiliados iguales y aversos al riesgo, con poder de negociación $\gamma \in [0,1]$. Cada uno trabaja o está desempleado, con probabilidad de ser contratado $p = \text{mín}\{L/N^s, 1\}$:

$$U^e = p \cdot U(w) + (1-p) \cdot U(\underline{w})$$

- *Empresa*: con *shocks* tecnológicos A_θ, su beneficio es $\pi = A_\theta f(L) - wL$, con CPO $A_\theta f'(L) = w$.

2.2.2. Modelo

Planteamiento: el salario maximiza el producto de Nash generalizado, sabiendo que la empresa elegirá después el empleo sobre su curva de demanda:

$$\max_w \ \Omega(w) = \pi(w)^{1-\gamma} \cdot [U^e(w) - U(\underline{w})]^\gamma \quad \text{s.a.:} \begin{cases} L = L^d(w) \leq N^s \\ w \geq \underline{w} \end{cases}$$

CPO: iguala la ganancia marginal del sindicato, más salario pero menos empleo, con la pérdida marginal de beneficio de la empresa, cada una ponderada por el poder de su parte. Se supone que el producto de Nash es estrictamente cuasicóncavo, y el salario es único. Con $\eta = -\frac{w}{L}\frac{dL^d}{dw}$, la elasticidad-salario de la demanda de trabajo, la condición en logaritmos es:

$$\gamma \left[\frac{U'(w)}{U(w) - U(\underline{w})} - \frac{\eta}{w} \right] = (1-\gamma)\frac{L}{\pi}$$

Solución: el salario pactado sube con el poder sindical y con el salario de reserva, $w^* = w^*(\gamma^+, \underline{w}^+)$. Como caso límite, con un sindicato neutral al riesgo y $f(L) = L^\alpha$, la demanda es isoelástica, con elasticidad $\eta = 1/(1-\alpha)$, y la ratio $wL/\pi = \alpha/(1-\alpha) = \eta - 1$ es constante. El salario es entonces un margen sobre el de reserva:

$$w^* = \underline{w}\left(1 + \frac{\gamma}{\eta - 1}\right) = \begin{cases} \underline{w} & \text{si } \gamma = 0 \\ \frac{\eta}{\eta-1}\,\underline{w} & \text{si } \gamma = 1 \end{cases}$$

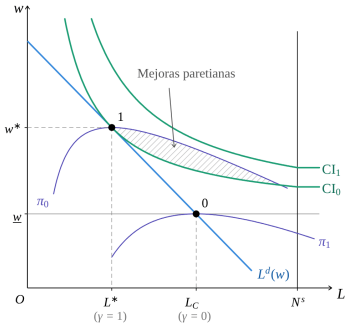
El margen crece con γ y decrece con η.

Equilibrio: depende del reparto del poder.

- $\gamma = 0$: el sindicato acepta $\underline{w}$ y se logra el equilibrio competitivo, con eficiencia asignativa y productiva.
- $\gamma = 1$: es la **solución de Dunlop**, el monopolio salarial, en la tangencia entre la curva de indiferencia del sindicato y la demanda. No es óptimo de Pareto y hay desempleo involuntario.
- $\gamma \in (0,1)$: equilibrios intermedios, con **contratos ineficientes** en el sentido de Pareto.

Así, el poder de negociación eleva los salarios sobre los competitivos y genera desempleo involuntario.

Desarrollo gráfico: en el plano empleo-salario, las curvas de indiferencia del sindicato son decrecientes y convexas, y horizontales a partir de N^s y en $\underline{w}$. Las isobeneficio son cóncavas, y la demanda une sus máximos. Con $\gamma = 1$ el equilibrio es el punto 1, y la lente entre CI_0 y π_0 recoge contratos que mejorarían a ambas partes.



Modelo *right to manage*: el acuerdo se sitúa sobre la demanda de trabajo y es ineficiente.

Estática comparativa: con demanda isoelástica, un *shock* que la desplace no altera su elasticidad, que es la que fija el salario. El salario es rígido y **el ajuste se hace vía cantidades**, con un empleo muy volátil.

2.2.3. Conclusiones

Aspectos positivos: anida como casos límite el monopolio de Dunlop y el equilibrio competitivo, y su regla salarial es la base de la curva del salario real negociado (SRN) del modelo de la NAIRU (Layard et al., 1991).

Limitaciones: el carácter secuencial hace la solución ineficiente siempre que $\gamma > 0$, porque el acuerdo queda sobre la curva de demanda y deja sin explotar ganancias mutuas.

2.3. Contratos eficientes (McDonald y Solow, 1981)

Idea: ambas partes acuerdan a la vez salario y empleo, y el acuerdo se sitúa sobre la **curva de contratos**, fuera de la demanda de trabajo, por lo que es eficiente en el sentido de Pareto, aunque el empleo puede diferir del competitivo.

2.3.1. Supuestos

Entorno: información perfecta y contexto estático. Las partes negocian a la vez todo el convenio, a diferencia del *right to manage*: el salario y el empleo en los contratos débiles, y el salario y una prestación para los afiliados desempleados en los fuertes. Si la negociación fracasa, la empresa obtiene beneficio nulo y el sindicato, $\underline{w}$.

Agentes: sindicato y empresa, los del *right to manage*, aunque ahora el sindicato puede tener cualquier perfil de riesgo.

2.3.2. Modelo

Planteamiento: cada punto de la curva de contratos maximiza el beneficio dada una utilidad mínima del sindicato:

$$\max_{w,\,L} \ \pi = A_\theta f(L) - wL \quad \text{s.a.: } U^e(w,L) \geq \bar{U}$$

CPO: la tangencia entre la curva de indiferencia del sindicato, de pendiente $-[U(w) - U(\underline{w})]/[L\,U'(w)]$, y la isobeneficio, de pendiente $(A_\theta f'(L) - w)/L$, da la curva de contratos:

$$\frac{U(w) - U(\underline{w})}{U'(w)} = w - A_\theta f'(L)$$

CSO: con curvas de indiferencia convexas e isobeneficio cóncavas, la tangencia es un máximo.

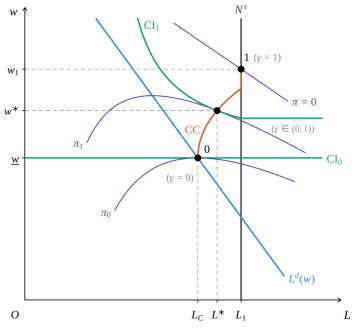
Solución: la pendiente de la curva de contratos depende del perfil de riesgo del sindicato: positiva con aversión, porque usa su poder para cubrir a sus miembros frente al desempleo y para subir el salario; vertical en el empleo competitivo con neutralidad, y negativa con propensión al riesgo. Con neutralidad, la condición se reduce a $A_\theta f'(L) = \underline{w}$; con aversión, $A_\theta f'(L) < \underline{w}$ y el empleo supera al competitivo.

En los contratos fuertes se negocia también un subsidio z que la empresa paga a los afiliados desempleados. Con seguro completo, $w = \underline{w} + z$, los afiliados obtienen lo mismo trabajen o no, y al sindicato le es indiferente el empleo. Para la empresa, con beneficio $\pi = A_\theta f(L) - wL - z(N^s - L)$, cada contratación cuesta el salario pero ahorra el subsidio:

$$A_\theta f'(L) = w - z = \underline{w}$$

Equilibrio: se sitúa sobre la curva de contratos, en el punto que fija γ.

- *Débiles*: con $\gamma = 0$ hay eficiencia asignativa y productiva; con $\gamma > 0$ y aversión al riesgo hay **sobreempleo**, eficiencia asignativa pero no productiva, por lo que los contratos son **débilmente eficientes**.
- *Fuertes*: las curvas de indiferencia son horizontales y la curva de contratos es vertical. El empleo es siempre el competitivo y el salario queda entre $\underline{w}$ ($\gamma = 0$) y w_1 ($\gamma = 1$). El poder negociador solo reparte la renta, sin crear ineficiencias, por lo que son **fuertemente eficientes**.



Contratos eficientes débiles: la curva de contratos, a la derecha de la demanda, da sobreempleo.

Estática comparativa: un aumento de γ desplaza el acuerdo sobre la curva de contratos. En los débiles, con aversión al riesgo, suben el salario y el empleo, y el sobreempleo crece con γ. En los fuertes solo sube el salario, y las rentas pasan de la empresa a los trabajadores.

2.3.3. Conclusiones

Aspectos positivos: el poder sindical no genera por sí mismo desempleo, porque si se negocia el empleo, el acuerdo es eficiente y el empleo puede superar al competitivo, y si además se asegura a los afiliados, el poder solo redistribuye. En EE. UU. algunos convenios negocian prestaciones complementarias, como en el automóvil.

Limitaciones: los convenios rara vez fijan el empleo, y un contrato fuera de la demanda es difícil de hacer cumplir, porque la empresa tiene incentivos a volver a ella. Además, los subsidios negociados no dan cobertura perfecta, y las negociaciones se dilatan.

2.4. *Insiders-outsiders*

Idea: Lindbeck y Snower (1984, 1988) distinguen a los *insiders*, que ya trabajan en la empresa y cuyos intereses recoge el convenio, de los *outsiders*, sin contrato. La asimetría de poder a favor de los primeros genera rigidez salarial, porque miran por su bienestar y no por el de los *outsiders*. Sus causas son la *descapitalización* de los parados de larga duración, unos *costes de rotación* elevados, que impiden a los *outsiders* competir por los puestos, las *rentas situacionales* por capital humano específico, la negativa a *cooperar* con los recién contratados (Lindbeck y Snower, 1988) y *factores institucionales*, como la afiliación sindical y la protección del empleo.

2.4.1. Supuestos

Entorno: modelo estático, con información perfecta y el trabajo como único factor: la empresa produce $Af(L)$, con $L = L_I + L_O$.

Agentes: los L_I *insiders*, ya empleados, son los únicos que negocian. Primero quieren conservar el empleo y después subir su salario, sin interés en crear empleo para los *outsiders*, cuya oferta es ilimitada al salario w_O. Ambos son igual de productivos, pero sustituir a un *insider* cuesta el despido, d, y la contratación, c, y los *outsiders* no pueden ofrecerse por menos (**salarios dependientes**).

2.4.2. Modelo

Planteamiento: en la primera etapa, los *insiders* fijan el salario más alto que no lleva a la empresa a despedirlos ni a sustituirlos por *outsiders*:

$$\max_{w_I} \; w_I \quad \text{s.a.:} \begin{cases} w_I \le Af'(L_I) + d \\ w_I \le w_O + c + d \end{cases}$$

En la segunda, la empresa decide cuántos *outsiders* contrata:

$$\max_{L_O \ge 0} \; \pi = Af(L_I + L_O) - w_I L_I - (w_O + c)L_O$$

CPO: los *insiders* saturan la restricción más exigente y se apropian de toda la renta que les conceden los costes de rotación. La empresa contrata *outsiders* mientras su productividad marginal cubra su salario más el coste de contratación:

$$w_I = \min\{Af'(L_I) + d,\; w_O + c + d\}$$
$$Af'(L_I + L_O) \le w_O + c, \quad \text{con igualdad si } L_O > 0$$

Solución: w_I crece con la productividad A hasta el techo $w_O + c + d$, y es tanto mayor cuanto mayores son los costes de rotación, fuente del poder de los *insiders*.

Equilibrio: $\text{Pmg}_L + d$, decreciente, alcanza el techo $w_O + d + c$ en L_1 e iguala $\underline{w}$ en L_2.

- *Caso A*, $L_I < L_1$: hay pocos *insiders*, que exigen el techo, y la empresa contrata además *outsiders*.
- *Caso B*, entre L_1 y L_2: el salario es $\text{Pmg}_L + d$, y la empresa ni despide ni contrata.
- *Caso C*, $L_I > L_2$: $\text{Pmg}_L + d < \underline{w}$, y la empresa despide hasta L_2.

Desarrollo gráfico: en el plano empleo de los *insiders*-salario, el techo es una horizontal y $\text{Pmg}_L + d$, decreciente, la corta en L_1.

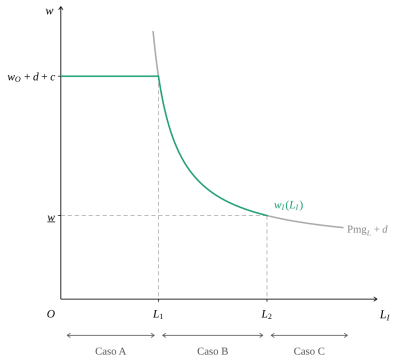
Estática comparativa: el salario es anticíclico a través de L_I. En una recesión la empresa despide *insiders*, y los que quedan, con mayor productividad marginal, ganan poder para subir el salario. En los auges, con más *insiders*, $f'(L_I)$ es menor y aceptan un salario más bajo a cambio de asegurar sus puestos.

Extensiones: predice la marginación de los *outsiders*, salarios anticíclicos bajo ciertas condiciones e histéresis. Recomienda reducir los costes de rotación, introducir cláusulas de descuelgue y formar a los parados.

2.4.3. Conclusiones

Aspectos positivos: explica el **desempleo involuntario**, porque hay quien trabajaría al salario de reserva y no puede; la **histéresis**, porque con menos *insiders* el salario sube e impide reabsorber a los despedidos, y la dualidad del mercado de trabajo.

Limitaciones: sin coste de despido, el convenio igualaría la productividad marginal al coste de oportunidad de los *insiders*; sin oferta externa ilimitada, el salario subiría, y sin salarios dependientes, el supuesto más fuerte, los *outsiders* se contratarían al salario de mercado.



Modelo *insiders-outsiders*: los internos fijan el salario que garantiza su empleo y dejan fuera a los externos.

En suma, el salario reparte una renta según el poder de cada parte, y el empleo depende de lo que se negocia: solo el salario da desempleo involuntario; salario y empleo, eficiencia, y la heterogeneidad, persistencia del paro.

3. Salarios de eficiencia

En los modelos de salarios de eficiencia la propia empresa paga un salario elevado, porque, además de costes, le reporta beneficios.

3.1. Concepto: causas y origen

Hipótesis: pagar más encarece la mano de obra, pero también reporta ganancias a la empresa. Hay varias líneas de argumentación (Yellen, 1984):

- *Nutricionistas* (Leibenstein, 1957): en países menos desarrollados, un mayor salario mejora la salud y la productividad.
- *Sociológicas* (Akerlof, 1982): un salario superior al de mercado eleva la moral, y los trabajadores lo corresponden, como en un intercambio de regalos, con más esfuerzo.
- *Rotación* (Stiglitz, 1974; Salop, 1979): fomenta la permanencia y reduce los costes de rotación.
- *Selección adversa* (Weiss, 1980): atrae a trabajadores más capacitados, cuya calidad la empresa no observa.
- *Riesgo moral* (Shapiro y Stiglitz, 1984): si el esfuerzo no es observable, hace costoso el despido y disuade de holgazanear.

Todas son incompatibles con el mercado walrasiano, donde la empresa toma el salario como dado y este vacía el mercado.

3.2. Condición de Solow (1979)

Idea: Solow formalizó la hipótesis en forma reducida. El esfuerzo es una función del salario que la empresa conoce, y la empresa elige el salario que minimiza el coste por unidad de trabajo efectivo.

3.2.1. Supuestos

Entorno: corto plazo, muchas empresas competitivas, información perfecta y precios normalizados. La producción depende del trabajo efectivo, $e(w) \cdot L$.

Agentes: el esfuerzo no puede fijarse por contrato, pero la empresa sabe cómo responde al salario: crece más que proporcionalmente para salarios bajos y menos para salarios altos.

3.2.2. Modelo

Planteamiento: la empresa elige salario y empleo:

$$\max_{w,L} \; \pi = f[e(w) \cdot L] - wL$$

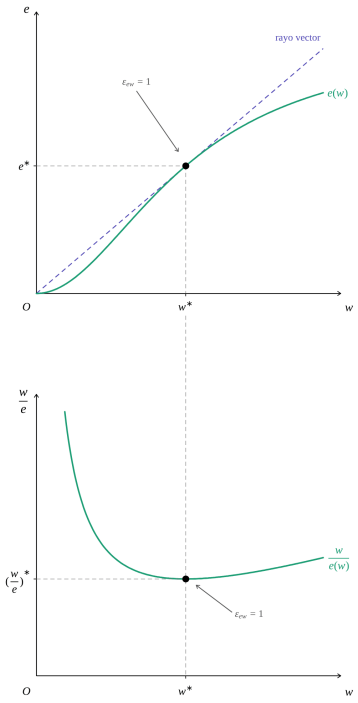
CPO: la empresa contrata hasta que la productividad marginal del trabajo efectivo iguala su coste, y el salario cumple la **condición de Solow**: la elasticidad-salario del esfuerzo es unitaria, sea cual sea la situación económica:

$$f'(eL) = \frac{w}{e}, \qquad \varepsilon_{ew} = \frac{w}{e} \cdot \frac{\partial e}{\partial w} = 1$$

La segunda combina la derivada respecto del salario, $f'(eL)\,e'(w) = 1$, con la primera.

CSO: la forma de $e(w)$, convexa y luego cóncava, garantiza que la elasticidad unitaria es un mínimo del coste unitario $w/e(w)$.

Desarrollo gráfico: en el panel superior, la función de esfuerzo es primero convexa y después cóncava, y la pendiente de cada rayo desde el origen es el esfuerzo por unidad de salario, e/w. En el inferior, el coste por unidad de trabajo efectivo, $w/e(w)$, es la inversa de esa pendiente. El equilibrio está donde la función de esfuerzo es tangente al rayo de máxima pendiente, con salario w^*, esfuerzo e^* y coste unitario mínimo. El coste unitario tiene forma de U, porque decrece mientras la elasticidad del esfuerzo supera la unidad y crece cuando es menor.



Condición de Solow: el salario de eficiencia hace unitaria la elasticidad del esfuerzo.

Extensiones: el salario de eficiencia puede quedar por encima del de vaciado y generar **desempleo involuntario**, porque la empresa contrata hasta igualar la productividad marginal a ese salario, al margen de la oferta. Además, ante un *shock* negativo no rebaja el salario, porque el menor esfuerzo costaría más que el ahorro, de modo que hay **rigidez salarial** y el ajuste es vía cantidades.

Gráficamente, en el plano empleo-salario, la empresa fija w^* por encima del salario de vaciado, w_C, y contrata sobre la demanda hasta L_0. A ese salario se ofrecen $L^s(w^*)$ trabajadores, y la diferencia es desempleo involuntario. Ante un *shock* negativo, $\Delta A<0$, la demanda se desplaza, el salario no cambia y el empleo cae de L_0 a L_1.

3.2.3. Conclusiones

Aspectos positivos: explica por qué, a corto plazo, las variaciones de la demanda de trabajo apenas mueven el salario real y se reflejan en el empleo.

Limitaciones: hay cuatro.

- *Miopía empresarial*: no deja a las empresas usar otros incentivos.
- *Permanencia*: el salario real quedaría inmóvil indefinidamente, salvo que el esfuerzo dependa también de los salarios de otras empresas y del paro.
- *Oferta laboral*: no se tiene en cuenta.
- *Forma reducida*: la principal es que $e(w)$ se postula sin explicar por qué el esfuerzo responde al salario.

3.3. *No-shirking* (Shapiro y Stiglitz, 1984)

Idea: el esfuerzo no es observable y **el desempleo actúa como mecanismo disciplinario**, porque la empresa paga más que el salario de vaciado para que perder el empleo sea costoso.

3.3.1. Supuestos

Entorno: se mantienen los supuestos de Solow salvo la información perfecta.

Agentes: hay riesgo moral, porque el esfuerzo no es observable y es discreto, $e=0$ o $e=e^*$, y hay conflicto de intereses: más esfuerzo da más ingresos a la empresa y más desutilidad al trabajador.

3.3.2. Modelo

Planteamiento: el trabajador determina la oferta de trabajo, y la empresa, la demanda.

- *Oferta*: el trabajador obtiene $w-e^*$ si trabaja y se esfuerza, w si holgazanea y $\bar{w}$ si está desempleado. Las transiciones siguen procesos de Poisson, con λ_R la salida exógena del empleo, λ_D la probabilidad de que el holgazán sea descubierto y despedido y a la salida del desempleo. Con la tasa de descuento r, las ecuaciones de valor del que se esfuerza, del que holgazanea y del desempleado son:

$$rU_T^e = w - e^* + \lambda_R(U_D^e - U_T^e)$$
$$rU_H^e = w + (\lambda_R+\lambda_D)(U_D^e - U_H^e)$$
$$rU_D^e = \bar{w} + a(U_T^e - U_D^e)$$

- *Demanda*: la empresa contrata hasta que la productividad marginal iguala el salario, y fija este de modo que trabajar con esfuerzo valga al menos lo que holgazanear, $U_T^e(w^*)\geq U_H^e$.

Salario de eficiencia: el valor de cada estado cumple una ecuación de valor de un activo, en la que el rendimiento iguala el flujo de utilidad más la ganancia o pérdida esperada de cambiar de estado. Igualando el valor de esforzarse y el de holgazanear, la renta del empleo debe igualar la desutilidad del esfuerzo dividida por la probabilidad de ser descubierto, $U_T^e-U_D^e=e^*/\lambda_D$, lo que da:

$$w^* = \bar{w} + e^* + \frac{e^*}{\lambda_D}(a+\lambda_R+r)$$

- Crece con la renta del desempleado, el esfuerzo, la facilidad para encontrar trabajo, la rotación, porque holgazanear arriesga menos, y la impaciencia; decrece con la probabilidad de detección.

NSC: en estado estacionario, las entradas y salidas del desempleo coinciden, $a(N-L)=\lambda_R L$, y se obtiene la **condición de no escaqueo**:

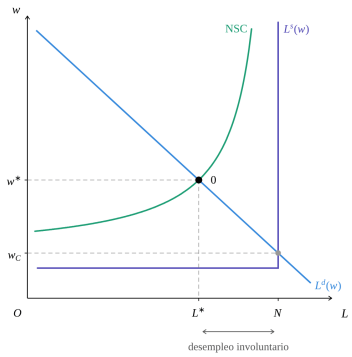
$$w = \bar{w} + e^* + \frac{e^*}{\lambda_D}\left(\frac{L}{N-L}\lambda_R + \lambda_R + r\right)$$

- Es creciente en el empleo, porque con menos paro es más fácil encontrar otro trabajo, y asintótica a la oferta. Con pleno empleo el despido no cuesta nada y ningún salario evita la holgazanería.

Equilibrio: con supervisión perfecta, la demanda corta a la oferta y hay pleno empleo walrasiano. Con información asimétrica, la NSC corta a la demanda con un salario superior al de vaciado y **desempleo involuntario**. Este equilibrio es un *second best*, porque supera la asimetría de información a costa de un desempleo improductivo.

Los parados trabajarían con esfuerzo por un salario menor, pero no pueden lograr que baje, porque las empresas saben que, con un salario inferior, no se esforzarían. Si todas pagasen el salario de vaciado, no habría desempleo y nadie se esforzaría, porque el despedido hallaría de inmediato otro empleo con igual salario. El desempleo es, así, el mecanismo de disciplina.

Desarrollo gráfico: la oferta es horizontal al salario $\bar{w}+e^*$, que compensa la renta del desempleado y la desutilidad del esfuerzo, hasta el pleno empleo, N, donde se vuelve vertical. La NSC corta a la demanda en el punto 0, con w^* y L^*, y $N-L^*$ es desempleo involuntario.



Shapiro y Stiglitz: la NSC corta a la demanda por encima de la oferta y genera paro de equilibrio.

Estática comparativa:

- *Shock negativo de demanda*: cae el salario de eficiencia, pero sigue por encima del de vaciado. El ajuste es sobre todo vía cantidades, y mayor cuanto más alto es el paro, porque la NSC es allí más plana.
- *Desplazamiento de la NSC*: más renta del desempleado, más rotación, más descuento o menor detección la suben, con más paro y salario mayor.
- *Caída de la población activa*: la oferta se traslada a la izquierda y la NSC se hace más empinada. Sube el salario y baja el empleo, pero el paro disminuye, porque, como $\frac{L}{N-L}\lambda_R+\lambda_R=\lambda_R\frac{N}{N-L}$, la NSC solo depende de la tasa de paro y un salario mayor exige una tasa menor.
- *Más esfuerzo requerido*: suben la NSC y la oferta, y aumenta el desempleo.

3.3.3. Conclusiones

Aspectos positivos: explica el desempleo involuntario como equilibrio con empresas maximizadoras, y predice que el paro sube con las prestaciones y baja con la capacidad de supervisión.

Limitaciones: fianzas o salarios crecientes con la antigüedad disuadirían la holgazanería sin desempleo (Lazear, 1979; Carmichael, 1985), y el paro de equilibrio depende de parámetros difíciles de medir, como la probabilidad de detección.

1. Fundamentos de la TNP, la NAIRU y la persistencia

El análisis arranca del debate entre la visión neoclásica y la keynesiana y continúa con la síntesis neoclásica, el monetarismo y los neokeynesianos del desequilibrio.

1.1. Visión neoclásica y keynesiana

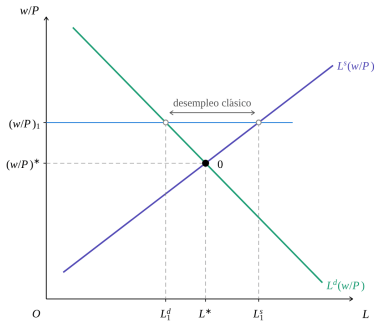
El debate fija los dos polos de toda la teoría posterior: el desempleo como fenómeno de oferta, corregible con flexibilidad, o como fenómeno de demanda, corregible con estabilización.

1.1.1. La visión neoclásica

Idea: para Marshall y Pigou (este último, 1933), el trabajo se compra y se vende como cualquier bien. Su precio, el salario real, refleja la escasez relativa de trabajo y, como se ajusta libremente, el mercado se vacía y hay pleno empleo.

Razonamiento: con agentes optimizadores, el equilibrio es único, estable y eficiente. Para que haya desempleo involuntario algo tiene que impedir el ajuste del salario (el *desempleo clásico* de Malinvaud). La receta son políticas de oferta que supriman esas rigideces, porque las de demanda no mueven las variables reales. Sin rigideces, el desempleo solo puede ser voluntario o friccional.

Desarrollo gráfico: en el plano empleo-salario real, la demanda de trabajo, decreciente, y la oferta, creciente, se cortan en $(w/P)^*$ y L^*, con pleno empleo. Con un salario rígido $(w/P)_1$ por encima del de equilibrio, la oferta supera a la demanda y la diferencia es desempleo clásico.



Mercado de trabajo neoclásico: con salario real flexible hay pleno empleo.

Conclusiones: es un marco coherente, con salario y empleo resultado de decisiones optimizadoras, y la referencia de largo plazo de la teoría posterior. Destaca el papel de las rigideces institucionales y de las políticas de oferta. Pero no puede explicar el paro masivo y persistente de la Gran Depresión.

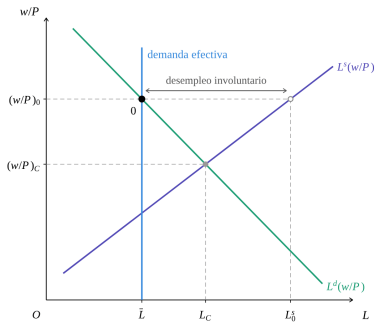
1.1.2. La visión keynesiana

Origen: en la Gran Depresión el paro de EE. UU. alcanzó el 24,9 % en 1933 (Lebergott, 1964), difícilmente atribuible a decisiones voluntarias. Keynes (*The General Theory of Employment, Interest and Money*, 1936) rompe con la idea de que el desempleo es solo voluntario.

Idea: el desempleo involuntario es posible aunque los salarios sean flexibles, porque el empleo lo determina la **demanda efectiva** en el mercado de bienes y no el salario.

Razonamiento: no se trata de ilusión monetaria. A los trabajadores les importa su salario relativo: se oponen a bajadas del salario nominal, que les harían perder terreno, pero aceptan el menor salario real que deja una subida general de precios. Una bajada salarial no eleva el empleo si no aumenta la demanda agregada, de modo que la flexibilidad ni basta ni es imprescindible para el pleno empleo.

Desarrollo gráfico: la demanda efectiva es una vertical en el empleo $\bar{L}$. Las empresas están sobre su demanda de trabajo, y al salario real resultante los trabajadores querrían ofrecer más: la diferencia es desempleo involuntario.



Mercado de trabajo keynesiano: la demanda efectiva fija $\bar{L}$ y aparece paro involuntario.

Conclusiones: explica el paro masivo por insuficiencia de demanda efectiva, y no por decisiones de los trabajadores, y fundamenta las políticas de demanda, que al estabilizar la demanda agregada estabilizan también el empleo. Pero la resistencia de los salarios nominales a la baja se justifica de manera informal, sin microfundamentos, y deja en segundo plano la oferta y la inflación, que incorporan la síntesis y el monetarismo.

1.2. Síntesis neoclásica

La síntesis acepta el equilibrio neoclásico a largo plazo y trata el desempleo keynesiano como un problema transitorio que las políticas de demanda pueden corregir.

1.2.1. La curva de Phillips (Phillips, 1958; Lipsey, 1960)

Origen: Phillips comprobó, con datos del Reino Unido de 1861 a 1957, que la inflación salarial era mayor cuanto menor era el desempleo, en una relación no lineal.

Razonamiento: Lipsey le dio fundamento con tres supuestos:

- La inflación salarial crece con el exceso de demanda de trabajo, $\dot{w}/w = \gamma Z_L$.
- El exceso de demanda decrece con el desempleo de forma no lineal, lo que explica la curvatura, y la relación linealizada es $Z_L = \nu(u - \bar{u})$, con $\nu < 0$ y $\bar{u}$ la tasa friccional a la que se anula.
- La productividad es constante, y la inflación salarial se traslada a los precios, $\dot{w}/w = \pi$.

Se obtiene la curva de Phillips sin expectativas:

$$\pi_t = -\kappa_u(u_t - \bar{u}), \qquad \kappa_u = -\gamma\nu > 0$$

Como Lipsey obtenía curvas por industria y las agregaba, la inflación depende también de la distribución del desempleo entre industrias. La curva establece una disyuntiva (*trade-off*) estable, que la ley de Okun traslada al plano producción-inflación, y Samuelson y Solow (1960) vieron en ella un **menú de política económica**: con un "ajuste fino" se podía comprar un paro bajo al precio de una inflación persistente.

Conclusiones: fue la primera relación cuantificable entre inflación y empleo y el instrumento central de los años sesenta. Pero desde 1970 la **estanflación**, con los *shocks* del petróleo, la caída de la productividad, el poder sindical y el fin de Bretton Woods, la hizo inestable. Además, solo se sostiene a cualquier plazo si los trabajadores sufren una ilusión monetaria permanente, con expectativas estáticas, $\pi_t^e = 0$.

1.2.2. La ley de Okun (1962)

Idea: la desviación del desempleo respecto a su nivel de equilibrio es proporcional a la brecha de producción, con ϕ el coeficiente de Okun:

$$y_t - \bar{y}_t = -\phi\,(u_t - \bar{u})$$

Razonamiento: en la estimación original de Okun, ϕ valía en torno a tres, es decir, cada punto de paro por encima del natural costaba unos tres puntos de producción. Es mayor que la participación del trabajo porque el empleo se ajusta junto con las horas, la participación y la productividad, por el atesoramiento de trabajo. En versión dinámica, $\Delta u_t = -\frac{1}{\phi}\,(g_{Y,t} - \bar{g}_Y)$: el desempleo solo cae si la economía crece por encima de su potencial.

Conclusiones: sigue siendo el puente entre la tasa natural y la brecha de producción que usa la política de estabilización. Pero no tiene fundamento teórico propio y su coeficiente difiere entre países, aunque dentro de la mayoría ha sido estable, también en la Gran Recesión (Ball et al., 2017). En España es bajo, porque el desempleo responde con mucha intensidad a la producción.

1.3. Monetarismo

El monetarismo acepta el marco de la síntesis, pero niega su resultado central, la disyuntiva estable, al introducir las expectativas en la curva de Phillips.

1.3.1. Curva ampliada y tasa natural (Friedman, 1968; Phelps, 1967)

Idea: no hay disyuntiva permanente entre inflación y desempleo, porque los agentes incorporan sus expectativas de inflación a la negociación salarial. A largo plazo, el desempleo vuelve a su **tasa natural**, determinada por factores reales.

Expectativas adaptativas: a los agentes les importa el poder adquisitivo de su salario, y Friedman supone que corrigen su expectativa en una fracción λ del error cometido, la velocidad de aprendizaje. Resolviendo recursivamente, la inflación esperada es una media ponderada de la pasada, con pesos geométricamente decrecientes:

$$\pi_{t+1}^e = \pi_t^e + \lambda(\pi_t - \pi_t^e) = \lambda\sum_{j=0}^{\infty}(1-\lambda)^j\pi_{t-j}, \qquad \lambda \in (0,1]$$

- Por eso, si la inflación sube de forma sostenida, los agentes la subestiman sistemáticamente.

Curva ampliada por las expectativas: la inflación es la esperada más el efecto de la presión del mercado de trabajo. La curva original es el caso $\pi_t^e = 0$:

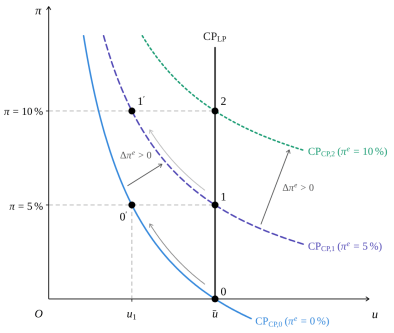
$$\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u}), \qquad \kappa_u > 0$$

- Si $\pi_t \neq \pi_t^e$, los salarios reales negociados resultan menores o mayores de lo previsto y el desempleo se aparta de $\bar{u}$: hay *trade-off*, pero solo mientras dura el error, y una curva de corto plazo por cada nivel de inflación esperada.
- Si $\pi_t = \pi_t^e$, el desempleo está en su tasa natural cualquiera que sea la inflación, y la curva de largo plazo es **vertical** sobre $\bar{u}$.

Tasa natural: es el desempleo del equilibrio de largo plazo, cuando nadie se equivoca. Abarca el paro friccional y el estructural y depende de factores reales e institucionales, como la eficiencia del emparejamiento, las instituciones laborales o la competencia en los mercados de bienes.

Desarrollo gráfico: en el plano desempleo-inflación, cada curva de corto plazo corresponde a una inflación esperada, del 0 %, el 5 % o el 10 %, y es decreciente, porque con las expectativas fijas un desempleo menor eleva los salarios nominales y los precios. La de largo plazo es vertical sobre $\bar{u}$, porque recoge las situaciones en que las expectativas se cumplen.

Dinámica: una expansión que reduce el desempleo a u_1 lleva la economía del punto 0 al 0 sobre la curva de corto plazo inicial, con una inflación del 5 %. Al revisarse las expectativas, la curva sube y el desempleo vuelve a $\bar{u}$ en el punto 1, con una inflación del 5 % ya esperada. Si la autoridad insiste en mantener u_1, el proceso se repite con una inflación cada vez mayor.



Curva de Phillips ampliada: curvas de corto plazo que se desplazan y curva vertical a largo plazo.

Conclusiones: la tasa natural sigue siendo la guía para medir el paro cíclico, y la teoría anticipó la ruptura de la curva antes de la estanflación. Pero con expectativas adaptativas los agentes cometen errores sistemáticos y predecibles, difíciles de conciliar con la racionalidad.

1.3.2. Teoría aceleracionista (Friedman, 1968; Phelps, 1967)

Idea: el Gobierno solo puede mantener el desempleo por debajo de la tasa natural **acelerando la inflación**, es decir, explotando permanentemente los errores de los agentes, lo que exige un crecimiento monetario cada vez mayor. Si deja de acelerarla, la economía vuelve a $\bar{u}$ con más inflación. Si persiste, el intento degenera en una espiral inflacionista.

Razonamiento: con un objetivo de desempleo inferior a la tasa natural, la autoridad ha de generar cada periodo una inflación superior a la esperada. Pero, con expectativas adaptativas, la esperada se ajusta a la pasada, de modo que sostener la sorpresa exige una inflación creciente.

Corolario: solo las políticas de oferta reducen la tasa natural de forma duradera, y la política monetaria debe renunciar a explotar la curva de Phillips.

Conclusiones: explica la inflación creciente de los años setenta. Pero descansa en las expectativas adaptativas, y con expectativas ancladas, como en EE. UU. desde los noventa, la curva vuelve a ligar el nivel de la inflación, y no su variación, con el desempleo (Blanchard, 2017).

1.4. Desequilibrio: Malinvaud (1977)

Idea: la escuela del desequilibrio (Clower, 1965; Leijonhufvud, 1968; Barro y Grossman, 1971) reinterpreta a Keynes como fallos de coordinación con precios rígidos y racionamiento. Malinvaud (1977), con la **hipótesis de decisión dual**, muestra que un mismo desempleo puede tener causas distintas y exigir remedios opuestos, según los excesos en los mercados de trabajo y de bienes:

1. *Desempleo clásico*: exceso de oferta de trabajo y de demanda de bienes, con un salario real superior al de vaciado, por ejemplo por la acción sindical. Se corrige bajando el salario real, que eleva la oferta de bienes y la demanda de trabajo. Un estímulo de demanda es aquí ineficaz, porque la producción está limitada por la oferta, y solo hace falta si la bajada salarial lleva la economía al desempleo keynesiano.
2. *Desempleo keynesiano*: exceso de oferta en ambos mercados. El problema no es el precio del trabajo, sino la demanda efectiva, y se corrige con una subida limitada del salario real o, sobre todo, con impulso fiscal.
3. *Inflación reprimida*: exceso de demanda en ambos, porque los controles impiden que precios y salarios suban. El desequilibrio aflora como escasez y racionamiento, como en las economías planificadas o de guerra. Se corrige liberalizando precios y salarios con una contracción de la demanda que evite una inflación abierta.
4. *Subconsumo*: exceso de demanda de trabajo y de oferta de bienes, lógicamente posible pero infrecuente. Se corrige elevando el salario real, que estimula el consumo y modera la demanda de trabajo.

Conclusiones: permite elegir la política según el régimen, frente a las recetas universales. Pero las rigideces se postulan sin microfundamentar y el régimen vigente es difícil de identificar, lo que explica el declive del enfoque tras las expectativas racionales.

En suma, los fundamentos dejan dos problemas abiertos, los errores sistemáticos de las expectativas adaptativas y las rigideces sin microfundamentos, que abordan las expectativas racionales y el modelo de la NAIRU.

2. Modelos de la tasa de desempleo de equilibrio

La tasa natural de Friedman es un concepto más que un modelo: no dice cómo se forman las expectativas, de qué depende la tasa ni por qué el paro europeo no volvió a su nivel tras cada recesión. Responden la nueva macroeconomía clásica, la NAIRU y la histéresis.

2.1. Nueva macroeconomía clásica

2.1.1. Crítica de Lucas (1976)

Idea: los modelos econométricos de la síntesis no sirven para evaluar políticas alternativas, porque sus parámetros dependen del régimen de política. Al cambiar este, cambian las reglas de decisión de los agentes, y la curva de Phillips estimada bajo un régimen no predice bajo otro.

Razonamiento: la crítica exige microfundamentar los modelos y modelizar explícitamente las expectativas, que es el programa de la nueva macroeconomía clásica y, después, de la nueva economía keynesiana.

Conclusiones: explica por qué la curva de los años sesenta dejó de funcionar cuando se trató de explotarla, y su exigencia de parámetros estructurales, invariantes a la política, es hoy un estándar. Pero su relevancia empírica es discutida, porque pocas veces cambian los parámetros con la política (Ericsson e Irons, 1995). Además, la invariancia de los parámetros de preferencias y tecnología es un supuesto difícil de contrastar.

2.1.2. Sustitución intertemporal (Lucas y Rapping, 1969)

Idea: los trabajadores eligen cuánto trabajar comparando el salario real actual con el esperado, de modo que las fluctuaciones del empleo son sustitución intertemporal voluntaria del ocio ante salarios percibidos como transitoriamente altos o bajos. Combinada con información imperfecta sobre los precios, genera fluctuaciones sin abandonar el vaciado continuo, y solo la tasa natural es desempleo en sentido propio.

Conclusiones: es la primera explicación de las fluctuaciones con mercados que se vacían. Pero exige una elasticidad intertemporal elevada que la evidencia no respalda (Altonji, 1986), y es difícil ver el paro de las recesiones como ocio elegido.

2.1.3. Expectativas racionales (Lucas, 1972; Sargent y Wallace, 1975)

Idea: con la **hipótesis de las expectativas racionales**, los agentes usan con eficiencia toda la información, incluido cómo funciona la economía y la política anunciada, y solo se equivocan al azar y sin sesgo. Desaparece así lo que en Friedman daba lugar al *trade-off* (unos agentes que erraban siempre en la misma dirección).

Razonamiento: en la parábola de las islas, un productor ve subir su precio sin saber si es un cambio relativo o general. Si atribuye parte de la subida a un cambio relativo, produce y emplea más. Al agregar y aplicar la ley de Okun, el desempleo solo se desvía de su tasa natural si el nivel de precios difiere del esperado:

$$u_t = \bar{u} - \frac{\zeta}{\phi}(p_t - p_t^e) + \varepsilon_t, \qquad \zeta, \phi > 0$$

- Con expectativas racionales, $p_t - p_t^e$ es un error impredecible con la información disponible, de modo que el desempleo solo baja de su tasa natural con sorpresas inflacionarias y vuelve a ella en cuanto se disipan.
- Una política monetaria sistemática se anticipa y no tiene efecto real ni a corto plazo, y la curva es así **vertical también a corto plazo**, salvo por las sorpresas.

Conclusiones: fue una revolución metodológica que obliga a modelizar las expectativas, y su implicación, que la estabilización sistemática es inútil, fijó el debate posterior. Pero confundir precios relativos y generales durante mucho tiempo es poco creíble cuando los precios se publican cada mes, y la sustitución intertemporal exige una elasticidad poco realista. La NMC evolucionó hacia el ciclo real, y la nueva economía keynesiana conservó las expectativas racionales pero abandonó el vaciado continuo.

2.2. La NAIRU (Layard y Nickell, 1985)

Idea: Layard y Nickell (1985) y Carlin y Soskice (1990), sistematizados por Layard, Nickell y Jackman (1991), aceptan las expectativas racionales, pero no un mercado de trabajo walrasiano: los salarios se negocian

y los precios los fijan empresas con poder de mercado. El desempleo de equilibrio es el que hace **compatibles las aspiraciones** de trabajadores y empresas sobre el reparto de la renta.

2.2.1. Supuestos

Economía cerrada con mercados de trabajo y de bienes imperfectos. El trabajo es el único factor variable, con rendimientos constantes, $Y = A_L \cdot L$, de modo que el coste marginal es el salario dividido por la productividad. Sindicatos y empresas tienen poder de mercado y no sufren ilusión monetaria: negocian en términos reales esperados.

- *Sindicatos*: fijan el salario real al que aspiran, que crece con la presión del mercado.
- *Empresas*: fijan el precio como un margen sobre el coste marginal, y ese margen determina el salario real que están dispuestas a pagar.

2.2.2. Modelo

Planteamiento: el desempleo es la variable que reconcilia lo que pretenden los trabajadores con lo que conceden las empresas. Si el salario al que aspiran los primeros supera al que pagan las segundas, la diferencia se traduce en una inflación superior a la esperada.

Salario real negociado (SRN): los trabajadores negocian un salario real esperado que decrece con el desempleo, porque reduce su poder, y crece con factores institucionales z, como prestaciones, salario mínimo o cotizaciones:

$$\frac{w}{P^e} = f(u^-, z^+)$$

Salario real pagado (SRP): cada empresa iguala el ingreso marginal al coste marginal, que con rendimientos constantes es el coste laboral unitario, w/A_L, y fija así un margen $\mu = 1/(\sigma - 1)$, con σ la elasticidad-precio de la demanda. El salario real que concede depende solo de la productividad y de la competencia, no del empleo:

$$\text{Img} = \text{Cmg} \iff P = \frac{\sigma}{\sigma - 1}\frac{w}{A_L} = (1+\mu)\frac{w}{A_L}$$

$$\frac{w}{P} = \frac{A_L}{1+\mu} = (1 - I_L)\,A_L, \qquad I_L = \frac{P - \text{Cmg}}{P} = \frac{1}{\sigma} \in (0,1)$$

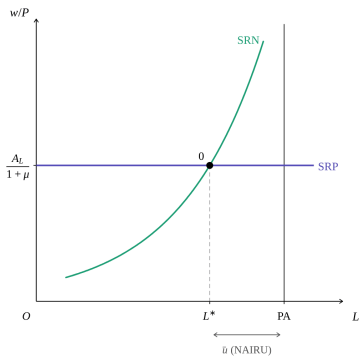
- Si el margen fuera anticíclico, porque la colusión cuesta más en los auges, el SRP tendría pendiente positiva.

Equilibrio: la **NAIRU** (*Non-Accelerating Inflation Rate of Unemployment*) es la tasa a la que SRN y SRP coinciden y se cumplen las expectativas, $P^e = P$. A ella la inflación no se acelera:

$$f(\bar{u}, z) = \frac{A_L}{1+\mu}$$

La NAIRU es única y estable sin histéresis, compatible con desempleo involuntario y engloba el estructural y el friccional. A diferencia de la tasa natural, concepto de vaciado del mercado, es un concepto de compatibilidad entre agentes con poder de mercado, y depende explícitamente de instituciones que el modelo competitivo no tiene. En la práctica, ambos términos se usan como sinónimos. Friedman, no obstante, incorporaba ya en la tasa natural las imperfecciones del mercado y los costes de información y movilidad.

Desarrollo gráfico: en el plano empleo-salario real, el SRN es creciente y muy inclinado cerca de la población activa, y el SRP es horizontal en $A_L/(1+\mu)$. Su corte da L^*, y la distancia hasta la población activa es la NAIRU.



NAIRU: el SRN y el SRP se cortan en L^*; la distancia a la población activa es $\bar{u}$.

Estática comparativa:

- *Demanda*: una expansión reduce el paro por debajo de la NAIRU, eleva el SRN por encima del SRP y desata una espiral salarios-precios que devuelve la economía a la NAIRU con más inflación. Las políticas de demanda solo tienen, pues, efectos transitorios, y a costa de inflación.
- *Oferta*: son las únicas que alteran la NAIRU, porque desplazan una de las curvas. Más z sube el SRN y más margen μ baja el SRP, y ambos la elevan. Más productividad A_L la reduce, aunque solo de forma transitoria si las aspiraciones salariales se indexan a ella, y menos impuestos y cotizaciones reconcilian las aspiraciones a un menor desempleo. Las políticas óptimas eliminan las rigideces de ambos mercados.

Curva de Phillips: sin imponer el cumplimiento de las expectativas y con $f(u,z) = z\,e^{-\kappa_u u}$, el SRN y el SRP dan una curva ampliada fundada en la **puja distributiva**:

$$\pi_t = \pi_t^e + \alpha - \kappa_u u_t = \pi_t^e - \kappa_u(u_t - \bar{u})$$

$$\bar{u} = \frac{\alpha}{\kappa_u}, \qquad \alpha = \ln(1+\mu) + \ln z - \ln A_L$$

- Coincide con la de Friedman y Phelps, pero la NAIRU ya no es un dato, sino una función de los parámetros estructurales que crece con el margen y con z y decrece con la productividad.

Evidencia: en la OCDE, la NAIRU es mayor con prestaciones indefinidas sin búsqueda activa, sindicación sin coordinación y una cuña fiscal elevada (Nickell, 1997). La protección del empleo afecta más a la duración del paro que a su nivel. Blanchard y Wolfers (2000) muestran que ni las perturbaciones ni las instituciones bastan por sí solas. Lo explica su interacción, porque las mismas perturbaciones adversas elevaron más el paro en los países con instituciones más rígidas.

2.2.3. Conclusiones

Aspectos positivos: funda el desempleo de equilibrio en las instituciones de los mercados de trabajo y de bienes, lo que permite analizar las políticas estructurales, y da a la curva de Phillips un fundamento en la puja distributiva y no en la información imperfecta.

Limitaciones: la NAIRU no es observable, sus estimaciones son inestables y, al ser estática, no explica por qué el paro europeo siguió alto tras las perturbaciones de los setenta.

2.3. Histéresis (Blanchard y Summers, 1986)

Idea: la NAIRU estimada parecía seguir al desempleo observado en lugar de anclarlo. Hay **histéresis** cuando la tasa de equilibrio depende del desempleo pasado. Es parcial si el paro vuelve muy despacio y pura si no vuelve.

2.3.1. Mecanismos

Blanchard y Summers identifican tres mecanismos:

1. *Menor acumulación de capital*: la caída del capital reduce aún más la demanda de trabajo.
2. *Pérdida de capital humano*: el paro de larga duración deprecia las cualificaciones y genera desánimo, lo que dificulta la reinserción (Phelps, 1972), con mucho respaldo empírico.
3. *Insiders-outsiders*: los empleados fijan el salario atendiendo solo a su empleo, y en las recesiones no baja lo suficiente para reabsorber a los despedidos.

2.3.2. En el modelo de la NAIRU

Razonamiento: la NAIRU depende del desempleo pasado, con $\bar{u}^*$ la NAIRU estructural y h el grado de histéresis, y se sustituye en la curva de Phillips:

$$\bar{u}_t = (1-h)\,\bar{u}^* + h\,u_{t-1}, \qquad h \in [0,1]$$

$$\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u}_t)$$

- Con $h = 0$ la NAIRU es constante. Con $0 < h < 1$ hay persistencia, pues tras una perturbación vuelve gradualmente a $\bar{u}^*$, y la de largo plazo es única, aunque a corto o medio plazo puede haber varias.
- Con $h = 1$, histéresis pura, el desempleo es un paseo aleatorio sin NAIRU única, y la inflación depende de su variación, $\pi_t = \pi_t^e - \kappa_u\,\Delta u_t$.

Desarrollo gráfico: tras una contracción, los *insiders* restantes sostienen sus salarios. El SRN se desplaza a la izquierda y, con el SRP igual, el empleo cae y la NAIRU sube. Cuando la coyuntura mejora, el paro ya no vuelve a su nivel inicial.

Corolario: las políticas activas de empleo, como la formación, la intermediación o la recualificación de los parados de larga duración, favorecen el retorno al nivel inicial, porque frenan la pérdida de capital humano, que el apartado 3.1 reinterpreta como pérdida de eficiencia del emparejamiento.

2.3.3. Conclusiones

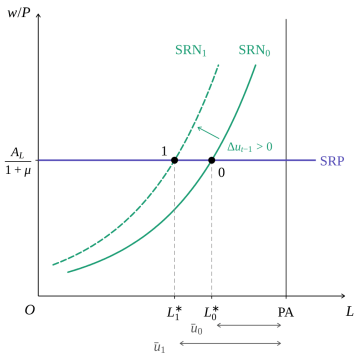
Aspectos positivos: explica la persistencia europea, con España como ilustración destacada, y devuelve a la demanda efectos a largo plazo.

Limitaciones: la histéresis pura es extrema, porque los *outsiders* influyen algo en la negociación, los *insiders* también arriesgan el paro y las empresas pueden amenazar con sustituirlos. Hay persistencia elevada, no histéresis pura.

En suma, con expectativas racionales solo las sorpresas apartan el paro de su tasa. En la NAIRU la tasa depende de instituciones y solo la oferta la altera, y la histéresis le suma el desempleo pasado.

3. Desarrollos actuales y convergencia

La teoría actual incorpora las fricciones de búsqueda y la curva de Phillips neokeynesiana y afronta el problema práctico de estimar la NAIRU.



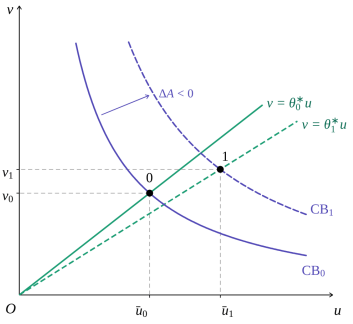
Histéresis: un *shock* transitorio desplaza el SRN y eleva la NAIRU.

3.1. Búsqueda y emparejamiento

Aportación: el modelo de búsqueda y emparejamiento microfundamenta el desempleo de equilibrio como resultado de un encuentro costoso entre trabajadores y vacantes.

- *Nuevos parámetros*: la tasa depende de la eficiencia del emparejamiento, la destrucción de empleo, el poder negociador y el coste de la vacante, lo que da contenido a los factores z.
- *Perturbaciones*: las coyunturales mueven el equilibrio sobre la curva de Beveridge y las estructurales la desplazan, lo que separa el paro cíclico del estructural.
- *Persistencia*: la explica como pérdida de eficiencia del emparejamiento tras las recesiones, por pérdida de capital humano y desajuste, de modo que la histéresis es también un fenómeno de la búsqueda y no solo de la negociación. Las políticas activas son el instrumento para acercar la curva al origen. Ha permitido, además, integrar las fricciones de búsqueda en los modelos del ciclo.

Desarrollo gráfico: en el plano desempleo-vacantes, la curva de Beveridge es decreciente y convexa, y la recta de creación de empleo parte del origen con pendiente θ^*. Una caída de la eficiencia, $\Delta A < 0$, desplaza la curva hacia fuera y gira la recta, y la tasa de equilibrio sube de $\bar{u}_0$ a $\bar{u}_1$. La curva es decreciente porque con más vacantes los desempleados encuentran trabajo antes, y la pendiente de la recta es la estrechez que fija la condición de creación de empleo.



Caída de la eficiencia del emparejamiento: la CB se desplaza hacia fuera y sube $\bar{u}$.

3.2. Curva de Phillips de la NEK2G

Origen: la nueva síntesis neoclásica deriva la curva de la fijación escalonada de precios a la Calvo (1983) con expectativas racionales.

3.2.1. Supuestos

Modelo dinámico, en tiempo discreto y con horizonte infinito; hogares idénticos que eligen consumo, trabajo y bonos, y un continuo de empresas en competencia monopolística que fijan el precio de su variedad, todos con expectativas racionales. Cada periodo solo una fracción $1-\xi$ de las empresas, elegida al azar, puede reajustar su precio, y el nivel de precios es una media ponderada del precio pasado y del óptimo de reajuste:

$$P_t = \left[\xi P_{t-1}^{1-\sigma} + (1-\xi)(P_t^*)^{1-\sigma}\right]^{1/(1-\sigma)}, \qquad \xi \in [0,1)$$

3.2.2. Modelo

Planteamiento: los hogares maximizan su utilidad intertemporal sujeta a su restricción presupuestaria, y cada empresa elige el precio de su variedad para maximizar el valor esperado de sus beneficios:

$$\max_{P_{i,t}} \mathbb{E}_0\left[\sum_{t=0}^{\infty} \beta^t \Pi(P_{i,t}, Y_{i,t}, C_t)\right] \quad \text{s.a.:} \begin{cases} Y_{i,t} = f(\ell_{i,t}) \\ Y_{i,t} = \left(P_{i,t}/P_t\right)^{-\sigma} C_t \\ P_{i,t} = P_{i,t-1} \text{ con probabilidad } \xi, \qquad \xi \in [0,1) \end{cases}, \qquad \beta \in (0,1)$$

CPO: las de los hogares son la ecuación de Euler, que loglinealizada da la *IS dinámica*, y la condición intratemporal, que iguala la relación marginal de sustitución consumo-ocio al salario real. El precio óptimo de la empresa aplica el margen $1+\mu = \sigma/(\sigma-1)$ a una media de los costes marginales actual y esperados, ponderados por $(\beta\xi)^k$, porque no podrá cambiarlo durante un tiempo. De ahí que la inflación dependa de las expectativas.

Equilibrio: agregando y loglinealizando, y con la ley de Okun, se obtiene la curva de Phillips neokeynesiana en términos de desempleo:

$$\pi_t = \beta \mathbb{E}_t[\pi_{t+1}] - \kappa_u\,(u_t - \bar{u}_t) + \varepsilon_t^{\pi}, \qquad \kappa_u > 0,\ \beta \in (0,1)$$

- Es **prospectiva**, porque la inflación de hoy depende de la esperada, y **microfundamentada**, con parámetros profundos, robusta a la crítica de Lucas si la frecuencia de ajuste no cambia con la política.
- La brecha mueve la economía sobre la curva, mientras que la inflación esperada y los *shocks* de oferta, ε_t^{π}, la desplazan.
- Antes de aplicar la ley de Okun, $\tilde{y}_t \approx -\phi(u_t - \bar{u}_t)$, la curva se escribe con la brecha de producción, porque el coste marginal real se mueve con ella:

$$\pi_t = \beta \mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t + \varepsilon_t^{\pi}, \qquad \kappa > 0,\ \beta \in (0,1)$$

Extensiones: Galí (2011) recupera la curva salarial original, microfundamentada y con expectativas. Su importancia es teórica, porque la fijación escalonada de salarios implica una curva salarial en términos de desempleo y cierra el círculo iniciado por Phillips (1958), y empírica, porque se ajusta razonablemente bien a los datos de EE. UU. de las últimas décadas, incluso con una tasa natural constante:

$$\pi_t^w = \beta\,\mathbb{E}_t[\pi_{t+1}^w] + \kappa_w\,(\bar{u}_t - u_t)$$

3.2.3. Conclusiones

Aspectos positivos: es microfundamentada y prospectiva y, sin *shocks* de costes, presenta la **divina coincidencia** (Blanchard y Galí, 2007): estabilizar la inflación equivale a estabilizar la brecha.

Limitaciones: en su versión pura no genera la persistencia observada de la inflación, lo que obliga a versiones híbridas (Galí y Gertler, 1999), y su estimación es sensible a la medida de la brecha.

3.3. Estimación de la NAIRU

Métodos: la NAIRU separa el paro cíclico del estructural, y se estima con métodos estructurales, a partir de un modelo de salarios y precios; estadísticos, con filtros, y semiestructurales, con una curva de Phillips de NAIRU variable (Gordon, 1997; Staiger et al., 1997).

Conclusiones: la incertidumbre es grande. Los intervalos de confianza abarcan varios puntos, las revisiones son grandes y las estimaciones en tiempo real difieren mucho de las posteriores. Por eso los bancos centrales la tratan como una referencia entre otras y no como un objetivo. En la zona del euro y en España la NAIRU estimada sube tras cada recesión y baja despacio, lo que es coherente con la histéresis, aunque en parte refleja que los métodos estadísticos y semiestructurales siguen, por construcción, al paro observado.

3.4. Convergencia del pensamiento

La teoría actual conserva de la nueva macroeconomía clásica las expectativas racionales, los microfundamentos y una tasa de equilibrio determinada por factores reales, y de la nueva economía keynesiana, las rigideces y los mercados imperfectos. La NAIRU y la curva de Phillips neokeynesiana son el punto de encuentro: la política monetaria no puede explotar esa relación de forma sistemática, pero encuentra en ella margen para estabilizar.

1. Fundamentos de la elección intertemporal

La elección intertemporal introduce el tiempo en la decisión del consumidor, con tres novedades: los bienes son consumos fechados, la dotación es una corriente de rentas y el precio relativo entre el consumo de hoy y el de mañana es $1+r$.

1.1. Conjunto de elección y preferencias

1.1.1. El conjunto de elección

Definición: los objetos de elección son **sendas de consumo**, $\mathbf{c} = (c_0, c_1, \dots, c_T) \in \mathbb{R}_+^{T+1}$, con c_t el consumo del periodo t. El tiempo convierte cada bien físico en tantas mercancías como periodos.

Propiedades: el conjunto tiene tres rasgos que la teoría estática no tiene.

- *Dimensión*: la fija el horizonte, finito en los modelos de ciclo vital e infinito en los de dinastías, que exigen condiciones adicionales.
- *Dotación*: es una corriente de rentas, $\mathbf{y} = (y_0, \dots, y_T)$, que solo se transforma en consumo de otros periodos mediante el ahorro y el endeudamiento.
- *Orden*: la senda $(1, 2)$ no es la $(2, 1)$, porque el tiempo se vive en una dirección.

1.1.2. La ordenación de las preferencias

Axiomas: la relación de preferencias cumple los axiomas estáticos, reinterpretados en el tiempo: completitud, transitividad, continuidad y monotonía, sin saciedad en ningún periodo. La **convexidad** es en el tiempo preferencia por la **suavización** y origen del ahorro por ese motivo.

Propiedades intertemporales: se añaden tres, que Koopmans (1960) axiomatizó y que llevan a la función de Samuelson.

- *Separabilidad*: la valoración del consumo de un periodo no depende del de los demás, lo que excluye hábitos y bienes duraderos.
- *Estacionariedad*: la ordenación entre dos sendas no cambia si ambas se desplazan un periodo, de modo que la utilidad de cada periodo es la misma y el descuento entre periodos consecutivos, constante.
- *Impaciencia*: el futuro se descuenta a una tasa positiva.

1.2. Utilidad y curvas de indiferencia

1.2.1. Función de utilidad (Samuelson, 1937)

Definición: los axiomas permiten representar las preferencias con la **función de utilidad descontada de Samuelson**, donde β es el factor de descuento, $\rho > 0$ la tasa de descuento subjetiva, que refleja la preferencia por el presente (mortalidad, impaciencia, sesgo de visibilidad), y $u(\cdot)$ la utilidad instantánea o función de "felicidad":

$$U_d = \sum_{t=0}^{T} \beta^t u(c_t), \quad \beta = \frac{1}{1+\rho} \in (0,1)$$

Propiedades: es continua y dos veces diferenciable, creciente, $u' > 0$, y estrictamente cóncava, $u'' < 0$, de modo que las curvas de indiferencia son convexas.

- La representación aditiva exige que la utilidad instantánea sea **cardinal**, única salvo transformaciones afines positivas, como la de Bernoulli.
- Es **recursiva**: la utilidad de la senda es la de hoy más la descontada de la senda que empieza mañana, lo que permite resolver hacia atrás y garantiza la consistencia dinámica.

1.2.2. Curvas de indiferencia intertemporales

Pendiente: su valor absoluto es la **relación marginal de sustitución intertemporal**, el precio subjetivo del consumo presente en unidades de consumo futuro. Es mayor cuanto más escaso es el consumo presente y cuanto mayor es la impaciencia, y a lo largo de una senda constante vale $1+\rho$:

$$\mathrm{RMSI}_{t+1,t} = \frac{1}{\beta} \cdot \frac{u'(c_t)}{u'(c_{t+1})}$$

Convexidad: la mide la **elasticidad de sustitución intertemporal** (ESI), la respuesta del consumo relativo entre dos periodos a su precio relativo:

$$\sigma = \frac{\partial \ln(c_{t+1}/c_t)}{\partial \ln(1+r)}$$

- Con σ alta, las curvas son poco curvadas y el ahorro responde mucho al tipo de interés. Con σ baja, el consumidor prefiere suavizar a aprovechar la recompensa.
- Con $\sigma \to 0$ las curvas son de ángulo recto. Con $\sigma \to \infty$ son rectas y el consumo se concentra en el periodo más barato.
- Con la utilidad isoelástica, $u(c) = c^{1-\theta}/(1-\theta)$, la ESI es $1/\theta$, el inverso de la aversión relativa al riesgo. La misma curvatura gobierna el riesgo y la sustitución en el tiempo.

1.3. Restricción intertemporal de recursos

1.3.1. Restricción intertemporal

Ecuación de estado: en cada periodo, el consumo más la acumulación de activos financieros, a_t, positivos si presta y negativos si se endeuda, no puede superar la renta más los intereses, con r el tipo de interés real constante:

$$c_t + (a_{t+1} - a_t) = y_t + r\, a_t$$

Condiciones de frontera: la inicial fija la riqueza de partida, a_0, y la terminal, con horizonte finito, exige $a_{T+1} = 0$: no se dejan activos sin consumir ni deudas sin pagar.

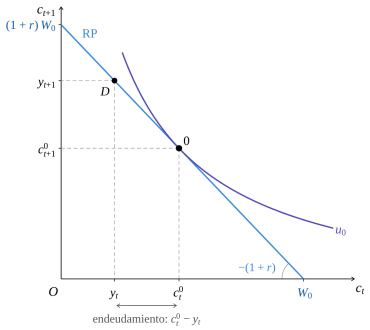
Restricción en valor presente: encadenando todos los periodos, el valor presente del consumo no puede superar la **riqueza**, W_0, suma de los activos iniciales y del valor presente de las rentas:

$$\sum_{t=0}^{T} \frac{c_t}{(1+r)^t} \le (1+r)\, a_0 + \sum_{t=0}^{T} \frac{y_t}{(1+r)^t} \equiv W_0$$

Gráficamente, con dos periodos y $a_0 = 0$, la restricción es la recta de riqueza:

$$c_{t+1} = y_{t+1} + (1+r)\,(y_t - c_t)$$

que pasa por la dotación $D = (y_t, y_{t+1})$, tiene pendiente $-(1+r)$, el coste de oportunidad del consumo presente, y corta los ejes en W_0 y $(1+r)\,W_0$. A la izquierda de la dotación el consumidor ahorra y a la derecha se endeuda.

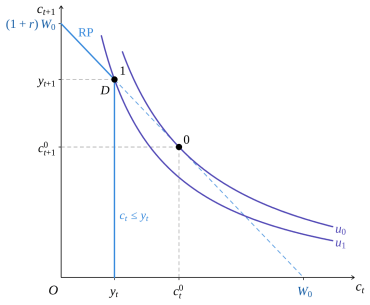


Restricción presupuestaria intertemporal: la recta pasa por la dotación y tiene pendiente $-(1+r)$.

1.3.2. Restricciones de liquidez

Definición: la restricción anterior supone que se puede prestar y endeudarse sin límite al mismo tipo r. Si el crédito falta o tiene un tope, las **restricciones de liquidez** dejan fuera del conjunto asequible el tramo de la recta a la derecha de la dotación.

Gráficamente, a la izquierda de D la frontera sigue siendo la recta de riqueza, porque se puede ahorrar al tipo r. En D se hace vertical, $c_t \le y_t$. La restricción solo es efectiva si el óptimo sin restricción, 0, queda a la derecha de la dotación. El consumidor se queda en ella, punto 1, sobre una curva más baja, con $\mathrm{RMSI} > 1+r$.



Restricción de liquidez: la recta se quiebra en la dotación y el consumidor se queda en ella (punto 1).

1.4. Consumo intertemporal (Fisher, 1930)

Idea: Fisher trasladó al tiempo la teoría del consumidor. El consumo de cada periodo depende de la riqueza total, no de la renta corriente, y de la relación entre el tipo de interés y la impaciencia.

1.4.1. Modelo

Planteamiento: el consumidor elige la senda que maximiza su utilidad sujeta a que el valor presente del consumo iguale la riqueza:

$$\max_{\{c_t\}} \sum_{t=0}^{T} \beta^t u(c_t) \quad \text{s.a.:} \quad \sum_{t=0}^{T} \frac{c_t}{(1+r)^t} = W_0, \qquad \beta \in (0,1)$$

CPO: la RMSI iguala la rentabilidad bruta del ahorro, $1+r$, lo que, escrito de otro modo, es la **ecuación de Euler**:

$$u'(c_t) = \beta\,(1+r)\,u'(c_{t+1}) = \frac{1+r}{1+\rho}\,u'(c_{t+1})$$

– En el óptimo, renunciar hoy a una unidad de consumo cuesta lo mismo que rinde ahorrarla y gastarla mañana, descontada la impaciencia. La concavidad de u hace la condición suficiente.
– La misma ecuación sale de la **ecuación de Bellman** (Bellman, 1957), $V_t(a_t) = \max_{c_t}\{u(c_t) + \beta V_{t+1}(a_{t+1})\}$, con su CPO y la condición de la envolvente. La programación dinámica resuelve el problema hacia atrás, periodo a periodo, y se extiende sin cambios a la incertidumbre.

Solución: la ecuación de Euler fija el **perfil** de la senda y la restricción, su **nivel**.

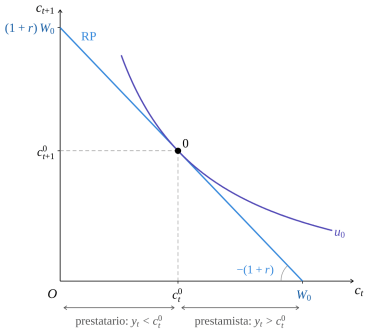
– Si $r = \rho$, el consumo es constante; si $r > \rho$, la senda es creciente y, con renta constante, el individuo ahorra; si $r < \rho$, es decreciente y demanda capital.
– El individuo desea sendas estables, la **suavización del consumo**, lo que exige $u'' < 0$ y mercados financieros.

Con $r = \rho$, el consumo es la **anualidad** que la riqueza puede financiar durante el horizonte, y no una fracción de la renta corriente:

$$c_t = \frac{r}{1+r-(1+r)^{-T}}\, W_0 \xrightarrow{T\to\infty} \frac{r}{1+r}\, W_0$$

– Esta expresión es la forma general de las teorías del consumo. La renta permanente (Friedman, 1957) llama así a esa anualidad, y el ciclo vital (Modigliani y Brumberg, 1954) usa la vida restante.

Desarrollo gráfico: el óptimo es la tangencia entre la curva de indiferencia más alta y la recta de riqueza, donde RMSI $= 1 + r$.



Elección intertemporal de Fisher: tangencia de la curva de indiferencia con la recta de pendiente $-(1+r)$.

Dotación: un aumento de renta desplaza la recta hacia fuera. Con $r = \rho$, un aumento permanente eleva el consumo de cada periodo en la misma cuantía, y uno transitorio se reparte entre todos los periodos y eleva poco el consumo corriente.

Tipo de interés: el efecto sobre el consumo presente se descompone con la ecuación de Slutsky con dotaciones:

$$\frac{\partial c_t}{\partial r} = \underbrace{\left.\frac{\partial c_t}{\partial r}\right|_{u_0}}_{\text{ES}<0} + \underbrace{(y_t - c_t)\frac{\partial c_t}{\partial y}}_{\text{ERT}}$$

– El **efecto sustitución** es inequívoco, porque una subida de r encarece el consumo presente y eleva el ahorro.
– El **efecto renta total** suma un efecto renta ordinario negativo y un efecto renta dotación positivo. Para un prestamista es positivo y el efecto total, ambiguo, mientras que para un prestatario ambos son negativos.

Desarrollo gráfico: una subida de r hace pivotar la recta sobre la dotación, más inclinada, lo que reduce las posibilidades de quien se endeuda y amplía las de quien presta. Para un prestamista, el efecto sustitución lleva de 0 a 0 sobre la curva inicial y la recta compensada, y el efecto renta total, de 0 a 1. Si domina el primero, el consumo presente cae.

La evidencia no es concluyente: la ESI estimada es baja y el ahorro agregado responde poco al tipo de interés.

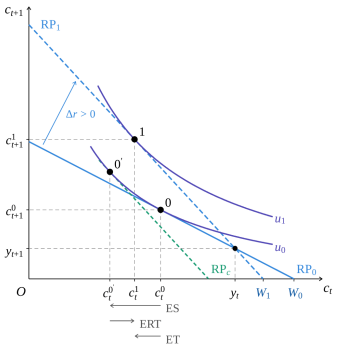
1.4.2. Paseo aleatorio (Hall, 1978)

Idea: con incertidumbre, expectativas racionales, utilidad cuadrática y $r = \rho = 0$, el consumo de hoy ya incorpora todo lo que se sabe sobre la renta futura, y solo la información nueva lo altera. La ecuación de Euler esperada da:

$$\mathbb{E}_t[c_{t+1}] = c_t \quad \Rightarrow \quad c_{t+1} = c_t + \varepsilon_{t+1}, \qquad \mathbb{E}_t[\varepsilon_{t+1}] = 0$$

– La variación del consumo es imprevisible con información pasada, y la política solo le afecta si genera sorpresas o altera la riqueza esperada.

Conclusiones: convierte la renta permanente en una restricción contrastable, la ecuación de Euler, punto de partida de la literatura empírica moderna. Pero la evidencia muestra exceso de sensibilidad a la renta corriente y exceso de suavidad.



Subida del tipo de interés para un prestamista: efecto sustitución ($0 \to 0'$) y efecto renta total ($0' \to 1$).

1.4.3. Restricciones de liquidez (Deaton, 1991)

Idea: con un límite de endeudamiento, $a_{t+1} \geq -\bar{a}$, si el límite opera la igualdad de Euler se rompe, $u'(c_t) > \beta(1+r)\,u'(c_{t+1})$. El consumidor querría adelantar consumo y no puede, de modo que consume todos sus recursos y su consumo sigue a la renta corriente. Si no opera pero puede hacerlo en el futuro, el consumidor prudente acumula un colchón de activos, el ahorro **buffer-stock**, y su propensión a consumir la renta transitoria es mayor que en Fisher.

Conclusiones: explican que el consumo siga a la renta, que la política fiscal transitoria tenga efectos reales y que la distribución de la riqueza importe para el multiplicador. Pero el problema solo se resuelve numéricamente.

1.4.4. Bienes duraderos (Mankiw, 1982)

Idea: un bien duradero da utilidad mientras dura, de modo que en la utilidad entra el flujo de servicios del *stock*, no el gasto. La CPO iguala la utilidad marginal de esos servicios al **coste de uso**, $p_d\,(r + \delta - \pi_d)$, de la misma forma que el del capital de la empresa. Se suaviza el *stock*, y el gasto, que es su variación, es mucho más volátil que el de no duraderos y responde con fuerza al tipo de interés y al crédito. Con utilidad cuadrática y expectativas racionales, el *stock* sigue un paseo aleatorio, y la primera diferencia del gasto, con $x_t = d_t - (1-\delta)\,d_{t-1}$, un proceso de medias móviles de orden uno:

$$\Delta x_t = \varepsilon_t - (1-\delta)\,\varepsilon_{t-1}$$

Conclusiones: explica la parte más volátil del consumo y unifica las teorías del consumo y de la inversión. Pero el gasto en duraderos se comporta como si la depreciación fuera casi total.

En suma, el modelo de Fisher es la base de las teorías modernas del consumo, aunque supone mercados de capitales perfectos y preferencias separables.

2. Decisiones intertemporales de la empresa

La decisión de inversión tiene la misma estructura que la de consumo, pero la empresa maximiza su valor, descontado al tipo de mercado, y su restricción es una tecnología que convierte la inversión de hoy en capital de mañana. El consumidor elige cuándo consumir una riqueza dada; la empresa, **cuánta riqueza crear**.

2.1. Conjunto de elección y objetivo

Tiempo: a diferencia del caso estático, el capital no se alquila cada periodo, sino que es un *stock* que se acumula, se deprecia y cuesta ajustar, de modo que la decisión de hoy condiciona las de mañana y la empresa ha de anticipar toda la senda futura de precios y productividades.

Variables: las de *control* son los flujos que decide cada periodo, la inversión I_t y el empleo L_t. La de *estado* es el capital K_t, que la empresa solo elige a través de la inversión acumulada.

Objetivo: la empresa maximiza su valor, el valor presente de sus beneficios futuros. Por el teorema de separación de Fisher, con mercados de capitales perfectos todos los propietarios lo quieren, porque maximiza su riqueza, y la tasa de descuento es el tipo de interés de mercado.

2.2. Función de valor y precio sombra

Función de valor: es la de Samuelson con el beneficio, neto del gasto en inversión, en lugar de la utilidad y $1/(1+r)$ en lugar de β:

$$V_0 = \sum_{t=0}^{T} \frac{\Pi_t}{(1+r)^t}$$

– Evaluada en la senda óptima, depende solo del capital heredado y de los precios, $V_t = V_t(K_t)$, y es cóncava en el capital con beneficios cóncavos y costes de ajuste convexos.
– Con mercados perfectos coincide con el valor de mercado de la empresa, que es observable.

Precio sombra del capital: el equivalente del valor marginal de la riqueza del consumidor es $q_t = \partial V_t/\partial K_t$, cuánto aumenta el valor de la empresa una unidad más de capital. Resume la información sobre el futuro que necesita la decisión de invertir. En Jorgenson es el precio del bien de capital y en Tobin la q marginal.

2.3. Restricción tecnológica

Ley de acumulación: es la ecuación de estado de la empresa, con δ la tasa de depreciación, y liga la decisión de hoy con el estado de mañana:

$$K_{t+1} = (1-\delta)\,K_t + I_t \qquad \text{o} \qquad \dot{K} = I - \delta K$$

- Los modelos difieren en la **reversibilidad** del capital y en la existencia de **costes de ajuste**, que separan a Jorgenson de la q de Tobin.
- Con horizonte infinito, la condición de transversalidad, $\lim_{t\to\infty} e^{-rt} q(t) K(t) = 0$, excluye acumular capital cuyo valor nunca se recupera.

2.4. Decisión de inversión

2.4.1. Modelo

Planteamiento: es un problema de **control óptimo**, resuelto con el principio del máximo de Pontryagin: maximizar $V_0 = \int_0^\infty e^{-rt}\,\Pi(K, I, L)\,dt$ sujeto a $\dot{K} = I - \delta K$, con el hamiltoniano en valor corriente:

$$H = \Pi(K, I, L) + q\,(I - \delta K)$$

CPO: la de control iguala el coste marginal de invertir al precio sombra q; la de estado, $\partial H/\partial K = rq - \dot{q}$, es la ecuación de Euler de la empresa, y la de transversalidad fija el nivel de la senda. Son suficientes con beneficios cóncavos y costes de ajuste convexos.

2.4.2. Consumidor-productor (Fisher, 1930; Hirshleifer, 1958)

Idea: Fisher extendió su análisis al caso en que la renta resulta de una decisión de producción. Un Robinson Crusoe, dueño de los factores y de la tecnología, decide cuánto consumir e invertir, con un capital inicial k_0 y una frontera de posibilidades de producción intertemporal cóncava, porque $f'' < 0$.

Sin mercado de capitales: la única forma de trasladar consumo al futuro es invertir, y la RMSI iguala la **relación marginal de transformación intertemporal**, el rendimiento bruto de invertir. Consumo e inversión quedan ligados:

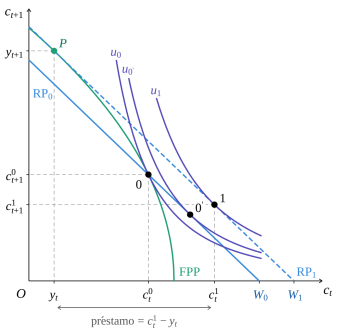
$$\text{RMSI}_{t+1,t} = 1 + f'(k_{t+1}) - \delta \equiv \text{RMTI}$$

Con mercado de capitales: las condiciones se separan. Se invierte hasta que el rendimiento neto del capital iguala el tipo de interés, con independencia de las preferencias, y el consumo se elige después como en el modelo de Fisher:

$$f'(k_{t+1}) - \delta = r \qquad \text{y} \qquad \text{RMSI}_{t+1,t} = 1 + r$$

- Según el **teorema de separación de Fisher**, la inversión, que maximiza la riqueza, es independiente del consumo, que la reparte en el tiempo, y de cómo se financie.

Desarrollo gráfico: la recta de riqueza es tangente a la frontera en el óptimo de producción P, donde RMTI $= 1 + r$, y a la curva u_1 en el óptimo de consumo 1. La distancia entre ambos se financia en el mercado. Frente a la autarquía, el mercado da una **ganancia de intercambio**, con la producción fija, y una **ganancia de producción**, al llevarla a P.



Separación de Fisher: la inversión maximiza el valor presente y el consumo se elige sobre la recta de riqueza.

Conclusiones: fundamenta la valoración de proyectos por su valor actual neto al tipo de mercado y que la empresa maximice su valor. Pero exige mercados de capitales perfectos, sin restricciones de crédito.

2.4.3. Sin costes de ajuste (Jorgenson, 1963)

Idea: la empresa quiere el capital cuya productividad marginal iguala su coste de uso, y la inversión es el ajuste hacia ese *stock*. Con horizonte infinito, tiempo continuo, tipo de interés real exógeno, oferta de bienes de capital perfectamente elástica y sin costes de ajuste, $q = P_K$, y las CPO se reducen a dos condiciones estáticas:

$$Pf'_L = w \qquad \text{y} \qquad Pf'_K = P_K\,(r + \delta) \equiv r_K$$

- El **coste de uso del capital** suma el precio de adquisición, el coste de oportunidad de los fondos y la depreciación. La inversión neta es positiva si el capital deseado supera al instalado. Si coinciden, la empresa solo repone lo que se deprecia, y si el deseado queda por debajo, desinvierte.

Extensiones: con inflación del capital, $r_K = P_K(r + \delta - \pi_K)$, y con imposición, $r_K = \chi_t P_K(r + \delta)$ (Hall y Jorgenson, 1967). La política monetaria actúa a través del tipo de interés real. La fiscal es favorable si usa deducciones, que reducen el coste de uso, y adversa si el gasto eleva el tipo de interés.

Conclusiones: es el modelo canónico de la inversión. Pero sin costes de ajuste el futuro no condiciona la decisión, la CPO es estática y, ante un cambio discreto de r, la inversión daría saltos que la producción agregada, limitada, no permite.

2.4.4. Costes de ajuste: q de Tobin (Tobin, 1969; Hayashi, 1982)

Idea: Tobin propuso como determinante de la inversión la ratio entre el valor de mercado del capital y su coste de reposición, y Hayashi le dio fundamento. Con costes de ajuste internos y convexos, $\varphi(I)$, con $\varphi(0) = \varphi'(0) = 0$ y $\varphi'' > 0$, que afectan también a la desinversión, la empresa se acerca gradualmente al *stock* deseado. Hay un **ritmo óptimo de inversión**, y no un nivel óptimo de capital.

Planteamiento: la empresa elige la senda de inversión que maximiza su valor neto de los costes de ajuste, con k su capital y K el de la industria:

$$\max_{\{I\}} V = \int_0^\infty e^{-rt}[\Pi(K)k - I - \varphi(I)]\,dt \quad \text{s.a.:} \begin{cases} \dot{k} = I \\ k_0 \text{ dado} \end{cases}$$

CPO: el precio sombra es la **q marginal de Tobin**. Se invierte hasta que una unidad más de capital vale lo que cuesta adquirirla e instalarla, y, sin depreciación, el ingreso marginal del capital, $\Pi(K)$, iguala su coste de uso:

$$1 + \varphi'(I) = q, \qquad \Pi(K) = rq - \dot{q}$$

- Los costes de ajuste suavizan la inversión y dan una relación continua y decreciente con el tipo de interés, a diferencia de la discontinua de Jorgenson.
- La segunda es la ecuación de Euler de la empresa: el coste de uso es el rendimiento de invertir q al tipo de interés menos la ganancia de capital esperada.

Conclusiones: da a la inversión una dinámica propia y un vínculo con los mercados financieros, porque la q resume las expectativas de rentabilidad. Pero la q marginal no es observable y solo coincide con la media bajo competencia perfecta y rendimientos constantes, y la inversión responde al flujo de caja, sobre todo en las empresas con difícil acceso a la financiación externa.

En suma, la empresa comparte la estructura del consumidor, pero la separación de Fisher exige mercados financieros perfectos, la regla de Jorgenson carece de dinámica propia y la q marginal no es observable.

3. Horizonte infinito y generaciones solapadas

Los modelos intertemporales microfundamentan la macroeconomía, un interés que afianza la crítica de Lucas (1976), con un **agente representativo**: un agente cualquiera, si todos son idénticos; un agente medio, o un agente ficticio cuyo comportamiento coincide con el colectivo. Permiten comparar políticas en términos de bienestar, y los dos marcos difieren solo en la duración de la vida del agente.

3.1. Horizonte infinito (MHI)

Origen: Ramsey (1928) planteó el ahorro óptimo con horizonte infinito, y Cass y Koopmans (1965) lo convirtieron en el modelo de crecimiento óptimo, del que parten el ciclo real y los modelos DSGE.

Idea: el agente representativo vive indefinidamente, lo que se justifica por las **dinastías** (Barro, 1974), familias unidas por altruismo intergeneracional con herencias operativas, equivalentes a un único agente inmortal, o por la **muerte de Poisson** (Yaari, 1965; Blanchard, 1985), una vida finita de duración desconocida que eleva la tasa de descuento. En este caso hay recambio de individuos aunque la población sea constante.

3.1.1. Supuestos

Tiempo discreto e infinito, mercados financieros perfectos y un agente representativo con utilidad $\sum_{t=0}^{\infty} \beta^t u(c_t)$, rentas y_t y activos a_t. Las restricciones de cada periodo se encadenan en una única restricción intertemporal con la **condición de no-Ponzi**, que impide pagar indefinidamente la deuda con deuda nueva. Sin ella el consumo podría ser infinito:

$$\lim_{T\to\infty} \frac{a_T}{(1+r)^T} \geq 0$$

3.1.2. Modelo

Planteamiento: es el problema de Fisher con infinitos periodos y la condición de no-Ponzi:

$$\max_{\{c_t\}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad \text{s.a.:} \begin{cases} \sum_{t=0}^{\infty} \frac{c_t}{(1+r)^t} = a_0 + \sum_{t=0}^{\infty} \frac{y_t}{(1+r)^t} \\ \lim_{T\to\infty} \frac{a_T}{(1+r)^T} \geq 0 \end{cases} , \qquad \beta \in (0,1)$$

CPO: la ecuación de Euler rige el perfil, como en Fisher, y la **condición de transversalidad** fija el nivel: el valor presente de los activos sin consumir debe ser nulo, y selecciona la única senda compatible con la restricción:

$$\lim_{T\to\infty} \beta^T u'(c_T)\, a_T = 0$$

Solución: con $r = \rho$ la senda es constante, y el consumo es la anualidad de la riqueza, la renta permanente en su forma más pura:

$$c = \frac{r}{1+r}\left[a_0 + \sum_{t=0}^{\infty} \frac{y_t}{(1+r)^t}\right]$$

Equilibrio: en una economía cerrada, la riqueza es el capital, $a_t = k_t$, y $r = f'(k) - \delta$. En el estado estacionario, $r = \rho$ y rige la **regla de oro modificada**, con un capital inferior al de la regla de oro, $f'(k_{\text{oro}}) = n+\delta$, si $\rho > n$ (A43, apartado 3.2):

$$f'(k^*) = \rho + \delta$$

Estática comparativa: con r dado, un aumento permanente de la renta eleva el consumo en toda su cuantía, y uno transitorio solo en su anualidad. Ante una subida del tipo de interés, cobra especial peso, dentro del efecto renta, un efecto riqueza negativo, porque reduce el valor presente de las rentas futuras.

Bienestar: como el agente internaliza el bienestar de todas las generaciones, el equilibrio competitivo es **Pareto-eficiente** y coincide con el del planificador, sin ineficiencia dinámica.

3.1.3. Conclusiones

Aspectos positivos: son casi estructurales y, por ello, en gran medida inmunes a la crítica de Lucas, y su eficiencia sirve de referencia. Se aplican a la oferta de trabajo, el consumo, la inversión, la demanda de dinero, el crecimiento y el ciclo.

Limitaciones: la vida infinita es un supuesto fuerte, desaparecen la heterogeneidad por edad y los conflictos entre generaciones, y suelen exigir métodos numéricos.

3.2. Generaciones solapadas (MGS)

Origen: Allais (1947) y Samuelson (1958) introdujeron la estructura, y Diamond (1965) la llevó a un modelo con capital.

Idea: cada agente vive un número finito de periodos y cada generación llega antes de que se extinga la anterior. En cada periodo conviven jóvenes, que ahorran, y ancianos, que desahorran, y esa **heterogeneidad por edad** es la razón de ser del marco.

3.2.1. Supuestos

La economía es infinita, la población crece a la tasa n y cada agente vive dos periodos. De joven trabaja, cobra w_t, consume y ahorra s_t, y de anciano consume su ahorro capitalizado, sin herencias ni altruismo.

3.2.2. Modelo

Empresas: con una función neoclásica de rendimientos constantes, $y = f(k)$, maximizan cada periodo su beneficio, $F(K_t, L_t) - w_t L_t - r_t K_t$, con salario y tipo de interés dados. Cada factor cobra su productividad marginal, sin depreciación, y el beneficio es nulo:

$$w_t = f(k_t) - k_t f'(k_t), \qquad r_{t+1} = f'(k_{t+1})$$

Hogares: el joven maximiza su utilidad de los dos periodos, y la ecuación de Euler de dos periodos da una función de ahorro $s_t = s(w_t, r_{t+1})$, creciente en el salario y de signo ambiguo respecto del tipo de interés. Con $c_{1t} = w_t - s_t$ y $c_{2t+1} = (1 + r_{t+1})\, s_t$, la ecuación de Euler es:

$$u'(c_{1t}) = \frac{1 + r_{t+1}}{1+\rho}\, u'(c_{2t+1})$$

Equilibrio: como los ancianos consumen todo su capital, el capital del periodo siguiente es el ahorro de los jóvenes:

$$(1+n)\, k_{t+1} = s(w_t, r_{t+1})$$

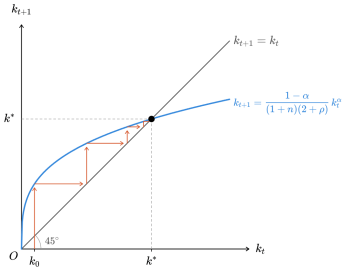
Solución: con utilidad logarítmica y Cobb-Douglas, $y = k^\alpha$, el joven ahorra la fracción $1/(2+\rho)$ de su salario. El capital sigue así una ecuación en diferencias que converge monótonamente, como en Solow pero con un ahorro determinado por las preferencias, al estado estacionario:

$$k_{t+1} = \frac{1-\alpha}{(1+n)(2+\rho)}\, k_t^\alpha$$

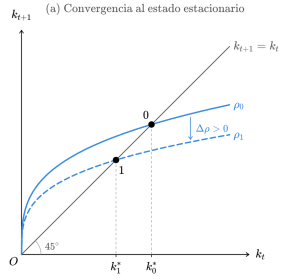
$$k^* = \left[\frac{1-\alpha}{(1+n)(2+\rho)}\right]^{1/(1-\alpha)} \qquad r^* = f'(k^*)$$

Estática comparativa:

- *Impaciencia*: un aumento de ρ reduce la fracción ahorrada, $1/(2+\rho)$, y con ella k^*, y el tipo de interés de equilibrio, $r^* = f'(k^*)$, sube.
- *Población*: un aumento de n reparte el ahorro de cada generación entre más trabajadores, reduce k^* y eleva r^*.
- *Transferencias*: una transferencia de jóvenes a ancianos, con deuda o pensiones de reparto, reduce el ahorro y k^*, y solo mejora a todas las generaciones si la economía es dinámicamente ineficiente.



(a) Convergencia al estado estacionario



(b) Aumento de la tasa de descuento

Dinámica del capital en el modelo de Diamond.

Bienestar: nada garantiza que k^* esté por debajo de la regla de oro, porque cada generación ahorra para su propio ciclo vital. Hay **ineficiencia dinámica** (condición de Malinvaud-Cass) si:

$$r^* < n$$

- Con depreciación, la condición es $f'(k^*) - \delta < n$. La economía ha *sobreacumulado* capital: reducir permanentemente el ahorro elevaría el consumo de todas las generaciones. Se debe a las externalidades pecuniarias intergeneracionales y a la infinitud de agentes y bienes, y la deuda o las pensiones de reparto acercan el capital a la regla de oro.

3.2.3. Conclusiones

Aspectos positivos: son el instrumento para todo lo que depende de la edad, como el ciclo vital, las pensiones, la deuda pública o la demografía, y muestran que la eficiencia no es una propiedad general de las economías dinámicas, sino del supuesto de vida infinita.

Limitaciones: el equilibrio puede no existir o estar indeterminado, con burbujas como el dinero fiduciario; el ciclo vital se reduce a dos periodos, y las economías avanzadas no parecen dinámicamente ineficientes (Abel et al., 1989).

3.3. Comparación entre ambos marcos

Diferencias: ambos marcos comparten agentes optimizadores y microfundamentos. En los MHI el equilibrio es eficiente y se cumple la **equivalencia ricardiana**: una rebaja de impuestos de suma fija financiada con deuda no altera el consumo, porque se ahorra para pagar los impuestos futuros. En los MGS puede haber ineficiencia, y la equivalencia falla, porque quienes reciben la rebaja no son quienes la pagarán.

Juventud perpetua (Blanchard, 1985): tiende un puente entre ambos. Con probabilidad de muerte nula es un MHI, y con probabilidad positiva se comporta como un MGS, y la equivalencia falla tanto más cuanto mayor es esa probabilidad.

Elección: para el ciclo, el crecimiento y la política monetaria, donde la edad es secundaria, el MHI, cuya eficiencia sirve de referencia; para las pensiones, la deuda, la demografía y cualquier cuestión distributiva entre generaciones, el MGS.

En conjunto, la modelización dinámica sustituye las funciones *ad hoc* por decisiones explícitas y permite evaluar las políticas en bienestar. Pero el agente representativo esconde la heterogeneidad, que los modelos HANK reintroducen, y exige supuestos funcionales fuertes.

1. Fundamentos de la contabilidad nacional

La contabilidad nacional descansa en un fundamento teórico, el flujo circular de la renta, que el SEC-2010 concreta en un sistema normalizado, con su concepto y base normativa, su contenido y los principios de su marco contable.

1.1. Fundamento: el flujo circular de la renta

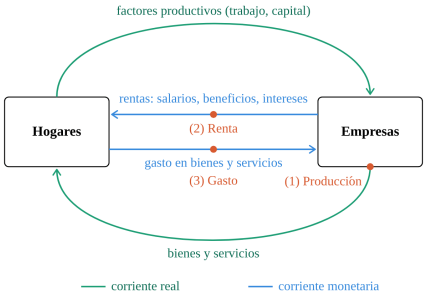
Origen: el fundamento teórico procede de tres siglos. En el XVIII, el *Tableau économique* (1758) del fisiócrata Quesnay avanzó la idea del flujo circular del producto neto; en el XIX, Walras (1874) formalizó la interdependencia general entre los mercados, y en el XX, Knight (1933) dibujó el primer diagrama del flujo circular. Hicks, Leontief, Kuznets y Stone sentaron después las bases de la contabilidad nacional moderna.

Idea: cada unidad física de producto tiene su correspondencia en una unidad monetaria de ingreso, y cada unidad de ingreso termina como gasto en un mercado.

Supuestos: una economía simplificada y cerrada, sin sector público, con dos agentes de doble función: las empresas ofrecen productos y demandan factores, y los hogares demandan bienes y ofrecen factores.

Razonamiento: entre ambos nacen dos corrientes de sentido opuesto: una real, de bienes, servicios y factores entendidos como entidades físicas, y otra monetaria, el ingreso que se obtiene en las transacciones y el gasto, que es su imagen especular.

Desarrollo gráfico: la corriente real lleva los factores de los hogares a las empresas y los bienes y servicios de vuelta, y la monetaria, las rentas hacia los hogares y su gasto hacia las empresas.



Flujo circular de la renta: corriente real y corriente monetaria entre hogares y empresas, y los tres puntos de medida del PIB.

Intuiciones: el modelo desvela dos.

1. Las interrelaciones entre los sectores son **identidades contables**, que se cumplen siempre por construcción y no implican equilibrio económico.
2. La producción, la renta y el gasto son tres medidas complementarias de un mismo flujo y los tres puntos del circuito en los que se sitúan las cuentas.

Representa de forma sencilla la interdependencia entre hogares y empresas y fundamenta la equivalencia entre producción, renta y gasto en la que descansa la medición del PIB. Pero en su versión básica omite el ahorro, la inversión, el sector público y el resto del mundo, y, como sus relaciones son identidades, describe cómo circula la renta, pero no explica qué determina su nivel ni sus fluctuaciones.

1.2. Concepto y base normativa

Definición: la **contabilidad nacional** es el registro sistemático y estadístico del conjunto de la actividad macroeconómica de un país durante un periodo. Su objetivo es un marco estadístico completo, coherente y flexible para el análisis económico comparado y la formulación de la política económica. Es una técnica de representación que da una descripción cuantitativa y simplificada de la actividad y muestra cómo se alcanza el equilibrio contable entre las magnitudes agregadas.

SCN: en el ámbito internacional, un grupo de expertos presidido por Richard Stone elaboró en 1953 para la ONU el primer Sistema de Cuentas Nacionales, actualizado en 1968, 1993, 2008 y 2025. El SCN-2008, concebido para todos los países, está armonizado con el Manual de Balanza de Pagos (MBP6) del FMI. En marzo de 2025 la ONU aprobó el SCN-2025 y el FMI publicó el MBP7, que la UE incorporará en una futura revisión del SEC.

SEC-2010: en la Unión, el Sistema Europeo de Cuentas Nacionales y Regionales sustituye al SEC-95 y es un reglamento vinculante, con reglas, definiciones, nomenclaturas y normas contables comunes. De él dependen el recurso propio basado en la renta nacional bruta (RNB), el procedimiento de déficit excesivo o los fondos estructurales.

Cambios respecto al SEC-95: la estructura apenas varía. Con impacto sobre el PIB, los gastos en I+D y los sistemas de armamento pasan a ser formación bruta de capital fijo (FBKF). Sin impacto, se detalla la delimitación del sector público y de las instituciones financieras y se armoniza con el MBP6 el registro de los flujos con el resto del mundo, como el comercio triangular o los bienes enviados al extranjero para su transformación.

1.3. Contenido: unidades, flujos y *stocks*

El sistema describe la economía a partir de las unidades que realizan operaciones y poseen activos y pasivos, y de los precios a los que se valora la producción.

1.3.1. Unidades y sectores institucionales

Unidades institucionales: entes capaces de ser propietarios de bienes y activos, contraer pasivos y operar con otras unidades en nombre propio. Se agrupan, por el tipo de productor y su función principal, en cinco sectores mutuamente excluyentes, más el resto del mundo:

- Sociedades no financieras (S.11) e instituciones financieras (S.12), productores de mercado privados y públicos, incluido el banco central.
- Administraciones públicas (S.13), productores de no mercado que además redistribuyen la renta y la riqueza, y comprenden la Administración central, la regional, la local y los fondos de la Seguridad Social.
- Hogares (S.14), consumidores y empresarios individuales, e instituciones sin fines de lucro al servicio de los hogares (ISFLSH, S.15), como ONG, sindicatos, partidos, fundaciones o iglesias.
- *Resto del mundo (S.2)*: las unidades no residentes y las instituciones de la UE.

Mercado y no mercado: la distinción separa las empresas públicas de las AA. PP. Una unidad pública es productor de mercado si, entre otros criterios, sus ventas cubren al menos el 50 % de sus costes de producción durante varios años. Para analizar el proceso productivo, las unidades de actividad económica locales (UAE) agrupan la producción por ramas, la base de las tablas de origen y destino.

1.3.2. Flujos y *stocks*

Flujos: crean, transforman, intercambian, transfieren o extinguen valor económico y modifican lo que valen los activos o pasivos de la unidad afectada. Se dividen en *operaciones*, de bienes y servicios, distributivas, financieras y otras, y *otras variaciones de activos*, como catástrofes, descubrimientos o revalorizaciones.

Stocks o **posiciones**: son los activos y pasivos en un momento concreto. Los activos, entidades sobre las que se ejercen derechos de propiedad y de las que se derivan beneficios económicos, son no financieros, producidos, como el capital fijo y las existencias, o no producidos, como la tierra, o financieros, con un pasivo de otra unidad como contrapartida. La diferencia entre activos y pasivos es el *patrimonio neto*, que se registra en los balances.

1.3.3. Precios básicos y de adquisición

Precio básico: lo que retiene el productor por unidad del producto j, el coste de los consumos intermedios y de los factores más los otros impuestos netos sobre la producción, $p_j^b = c_j + t_{p,j}$. Entre esos otros impuestos están los que gravan los inmuebles de la empresa o las licencias de actividad.

Precio de adquisición: lo que paga el comprador, que añade los impuestos netos sobre los productos y los márgenes de distribución y transporte, $p_j^a = p_j^b + t_{c,j} + \mu_{c,j}$. Los impuestos sobre los productos son el IVA no deducible, los impuestos especiales y los derechos de importación.

Agregados: el valor añadido bruto de cada rama es su producción a precios básicos menos los consumos intermedios, valorados a precios de adquisición, $\mathrm{VAB}_i^{\mathrm{pb}} = \mathrm{VP}_i - \mathrm{CI}_i$. Al agregar, los márgenes se cancelan, porque son producción de las ramas de comercio y transporte, y solo hay que añadir los impuestos sobre los productos, que no son valor añadido de ninguna rama:

$$\mathrm{PIB}_{\mathrm{pm}} = \sum_i \mathrm{VAB}_i^{\mathrm{pb}} + T_c$$

- La recaudación es $T_c = \sum_j t_{c,j}\, x_j$, con x_j las unidades vendidas del producto j. Deduciendo los otros impuestos netos sobre la producción se obtiene el *valor añadido a coste de los factores*, $\mathrm{VAB}_i^{\mathrm{cf}} = \mathrm{VAB}_i^{\mathrm{pb}} - T_{p,i}$, con $T_{p,i}$ los de la rama, que el SEC-2010 no utiliza formalmente.

1.4. Principios del marco contable

Cinco decisiones metodológicas fijan qué se contabiliza, cómo se registra, cuándo, con qué criterio de valoración y en qué unidad.

1.4.1. Delimitación: residencia

Criterio de residencia, y no de nacionalidad: es residente la unidad cuyo centro de interés económico predominante está en el territorio económico, lo que en la práctica exige actividad allí durante un año como mínimo. La economía nacional agrupa a los residentes, y el resto del mundo, a los no residentes que operan con ellos.

1.4.2. Registro: partida doble

Partida doble: cada operación se anota dos veces con el mismo valor para cada unidad, como **recurso**, si aumenta su valor económico, o variación de pasivos, y como empleo, si lo reduce, o variación de activos. Como cada operación afecta a dos unidades, el sistema es en realidad de *partida cuádruple*, lo que garantiza la coherencia. Las cuentas se

presentan en T, con los empleos a la izquierda y los recursos a la derecha, y se cierran con un **saldo contable** que pasa como recurso a la cuenta siguiente. Son recursos, entre otros, la producción o las rentas recibidas, y empleos, los consumos intermedios o las rentas pagadas.

1.4.3. Momento del registro: devengo

Principio de devengo: los flujos se anotan cuando se crea, transforma o extingue el valor, o cuando nacen o se cancelan derechos y obligaciones, al margen de cobros y pagos. La producción se registra cuando se obtiene, de ahí la variación de existencias; los activos, cuando se transfiere su propiedad económica, y los intereses, cuando se devengan. La única desviación son los impuestos y otros flujos de las AA. PP., con un registro por caja **ajustado en el tiempo**.

1.4.4. Valoración

Criterios: el SEC-2010 aplica tres.

1. *Precios de mercado*, la referencia general: el valor al que los flujos y *stocks* se cambian o podrían cambiarse por efectivo, con la producción a precios básicos y los empleos a precios de adquisición.
2. *Costes de producción*, para los servicios no de mercado de las AA. PP. y las ISFLSH, cuando no hay precios de mercado propios ni análogos: consumos intermedios, remuneración de asalariados, consumo de capital fijo y otros impuestos netos sobre la producción.
3. *Valor franco a bordo* (FOB) para el comercio de bienes, lo que obliga a ajustar las importaciones aduaneras, declaradas a valor CIF. El valor FOB es el de la frontera del país exportador, sin fletes ni seguros.

Los agregados se valoran a **precios corrientes**. El volumen, el "PIB real", no es otra valoración, sino una descomposición posterior de la variación nominal en precios y cantidades, con índices encadenados.

1.4.5. Unidad de cuenta

La unidad de cuenta es la moneda nacional. Para comparar países se usan tipos de cambio o, preferiblemente, **paridades de poder adquisitivo** (PPA), que corrigen las diferencias de nivel de precios.

En suma, el flujo circular da a la contabilidad nacional su fundamento, y el SEC-2010, con sus sectores, sus flujos y *stocks*, sus precios y sus cinco principios, hace las cuentas coherentes y comparables.

2. Las cuentas nacionales

El PIB, que puede medirse por tres vías, es el primero de los saldos de una secuencia de cuentas, no financieras y financieras, cuya lectura económica se hace en términos de ahorro e inversión.

2.1. El PIB y el triple enfoque

Definición: el **producto interior bruto** a precios de mercado es el valor de los bienes y servicios finales producidos por las unidades residentes durante un periodo. Es un flujo, medido en dinero para sumar bienes heterogéneos, y solo computa bienes finales para no contar más de una vez los intermedios. Por eso, el concepto relevante es el **valor añadido**.

Triple enfoque: se mide por tres vías que coinciden, los tres puntos del flujo circular:

$$\mathrm{PIB}_{\mathrm{pm}} = \underbrace{\sum_i \mathrm{VAB}_i^{\mathrm{pb}} + T_c}_{\text{Oferta}} = \underbrace{\mathrm{RA} + T + \mathrm{EBE} + \mathrm{RM}}_{\text{Rentas}} = \underbrace{C + I + G + (X - M)}_{\text{Demanda}}$$

- La oferta, saldo de la cuenta de producción, suma el valor añadido de las ramas y los impuestos sobre los productos.
- Las rentas, reparto de la cuenta de explotación, suman la remuneración de los asalariados, los impuestos netos sobre la producción y las importaciones, el excedente bruto de explotación y la renta mixta.
- La demanda, los empleos finales, suma el consumo final de hogares e ISFLSH, la formación bruta de capital, el consumo final de las AA. PP. y las exportaciones netas. La formación bruta de capital comprende la FBKF, la variación de existencias y las adquisiciones menos cesiones de objetos valiosos. El consumo de los hogares ronda el 55-60 % del PIB en la OCDE y es estable y procíclico, y el de las AA. PP., una quinta parte en Europa.

Interior y nacional: del PIB se pasa a la RNB y a la renta nacional bruta disponible (RNBD) con las rentas y transferencias con el resto del mundo. La diferencia no es de nacionalidad, sino de residencia: el PIB atiende a la de quien produce, y la RNB, a la de quien percibe las rentas.

Bruto y neto: las magnitudes netas descuentan el consumo de capital fijo, $\mathrm{PIN} = \mathrm{PIB} - \mathrm{CCF}$, y la inversión neta es la que amplía el *stock* de capital, $I_N = K_{t+1} - K_t$, igual a la bruta menos el consumo de capital fijo, $I_N = I - \mathrm{CCF}$.

2.2. Cuentas no financieras

Se elaboran para cada sector y para la economía total, de la producción a la acumulación y de esta a los balances; aquí, las de la economía total.

2.2.1. Cuentas corrientes

Las **cuentas corrientes** registran la producción, la generación y distribución de la renta y su utilización en consumo y ahorro.

Cuenta de producción (C.I): los recursos son la producción a precios básicos y, para la economía total, los impuestos netos sobre los productos; los empleos, los consumos intermedios. El saldo es el *valor añadido bruto* de cada sector y, para el total, el PIB, que se obtiene agregando valores añadidos, no producciones.

Cuenta de explotación (C.II.1.1): parte del PIB y le resta la remuneración de los asalariados interior y los impuestos netos sobre la producción y las importaciones. El saldo es el *excedente bruto de explotación* (EBE) y la *renta mixta bruta* de los empresarios individuales, que lleva implícita la remuneración de su trabajo. El EBE es el excedente antes de deducir intereses, rentas de la tierra y otras rentas de la propiedad.

Asignación de la renta primaria (C.II.1.2): adopta la óptica de los residentes que perciben la renta. Sus recursos son el EBE y la renta mixta, la remuneración que reciben los hogares residentes, de empleadores residentes o no, los impuestos netos que reciben las AA. PP. y las rentas de la propiedad recibidas, y sus empleos, las rentas de la propiedad pagadas, también al exterior. El saldo es la **renta nacional bruta**, que difiere del PIB en el saldo de rentas primarias con el resto del mundo, SBR_1, que suele ser negativo en los países deudores netos por el pago de intereses y dividendos.

Distribución secundaria (C.II.2): redistribuye la renta con transferencias corrientes, como los impuestos sobre la renta y el patrimonio y las cotizaciones y prestaciones sociales. Para el total solo quedan las cruzadas con el exterior, el saldo de rentas secundarias, SBR_2, y el saldo es la **renta nacional bruta disponible**:

$$\mathrm{RNB} = \mathrm{PIB} + \mathrm{SBR}_1, \qquad \mathrm{RNBD} = \mathrm{RNB} + \mathrm{SBR}_2$$

- Las rentas primarias con el exterior son la remuneración de asalariados, las rentas de la propiedad y los impuestos netos con las instituciones de la UE, y las secundarias, p. ej. las remesas, la cooperación al desarrollo y los recursos propios IVA y RNB de la UE.

Redistribución de la renta en especie (C.II.3): traslada a los hogares la educación y la sanidad que financian las AA. PP. y las ISFLSH. Su saldo, la *renta disponible ajustada*, no altera el total, pero sí su reparto entre sectores.

Utilización de la renta disponible (C.II.4.1): reparte la RNBD entre consumo final y ahorro. Solo consumen los hogares, las ISFLSH y las AA. PP. El consumo *individual* es el de hogares e ISFLSH y parte del de las AA. PP., como la educación o la sanidad; el *colectivo*, el resto del de las AA. PP., bienes públicos puros como la defensa o la justicia. Con el ajuste por la variación de los derechos de pensión, el saldo es el **ahorro nacional bruto**.

2.2.2. Cuentas de acumulación

Registran las variaciones de activos, pasivos y patrimonio neto: las cuentas de capital y financiera, las debidas a operaciones, y la de otras variaciones, las restantes.

Cuenta de capital (C.III.1): en la primera subcuenta, el ahorro neto más las transferencias de capital netas recibidas, como los fondos estructurales o las condonaciones de deuda, da la *variación del patrimonio neto debida al ahorro y a las transferencias de capital* (B.10.1). En la segunda, ese saldo financia la FBKF, incluida la vivienda nueva, la variación de existencias, los objetos valiosos y los activos no producidos, netos del consumo de capital fijo. El saldo es la **capacidad (+) o necesidad (−) de financiación** (CIN).

La FBKF ronda el 20 % del PIB y es mucho más volátil y procíclica que el consumo. Es coincidente o ligeramente retrasada respecto al ciclo, y solo la inversión residencial suele adelantarse. La variación de existencias es pequeña, en torno al 0,5-1 % del PIB, pero de gran varianza.

CIN: es el saldo central del sistema. Si es positiva, la nación ha ahorrado, sumadas las transferencias netas de capital, más de lo que ha invertido y presta al resto del mundo, y si es negativa, ha necesitado financiación exterior. Como es la suma de la de los sectores, el saldo exterior refleja los desequilibrios internos. En varias economías del euro, la elevada necesidad de financiación previa a la Gran Recesión procedía de hogares y sociedades no financieras, y el ajuste posterior descansó en el aumento del ahorro privado, mientras crecía la necesidad de las AA. PP.

Otras variaciones de los activos (C.III.3): las de volumen (OVV), como descubrimientos, catástrofes o cancelaciones unilaterales de deuda, y las revalorizaciones (REV) por cambios de precio de los activos existentes.

2.2.3. Balances

Balances: recogen el patrimonio neto, los activos y los pasivos al inicio y al cierre del ejercicio, y sus variaciones son exactamente las registradas en las cuentas de acumulación:

$$\mathrm{PN}_t = \mathrm{PN}_{t-1} + \mathrm{B.10.1} + \mathrm{OVV} + \mathrm{REV}$$

Para la economía total, el patrimonio neto es la riqueza nacional.

2.3. Cuentas financieras

Lema: toda operación, real o financiera, tiene una contrapartida financiera, salvo el trueque, las operaciones en especie y las internas. De ahí tres implicaciones:

1. *Doble registro*: la venta de un bien se cobra en efectivo o a crédito, y un préstamo genera un depósito.
2. *Saldo real y financiero*: el saldo de las cuentas reales de cada unidad iguala el de sus cuentas financieras, porque quien gasta más de lo que ingresa contrae pasivos o se desprende de activos financieros.
3. *Economía total*: la suma de las capacidades y necesidades de los sectores iguala la de sus saldos financieros.

Cuenta financiera (C.III.2): muestra en qué instrumentos, de F.1 a F.8, del oro monetario y los DEG a los derivados y los créditos comerciales, se materializa la CIN. Su saldo es conceptualmente idéntico al de la cuenta de capital, y la diferencia estadística entre ambos es el ajuste entre las cuentas real y financiera:

$$\mathrm{SB_{CF}} \equiv \Delta \mathrm{AF} - \Delta \mathrm{PF} = \mathrm{CIN}$$

Para la economía total, los activos y pasivos entre residentes se cancelan, y el saldo coincide conceptualmente con el de la cuenta financiera de la balanza de pagos:

$$\mathrm{SB_{CF}^{CN}} \simeq \mathrm{SB_{CF}^{BP}}$$

COVAPF: la cuenta de otras variaciones de activos y pasivos financieros recoge las revalorizaciones y las otras variaciones de volumen, y concilia la balanza de pagos con la variación de la posición de inversión internacional neta, $\Delta \mathrm{PIIN} = \mathrm{SB_{CF}^{BP}} + \mathrm{COVAPF}$. Sus componentes son los efectos de valoración, por tipo de cambio o precio de los activos, y las otras variaciones de volumen, como cancelaciones o reclasificaciones:

$$\mathrm{COVAPF} = \mathrm{REV} + \mathrm{OVV}$$

2.4. Identidades ahorro-inversión

Con el ahorro definido como la renta disponible no consumida, $S = \mathrm{RNBD} - C - G$, los saldos se leen en términos de ahorro e inversión, según los agentes considerados, de la economía cerrada sin sector público a la abierta:

$$\begin{aligned} \mathrm{RNBD} &= C + I & S &= I \\ \mathrm{RNBD} &= C + I + G & (S_P - I_P) + (S_G - I_G) &= 0 \\ \mathrm{RNBD} &= C + I + G + \mathrm{SB_{CC}} & (S_P - I_P) + (S_G - I_G) &= \mathrm{SB_{CC}} \end{aligned}$$

Economía cerrada: sin sector público, la renta disponible coincide con el PIB y el ahorro privado financia toda la inversión privada. Con sector público, lo que un sector invierte por encima de su ahorro lo cubre el ahorro sobrante del otro.

Economía abierta: si la inversión supera al ahorro nacional, la diferencia se cubre con financiación exterior y aparece un déficit por cuenta corriente. El saldo corriente suma los de bienes y servicios, $X-M$, rentas primarias y rentas secundarias.

Sumando a ambos lados el saldo de la cuenta de capital de la balanza de pagos, $\mathrm{SB_{CK}}$, se obtiene la capacidad o necesidad de financiación de la nación, con la que se cierra la cuenta de capital:

$$S + \mathrm{SB_{CK}} - I = \mathrm{SB_{CC}} + \mathrm{SB_{CK}} \equiv \mathrm{CIN}$$

– El $\mathrm{SB_{CK}}$ recoge las transferencias netas de capital recibidas del exterior, ahorro aportado desde fuera sin contrapartida, y las cesiones netas al exterior de activos no financieros no producidos. La cuenta financiera muestra cómo se materializa la CIN.

Precisiones: dos.

– El saldo público, $S_G - I_G$ más las transferencias netas de capital, es la capacidad o necesidad de financiación de las AA. PP., el saldo presupuestario, que no debe confundirse con el **ahorro público**.
– La identidad se cumple *ex post* por definición, mientras que el equilibrio exige que los planes de ahorro e inversión coincidan *ex ante*, y la diferencia se manifiesta en la variación no deseada de existencias.

En suma, la secuencia de cuentas describe la economía de forma completa y coherente, de la producción a la riqueza, con saldos encadenados, y las cuentas económicas integradas muestran los desequilibrios sectoriales y su reflejo exterior. Sus límites: deja fuera el trabajo doméstico no remunerado y las externalidades, capta de forma imperfecta la economía sumergida, valora la producción de no mercado por sus costes y mide actividad, no bienestar. El SEC-2010 exige exhaustividad, y los institutos incluyen estimaciones de la economía no observada, desde 2014 también de ciertas actividades ilegales.

3. Otras cuentas, indicadores y reflexiones

Junto al marco central, el sistema comprende otras cuentas, y sobre él se construyen los principales indicadores macroeconómicos. El apartado se cierra con el PIB como medida del bienestar.

3.1. Otras cuentas

3.1.1. Cuenta del resto del mundo (S.2)

Resto del mundo: registra las operaciones entre residentes y no residentes y se corresponde con las cuentas internacionales del MBP6, pero desde la óptica del no residente, de modo que los signos se invierten. Una exportación, crédito en la balanza de pagos, es aquí un empleo. Agrupa a los no residentes en una unidad que no es un sector institucional pero desempeña un papel análogo, con las cuentas de intercambios exteriores de bienes y servicios, de rentas primarias y transferencias corrientes, de capital y financiera.

3.1.2. Contabilidad regional (CRE)

CRE: el INE distribuye el PIB, el valor añadido, el empleo y la renta de los hogares entre comunidades autónomas y provincias. De ella depende la elegibilidad para los fondos estructurales como región menos desarrollada, con PIB per cápita por debajo del 75 % de la media de la UE, y alimenta el debate sobre balanzas fiscales y financiación autonómica.

3.1.3. Tablas *input-output* y cuentas satélite

Tablas de origen y destino e *input-output*: describen las relaciones intersectoriales, son la base del marco de Leontief y dan los multiplicadores de producción y empleo de cada rama y el contenido importado de la demanda final.

Cuentas satélite: añaden detalle sin alterar el marco central, como las de la agricultura, el medio ambiente, la salud, la producción de los hogares, el turismo, la I+D o la economía digital.

3.2. Crecimiento, renta y ciclo

3.2.1. Índices de volumen encadenados

Ponderaciones fijas: tradicionalmente, el PIB real valoraba la producción a los precios de un año base, que perdía representatividad con el tiempo, más aún con unos precios relativos que la innovación cambia deprisa. Separaba precios y cantidades, pero obligaba a actualizar la base cada cinco o diez años.

Índices encadenados: desde mediados de los noventa, en Estados Unidos en 1996 y en la UE en la década de 2000, se valora la producción de cada año a los precios del anterior y se encadenan los eslabones:

$$\mathrm{EV}_t = \frac{\sum_i p_{i,t-1}\, x_{i,t}}{\sum_i p_{i,t-1}\, x_{i,t-1}} \qquad \mathrm{IV}_t = \mathrm{IV}_{t-1} \cdot \mathrm{EV}_t$$

La base se actualiza cada año, lo que resuelve la representatividad y reduce las revisiones. Pero no son aditivos, es decir, la suma de los componentes del PIB en volumen no coincide con el PIB en volumen, salvo en el año de referencia y el siguiente, y por eso se difunden las *contribuciones al crecimiento*, que sí lo son en términos anuales.

3.2.2. Renta y riqueza

Riqueza nacional: es el patrimonio neto de la economía total, igual a los activos no financieros situados en el país, K, más la posición de inversión internacional neta, diferencia entre los activos, $\mathrm{AF^{ext}}$, y los pasivos, $\mathrm{PF^{ext}}$, de los residentes frente al exterior, porque los activos financieros entre residentes se cancelan:

$$\mathrm{AN} = K + \mathrm{PIIN} = K + (\mathrm{AF^{ext}} - \mathrm{PF^{ext}})$$

– Una PIIN deudora reduce la riqueza nacional respecto a los activos situados en el país.

PIB per cápita: es el indicador habitual del nivel de desarrollo y se compara en PPA. Eurostat lo publica en estándares de poder adquisitivo (PPS), la referencia de la política de cohesión. Los PPS son una unidad monetaria artificial que elimina las diferencias de nivel de precios entre países.

Distribución: como el PIB per cápita es una media, se usan el *índice de Gini*, la *ratio S80/S20* y la *tasa de riesgo de pobreza*, con el umbral en el 60 % de la mediana. En Europa, la desigualdad de la renta de mercado es muy superior a la de la renta disponible, por las transferencias, sobre todo las pensiones. El índice de Gini es el cociente entre el área comprendida entre la curva de Lorenz y la diagonal y el área bajo la diagonal, entre 0 y 1, y la ratio S80/S20, la renta del quintil superior sobre la del inferior.

3.2.3. Ciclo

Indicadores de confianza: encuestas a consumidores y empresas sobre la situación y las expectativas, como el indicador de sentimiento económico (ESI) de la Comisión Europea o los PMI, con su umbral de 50. Son rápidos, pero miden opiniones y no operaciones. Preguntan por la situación financiera del hogar, la intención de compra de duraderos, la cartera de pedidos o el empleo previsto.

Indicadores adelantados: índices compuestos de variables que anticipan los puntos de giro, como los pedidos industriales, los permisos de construcción o la pendiente de la curva de tipos. Destacan los de la OCDE y el *nowcasting* de los bancos centrales.

3.3. Empleo y precios

3.3.1. Mercado de trabajo

Encuesta a los hogares: la EPA en España, o la CPS en EE. UU., clasifica a la población en edad de trabajar en ocupados, que trabajaron

al menos una hora en la semana de referencia; parados, sin empleo, disponibles y que han buscado activamente empleo en las cuatro semanas anteriores, e inactivos:

$$u = \frac{U}{N_A} \qquad a = \frac{N_A}{N_{16+}} \qquad e = \frac{E}{N_{16+}}$$

La tasa de empleo es la referencia de la Estrategia Europea de Empleo, porque no depende de los movimientos entre paro e inactividad. Son inactivos los estudiantes, los jubilados, las personas dedicadas al hogar y los trabajadores desanimados.

Registros administrativos: el paro registrado en el SEPE y la afiliación a la Seguridad Social pueden divergir de la encuesta, porque autónomos y pluriempleados se cuentan de otro modo y no todo parado se inscribe, ni todo inscrito es parado en sentido EPA, además de los errores de estimación.

Productividad aparente: valor añadido en volumen por unidad de empleo. Se denomina "aparente" porque atribuye todo el valor añadido al trabajo. Con la remuneración por asalariado da los *costes laborales unitarios* (CLU), indicador central de competitividad. El empleo puede medirse en personas, en puestos equivalentes a tiempo completo o en horas trabajadas.

3.3.2. Precios e inflación

IPC: compara lo que cuesta una cesta fija de consumo con lo que costaba en el periodo base. Es un índice de *Laspeyres*, y el IPC armonizado es la referencia del BCE:

$$P_L = \frac{\sum_{i=1}^{n} p_{it}\, x_{i0}}{\sum_{i=1}^{n} p_{i0}\, x_{i0}}$$

Tiene tres limitaciones:

1. *Sesgo inflacionista*: ignora el efecto sustitución y no recoge plenamente la calidad ni los nuevos productos. La Comisión Boskin (1996) lo estimó en 1,1 puntos anuales para EE. UU., y estimaciones posteriores, en 0,9.
2. *Índice plutocrático*: pondera a los hogares por su gasto, de modo que los ricos pesan más.
3. *Inflación importada*: recoge perturbaciones externas de la energía y los alimentos. Por eso se usa la *inflación subyacente*, que excluye la energía y los alimentos no elaborados.

Deflactor del PIB: el cociente entre el PIB nominal y el real es un índice de *Paasche*:

$$P = \frac{\sum_{i=1}^{n} p_{it}\, x_{it}}{\sum_{i=1}^{n} p_{i0}\, x_{it}}$$

Difiere del IPC en la cobertura, porque mide toda la producción y no solo el consumo de los hogares; en las importaciones, que no le afectan directamente, y en las ponderaciones. El IPC tiende a sobrestimar la subida del coste de la vida, y el deflactor, a subestimarla. El índice de **Fisher**, media geométrica de ambos, es el ideal.

3.4. Indicadores sectoriales

3.4.1. Sector público

Gasto e ingreso: el gasto no financiero sobre el PIB mide el tamaño del sector público, en torno al 45 % en España. La **presión fiscal** es la recaudación efectiva de impuestos y cotizaciones sobre el PIB, y el *esfuerzo fiscal* o índice de Frank la relativiza por el PIB per cápita:

$$\text{PFis} = \frac{T_{\text{tot}} + \text{CS}}{\text{PIB}} \qquad \text{EF} = \frac{\text{PFis}}{\text{PIB}_{\text{pc}}}$$

– La presión fiscal mide lo que efectivamente se paga, de modo que una mayor evasión la reduce y los tipos legales sobrestiman la carga efectiva.

Saldo y deuda: el saldo presupuestario, sus componentes y la deuda sobre el PIB sustentan los criterios de Maastricht, del 3 % y el 60 %, y el procedimiento de déficit excesivo. Como se elaboran según el SEC-2010, el sistema contable tiene un papel jurídico directo, porque la delimitación del sector AA. PP. y el registro por devengo determinan qué unidades y operaciones computan en el déficit y la deuda.

3.4.2. Sector financiero

Indicadores: los agregados monetarios M1, M2 y M3; la ratio de crédito privado sobre el PIB y su brecha (*credit-to-GDP gap*), referencia del ciclo financiero y del colchón anticíclico; el €STR, el euríbor, la curva de tipos y la prima de riesgo frente al *Bund*, y las ratios CET1, de morosidad, LCR y NSFR de las entidades. Los diferenciales soberanos y corporativos miden el riesgo percibido y la fragmentación financiera, y el endeudamiento de hogares y sociedades sobre la renta o el PIB, la vulnerabilidad de sus balances.

3.4.3. Sector exterior

Tipo de cambio efectivo real: es el nominal, media ponderada de los bilaterales expresados en moneda extranjera por unidad nacional, deflactado por los precios o costes relativos, $\text{TCER} = \text{TCEN} \cdot P/P^*$. Su aumento es una apreciación real, es decir, una pérdida de competitividad-precio. Cada moneda pondera según el peso del país en el comercio, con doble ponderación para recoger la competencia en terceros mercados.

Indicadores *ex post*: la cuota de mercado, la tasa de cobertura, la propensión exportadora y el grado de apertura y, con las cadenas globales de valor, el contenido importado de las exportaciones. Una mayor integración en esas cadenas eleva la productividad, pero también la exposición a los *shocks* internacionales. Los cuatro primeros son, por ese orden:

$$s^j = \frac{X^j}{\sum_i M^i} \qquad \frac{X}{M} \qquad \frac{X}{\text{PIB}} \qquad \frac{X+M}{\text{PIB}}$$

Reservas: se juzgan por tres meses de importaciones, por la regla de Guidotti-Greenspan, que pide cubrir el 100 % de la deuda externa a corto plazo, o por el 20 % de M2. Miden la capacidad de las autoridades para hacer frente a salidas de capital.

3.5. El PIB como medida del bienestar

El PIB es la variable flujo más importante, pero mide actividad, no bienestar, como advirtió Kuznets (1934): "el bienestar de una nación difícilmente puede inferirse de una medida de la renta nacional". Aun así, indica en qué medida una sociedad cubre sus necesidades materiales con los recursos de que dispone.

Paradoja de Easterlin (1974): dentro de un país, los más ricos declaran más felicidad, pero con el tiempo el crecimiento del PIB per cápita no eleva apreciablemente la felicidad media, lo que apunta a la renta relativa y a la adaptación. Estudios posteriores la matizan: la relación es positiva también entre países y en el tiempo, con rendimientos decrecientes.

Comisión Stiglitz-Sen-Fitoussi (2009): recomendó complementar el PIB con medidas de la renta y el consumo de los hogares, de la distribución, de las actividades no de mercado, de la calidad de vida y de la sostenibilidad. Esta agenda refuerza la iniciativa Beyond GDP de la Comisión Europea (2007) e inspira el *Better Life Index* de la OCDE (2011) y los indicadores de la Agenda 2030. La Comisión, creada a instancias del Gobierno francés, antepuso la óptica de los hogares a la de la producción y propuso medir la sostenibilidad por los *stocks* de capital físico, humano y natural.

Índices compuestos: el IDH del PNUD combina renta per cápita, esperanza de vida y educación, con variantes por desigualdad y por género. Sus ponderaciones son arbitrarias y sus componentes, muy correlacionados con el PIB per cápita.

Límites del PIB: excluye el trabajo no remunerado y el ocio, contabiliza como positivo el gasto defensivo, ignora la depreciación del capital natural y no informa del reparto. La respuesta de la contabilidad nacional ha sido ampliar el marco, con cuentas satélite, de distribución de la renta de los hogares y del capital natural, antes que sustituir el agregado.

A32. El modelo de oferta y demanda agregada: determinación de renta e inflación en una economía abierta. Análisis de las políticas monetaria y fiscal, y de los shocks y políticas de oferta

1. El modelo IS-LM-BP (Fleming, 1962; Mundell, 1963)

Idea: el modelo de Fleming (1962) y Mundell (1963), canónico de la economía abierta a corto plazo, extiende el IS-LM de la síntesis neoclásica con un tercer mercado, el sector exterior, y con él dos elementos nuevos: el tipo de cambio y el régimen cambiario.

1.1. Supuestos

Entorno: economía abierta y pequeña, con P^*, i^* e Y^* exógenos, y corto plazo con precios fijos e inflación nula, de modo que $i = r$ y los tipos de cambio nominal y real son proporcionales. La demanda determina la producción. El tipo de cambio, E, se define en unidades de moneda nacional por unidad de moneda extranjera, y su aumento es una depreciación.

Agentes: son neutrales al riesgo, de modo que no exigen prima por mantener activos en distintas monedas, y tienen expectativas estáticas sobre el tipo de cambio. Los flujos de capital responden al diferencial de tipos con una intensidad γ, que mide la movilidad del capital, y en el límite, con bonos nacionales y extranjeros sustitutivos perfectos, se cumple la paridad de intereses.

Sector exterior: la cuenta corriente es la balanza comercial y se cumple la condición Marshall-Lerner, $\varepsilon_X + |\varepsilon_M| > 1$, de modo que una depreciación mejora la balanza, porque los efectos volumen dominan el efecto valoración. Con cambio fijo, la variación de reservas cierra la balanza de pagos.

1.2. Modelo

1.2.1. Mercado de bienes: curva IS

Definición: la **curva IS** reúne los pares (Y, r) que equilibran el mercado de bienes, $Y = C + I + G + \text{XN}$, con un consumo keynesiano, $C = C_0 + c(1-\tau)Y$, una inversión $I = I_0 - v\,r$, un gasto exógeno, G_0, y una balanza comercial $\text{BC} = x\,Y^* - m\,Y + \phi\,e$, donde $e = EP^*/P$ es el tipo de cambio real y $\phi > 0$ por Marshall-Lerner:

$$Y = \frac{1}{1 - c(1-\tau) + m}\left[C_0 + I_0 + G_0 - v\,r + x\,Y^* + \phi\,e\right]$$

Multiplicador: el de la economía abierta, $1/[1-c(1-\tau)+m]$, es menor que el de la cerrada, $1/[1-c(1-\tau)]$, porque a la filtración fiscal se suma la exterior, m, la parte del gasto adicional que se dirige a importaciones y no genera renta nacional. Aumenta con c y disminuye con τ y con m.

Pendiente: es negativa, $\partial r/\partial Y = -[1-c(1-\tau)+m]/v$, porque una bajada del tipo de interés estimula la inversión y, a través del multiplicador, la renta. Es más inclinada cuanto menores son v y el multiplicador, y el efecto filtración la hace más inclinada que la de la economía cerrada, de modo que una misma bajada del tipo eleva menos la renta. La desplazan a la derecha los aumentos de C_0, I_0, G_0 e Y^* y una depreciación real, y una subida de τ la hace rotar hacia dentro.

1.2.2. Mercado de dinero: curva LM

Definición: la **curva LM** reúne los pares (Y, i) que igualan la oferta real de dinero y la demanda de saldos reales, creciente en la renta, por k, y decreciente en el tipo de interés, por h:

$$\frac{M_0^s}{\bar{P}} = k\,Y - h\,i$$

– Despejando el tipo de interés, la LM es $i = (k\,Y - M_0^s/\bar{P})/h$.

Pendiente: es positiva, k/h, y la apertura no la altera, porque solo cambia el control de la oferta. Con tipo flexible, M^s es exógena, y con fijo tiene un componente endógeno, porque un superávit expande la base monetaria, al comprar divisas el banco central, y un déficit la contrae, salvo esterilización.

1.2.3. Sector exterior: curva BP

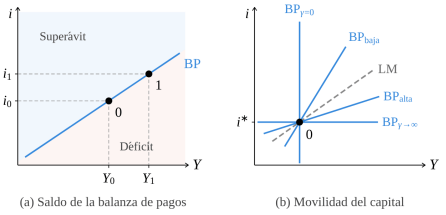
Definición: la **curva BP** reúne los pares (Y, i) compatibles con el equilibrio externo, un saldo nulo de la cuenta corriente más la cuenta financiera:

$$\text{BP} = (x\,Y^* - m\,Y + \phi\,e) + \gamma\,(i - i^*) = 0$$

– Despejando, la BP es $i = [m\,Y - x\,Y^* - \phi\,e]/\gamma + i^*$, con las entradas netas de capital crecientes en el diferencial según la movilidad, $\gamma \geq 0$.

Pendiente: es $m/\gamma \geq 0$. Como las importaciones crecen con la renta, solo un tipo de interés más alto, que atraiga capital, restablece el equilibrio. Por encima de la curva hay superávit, y por debajo, déficit.

Movilidad: la BP es vertical con movilidad nula. Con movilidad imperfecta tiene pendiente positiva, y la movilidad es alta si la BP es más plana que la LM, $m/\gamma < k/h$, y baja si es más inclinada. Con movilidad perfecta es horizontal en $i = i^*$. Una depreciación real o un aumento de Y^* la desplazan a la derecha, y una subida de i^*, hacia arriba. Las medidas sobre la cuenta corriente, como aranceles o depreciaciones, desplazan a la vez la BP y la IS. Una mayor movilidad aplana la BP, y una mayor propensión a importar la hace más rígida.



(a) Saldo de la balanza de pagos (b) Movilidad del capital

Curva BP: superávit por encima y déficit por debajo (a) y aplanamiento al crecer la movilidad del capital (b).

1.2.4. Equilibrio interno y externo

Equilibrio: las tres curvas se cortan y fijan Y, i y el tipo de cambio, si flota, o la oferta monetaria, si es fijo. Se cumplen a la vez el **equilibrio interno** y el externo. Si la IS y la LM se cortan por debajo de la BP, hay equilibrio interno con déficit.

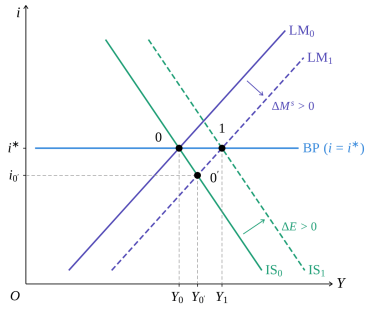
Dinámica: con tipo flexible, un déficit incipiente deprecia la moneda y desplaza la IS y la BP a la derecha hasta el equilibrio. Con tipo fijo, el ajuste recae en la oferta monetaria: el déficit se financia vendiendo reservas, lo que desplaza la LM a la izquierda, o con endeudamiento exterior, o se corrige con medidas de ajuste sobre la balanza comercial, que reducen la absorción o desvían el gasto, o sobre la cuenta financiera.

1.3. Políticas con precios fijos

Dos políticas, dos regímenes y cuatro grados de movilidad dan dieciséis casos. Los de referencia suponen movilidad perfecta, la más próxima a la realidad de las economías avanzadas.

1.3.1. Tipo de cambio flexible

Política monetaria expansiva: es plenamente **eficaz**, por la depreciación que induce. El aumento de M^s desplaza la LM y reduce el tipo de interés por debajo de i^* (punto 0). La salida de capital deprecia la moneda, la balanza comercial mejora y la IS se desplaza hasta que el tipo vuelve a i^*, en el punto 1. La inversión recupera su nivel y todo el aumento de la renta descansa en las exportaciones netas.



Mundell-Fleming con tipo flexible: la expansión monetaria deprecia la moneda y es plenamente eficaz.

Política fiscal expansiva: es ineficaz. El mayor gasto desplaza la IS y eleva el tipo por encima de i^*. La entrada de capital aprecia la moneda, la balanza comercial se deteriora y la IS retrocede por completo. La renta vuelve a su nivel, porque el gasto público desplaza íntegramente a las exportaciones netas, y el efecto expulsión exterior es pleno.

1.3.2. Tipo de cambio fijo

Los resultados se invierten, porque el ajuste de la balanza de pagos recae ahora sobre la oferta monetaria.

Política monetaria expansiva: es ineficaz. La bajada del tipo provoca una salida de capital. Para defender la paridad, el banco central vende reservas y contrae la oferta hasta que la LM vuelve a su posición. Solo se pierden reservas, y el banco central pierde el control de la oferta monetaria.

Política fiscal expansiva: es plenamente eficaz. La subida del tipo atrae capital. Para evitar la apreciación, el banco central compra divisas y expande la oferta, de modo que el tipo no sube, no hay expulsión de la inversión y la renta aumenta en todo el efecto multiplicador.

1.3.3. Movilidad imperfecta: los dieciséis casos

Saldo exterior: tras una expansión fiscal hay superávit si la BP es más plana que la LM, porque domina la entrada de capital, y déficit si es más inclinada, porque domina el deterioro comercial. Una expansión monetaria genera déficit con cualquier movilidad.

Tipo flexible: la política monetaria es siempre eficaz, y tanto más

cuanto mayor es la movilidad, por la depreciación que induce. La fiscal pierde eficacia con la movilidad. Con movilidad baja, el déficit comercial deprecia la moneda y refuerza el estímulo, y con alta, la entrada de capital la aprecia y lo compensa en parte.

Tipo fijo: la política monetaria es ineficaz con cualquier movilidad, salvo esterilización transitoria. La fiscal es eficaz si hay alguna movilidad, y más cuanto mayor es. Con alta movilidad, el superávit expande la oferta monetaria y refuerza el estímulo, y con baja, el déficit la contrae y lo compensa en parte. Con movilidad nula, la BP es vertical, la renta queda anclada por el equilibrio comercial y la contracción monetaria inducida anula el estímulo.

En suma, el régimen cambiario decide qué política es eficaz y la movilidad del capital, cuánto. Con perfecta movilidad, además, ambas políticas son alternativas y no complementarias.

1.3.4. País grande y trilema de Mundell-Fleming

País grande: aparecen efectos desbordamiento, positivos en la expansión fiscal, el *efecto locomotora*, y negativos en la monetaria con tipo flexible, el *empobrecimiento del vecino* por depreciación competitiva, que puede desembocar en guerras de divisas.

Trilema: es imposible lograr a la vez un tipo de cambio fijo, la perfecta movilidad del capital y una política monetaria independiente, y cada combinación renuncia al tercer vértice: Hong Kong renuncia a la política monetaria, Estados Unidos al tipo fijo y China, durante décadas, a la libre movilidad. Está en línea con la regla de Tinbergen, según el cual el número de objetivos no puede superar el de instrumentos independientes.

Su simplicidad lo hace muy útil pedagógicamente, y sigue siendo la referencia para razonar la elección del régimen cambiario, las crisis de balanza de pagos y la integración monetaria. Pero es estático e incompleto, está sujeto a la crítica de Lucas y prescinde de la oferta agregada y de las expectativas. La evidencia de panel respalda su predicción fiscal, aunque su magnitud depende de factores que ignora, como el nivel de desarrollo o de deuda pública.

2. El modelo OA-DA en economía abierta

Idea: el esquema de oferta y demanda agregadas levanta el supuesto de precios fijos y determina a la vez la renta y el nivel de precios, con una oferta que se fundamenta en la NMC o en la NEK1G.

2.1. Supuestos

Entorno: se mantienen los supuestos del IS-LM-BP salvo los precios fijos, con horizonte de medio plazo, precios y salarios flexibles y una movilidad del capital alta, pero no perfecta, de modo que el tipo de interés nacional puede separarse transitoriamente del exterior. Las empresas solo emplean *inputs* nacionales, lo que aísla el efecto del tipo de cambio sobre la oferta de trabajo. Coexisten dos índices de precios. El deflactor del PIB, P, solo recoge bienes nacionales. El IPC, P_c, incluye también los importados, sustitutivos imperfectos, con un peso $1-\sigma$, y su valor esperado rige la oferta de trabajo:

$$P_c = P^{\sigma}\,(E\,P^{*})^{1-\sigma}, \qquad \sigma \in (0,1)$$
$$P_c^{e} = (P^{e})^{\sigma}\,(E^{e}\,P^{*e})^{1-\sigma}$$

Agentes: las empresas contratan según el deflactor, porque venden bienes nacionales, y los trabajadores negocian y ofrecen trabajo según el IPC esperado, porque consumen también importados. Hay expectativas racionales con extracción de señales en la NMC e inercia nominal en la NEK1G.

2.2. Modelo

2.2.1. Demanda agregada

Definición: la **demanda agregada** reúne los pares (Y,P) que equilibran a la vez los mercados de bienes y de dinero. Se obtiene introduciendo la LM en la IS, con P endógeno, que entra por los saldos reales, el tipo de cambio real y el tipo de interés real:

$$Y^{d} = \mathrm{DA}(A_0^{+},\, M_0^{s+},\, P^{e+},\, \tau^{-},\, P^{-};\, Y^{*+},\, E^{+},\, P^{*+})$$

– El nivel de precios esperado la aumenta por el efecto Fisher, porque reduce el tipo real esperado para un tipo nominal dado.

Pendiente: es negativa por dos canales. En el del tipo de interés, o efecto Keynes, común a la economía cerrada, un mayor P reduce la oferta real de dinero, eleva el tipo de interés y contrae la inversión. En el del tipo de cambio real, propio de la abierta, un mayor P aprecia el tipo de cambio real y deteriora la balanza comercial, de modo que la DA abierta es más plana que la cerrada, y más cuanto menor es σ. La pendiente depende también del multiplicador, v, h y k.

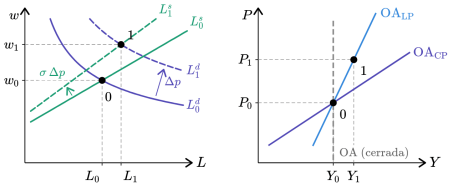
2.2.2. Oferta agregada de la NMC (Lucas, 1972-1975)

Definición: en economía cerrada, $y_t = \bar{y}_t + \zeta(p_t - p_t^{e}) + \varepsilon_t^{s}$: solo las sorpresas de precios afectan al *output*, y la OA de largo plazo es vertical. En la abierta, la oferta depende del deflactor relativo al IPC esperado:

$$Y^{s} = \mathrm{OA}(K^{+},\, P^{+},\, P^{e-};\, E^{e-},\, P^{*e-}) \;\Leftrightarrow\; Y^{s} = \mathrm{OA}(K^{+},\, (P/P_c^{e})^{+})$$

Razonamiento: ante una subida anticipada de los precios nacionales, la demanda de trabajo sube en Δp, pero el IPC esperado solo sube $\Delta p_c^{e} = \sigma\,\Delta p^{e} < \Delta p$, porque los importados mantienen su precio. La oferta de trabajo retrocede menos, y el empleo y la producción aumentan aun sin sorpresas. Las variables esperadas, en cambio, reducen la oferta, porque elevan el salario nominal que exigen los trabajadores.

– A largo plazo hay tantos niveles de producción como de tipo de cambio real, y la **OA de largo plazo no es vertical**.
– A corto plazo, por el problema de extracción de señales, la OA es más plana que la de largo plazo.



(a) Mercado de trabajo (b) Oferta agregada

Oferta agregada de la NMC en economía abierta: la demanda de trabajo sube en Δp y la oferta solo en $\sigma\,\Delta p$, y la OA de largo plazo tiene pendiente positiva.

2.2.3. Oferta agregada de la NEK1G (Layard y Nickell, 1985)

Definición: el salario real negociado, SRN, se fija sobre el IPC esperado, y el salario real pagado, SRP, sobre el deflactor, de donde se obtiene la oferta agregada:

$$\mathrm{SRN}:\; w/P_c^{e} = f(u^{-}, z^{+}) \qquad \mathrm{SRP}:\; w/P = A_L/(1+\mu)$$
$$Y^{s} = \mathrm{OA}(A_L^{+},\, P^{+},\, P^{e-},\, \mu^{-},\, z^{-};\, E^{e-},\, P^{*e-})$$

Razonamiento: la NEK incorpora poder de mercado, μ y z, y rigideces reales, mientras que la NMC supone competencia perfecta. A corto plazo, la inercia nominal da pendiente positiva, que la asimetría deflactor/IPC mantiene a largo. El SRP es horizontal, porque las empresas trasladan íntegramente el salario a los precios, y la subida de P reduce P_c^{e}/P y desplaza hacia abajo el SRN en términos del deflactor, de modo que ese salario compra más en términos del IPC y los trabajadores aceptan más empleo. A diferencia de la NMC, el margen y los factores institucionales del mercado de trabajo desplazan la oferta, lo que da cabida a las políticas de oferta.

2.2.4. Propiedades y equilibrio

Propiedades: son comunes a ambas escuelas.

1. Pendiente positiva a largo plazo, aunque más inclinada que a corto.
2. *Efecto filtración*: la apertura aplana la OA, y más cuanto mayor es $1-\sigma$.
3. La oferta se desplaza con K en la NMC y con A_L en la NEK, que canalizan las perturbaciones y las políticas de oferta.

Equilibrio: la intersección de la DA y la OA fija la renta y los precios y, con ellos, el tipo de interés, el empleo, el salario real y el tipo de cambio real. Con tipo fijo, la balanza de pagos altera la oferta monetaria y desplaza la DA. Con flexible, el tipo de cambio desplaza también la OA, a través del IPC esperado. Cada ejercicio exige seguir dos rondas, la del desplazamiento inicial de la curva afectada y la del inducido por el ajuste exterior.

2.3. Políticas monetaria y fiscal

2.3.1. Tipo de cambio flexible

Política monetaria expansiva: a corto plazo, la depreciación refuerza el estímulo, pero eleva el IPC esperado y desplaza la OA a la izquierda. A largo plazo, la expansión se traslada en el mismo porcentaje a los precios y al tipo de cambio nominal, y se reproduce la neutralidad de la economía cerrada. Ni el tipo de cambio real ni los saldos reales ni la renta varían a largo plazo.

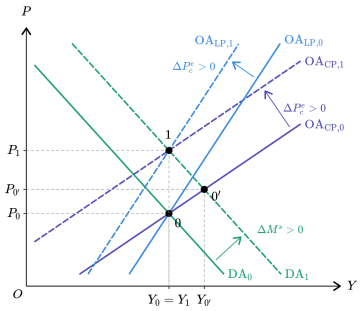
Política fiscal expansiva: con alta movilidad aprecia la moneda, pero la apreciación reduce el IPC esperado y desplaza la OA a la derecha. A largo plazo tiene efectos expansivos, mayores que en la economía cerrada. Frente al IS-LM-BP con movilidad imperfecta, no puede afirmarse *a priori* si es más o menos eficaz, e incluso con movilidad perfecta deja de ser ineficaz, porque el desplazamiento favorable de la oferta compensa en parte el efecto expulsión exterior.

Precios internacionales: ante su subida, la moneda se aprecia en la misma proporción, y el tipo de cambio real, el IPC y la renta no varían, de modo que la economía queda aislada de la inflación importada.

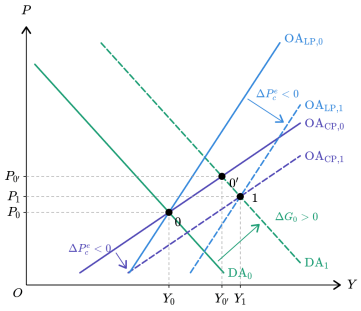
2.3.2. Tipo de cambio fijo

Política monetaria expansiva: sigue siendo ineficaz sobre la renta y los precios, en cualquier horizonte y para ambas escuelas.

Política fiscal expansiva: la DA se desplaza en dos rondas, por el



(a) Política monetaria expansiva



(b) Política fiscal expansiva con alta movilidad del capital

Tipo de cambio flexible en el modelo OA-DA de economía abierta.

estímulo y por la expansión monetaria endógena del superávit. El ajuste de la oferta difiere por escuela, pero en ambas la renta y el empleo aumentan también a largo plazo, con una eficacia mayor que en la economía cerrada y menor que en el IS-LM-BP, porque los precios más altos aprecian el tipo de cambio real y restan parte del estímulo.

Devaluación: desplaza la DA a la derecha, pero también las OA a la izquierda, porque encarece los importados, el IPC esperado y las demandas salariales. Sus efectos expansivos son transitorios, por la inercia nominal en la NEK1G y, en la NMC, solo si no es anticipada. A largo plazo, los precios suben en la misma medida, el tipo de cambio real y la renta vuelven a su nivel, y la devaluación **no tiene efectos duraderos**, salvo histéresis, frente a lo que concluía el IS-LM-BP con precios rígidos.

Precios internacionales: su subida equivale a una devaluación y a largo plazo se traslada por completo a los precios nacionales.

2.4. *Shocks* y políticas de oferta

Las perturbaciones de oferta adversas plantean un dilema, porque mueven la renta y los precios en direcciones opuestas.

2.4.1. *Shock* de precios energéticos

Planteamiento: con un factor importado, O, de precio real en moneda nacional $p_o = EP_o^*/P$, la empresa resuelve:

$$\max_{L,O} \Pi = F(K, L, O) - \frac{w}{P} L - \frac{E\,P_o^*}{P} O$$

CPO: cada factor se remunera según su productividad marginal, $F_L = w/P$ y $F_O = p_o$. Con factores complementarios, $F_{LO} > 0$, la demanda de cada uno cae con el precio real del otro. Ese **efecto cruzado** transmite el encarecimiento de la energía al empleo.

CSO: con F estrictamente cóncava en (L, O), las CPO son necesarias y suficientes para un máximo único.

Solución: las demandas de factores son:

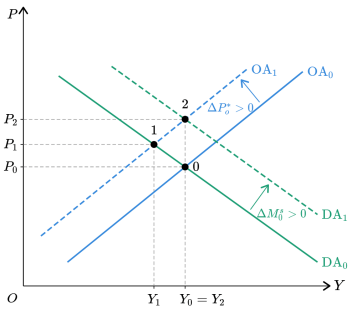
$$L^d = L(w/P^-,\, p_o^-), \qquad O^d = O(p_o^-,\, w/P^-)$$

Estática comparativa: una subida de P_o^*, o una depreciación que encarezca los *inputs* importados, reduce la demanda de energía, la productividad marginal del trabajo y el empleo. La OA se desplaza a la izquierda y hay **estanflación**. Con tipo fijo, la contracción monetaria endógena agrava la recesión. Con flexible, la depreciación encarece aún más los *inputs*. Acomodar con una expansión de la DA estabiliza la renta a corto plazo, pero consolida la subida de precios. No acomodar modera los precios a costa de una recesión mayor.

2.4.2. *Shocks* de productividad y salariales

***Shock* de productividad**: desplaza la OA a la derecha, sube la renta y bajan los precios, lo que deprecia el tipo de cambio real. Con tipo fijo, el superávit amplifica el efecto. Con flexible, la apreciación nominal absorbe parte de la ganancia de competitividad. La perturbación, un aumento permanente de A_L en la NEK, o de K o del progreso técnico en la NMC, no plantea dilema.

***Shock* salarial**: un aumento de z, canal que no existe en la NMC, desplaza el SRN hacia arriba y la OA a la izquierda, con estanflación. El dilema es acomodar, con la espiral precios-salarios, o no acomodar, con desempleo.



Shock energético: la OA se desplaza a la izquierda (de 0 a 1), y acomodarlo devuelve la renta a costa de más precios (de 1 a 2).

2.4.3. Políticas de oferta

Definición: las **políticas de oferta** desplazan la OA a la derecha, con más renta y menos precios de forma permanente. En la NEK1G son de tres tipos:

- *De mercado de trabajo*: reducen z, como la reforma de las prestaciones o la reducción de la cuña fiscal, y bajan la SRN y la NAIRU.
- *De competencia*: reducen μ con liberalización, apertura comercial o regulación de la competencia, y elevan el SRP.
- *De productividad*: elevan A_L o K con inversión en capital físico, humano y tecnológico e infraestructuras.

En economía abierta deprecian el tipo de cambio real, y con tipo fijo son la **devaluación interna**, sustituto de la nominal en una unión monetaria. A diferencia de esta, alteran la NAIRU y la renta de equilibrio.

En suma, el modelo determina precios y renta, analiza la oferta y distingue políticas anticipadas y no anticipadas, y su OA de largo plazo con pendiente positiva da efectos permanentes a la política fiscal y no a la devaluación. Pero es estático, determina el nivel de precios y no la inflación y trata la política monetaria como un control de la cantidad de dinero, y su oferta depende de la escuela.

3. El modelo IS-PC-MR en economía abierta

Idea: es la versión dinámica del análisis de oferta y demanda agregadas y el modelo canónico de la NEK, condensado por Clarida, Galí y Gertler (1999) y extendido a la economía pequeña y abierta por Galí y Monacelli (2005). Sustituye el nivel de precios por la inflación, la renta por la brecha de producción y la cantidad de dinero por una regla de tipos.

3.1. Supuestos

Entorno: competencia monopolística y rigideces nominales *à la* Calvo. En la versión abierta, economía pequeña, con tipo flexible como referencia y el fijo como caso límite, perfecta movilidad del capital, paridad descubierta de intereses y bienes nacionales y extranjeros sustitutivos imperfectos. Una fracción de la cesta de consumo es importada, y con Calvo solo una fracción de las empresas revisa su precio en cada periodo.

Agentes: expectativas racionales y una autoridad monetaria creíble, con un objetivo π^T, que fija el tipo de interés, de modo que la oferta monetaria es endógena.

Tecnología: la producción usa solo trabajo, y sus perturbaciones definen el *output* natural, $\bar{y}_t$, y el tipo de interés real natural, $\bar{r}_t$, los del equilibrio con precios flexibles.

3.2. Modelo en economía cerrada

Curva IS dinámica: procede de la ecuación de Euler. La brecha depende de la esperada y de la desviación del tipo real respecto del natural:

$$\tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(i_t - \mathbb{E}_t[\pi_{t+1}] - \bar{r}_t) + \varepsilon_t^d$$

- La sensibilidad, $1/\theta$, es la elasticidad de sustitución intertemporal del consumo, y ε_t^d recoge los *shocks* de demanda, como el gasto público o las preferencias.

Curva de Phillips de la NEK: con precios *à la* Calvo, la inflación depende de la esperada y de la brecha, con κ decreciente en la rigidez y ε_t^π los *shocks* de costes:

$$\pi_t = \beta\,\mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t + \varepsilon_t^\pi, \qquad \kappa > 0,\ \beta \in (0,1)$$

Regla de política monetaria: el banco central fija el tipo nominal según las desviaciones de la inflación y del *output*:

$$i_t = \bar{i}_t + \alpha_\pi\,\tilde{\pi}_t + \alpha_y\,\tilde{y}_t + \varepsilon_t^m, \qquad \alpha_\pi > 1$$

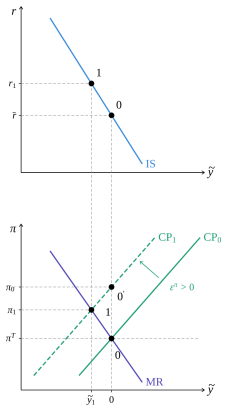
- El **principio de Taylor**, $\alpha_\pi > 1$, hace que el tipo real suba con la inflación y garantiza la estabilidad y unicidad del equilibrio (Blanchard-Kahn). Con $\alpha_\pi < 1$, la inflación se validaría a sí misma.

– Sustituida en la IS, la regla da una relación decreciente en $(\tilde{y}, \pi)$ que hace las veces de curva MR.
– En la regla, $\bar{i}_t = \bar{r}_t + \pi^T$ es el tipo nominal neutral, $\tilde{\pi}_t = \pi_t - \pi^T$ la desviación de la inflación respecto del objetivo, $\alpha_y \geq 0$, y ε_t^m recoge las desviaciones discrecionales.

Desarrollo gráfico: en los planos $(\tilde{y}, r)$ y $(\tilde{y}, \pi)$, la IS es decreciente en el tipo real, la curva de Phillips es creciente, con pendiente κ, y la MR es decreciente, porque el banco central responde a más inflación con un tipo real más alto. En el equilibrio, la curva de Phillips y la MR se cortan en $(0, \pi^T)$, y la IS asocia a esa brecha el tipo natural, $\bar{r}$.

Estática comparativa: sin perturbaciones, la economía está en su *output* natural con la inflación en el objetivo, $\tilde{y}_t = 0$, $\pi_t = \pi^T$ e $i_t = \bar{i}_t$, que es el equilibrio de precios flexibles.

– Un *shock* de demanda eleva la brecha y la inflación, tanto menos cuanto mayores son α_π y α_y, y el banco central los corrige sin conflicto entre objetivos.
– Un *shock* de productividad cambia el *output* y el tipo naturales. Acompañar el tipo natural estabiliza a la vez la inflación y la brecha (la **divina coincidencia**).
– Un *shock* de costes desplaza la curva de Phillips hacia arriba. Con brecha nula, la inflación subiría hasta $\pi_{0'}$, y contenerla exige subir el tipo real y abrir una brecha negativa, tanto mayor cuanto más pesa α_π respecto de α_y. La divina coincidencia se rompe, y esa disyuntiva hace de estos *shocks* el problema central de la estabilización.



Modelo IS-PC-MR: respuesta del banco central a un *shock* inflacionista.

3.3. Economía abierta (Galí y Monacelli, 2005)

3.3.1. Modificaciones al modelo base

Galí y Monacelli mantienen las tres ecuaciones, modifican cada una y añaden la paridad de intereses y el tipo de cambio natural. Con $p_{o,t}$ se denota el precio relativo de los importados, la inversa de la relación real de intercambio o, para los *inputs* energéticos, su precio real en moneda nacional, y con $\tilde{E}_t$, la desviación del tipo de cambio nominal respecto de su nivel natural.

Curva IS dinámica: entra el precio de los importados, y la sensibilidad al tipo, $1/\theta_\alpha$, depende de la apertura, porque los consumidores sustituyen entre bienes nacionales y extranjeros, no solo entre presente y futuro. El tipo natural depende del *output* exterior, de modo que la economía responde a las perturbaciones de productividad de sus socios:

$$\tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta_\alpha}(i_t - \mathbb{E}_t[\pi_{t+1}] - \bar{r}_t) + \Gamma_y\, p_{o,t} + \varepsilon_t^d$$

– El signo de Γ_y es indeterminado, porque el efecto sustitución traslada demanda a los bienes nacionales y es expansivo, y el efecto renta reduce la renta real y la demanda. Con bienes de consumo sustitutivos domina el primero, y con *inputs* energéticos, el segundo.

Curva de Phillips de la NEK: la inflación relevante es la del IPC, y el término $\Gamma_\pi p_{o,t}$, con $\Gamma_\pi > 0$, es el canal de la **inflación importada**: una depreciación o una subida de los precios exteriores eleva la inflación aunque la brecha no varíe. La pendiente, κ_α, depende también de la apertura:

$$\pi_t = \beta\, \mathbb{E}_t[\pi_{t+1}] + \kappa_\alpha\, \tilde{y}_t + \Gamma_\pi\, p_{o,t} + \varepsilon_t^\pi, \qquad \Gamma_\pi > 0, \qquad \beta \in (0,1)$$

Regla de Taylor ampliada: añade la desviación del tipo de cambio respecto de su nivel natural, $\alpha_E \tilde{E}_t$, con flotación pura si $\alpha_E = 0$, dirigida si $\alpha_E > 0$ y tipo fijo en el límite, $\alpha_E \to \infty$. Con flotación dirigida, el banco central añade un objetivo de estabilidad externa a los internos, y con tipo fijo renuncia a la regla sobre inflación y *output*. La regla es:

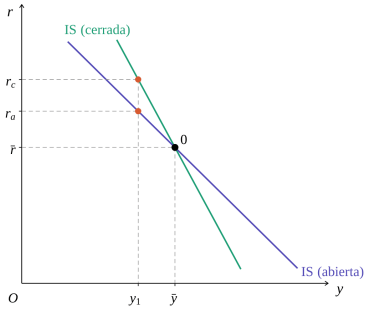
$$i_t = \bar{i}_t + \alpha_\pi\, \tilde{\pi}_t + \alpha_y\, \tilde{y}_t + \alpha_E\, \tilde{E}_t + \varepsilon_t^m, \qquad \alpha_E \geq 0$$

Paridad descubierta: $i_t - i_t^* = \mathbb{E}_t[\Delta s_{t+1}]$, con $s_t = \ln E_t$. Una subida del tipo aprecia de inmediato la moneda hasta que la depreciación esperada compensa el diferencial. El **tipo de cambio natural** es el compatible con el *output* y el tipo de interés naturales y con el equilibrio exterior, y la regla ampliada puede intentar corregir sus desviaciones, $\tilde{E}_t$.

3.3.2. Equilibrio y transmisión

Transmisión: sin perturbaciones, todas las variables, incluido el tipo de cambio, están en sus niveles naturales. Una subida del tipo contrae el consumo, el canal del tipo de interés, y aprecia la moneda, lo que reduce las exportaciones netas y abarata las importaciones, los canales del tipo de cambio sobre la demanda y sobre la inflación del IPC. La política monetaria es así más potente sobre la inflación, y su efecto se reparte entre el sector interior y el exterior.

Desarrollo gráfico: para llevar la producción de $\bar{y}$ a y_1, en economía cerrada el tipo debe subir hasta r_c. En la abierta basta $r_a < r_c$, porque la IS abierta, que recoge el efecto de la apreciación inducida sobre la demanda, es más plana.



IS-PC-MR en economía abierta: para cerrar la misma brecha basta mover menos el tipo de interés que en economía cerrada.

3.3.3. Perturbaciones

Demanda: con una expansión fiscal suben la brecha y la inflación, el banco central eleva el tipo, la moneda se aprecia y la demanda se contrae por los dos canales, mientras la apreciación modera la inflación del IPC. El efecto sobre el *output* es menor que en economía cerrada, el efecto expulsión de Mundell-Fleming, ahora con precios rígidos, pero no fijos. Resistir la apreciación da más estímulo y más inflación.

Energética, como la subida del gas en 2022: la inflación sube por Γ_π, y la menor demanda de energía reduce, si la energía es complementaria del trabajo, la demanda de trabajo y el *output* natural. La brecha se reduce por el efecto renta y, si los hogares sostienen su consumo, el deterioro exterior presiona hacia la depreciación. Subir el tipo contiene la inflación y sostiene la moneda, pero profundiza la caída del *output*, y no subirlo tolera la depreciación y la inflación importada. Si la perturbación es transitoria, se acomoda el efecto de primera ronda y se evitan los de segunda ronda sobre salarios.

Financiera (Rey, 2013): el ciclo financiero global provoca salidas de capital, presión hacia la depreciación y endurecimiento del crédito con cualquier régimen. Con libre movilidad, la política monetaria no es independiente ni con tipo flexible, porque los tipos a largo, las primas de riesgo y el crédito siguen el ciclo global y no el tipo oficial, de modo que el trilema se convierte en un **dilema**, aunque la flotación conserva una autonomía parcial. En el modelo, la perturbación entra como un *shock* a $\bar{r}_t$ y a $\tilde{E}_t$ que la regla de Taylor estándar no puede compensar sin desestabilizar la brecha.

3.3.4. Implicaciones de política

Régimen óptimo: bajo ciertas condiciones, es el objetivo de inflación interna con flotación pura, que replica la asignación de precios flexibles. El objetivo de inflación del IPC y el tipo fijo son inferiores en bienestar. Hay así una disyuntiva entre estabilización interna y externa, aunque la sobrerreacción de Dornbusch (1976) alimenta la tentación de un $\alpha_E > 0$. El régimen óptimo deja que el tipo de cambio absorba las perturbaciones, mientras que los otros estabilizan el tipo de cambio a costa de una volatilidad innecesaria de la brecha y de la inflación interna.

En suma, el modelo, en línea con la práctica de los objetivos de inflación, está microfundamentado, es en principio más robusto frente a la crítica de Lucas, sigue la trayectoria de las perturbaciones y muestra el tipo de cambio como canal de transmisión y como fuente de disyuntivas. Pero descansa en un agente representativo, sin capital, sector financiero ni primas de riesgo, su curva de Phillips explica mal la persistencia de la inflación y el límite inferior de los tipos exige extensiones.

A33. Teorías de la demanda de consumo corriente: ciclo vital y renta permanente. La demanda de bienes de consumo duradero. Evidencia empírica e implicaciones de política económica

1. Antecedentes de las teorías modernas

Los antecedentes de las teorías modernas del consumo son la aportación neoclásica, que trata el consumo como un residuo del ahorro, y la keynesiana.

1.1. Aportación neoclásica

Origen: Fisher (1930) proporcionó su microfundamentación con el modelo de elección intertemporal de dos periodos. Pigou (1943) introdujo el efecto saldos reales, y Modigliani (1944) y Patinkin (1956) sistematizaron la función de consumo agregada neoclásica, como contrapunto a la keynesiana.

Idea: el consumo es solo un gasto para satisfacer las necesidades corrientes y apenas un residuo del ahorro.

Razonamiento: hay dos formas de construir la función agregada.

- *Fisher*: el consumo real crece con la renta disponible y, si el efecto sustitución domina al efecto renta, decrece con el tipo de interés real, $C = f(Y^+, r^-)$. Se cumplen la dicotomía clásica y la neutralidad del dinero.
- *Pigou*: con el **efecto saldos reales**, el consumo depende también de la riqueza real, $C = f(Y^+, r^-, W^+)$, con $W = M/P + B/P$. Rompe la dicotomía clásica, pero no necesariamente la neutralidad, y busca dar eficacia al ajuste automático ante cambios de precios y salarios.

Extensiones: las políticas de demanda operan a través del tipo de interés real, cuya flexibilidad garantiza la ley de Say. La política fiscal expansiva eleva el tipo de interés y desplaza totalmente el gasto privado, con un *crowding out* total, lo que justifica la regla del equilibrio presupuestario y su excepción, la *regla de oro de las finanzas públicas*. Con efecto saldos reales, el tipo sube menos y el desplazamiento, que sigue siendo total, recae más sobre el consumo. La monetaria expansiva reduce el tipo real y estimula el consumo y la inversión hasta que la subida de precios reequilibra la economía, y con efecto saldos reales y un *stock* nominal de bonos que no crece como el dinero se rompe la neutralidad.

Fundamenta el consumo en la elección intertemporal e incorpora el tipo de interés real y la riqueza, que recogerán las teorías de la renta normal. Pero relega el consumo a un papel pasivo, confía en la flexibilidad de precios y tipos, por lo que no explica caídas persistentes de la demanda, y el efecto saldos reales es débil y exige deflaciones intensas, que elevan la carga real de las deudas.

1.2. Aportación keynesiana

La aportación keynesiana hace depender el consumo de la renta corriente, absoluta en Keynes y relativa en Duesenberry.

1.2.1. Renta corriente absoluta (Keynes, 1936)

Idea: Keynes elabora en la *Teoría general* (1936) la primera función de consumo agregada: el consumo depende de la renta corriente disponible a través de una propensión marginal a consumir estable y menor que la unidad, de la que se sigue el multiplicador.

Razonamiento: la producción tiene como destino final el consumo; las decisiones de consumo dependen de la renta corriente, y el tipo de interés pasa a segundo plano; y los planes de consumo son estables y se manifiestan en las propensiones.

- La **propensión marginal a consumir**, $c \equiv \partial C/\partial Y \in (0,1)$, es una *ley psicológica fundamental*: de cada unidad adicional de renta se gasta una parte y se ahorra otra. Es estable y fundamenta el multiplicador (Kahn, 1931; Keynes, 1936), $\partial Y/\partial A_0 = 1/(1-c)$.
- Keynes suponía que la propensión media, $C/Y = C_0/Y + c$, cae al crecer la renta, $d(C/Y)/dY = -C_0/Y^2 < 0$, porque veía en el ahorro un bien de lujo. La propensión marginal a ahorrar es $s \equiv \partial S/\partial Y = 1-c$, y el multiplicador solo exige que esta sea constante, no la tasa de ahorro.

Solución: la función de consumo es lineal y estable en la renta disponible, con un componente autónomo, el consumo de subsistencia, y otro endógeno:

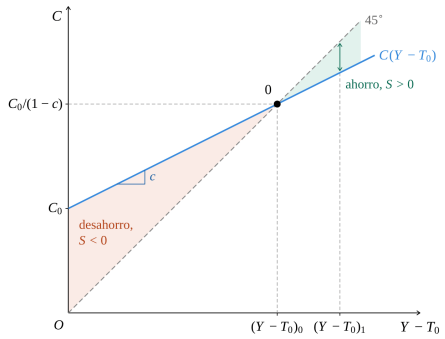
$$C = C_0 + c(Y - T_0), \quad C_0 > 0,\ c \in (0,1)$$

Desarrollo gráfico: la curva de consumo corta el eje vertical en C_0 y tiene pendiente c. Donde corta la bisectriz, el consumo agota la renta disponible y es $C_0/(1-c)$. A la derecha de ese punto los hogares ahorran, y a la izquierda desahorran o se endeudan.

Extensiones: el Gobierno debe tener un papel activo. La política monetaria es poco eficaz sobre la renta, por los *animal spirits* y la baja sensibilidad de la inversión al tipo de interés. La fiscal sí lo es, a través del multiplicador, y el gasto público es más eficaz que una rebaja de impuestos de la misma cuantía, porque parte de esta se ahorra:

$$\frac{\partial Y}{\partial G_0} = \frac{1}{1-c} > 1 \qquad \frac{\partial Y}{\partial T_0} = -\frac{c}{1-c} < 0$$

Las medidas redistributivas hacia los hogares de menor renta, con mayor propensión marginal, también estimulan el consumo.



Función de consumo keynesiana: a la derecha del corte con la bisectriz los hogares ahorran y a la izquierda desahorran.

Paradoja del ahorro (Keynes, 1936): un aumento del ahorro autónomo no eleva necesariamente el ahorro privado, porque la caída del consumo reduce, vía multiplicador, la producción y la renta, y así el ahorro puede obstaculizar el pleno empleo. Con inversión, gasto e impuestos exógenos, el equilibrio $I_0 = S + (T_0 - G_0)$ impide que varíe el ahorro privado, y la renta cae hasta que este vuelve a su nivel inicial:

$$\Delta Y = \frac{\Delta C_0}{1-c} < 0$$

Estancamiento secular (Hansen, 1938): si la propensión media al consumo cae al crecer la renta y las oportunidades de inversión se agotan, por el menor crecimiento de la población, el fin de la expansión territorial y un progreso técnico menos intensivo en capital, se llega a una depresión indefinida que solo evita la inversión pública. Contablemente, como $1 = C/Y + I/Y + G/Y$, si C/Y cae e I/Y no aumenta, G/Y debe aumentar para mantener el pleno empleo. Summers (2014) recuperó la idea tras la Gran Recesión.

Su sencillez le dio una acogida inmediata, y los presupuestos familiares y las series cortas mostraban una propensión media decreciente. Pero Kuznets (1946), con series largas de EE. UU. (1869-1938), halló una propensión media estable a largo plazo, $C/Y = k$, mayor que c. Tampoco se cumplió la predicción de Hansen. Tras la guerra, el consumo subió aunque la renta disponible cayó, porque los hogares liquidaron los bonos públicos acumulados durante el racionamiento, un efecto riqueza que mostró que la riqueza debía entrar en la teoría.

1.2.2. Renta corriente relativa (Duesenberry, 1949)

Idea: ante el **enigma de Kuznets**, James Duesenberry (1949) añade fundamentos psicológicos: cada hogar mide su consumo frente al de su entorno y frente al que él mismo tuvo, de modo que gobiernan el consumo su posición relativa y la renta más alta alcanzada.

Supuestos: dos efectos.

1. *Efecto demostración*: un hogar de renta relativamente baja procura no quedarse por debajo del consumo medio, aunque para ello se endeude o desahorre, lo que explica los datos de sección cruzada. Como el ahorro de cada hogar depende de su posición en la distribución, la tasa de ahorro agregada no aumenta al enriquecerse la economía mientras esa distribución no cambie.
2. *Efecto trinquete*: los hábitos hacen costoso renunciar a un nivel de vida ya alcanzado, lo que explica las series cortas. Con una renta que crece según su tendencia, ambos efectos frenan la reacción del consumo a corto plazo, que es menor que a largo.

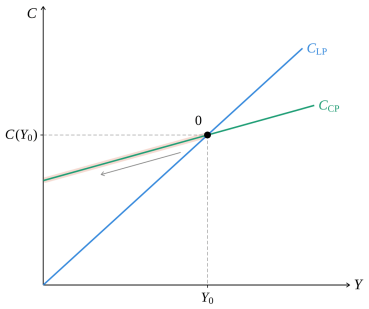
Desarrollo gráfico: la función de largo plazo pasa por el origen, con propensión media constante, y la de corto plazo es más plana y la corta en el punto 0, con la renta máxima alcanzada, Y_0. Si la renta cae, el consumo baja por la curva de corto plazo, desahorrando. En la recuperación vuelve al punto 0 y retoma el avance sobre la de largo plazo. La respuesta es asimétrica, porque el trinquete actúa en las caídas de renta y no en las subidas.

Extensiones: por el trinquete, la propensión media sube en las recesiones y amortigua la caída de la demanda.

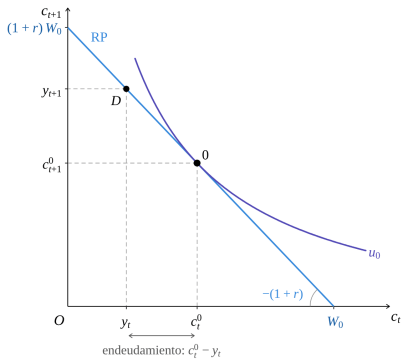
Aporta la suavización del consumo y la idea de una renta distinta de la del periodo, y reconcilia la propensión media decreciente a corto plazo con la constante a largo. Pero no explica con rigor cómo vuelve el consumo a su tendencia tras la recesión, ni incorpora dinámica. Además, su suavización mira hacia el pasado, por los hábitos, y no hacia el futuro.

En suma, Keynes liga el consumo a la renta corriente y Duesenberry a la relativa y pasada, cuando los hogares deciden mirando también a su renta futura.

2. El consumo como función intertemporal



Hipótesis de la renta relativa: función de corto plazo con efecto trinquete y función de largo plazo proporcional a la renta.



Consumo intertemporal en dos periodos: restricción presupuestaria y elección óptima.

En los años cincuenta surgen las **teorías de la renta normal**, que parten de Fisher (1930) y resuelven el enigma de Kuznets haciendo depender el consumo de la riqueza. Difieren en el horizonte, en la renta relevante y en el tratamiento de la incertidumbre.

2.1. Consumo intertemporal (Fisher, 1930)

2.1.1. Supuestos comunes

Entorno: mercados de capitales eficientes y completos, en los que se presta y se pide prestado al mismo tipo de interés, lo que hace posible suavizar el consumo.

Agentes: consumidores racionales y previsores (*forward looking*), que tienen en cuenta sus rentas futuras. Tienen una utilidad intertemporal aditivamente separable, con impaciencia y homotética, y una utilidad instantánea estrictamente cóncava, que expresa el deseo de suavizar el consumo.

2.1.2. Modelo

Planteamiento: el consumidor elige la senda que maximiza la utilidad intertemporal de Samuelson sujeto a su riqueza, los activos iniciales más el valor presente de las rentas:

$$\max_{\{c_t\}} \sum_{t=0}^{T} \beta^t u(c_t) \quad \text{s.a.:} \quad \sum_{t=0}^{T} \frac{c_t}{(1+r)^t} = A_{-1} + \sum_{t=0}^{T} \frac{y_t}{(1+r)^t} \equiv W_0, \qquad \beta \in (0,1)$$

CPO: la **ecuación de Euler** fija el perfil de la senda comparando lo que el mercado paga por esperar, r, con lo que cuesta esperar, ρ:

$$u'(c_t) = \frac{1+r}{1+\rho}\, u'(c_{t+1})$$

- Con $r = \rho$, el consumo es constante; con $r > \rho$, creciente, y con $r < \rho$, decreciente.
- En todo caso hay **suavización del consumo**: por la utilidad marginal decreciente y los mercados completos, el consumo no sigue las oscilaciones de la renta.

CSO: con utilidad estrictamente cóncava y restricción lineal, la ecuación de Euler y la restricción son necesarias y suficientes.

Solución: la restricción fija el nivel de la senda, y el consumo es una fracción de la riqueza, con una propensión κ que depende del tipo de interés, la impaciencia y el horizonte, pero no de la renta corriente. Con una distribución estable de la población por edades y de la riqueza, la función agregada es $C_t = \kappa W_t$ y, en componentes observables:

$$C_t = a\, A_{t-1} + b\, Y_t$$

- El componente predeterminado, aA_{t-1}, ocupa el lugar del consumo autónomo keynesiano, pero varía con la riqueza no humana. El inducido, bY_t, aproxima la riqueza humana, no observable, con b menor que la propensión marginal keynesiana. La hipótesis del ciclo vital (HCV) y la de la renta permanente (HRP) son dos formas de hacer operativa esta función, la primera acotando el horizonte al ciclo vital y la segunda definiendo la renta permanente como la anualidad de la riqueza.

Desarrollo gráfico: la restricción pasa por la dotación de rentas, con pendiente $-(1+r)$, y el óptimo es la tangencia con la curva de indiferencia más alta. Si queda a la derecha de la dotación, el consumidor se endeuda, y si queda a la izquierda, ahorra. La restricción corta los ejes en W_0 y $(1+r)\,W_0$, y la curva de indiferencia es convexa por la concavidad de la utilidad instantánea.

2.1.3. Conclusiones

El modelo da fundamento microeconómico al consumo, sustituye la renta corriente por la riqueza y explica con un solo mecanismo la evidencia de Kuznets. A corto plazo la renta fluctúa más que la riqueza y la propensión media cae en los auges. A largo plazo crecen a la par y la propensión media es constante. Pero supone certidumbre, mercados perfectos y preferencias homotéticas. Sin ellos, el consumo vuelve a depender de la renta corriente y tampoco vale la agregación. Las teorías siguientes los mantienen, salvo Hall, que introduce incertidumbre y utilidad cuadrática.

2.2. Ciclo vital (Modigliani y Brumberg, 1954)

Idea: la hipótesis del ciclo vital, de Modigliani y Brumberg (1954) y Ando y Modigliani (1963), pone el foco en el papel del ahorro en el plan de consumo óptimo.

2.2.1. Supuestos

Entorno: la vida es finita, de T periodos conocidos, sin herencias, y la renta laboral es predecible y sigue un perfil de joroba: baja en la juventud, elevada en la etapa central y nula en la jubilación.

2.2.2. Modelo

Planteamiento: es el problema anterior con horizonte igual a la vida.

CPO: por la ecuación de Euler, el consumo es constante o ligeramente creciente, y su perfil es independiente del de la renta.

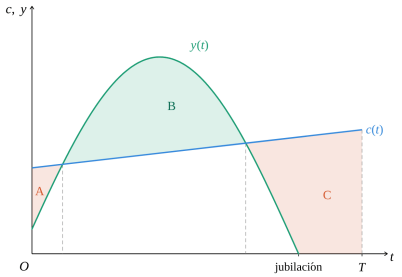
Solución: cada individuo consume una fracción de su riqueza vital, fracción que aumenta con la edad al acortarse el horizonte:

$$c^h_t = \kappa^h_t W^h_t, \qquad W^h_t = A^h_{t-1} + y^h_t + \sum_{j=1}^{T-t} \frac{y^{e,h}_{t+j}}{(1+r)^j}$$

Con una pirámide de edades y un reparto de la riqueza estables, la función agregada se expresa en propensiones medias, $C_t/Y_t = a(A_{t-1}/Y_t) + b$. Ando y Modigliani obtienen sus coeficientes aproximando la renta esperada media como proporcional a la corriente, $\bar{y}^{e,h}_t = \omega\, y^h_t$:

$$c^h_t = \kappa^h_t \left[A^h_{t-1} + [1 + \omega(T-t)]\, y^h_t\right] \equiv a^h A^h_{t-1} + b^h y^h_t$$

Desarrollo gráfico: para mantener un consumo regular, el individuo se endeuda en la juventud (área A), ahorra en los años centrales (área B) y consume lo acumulado en la jubilación (área C). La propensión media al consumo sigue así un patrón inverso al de la renta. Desciende en la etapa central de la vida y aumenta en la juventud y la vejez, y el consumo es ligeramente creciente porque se supone $r > \rho$.



Hipótesis del ciclo vital: endeudamiento (A), ahorro (B) y desahorro (C) a lo largo de la vida.

Estática comparativa: a corto plazo los activos apenas varían, de modo que la propensión media cae en los auges y sube en las recesiones, porque un aumento de la renta se reparte entre varios periodos. A largo plazo, la acumulación de activos hace que la propensión media y la marginal coincidan y sean estables.

Extensiones:

- *Política económica*: las políticas activas tienen efectos a corto plazo, pero a largo la economía tiene mayor capacidad de ajuste por sí misma.

2.2.3. Conclusiones

Es ampliamente aceptada, incorpora la demografía y explica la propensión media decreciente a corto plazo y en sección cruzada y constante a largo plazo. Pero ignora las restricciones de liquidez, la evidencia no

confirma el desahorro en la vejez y supone información perfecta sobre las rentas, sin rentas transitorias. De modo casi paralelo e independiente, Friedman desarrolla una teoría compatible, con un tratamiento más formal de la renta y de las expectativas.

2.3. Renta permanente (Friedman, 1957)

Idea: Friedman (1957) sostiene que los hogares consumen una fracción estable de su **renta permanente**, la anualidad de su riqueza total, y no de su renta corriente, cuyas desviaciones transitorias se ahorran.

2.3.1. Supuestos

Entorno: la renta y el consumo se dividen en un componente permanente y otro transitorio no correlacionados, $Y_t = Y_t^P + Y_t^T$ y $C_t = C_t^P + C_t^T$.

- La renta permanente es la que puede gastarse sin mermar la riqueza que la genera; la transitoria, la desviación aleatoria, como un premio de lotería o un periodo de desempleo. La riqueza suma la no humana, bonos, dinero y capital real, y la humana, el valor descontado de las rentas laborales esperadas:

$$W_t = \underbrace{B_t/r + M_t + K_t}_{A_t} + \underbrace{\sum_{j=0}^{\infty} \frac{y^e_{t+j}}{(1+r)^{j+1}}}_{W_H}$$

- El consumo transitorio no está correlacionado con la renta transitoria, de modo que la **propensión a consumir la renta transitoria es nula**: un aumento transitorio se ahorra y una caída se absorbe con desahorro.

Agentes: como la renta permanente no es observable, se estima con expectativas adaptativas, una media ponderada de las rentas pasadas, $Y_t^P = \lambda \sum_{j=0}^{J} (1-\lambda)^j Y_{t-j}$. Friedman estimó un horizonte de unos tres años, de modo que los cambios de la renta que persisten al menos ese tiempo se tratan como permanentes.

2.3.2. Modelo

Planteamiento: con horizonte indefinido y utilidad isoelástica (θ, aversión relativa al riesgo), la ecuación de Euler fija la tasa de crecimiento del consumo, $1+g_C = \left(\frac{1+r}{1+\rho}\right)^{1/\theta}$, con $g_C \approx (r-\rho)/\theta$, y la riqueza, humana y no humana, su nivel. Con $r = \rho$, el consumidor consume cada periodo la anualidad de su riqueza, $Y_t^P = rW_t$:

$$C_t^P = \kappa W_t = (\kappa/r)\, Y_t^P \equiv k_P\, Y_t^P$$

- Friedman postula los determinantes de k_P: decrece con el tipo de interés, coste de oportunidad de consumir, y con el peso de la riqueza humana, más insegura y peor aval, y depende de preferencias, demografía e incertidumbre, pero no del nivel de la renta permanente. Con z esos otros factores, la propensión es:

$$k_P = f(r^-, (W_H/W)^-, z^{\pm}) \in (0, 1]$$

Solución: al agregar, $C_t = k_P\, Y_t^P + C_t^T$, y la función depende del horizonte:

$$\text{Corto plazo:}\quad C_t = C_0 + b\, Y_t, \qquad b = k_P \lambda < k_P$$
$$\text{Largo plazo:}\quad C_t = k_P\, Y_t^P$$

A corto plazo, con expectativas adaptativas, solo se consume la parte del aumento de la renta que se incorpora a la permanente. A largo plazo, el consumo transitorio se cancela y la renta transitoria ahorrada pasa a ser permanente, con propensión media constante.

Estática comparativa: un *shock* transitorio, si se identifica como tal, no altera el consumo; uno permanente lo altera en la proporción k_P, gradualmente; la riqueza eleva el consumo, y una subida de r lo reduce por la caída de k_P y de la riqueza descontada.

Extensiones:

- *Política económica*: un impulso monetario transitorio no anticipado expande el consumo por el efecto riqueza, y uno permanente acaba siendo neutral. Un impulso fiscal transitorio no anticipado se ahorra si se percibe como tal, y uno permanente eleva el consumo. Que lo compense la anticipación de los impuestos futuros no es un resultado de Friedman, sino de la equivalencia ricardiana de Barro (1974).

2.3.3. Conclusiones

Es compatible con el ciclo vital y explica la evidencia de Kuznets y los presupuestos familiares. La propensión media de los hogares de renta baja es mayor que la de los de renta alta y que k_P, y es anticíclica, porque en el auge la renta transitoria se ahorra. Pero supone certidumbre, es internamente inconsistente, porque suaviza mirando adelante y estima la renta mirando atrás, las expectativas adaptativas pueden dar errores sistemáticos y la renta permanente no se observa.

2.4. Paseo aleatorio (Hall, 1978)

Idea: Robert Hall (1978) muestra que, con incertidumbre, expectativas racionales y utilidad cuadrática, la HCV o la HRP implican que el consumo sigue un **paseo aleatorio**.

2.4.1. Supuestos

Entorno: el individuo no conoce sus rentas futuras, con $r = \rho = 0$. Basta $r = \rho$ para el paseo aleatorio, y $r = \rho = 0$ solo simplifica la restricción.

Agentes: las expectativas son racionales. La utilidad es cuadrática, con utilidad marginal lineal, $u'(c_t) = 1 - \eta c_t$.

Restricciones: rige la condición de no-Ponzi.

2.4.2. Modelo

Planteamiento: con utilidad cuadrática, el consumidor maximiza la utilidad esperada de su senda sujeto a su riqueza esperada:

$$\max_{c_{t+j}} \mathbb{E}_t\left[\sum_{j=0}^{T-t}\left(c_{t+j} - \frac{\eta}{2} c_{t+j}^2\right)\right] \quad \text{s.a.:} \begin{cases} \sum_{j=0}^{T-t} \mathbb{E}_t[c_{t+j}] \le a_t + \sum_{j=0}^{T-t} \mathbb{E}_t[y_{t+j}] \\ a_{T+1} = 0 \end{cases}$$

CPO: la ecuación de Euler esperada, por la linealidad de la utilidad marginal, se reduce a $c_t = \mathbb{E}_t[c_{t+1}]$: para prever el consumo de mañana basta el de hoy. Si se esperase un aumento, convendría consumir más hoy hasta que la expectativa de cambio desapareciese. Con utilidad marginal lineal, la condición es:

$$1 - \eta c_t = \mathbb{E}_t[1 - \eta c_{t+1}] \;\to\; c_t = \mathbb{E}_t[c_{t+1}]$$

Solución: el consumidor reparte por igual su riqueza esperada entre los periodos que le quedan. Solo importa la esperanza de la renta, no su varianza, por la **equivalencia cierta**, y la incertidumbre no afecta al consumo. La función es la del ciclo vital con renta esperada en lugar de cierta:

$$c_t = \frac{1}{T-t+1}\left(a_t + \sum_{j=0}^{T-t} \mathbb{E}_t[y_{t+j}]\right)$$

Dinámica: con expectativas racionales, el consumo sigue una martingala:

$$c_{t+1} = c_t + \varepsilon_{t+1}, \quad \mathbb{E}_t[\varepsilon_{t+1}] = 0$$

Las variaciones del consumo son imprevisibles: solo la información nueva sobre la riqueza esperada lo mueve, y su efecto es permanente. Con $r \neq \rho$, el consumo sigue un AR(1), con coeficiente $(1+\rho)/(1+r)$:

$$\mathbb{E}_t[c_{t+1}] = \frac{r-\rho}{\eta(1+r)} + \frac{1+\rho}{1+r}\, c_t$$

Estática comparativa: una innovación en la riqueza esperada altera el consumo en $1/(T-t+1)$ de su cuantía, de una vez y para siempre. Una variación anticipada de la renta no lo altera.

Extensiones:

- *Política económica*: solo es eficaz si no es anticipada. Una rebaja fiscal anunciada no altera el consumo cuando se hace efectiva, porque ya se incorporó en el anuncio. Solo una medida sorpresa desplaza el consumo, y de forma permanente.

2.4.3. Conclusiones

Cierra la inconsistencia de la HRP y ofrece una restricción contrastable: ninguna variable retardada debe predecir Δc_t. En el contraste de Hall, la renta retardada no lo hace, aunque las cotizaciones sí, ligeramente. Pero la utilidad cuadrática tiene una aversión absoluta al riesgo creciente, la equivalencia cierta excluye el ahorro precautorio, el agente representativo ignora la heterogeneidad y la evidencia rechaza el paseo aleatorio en dos direcciones.

En suma, las teorías de la renta normal convergen en la suavización del consumo y difieren en la renta relevante, el horizonte y las expectativas, y con ellas en la eficacia de la política fiscal.

3. Desarrollos recientes y bienes de consumo duradero

La literatura posterior relaja los supuestos de las teorías de la renta normal para explicar el exceso de sensibilidad y de suavidad, las decisiones de consumo y ocio, los hábitos o la demanda de bienes duraderos.

3.1. Exceso de sensibilidad y de suavidad

3.1.1. Exceso de sensibilidad (Flavin, 1981)

Idea: el consumo responde a cambios anticipados de la renta, como la devolución previsible de un impuesto, que no deberían alterarlo. La hipótesis del paseo aleatorio (HPA) implica que Δc_t es ortogonal a toda la información disponible en $t-1$. Flavin (1981) modela la renta como un proceso ARMA, separa su innovación de su componente anticipado y halla que este tiene un coeficiente positivo y significativo, cuando debería ser nulo. En su ecuación, α mide la sobrerreacción del consumo a la renta corriente:

$$\Delta c_t = \alpha\, \Delta y_t + v_t + \phi\, \varepsilon_t$$

Explicaciones: restricciones de liquidez; reglas simples, pues para Campbell y Mankiw (1989) los consumidores que gastan su renta corriente perciben en torno a la mitad de la renta agregada; ahorro precautorio, y bienes duraderos. Si el consumo responde a la renta corriente, la política fiscal anticipada recupera eficacia.

Es un contraste estructural que orienta los desarrollos posteriores, pero descansa en series agregadas y en un modelo concreto de la renta, y documenta la anomalía sin discriminar entre sus explicaciones, para lo que hace falta evidencia microeconómica.

3.1.2. Exceso de suavidad (Deaton, 1987)

Idea: el consumo reacciona menos de lo debido a las innovaciones de la renta. Si la renta es muy persistente, con raíz unitaria y crecimiento positivamente autocorrelacionado, una sorpresa eleva más la renta permanente que la corriente, y el consumo debería variar más que la renta, pero la varianza de Δc_t es muy inferior (la **paradoja de Deaton**).

Explicaciones: los hogares no distinguen bien lo permanente de lo transitorio, están restringidos o son prudentes, o tienen hábitos u ocio en sus preferencias. Campbell y Deaton (1989) muestran que ambos fenómenos son la misma anomalía, porque si el consumo reacciona a la información pasada, reacciona menos a la nueva, lo que devuelve espacio a las políticas estabilizadoras.

Muestra que la HPA no implica por sí misma un consumo suave y hace de la persistencia de la renta el centro del contraste. Pero la paradoja depende del proceso supuesto, porque con series cortas es difícil distinguir una raíz unitaria de un proceso estacionario en torno a una tendencia, y en este caso la HPA sí predice un consumo suave.

3.2. Entorno y agentes

3.2.1. Equivalencia ricardiana (Barro, 1974)

Idea: Barro recupera la proposición de Ricardo en un modelo de generaciones solapadas: con horizonte infinito, o finito con altruismo intergeneracional, los hogares internalizan la restricción del sector público y anticipan que la deuda exige impuestos futuros de igual valor actual y no la consideran riqueza neta, de modo que una rebaja fiscal financiada con deuda se ahorra íntegramente, y la forma de financiar una senda de gasto dada es irrelevante para el consumo. Obliga a tratar la deuda como impuestos diferidos, pero exige supuestos fuertes, el altruismo operativo entre generaciones, los mercados de capitales perfectos y los impuestos de cuantía fija, y basta que algunos hogares tengan restricciones de liquidez para que deje de cumplirse.

3.2.2. Restricciones de liquidez (Deaton, 1991)

Idea: en el mundo real, los mercados financieros son imperfectos. Las restricciones elevan el ahorro de dos maneras: cuando se activan, la ecuación de Euler se cumple con desigualdad, $u'(c_t) > u'(c_{t+1})(1+r)/(1+\rho)$, y el consumo queda por debajo del deseado; y la mera posibilidad de que se activen en el futuro deprime el consumo actual. Explican el exceso de sensibilidad, pero se quedan cortas para el ahorro observado y suelen combinarse con el motivo precautorio.

3.2.3. Racionalidad limitada

Idea: Simon (1957), Kahneman y Tversky (1979) y Thaler y Sunstein (2008) cuestionan el *homo economicus*. El diseño del contexto influye en la decisión, el *nudge*, por sesgos como el efecto marco, y los consumidores valoran más lo que poseen, el efecto dotación. Con el **descuento cuasihiperbólico** (Laibson, 1997), un factor $b < 1$ sobre todo el futuro hace inconsistentes los planes, lleva a demandar activos ilíquidos como compromiso y liga el consumo a la renta corriente, y con las **cuentas mentales** (Shefrin y Thaler, 1988) la propensión al consumo depende de la cuenta (renta corriente, activos o renta futura). Pueden explicar ambas anomalías, pero son un conjunto de sesgos más que un modelo unificado: sus predicciones dependen del sesgo elegido, son difíciles de distinguir de las de agentes racionales con restricciones y su traslado al consumo agregado no es inmediato.

3.3. Función de utilidad

3.3.1. Consumo-ocio (Lucas y Rapping, 1969)

Idea: la utilidad depende del consumo y del ocio, y la renta deja de ser exógena. Con utilidad separable, el consumo sigue un paseo aleatorio. Con utilidad no separable, el salario real entra en la ecuación de Euler y consumo, empleo y salario se mueven juntos, lo que ayuda a explicar ambas anomalías. Pero el signo y la cuantía de la no separabilidad son difíciles de estimar, y con desempleo involuntario la covariación entre consumo y empleo refleja una restricción y no las preferencias.

3.3.2. Hábitos (Deaton, 1992)

Idea: retoma el efecto trinquete de Duesenberry: la utilidad depende del consumo relativo a un nivel de referencia formado por el pasado, $u(c_t - \gamma\, c_{t-1})$, y ante una sorpresa el consumo se ajusta gradualmente, $\Delta c_t = \gamma\, \Delta c_{t-1} + \varepsilon_t$. La respuesta inicial pequeña y el ajuste predecible explican a la vez el exceso de suavidad y el de sensibilidad. Pero la referencia y la intensidad del hábito se postulan sin fundamento propio, y con datos de hogares los hábitos son más débiles que con datos agregados, en los que la persistencia puede deberse en parte a la agregación.

3.3.3. Ahorro precautorio (Leland, 1968)

Idea: los hogares ahorran también ante la incertidumbre de sus rentas. Con utilidad cuadrática no hay motivo precautorio. Si la utilidad marginal es convexa, $u''' > 0$, por la desigualdad de Jensen la incertidumbre eleva la utilidad marginal esperada futura y la ecuación de Euler solo se restablece con menos consumo hoy. La intensidad la mide la **prudencia**, que refleja la curvatura de la utilidad marginal, mientras que la aversión al riesgo refleja la de la utilidad:

$$\text{CPA} = -\frac{u'''(c_t)}{u''(c_t)} \qquad \text{CPR} = -\frac{u'''(c_t)}{u''(c_t)} \cdot c_t$$

En el modelo de colchón de Carroll (1997), el consumidor impaciente mantiene una riqueza reducida y su consumo sigue de cerca a la renta. El motivo precautorio solo domina a la impaciencia cuando la riqueza cae por debajo de su objetivo.

Explica que el consumo responda a la incertidumbre y no solo a la renta esperada. Pero su cuantía depende de una prudencia y una incertidumbre difíciles de medir, y sus predicciones se confunden con las de las restricciones de liquidez, que también ligan el consumo a la renta corriente.

3.4. Bienes de consumo duradero

Idea: Mankiw (1982) y Bernanke (1985) extienden el análisis a los duraderos, que dan utilidad mientras duran, no cuando se compran.

Planteamiento: el consumidor vive dos periodos y consume un bien no duradero, c_t, y los servicios, $s(d_t)$, de un *stock* de duraderos que se deprecia a la tasa δ y se repone con el gasto $x_t = d_t - (1-\delta)\, d_{t-1}$:

$$\max_{c_1, c_2, d_1, d_2} \; u[c_1, s(d_1)] + \beta\, u[c_2, s(d_2)]$$

$$\text{s.a.: } c_1 + x_1 + \frac{c_2 + x_2}{1+r} = W_0$$

CPO: la del duradero iguala la utilidad marginal de sus servicios a su **coste de uso**, el interés perdido más la depreciación, porque comprar una unidad más hoy cuesta una de consumo y devuelve $1-\delta$ mañana:

$$u_s(1)\, s'(d_1) = u_c(1)\, \frac{r+\delta}{1+r}$$

Con un precio relativo del duradero p_d y una revalorización esperada π_d, el coste de uso es $p_d\,(r + \delta - \pi_d)$, análogo al del capital de la empresa. La suavización se aplica así al *stock*, no al gasto, que es mucho más volátil que el de no duraderos y sensible al tipo de interés y al crédito. En Mankiw, con utilidad separable y cuadrática, el *stock* sigue un paseo aleatorio y el gasto un proceso ARMA(1,1), cuya primera diferencia es de medias móviles:

$$\Delta x_t = \varepsilon_t - (1-\delta)\, \varepsilon_{t-1}$$

En Bernanke, con costes de ajuste, el *stock* se ajusta parcialmente, y con costes fijos sigue una regla (S, s): el consumidor deja que se deprecie hasta un umbral y entonces lo repone de una vez. Con duraderos, el gasto no sigue un paseo aleatorio ni un AR(1), sea cual sea la función de utilidad.

Los duraderos ofrecen una explicación parcial del exceso de sensibilidad, pero la evidencia rechaza ambas versiones: en Mankiw, el coeficiente estimado implica $\delta \approx 1$, como si fueran no duraderos, y en Bernanke el gasto responde a la renta corriente más de lo previsto.

3.5. Evidencia y política económica

Balance empírico: a largo plazo, la propensión media es constante, lo que descarta la función keynesiana como relación estructural. A corto plazo, la ortogonalidad de Hall se cumple solo de forma aproximada. La evidencia microeconómica sobre transferencias fiscales confirma propensiones marginales elevadas y muy heterogéneas. El consumo responde a la renta anticipada y poco a las sorpresas, y las fricciones de los desarrollos recientes explican ambas anomalías.

Implicaciones: cuanto más fuertes son las fricciones, más depende el consumo de la renta corriente y más eficaces son las políticas, anticipadas o no.

- *Política monetaria*: actúa por el tipo de interés real, que inclina la senda de consumo, y por el efecto riqueza. Los hogares restringidos responden a través de la renta corriente.
- *Política fiscal*: con hogares previsores, una rebaja transitoria apenas altera el consumo, una anunciada no lo altera al hacerse efectiva y una financiada con deuda tampoco, si hay equivalencia ricardiana. Con restricciones de liquidez o reglas simples, las transferencias temporales a los hogares de menor renta recuperan eficacia.

En suma, una fracción de los hogares consume su renta corriente, lo que devuelve a la política fiscal parte de la eficacia que le negaban las teorías intertemporales y fundamenta los modelos TANK y HANK de la macroeconomía actual.

1. Fundamentos de la teoría moderna

La **inversión real** es el gasto en bienes de capital, que, a diferencia de los servicios de factores que se alquilan, generan un flujo de servicios productivos a lo largo del tiempo. Fisher y Keynes ponen el énfasis en la conducta de los agentes, y son más bien teorías de la demanda de capital; el acelerador, en la demanda agregada, y Hayek, en la inversión como fenómeno de desequilibrio. La inversión bruta suma la neta, la variación del *stock* de capital, y la de reposición:

$$I^b = I^n + \delta K, \qquad I^n = \dot{K}$$

1.1. Aportación neoclásica (Fisher, 1930)

Idea: Fisher (1930) es uno de los primeros en microfundamentar la inversión. Quien invierte compra un derecho sobre un flujo de retornos, y la **tasa de retorno sobre el coste**, la actual tasa interna de rentabilidad, debe igualar el tipo de interés real, con independencia de las preferencias del inversor.

Supuestos: dos periodos y mercados financieros perfectos, con un único tipo de interés real, r; un consumidor-inversor racional, y una tecnología con rendimientos decrecientes, en la que el capital k invertido rinde $f(k)$ y se deprecia a la tasa δ.

Planteamiento: como k solo entra a través de la riqueza, el problema se separa en dos etapas, por el **teorema de separación de Fisher**: primero, la inversión maximiza la riqueza; después, el consumo se elige sobre esa riqueza, que con rentas y_1 e y_2 es:

$$W(k) = y_1 - k + \frac{y_2 + f(k) + (1-\delta)k}{1+r}$$

CPO: se invierte hasta que el rendimiento marginal del capital neto de depreciación iguala el tipo de interés real:

$$f'(k) - \delta = r$$

- La condición no depende de las preferencias, lo que justifica separar la teoría de la inversión de la del consumo, y, como $f'' < 0$, es suficiente.
- El capital deseado, $k^* = k(r)$, decrece con el tipo de interés, y agregando se obtiene la función de inversión neoclásica, $I = I(r)$, con $I'(r) < 0$. Una subida de r exige una productividad marginal mayor y, por los rendimientos decrecientes, menos capital.

Fondos prestables: la inversión es la demanda de fondos y el ahorro, creciente en r, la oferta. El tipo de interés real los iguala y garantiza la ley de Say. Una mejora tecnológica eleva la inversión y el tipo, y un aumento del ahorro reduce el tipo y eleva la inversión. Con sector público, a la demanda de fondos se suma el déficit.

Extensiones:

- *Política económica*: la expansión monetaria crea un exceso de oferta de fondos prestables y reduce transitoriamente el tipo, pero los precios suben hasta restablecer el equilibrio y es neutral a largo plazo. La fiscal eleva el tipo de interés real y expulsa totalmente al consumo y la inversión privados.

Su mérito es microfundamentar la inversión y separarla de las preferencias. Pero se limita a dos periodos en su representación gráfica y supone pleno empleo, con un tipo de interés que iguala el ahorro y la inversión deseados, supuesto que Keynes cuestiona tras la Gran Depresión.

1.2. Aportación keynesiana (Keynes, 1936)

Idea: Keynes acepta la tasa de Fisher, que llama **eficiencia marginal del capital** (EMK), pero con rendimientos esperados bajo una incertidumbre no reducible a probabilidades, que dependen del estado de la confianza. Lo decisivo no es la productividad física, sino las expectativas, y la inversión, movida por ellas, determina el ahorro y no al revés. Extiende así la "inversión de la ley de Say" del consumo al ahorro y la inversión.

Supuestos: desempleo e incertidumbre, de modo que la inversión efectiva puede diferir de la planeada, con variación involuntaria de existencias; empresarios dominados por los *animal spirits*, y rendimientos decrecientes, que permiten ordenar los proyectos por su rentabilidad.

EMK: es el tipo de descuento que iguala la suma actualizada de los rendimientos esperados de un proyecto con su coste de reposición, el precio de oferta del bien:

$$C_K = \sum_{j=1}^{T} \frac{Q_j^e}{(1+\text{EMK})^j}$$

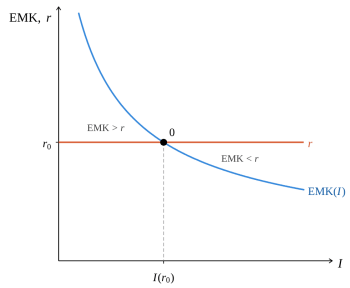
Regla: se invierte si EMK $> r$, es decir, si lo que se espera obtener supera lo que se deja de ganar en la alternativa. La función agregada, en términos reales, es:

$$I = I_0 + I(r), \qquad I'(r) < 0$$

- El componente exógeno, I_0, refleja los *animal spirits* y es grande, mientras que la sensibilidad al tipo de interés es baja, $I'(r) \approx 0$. En el agregado, la no reposición del capital depreciado y la desacumulación de existencias pueden hacer negativa la inversión neta. La función resulta de una decisión discreta sobre cada proyecto p, de coste C_p, cuyos escalones se funden en la curva de EMK cuando hay un continuo de proyectos:

$$I_p = \begin{cases} C_p & \text{si EMK}_p > r \\ 0 & \text{si EMK}_p \le r \end{cases}$$

Desarrollo gráfico: la curva de EMK ordena los proyectos de mayor a menor rentabilidad y es decreciente, porque el rendimiento esperado cae al aumentar la oferta de capital y la presión sobre las industrias que lo producen eleva su precio, a lo que se suma el riesgo creciente del endeudamiento (Kalecki, 1937). El tipo de mercado no depende de la inversión y es una recta horizontal, y la inversión llega hasta el punto en que la EMK lo iguala.



Eficiencia marginal del capital: se invierte hasta el proyecto cuya EMK iguala el tipo de interés.

Estática comparativa: una subida del tipo reduce poco la inversión, mientras que dominan los desplazamientos de la curva de EMK. Una mejora de las expectativas o de la tecnología la desplaza hacia la derecha y eleva la inversión para cualquier tipo de interés. Movida por los *animal spirits*, la inversión es errática e inestable, y con ella la demanda agregada.

Extensiones:

- *Política económica*: la baja sensibilidad al tipo de interés es el *primer eslabón débil*. Con la trampa de liquidez, explica la ineficacia de la política monetaria, y con la ilusión monetaria, que da rigidez salarial a la baja, explica también que la economía no vuelva sola al pleno empleo. La monetaria es poco eficaz sobre la renta y el empleo, porque solo actúa a través de una inversión poco sensible al tipo. La fiscal es muy eficaz sobre la renta y el empleo y apenas reduce la inversión, con una expulsión parcial.

Introduce las expectativas como determinante esencial. Pero explica cuánto capital se desea, no a qué ritmo se invierte. Los *animal spirits* no son observables, lo que dificulta estimar la EMK, y la subida del precio de oferta pierde fuerza si hay capacidad ociosa en las industrias que producen los bienes de capital, aunque para Keynes era la causa principal del descenso de la EMK a corto plazo.

1.3. El principio del acelerador

Idea: tras la *Teoría general*, el acelerador vuelve al primer plano para explicar a qué velocidad se acerca el capital a su nivel óptimo. Cuando el gasto supera lo producido, las empresas producen más en lugar de subir precios, y producir más exige más equipos. Su regreso se debió a que la parte endógena de la función keynesiana apenas explicaba la inversión, y sus versiones se apoyaban más en el ajuste a los datos que en la microeconomía.

1.3.1. Acelerador simple (J. M. Clark, 1917)

Supuestos: formulado por John Maurice Clark (1917) e incorporado al análisis del ciclo por Harrod (1936) y Samuelson (1939), supone que el capital efectivo coincide siempre con el deseado, que los agentes son miopes y que la tecnología es de coeficientes fijos, con $K_t^* = vY_t$, sin depreciación. La inversión neta coincide, así, con la bruta, y los agentes miopes esperan los niveles corrientes, $Y_t^e = Y_t$.

Solución: la inversión neta no depende del nivel de la demanda, sino de su variación:

$$I_t = v \cdot \Delta Y_t$$

- Un aumento permanente de Y da inversión solo en el periodo del cambio; uno transitorio, inversión y luego desinversión; un crecimiento constante, inversión constante, que solo cambia si Y se acelera, $\Delta I_t = v\,\Delta^2 Y_t$. Basta con que la producción crezca más despacio para que caiga la inversión (de ahí el nombre). El efecto es mayor cuanto mayor es v.

Multiplicador-acelerador (Samuelson, 1939): con consumo retardado e inversión inducida por la variación del consumo, la renta sigue una ecuación en diferencias:

$$\begin{aligned} C_t &= c(1-\tau)\, Y_{t-1} \\ I_t &= v\,(C_t - C_{t-1}) \\ Y_t &= c(1-\tau)(1+v)\, Y_{t-1} - c(1-\tau)\, v\, Y_{t-2} + G_0 \end{aligned}$$

A largo plazo, el efecto de G_0 es el multiplicador keynesiano, pero el acelerador amplía la respuesta transitoria, que puede oscilar, y la senda converge solo si $c(1-\tau)\,v < 1$. Solo si la inversión inducida depende del nivel de renta, $I = bY$, el *supermultiplicador*, $1/[1-c(1-\tau)-b]$, es permanentemente mayor, siempre que $c(1-\tau)+b<1$.

Extensiones:

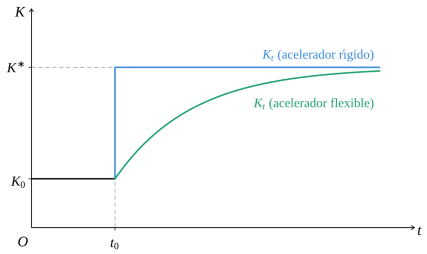
- *Política económica*: la monetaria pierde aún más eficacia, porque el tipo de interés no entra en la función. La fiscal mueve a corto plazo la renta y la inversión con más fuerza que en Keynes, sin expulsión, y de forma permanente con el supermultiplicador.

Explica bien la inversión en existencias. Pero supone que la empresa obtiene todo el equipo que quiere sin alterar su precio, un ajuste inmediato, agentes miopes y coeficientes fijos. Si el trabajo puede sustituir al capital, las empresas pueden responder a la demanda cambiando el uso de sus factores, sin invertir.

1.3.2. Acelerador flexible (Goodwin, 1951; Koyck, 1954)

Idea: Goodwin (1951), Chenery (1952) y Koyck (1954) introducen retardos: el capital cierra en cada periodo la fracción $1-\gamma$ de la brecha con el deseado, y la inversión responde a las variaciones pasadas de la demanda, con pesos decrecientes:

$$I_t = v\,(1-\gamma)\sum_{j=0}^{\infty}\gamma^j \Delta Y_{t-j} \quad \Leftrightarrow \quad K_t - K_{t-1} = (1-\gamma)\,(K_t^* - K_{t-1})$$



Acelerador flexible: el capital se ajusta gradualmente hacia su nivel deseado, frente al salto del acelerador rígido.

Un aumento permanente de la producción eleva el capital gradualmente: la inversión alcanza su máximo en el periodo del cambio y decrece después. Cuanto mayor es γ, más lento es el ajuste. La política monetaria sigue sin eficacia sobre la inversión, y la fiscal pierde algo de fuerza respecto al acelerador simple, porque la inversión responde de forma escalonada. Obtuvo buenos resultados empíricos, aunque la correlación no prueba causalidad. Pero ignora el tipo de interés, introduce el retardo sin explicar qué costes lo originan y nadie optimiza, por lo que es una regla de ajuste más que una teoría.

Conclusiones: no hay motivo para una ratio capital-output fija con independencia de los precios de los factores, y, como recuerda Haavelmo (1960), la demanda de inversión no se deduce sin más de la de capital: hay que explicar cuánto capital se desea, a qué ritmo se alcanza y cómo se repone.

1.4. Perspectiva austriaca (Hayek, 1941)

Idea: Hayek (1941) considera la inversión un fenómeno de desequilibrio, y la decisión de inversión, una decisión sobre la velocidad óptima de ajuste. Decidir que se necesita una fábrica es distinto de decidir a qué ritmo construirla. Por el orden espontáneo, la intervención del Estado debía ser mínima, para dejar actuar las velocidades óptimas de ajuste de los empresarios.

Anticipa los costes de ajuste, pero sin formalismo matemático, y los desarrollos posteriores eclipsaron su aportación.

En suma, ninguna de estas aportaciones une una decisión optimizadora y un ritmo de ajuste formalizado, ni responde a la vez a las tres preguntas de Haavelmo.

2. Teorías modernas de la inversión

Desde los años sesenta, la inversión se plantea como un problema de optimización dinámica (la maximización del valor descontado de los beneficios de la empresa).

2.1. Sin costes de ajuste (Jorgenson, 1963)

Idea: en el modelo de Dale Jorgenson (1963; con Hall, 1967), la empresa desea el capital que iguala su productividad marginal a su **coste de uso**, y la inversión es el ajuste hacia él. Así, los precios relativos de los factores, y no solo la demanda, determinan la inversión.

2.1.1. Supuestos

Entorno: horizonte infinito, tiempo continuo y precios constantes, con un tipo de interés real constante y exógeno, una oferta de bienes de capital perfectamente elástica, sin costes de ajuste ni impuestos distorsionantes.

Tecnología: una función de producción neoclásica, Cobb-Douglas, con la producción dada, y una depreciación constante, δ. La producción se toma como dada porque, con rendimientos constantes a escala y precio dado, las condiciones de óptimo solo fijarían la ratio K/L y no la escala de la empresa.

2.1.2. Modelo

Planteamiento: la empresa elige la inversión bruta y el empleo para maximizar el valor descontado de sus beneficios, sujeta a la acumulación del capital, con el capital inicial dado:

$$\max_{\{I,L\}} V = \int_0^{\infty} e^{-rt}\,[Pf(K,L) - wL - P_K I]\,dt \quad \text{s.a.:} \begin{cases} \dot{K} = I - \delta K \\ K_0 \text{ dado} \end{cases}$$

- I y L son las variables de control y K la de estado, que solo cambia con la inversión. La transversalidad, $\lim_{t\to\infty}\lambda_t K_t = 0$, excluye acumular capital indefinidamente o sendas explosivas. El hamiltoniano en valor presente u origen, con λ el precio sombra del capital, es:

$$H = e^{-rt}\,[Pf(K,L) - wL - P_K I] + \lambda\,(I - \delta K)$$

CPO: con el hamiltoniano y el principio del máximo de Pontryagin, el valor de la productividad marginal del trabajo iguala el salario, el precio sombra del capital iguala su precio descontado y, combinando esta con la ecuación de coestado, resulta la condición central, que con la Cobb-Douglas da el capital óptimo:

$$Pf'_K = P_K(r+\delta) \equiv r_K \;\Rightarrow\; K^* = \frac{\alpha PY}{r_K}$$

- El coste de uso desglosa el coste del capital en su precio de adquisición, su coste de oportunidad y su depreciación. El capital óptimo crece con el valor de la producción y decrece con el coste de uso.

CSO: el hamiltoniano es cóncavo, y las condiciones son suficientes. Sin costes de ajuste, el problema se reduce al estático de cada instante.

Solución: basta comparar el capital óptimo con el instalado: si coinciden, la empresa solo repone el desgaste; si el óptimo es menor, deja que el capital se deprecie o vende, y la inversión neta es negativa, y si es mayor, amplía el capital. Si el ajuste se completa en un periodo, la inversión bruta es la brecha entre el capital óptimo y el instalado más la reposición:

$$I = (K^* - K) + \delta K$$

Estática comparativa: el capital óptimo decrece con el tipo de interés real, el precio de los bienes de capital y la depreciación, y crece con el precio del producto y la productividad del capital. El modelo dice cuánto cambia K^*, pero no a qué ritmo se alcanza.

Extensiones: con inflación de los bienes de capital, $r_K = P_K(i+\delta-\pi_K)$, de modo que una revalorización esperada superior a la inflación general reduce el coste de uso. Con impuestos (Hall y Jorgenson, 1967), $r_K = \chi_t P_K(r+\delta)$: el impuesto sobre sociedades reduce el valor del producto marginal y las desgravaciones abaratan el coste de uso. Con $i = r+\pi^e$, el coste de uso con inflación equivale a $r_K = P_K[r+\delta-(\pi_K - \pi^e)]$.

- *Política económica*: la monetaria es eficaz, a través del tipo de interés real y el coste de uso; la fiscal depende de su forma: el gasto público puede elevar el tipo de interés, y las desgravaciones a la inversión reducen el coste de uso.

2.1.3. Conclusiones

Fue un paso notable en la microfundamentación. Pero da una predicción estática, sin sustitución intertemporal del capital. Fija cuánto capital se desea, pero no a qué ritmo se invierte, porque, ante una rebaja del tipo de interés que desplace K^*, el capital solo "saltaría" con una tasa de inversión infinita, plausible para una empresa, pero no en el agregado, y omite las expectativas, porque sin costes de ajuste la empresa solo atiende a las variables actuales. Para el ritmo, Jorgenson recurre a un retardo distribuido *ad hoc*, análogo al del acelerador flexible, con ponderaciones ω_j decrecientes:

$$I_t^n = \sum_{j=0}^{\infty}\omega_j\,(K_{t-j}^* - K_{t-j-1}^*)$$

2.2. Costes de ajuste y q de Tobin

2.2.1. Concepto de costes de ajuste

Definición: los **costes de ajuste**, noción debida a Eisner y Strotz (1963) y a Lucas (1967), son los que soporta la empresa al alterar su dotación de factores, mayores cuanto más rápido es el cambio. Son internos, como las mermas de producción al reemplazar máquinas, o externos, como las indemnizaciones o los costes de contratación.

Función: $\varphi(I)$ es no negativa, creciente en el valor absoluto de la inversión y convexa, con $\varphi(0) = \varphi'(0) = 0$ y $\varphi'' > 0$: invertir deprisa cuesta más que despacio. Con ella, la condición de contratación del capital incorpora, además del coste de uso, el coste marginal de ajuste, de modo que la empresa no salta al capital deseado, sino que elige un ritmo óptimo, y la inversión se suaviza en el tiempo. El coste marginal de ajuste tiene el signo de la inversión, $\varphi'(I) \gtrless 0$ según $I \gtrless 0$, porque desinvertir deprisa también cuesta.

2.2.2. La q de Tobin (Tobin, 1969)

Idea: Tobin (con Brainard, 1968; 1969) sostiene que la inversión es función de q, el valor del capital en el mercado en relación con su coste de reposición, $I = f(q)$, que recoge en un número las perspectivas futuras.

Versiones: la q marginal, lo que se pagaría por una unidad más de capital sobre su coste de reposición, gobierna la inversión, pero no se observa. La q media, el valor de mercado del capital instalado sobre su coste de reposición, se observa, y coincide con la marginal bajo las condiciones de Hayashi (1982), es decir, con rendimientos constantes a escala en la producción y en los costes de ajuste, competencia perfecta y capital homogéneo.

Interpretaciones: tres.

- *Financiera*: $q_{\text{me}} = V/(P_K K)$, la definición original de Tobin, con V la capitalización bursátil más la deuda, que en un mercado eficiente iguala el valor actual de los flujos esperados de la empresa.
- *Productiva*: $q_{\text{mg}} = (f'_K - \delta)/(i - \pi^e)$, la rentabilidad neta del capital sobre su coste de oportunidad. Si supera la unidad, el *stock* es inferior al óptimo y conviene invertir.
- *Contable*: el valor de mercado sobre el valor en libros, que, al no recoger la inflación, sobrevalora la q económica. No debe confundirse con el *price-to-book ratio*, que solo considera las acciones.

Regla: con $q > 1$ conviene invertir; con $q < 1$, desinvertir, y con $q = 1$ solo se repone la depreciación.

A diferencia de la EMK, la q media es observable y fácil de interpretar. Pero omite el capital humano, puede reflejar comportamientos de rebaño y burbujas y, en su formulación original, carece de microfundamentos. La omisión del capital humano, que el valor de mercado sí recoge, sesga la q al alza, y los microfundamentos llegan con los modelos con costes de ajuste.

2.3. Modelos q con costes de ajuste

Idea: Summers (1981), Abel (1982) y Hayashi (1982) integran los costes de ajuste en el problema de la empresa y demuestran que la q marginal es el **precio sombra del capital**, y la inversión es una función creciente de ella.

2.3.1. Supuestos

Entorno: los de Jorgenson, con costes de ajuste internos y sin depreciación.

Tecnología: el beneficio por unidad de capital, $\Pi(K)$, incorpora el empleo óptimo y decrece con el capital de la industria, porque más capital reduce el precio del producto. El trabajo se ajusta libremente, y los costes de ajuste entran de forma aditiva en la función objetivo.

2.3.2. Modelo

Planteamiento: la empresa elige su senda de inversión, con el precio del capital normalizado a la unidad:

$$\max_{\{I\}} V = \int_0^\infty e^{-rt}\left[\Pi(K)\,k - I - \varphi(I)\right] dt \quad \text{s.a.:} \begin{cases} \dot{k} = I \\ k_0 \text{ dado} \end{cases}$$

CPO: con el hamiltoniano en valor corriente, cuya variable de coestado es la q marginal:

$$H = \Pi(K)\,k - I - \varphi(I) + q\,I, \qquad q = e^{rt}\lambda$$

- *Demanda de inversión*: $1 + \varphi'(I) = q$. Se invierte mientras el precio más el coste marginal de ajuste quede por debajo del valor sombra, y la rapidez del ajuste depende de lo convexos que sean esos costes.
- *Arbitraje*: $\Pi(K) + \dot{q} = rq$. La rentabilidad del capital instalado, dividendo más ganancia de capital, iguala la de invertir su valor al tipo de interés. Leída como condición de equilibrio, se invierte hasta que el ingreso de una unidad más de capital iguala su coste de oportunidad, $rq - \dot{q}$: los intereses que no se ganan por no venderlo, menos la ganancia por conservarlo. Como ecuación de comportamiento, la valoración del capital se revisa con la bolsa, el tipo de interés y el ciclo. Con depreciación, la condición es $\Pi(K) + \dot{q} = (r + \delta)\,q$, y la rentabilidad del capital instalado iguala su coste de uso.

CSO: la convexidad de φ hace cóncavo el hamiltoniano, y las condiciones son suficientes.

Solución: invirtiendo la primera condición, la hipótesis de Tobin queda derivada del programa de la empresa:

$$I = (\varphi')^{-1}(q - 1) \;\to\; I = f(q^+)$$

Agregando sobre las N empresas idénticas, la inversión de la industria y la condición de arbitraje forman un sistema en (K, q) que fija el ritmo al que el capital se acerca a su estado estacionario:

$$\dot{K} = Nf(q) \qquad \dot{K} = 0 \;\Rightarrow\; q^* = 1$$
$$\dot{q} = rq - \Pi(K) \qquad \dot{q} = 0 \;\Rightarrow\; \Pi(K^*) = rq^*$$

Equilibrio: con $q^* = 1$, la condición de arbitraje fija el capital de la industria. Las empresas no quieren modificar su capital, y a los inversores les conviene conservarlo, porque no prevén ganancias ni pérdidas. Con depreciación, $q^* > 1$, porque hay inversión de reposición.

Desarrollo gráfico: la curva $\dot{K} = 0$ es horizontal en $q = 1$; la $\dot{q} = 0$, decreciente. El sistema tiene un **punto de silla**, y para cada capital heredado hay un único q sobre el sendero: q es una variable de salto y K, predeterminada. Las demás sendas las descarta la transversalidad: por encima del sendero, el capital crece mientras q solo se sostiene si se espera que siga subiendo, y ambos divergen, como en una burbuja especulativa. La curva $\dot{q} = 0$ es decreciente porque más capital reduce el beneficio por unidad, y por encima de ella el ingreso del capital no cubre su coste de oportunidad y q se revisa al alza.

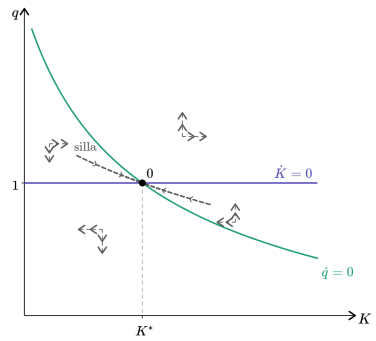
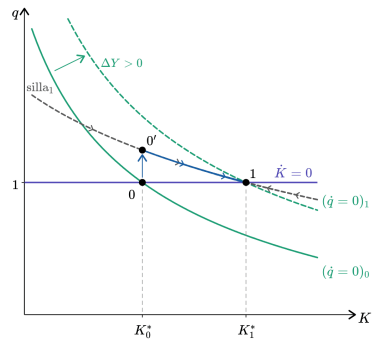


Diagrama de fases del modelo q: las curvas $\dot{K} = 0$ y $\dot{q} = 0$ y el sendero de silla hacia el estado estacionario 0.

Estática comparativa: cuatro perturbaciones.

- *Aumento permanente de la producción*: la curva $\dot{q} = 0$ sube, igual que con una rebaja del impuesto sobre sociedades. Cada unidad de capital rinde más, y a los inversores les basta una ganancia de capital menor: q salta al nuevo sendero y el capital crece gradualmente hasta un nivel mayor, mientras la inversión, tras el salto inicial, se diluye.



Aumento permanente de la producción: q salta al nuevo sendero de silla (0) y el capital crece hasta el nuevo equilibrio (1).

- *Aumento temporal*: q sube menos, porque la empresa prevé que tendrá que desmontar con costes parte del capital, y este empieza a reducirse antes de que la producción vuelva a la normalidad. La inversión es mayor cuando se espera una producción futura elevada (un acelerador, ahora microfundamentado). En T, cuando la producción vuelve a su nivel, la economía alcanza el sendero de silla inicial y regresa por él al equilibrio de partida.
- *Bajada del tipo de interés*: si es permanente y no anticipada, tiene efectos análogos a los del aumento permanente, pero lo que cuenta es toda la curva de tipos, y un tipo largo por debajo del corto anima la

inversión. Una *reducción salarial* eleva siempre el beneficio por unidad de capital y el capital de estado estacionario. En un marco estático, en cambio, solo lo eleva si el efecto producción cruzado supera al de sustitución.
- *Desgravación sobre el precio del capital*: equivale a una subvención por unidad de capital y desplaza la curva $\dot{K} = 0$ a $q = 1 - z$. Si es permanente, q baja de inmediato y el capital crece: se invierte más, a costa de beneficios menores y un capital que vale menos. Si es temporal, cuanto más cerca está su final, más se acelera la inversión, porque las empresas adelantan la inversión para aprovecharla, y el estímulo es mayor que con una permanente.

Extensiones:
- *Política económica*: la monetaria es más eficaz cuanto más permanente se perciba la bajada y más se traslade a los tipos largos; la fiscal, a través de incentivos, más si son temporales. Una expansión de la demanda eleva q y la inversión, pero la subida del tipo de interés la contrarresta. La inversión depende, así, de toda la senda esperada de beneficios y tipos. El ritmo de ajuste es finito y depende de la relevancia de los costes de ajuste.

2.3.3. Conclusiones

Es el modelo más microfundamentado y ofrece a la vez un *stock* deseado y un ritmo de ajuste. Pero las regresiones sobre la q observada dan coeficientes muy bajos, con costes de ajuste inverosímiles. Los costes cuadráticos y simétricos simplifican en exceso, y la equivalencia entre q marginal y media exige las condiciones de Hayashi. Los coeficientes bajos suelen atribuirse a la dificultad de medir la q marginal, y con costes asimétricos o irreversibilidad la respuesta de la inversión es asimétrica y aparece el valor de esperar.

En suma, Jorgenson determina el capital óptimo con el coste de uso, y los costes de ajuste añaden el ritmo y resumen en la q marginal las expectativas sobre beneficios y tipos.

3. Incertidumbre, irreversibilidad y financiación

Quedan dos aportaciones, la inversión como opción real y las decisiones de financiación, y las implicaciones de política económica.

3.1. Incertidumbre e irreversibilidad

Idea: el enfoque, que parte de Bernanke (1983) y formalizan Dixit y Pindyck (1994), trata la inversión como el ejercicio de una **opción real** de compra sobre un activo irreversible. Esperar tiene valor, porque permite decidir con más información.

Supuestos: tres cambios sobre Jorgenson: la incertidumbre sobre los rendimientos se despeja en parte con el tiempo; la empresa puede aplazar la decisión, que pasa a ser también de calendario, y las inversiones en capital fijo son irreversibles, del todo o en parte, porque deshacerlas supone perder valor.

Razonamiento: si la empresa invierte ya, cobra los rendimientos desde hoy, pero pierde la posibilidad de esperar; si aguarda, decide mejor informada. El criterio neoclásico se sustituye por un listón más alto:

$$Pf'_K \geq r_K + \Theta$$

- Θ es el **valor de la opción viva**, lo que aporta esperar neto de los rendimientos que se aplazan. La empresa invierte cuando hacerlo vale más que conservar la opción.

Estática comparativa: más incertidumbre eleva el valor de la opción y la productividad exigida, y la inversión se pospone. El efecto sobre el capital a largo plazo es ambiguo, porque la irreversibilidad impide también desinvertir. Por el **principio de las malas noticias** de Bernanke, la decisión de esperar depende solo de la gravedad de las malas noticias posibles, no de la magnitud de las buenas, y la inversión reacciona con retraso a las mejoras.

Extensiones:
- *Política económica*: pocas cosas pesan tanto en la inversión como la estabilidad macroeconómica, en sentido amplio: precios estables, tipos de cambio y de interés sin sobresaltos, derechos de propiedad respetados y una desigualdad que no alimente el conflicto. La autoridad económica puede, por tanto, animar la inversión si reduce la incertidumbre.

Explica que la incertidumbre frene con fuerza la inversión aunque los inversores sean neutrales al riesgo, y la lentitud y asimetría de la inversión tras las recesiones. Pero exige especificar el proceso de los rendimientos, supone irreversibilidad total y, como la q marginal, no es observable. La mayoría de las inversiones son, en cambio, parcialmente reversibles.

3.2. Decisiones de financiación

Los modelos anteriores olvidan que la empresa decide cómo financiar su inversión casi a la vez que decide invertir (con beneficios no repartidos, deuda o acciones).

3.2.1. Visión tradicional y Modigliani-Miller

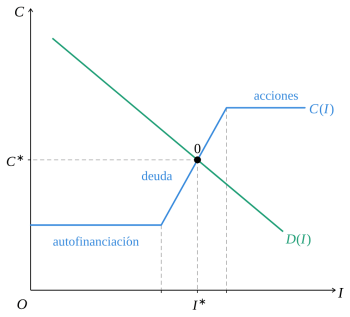
Idea: para la visión tradicional, asociada a Durand (1952), hay una estructura óptima del capital que minimiza su coste ponderado, y la inversión no es independiente de la financiación. Para Modigliani y Miller (1958), con mercados perfectos no la hay, y la decisión de inversión es independiente de la de financiación (proposición III). Su teorema es la referencia: al precisar cuándo la financiación es irrelevante, desplaza la pregunta hacia qué supuesto falla y justifica descontar con un único coste del capital, pero las imperfecciones de mercado lo hacen poco plausible. Entre ellas están los costes de transacción y de agencia, los impuestos distorsionantes y la inflación.

3.2.2. Racionamiento del crédito (Stiglitz y Weiss, 1981)

Idea: con información asimétrica, si el banco sube el tipo expulsa a los prestatarios más seguros y atrae a los más arriesgados, y, como su rendimiento esperado deja de crecer con el tipo, prefiere racionar la cantidad. El crédito se raciona en equilibrio, sin rigideces de precios, y la inversión depende de la disponibilidad de financiación y no solo de su coste. El racionamiento puede desaparecer si los contratos con garantías permiten la autoselección. El resultado exige, por tanto, que el banco no pueda separar a los prestatarios según su riesgo.

3.2.3. Jerarquía financiera (Myers y Majluf, 1984)

Idea: la empresa recurre primero a la autofinanciación, sin costes de emisión ni prima de información asimétrica; después, a la deuda, y por último a las acciones, cuya emisión los inversores leen como señal de sobrevaloración. La curva de costes de financiación tiene así tres tramos, y su corte con la demanda de inversión, decreciente en el coste, fija la inversión y su coste. Según el tramo del corte, la empresa tiene más o menos holgura, y la inversión depende de la liquidez interna y no solo del coste de uso. La demanda decrece porque al invertir más se acometen proyectos de menor rentabilidad. Todo ello se deriva de un solo supuesto, la información asimétrica entre directivos e inversores.



Jerarquía financiera: coste de los fondos creciente por tramos (autofinanciación, deuda, acciones) y equilibrio con la demanda de inversión.

Ordena las fuentes, pero no fija una estructura objetivo y prescinde de los impuestos y los costes de insolvencia que recoge la teoría del equilibrio estático.

3.2.4. El acelerador financiero

Idea: para Bernanke, Gertler y Gilchrist (1996), las restricciones de crédito ligadas al colateral amplifican los *shocks*: uno positivo eleva el valor de los activos y relaja el crédito, y uno negativo lo endurece y contrae la inversión más allá de los fundamentos. Convierte perturbaciones moderadas en fluctuaciones amplias y persistentes, explica la asimetría de las recesiones con crisis de crédito y se ha incorporado a los modelos de equilibrio general con fricciones financieras. Pero sitúa la fricción en el balance de las empresas prestatarias y deja fuera las restricciones de los propios intermediarios, que la crisis de 2008 puso en primer plano.

En suma, la inversión depende de la liquidez interna y del colateral, como documentan Fazzari, Hubbard y Petersen (1988) para las empresas con restricciones de acceso al crédito, aunque la sensibilidad al flujo de caja admite otras lecturas, porque el flujo de caja informa sobre la rentabilidad futura, y la jerarquía no siempre se observa en las emisiones.

3.3. Implicaciones de política económica

Por teorías: la eficacia de las políticas depende del determinante de la inversión.
- *Keynes*: la monetaria es poco eficaz, y la fiscal, muy eficaz sobre la renta, con una expulsión parcial de la inversión.
- *Acelerador*: la monetaria es ineficaz, y la fiscal, muy eficaz a corto plazo por el multiplicador-acelerador.
- *Jorgenson*: la monetaria es eficaz a través del tipo de interés; la fiscal, a través de las desgravaciones, pero adversa si eleva el tipo.
- *Modelo q*: la monetaria es eficaz, más si se percibe permanente y llega a los tipos largos, y la fiscal, más si es temporal.

Implicaciones: las opciones reales y las imperfecciones financieras añaden que la estabilidad macroeconómica, la certidumbre y un marco institucional que minimice las imperfecciones financieras son los principales determinantes de la inversión.

1. Fundamentos de la teoría sobre la demanda de dinero

Los fundamentos siguen el orden en que surgieron: la teoría cuantitativa, la preferencia por la liquidez, la síntesis neoclásica y el monetarismo.

1.1. Teoría cuantitativa del dinero

Idea: el paradigma clásico y neoclásico explica el sector monetario con la **teoría cuantitativa del dinero**, una teoría sobre las causas de los cambios del valor del dinero. A largo plazo, la cantidad de dinero y el nivel general de precios guardan una relación proporcional, ya anunciada por la escuela de Salamanca (Azpilcueta, 1556), Bodin (1568) y Hume (1752).

Supuestos comunes: los dos enfoques que la formalizan parten de la identidad contable del gasto y comparten tres supuestos.

- Pleno empleo, porque aceptan la ley de Say.
- Ausencia de atesoramiento, de modo que $M^s = M^d$ en todo momento.
- Oferta de dinero exógena, fijada por las autoridades y correspondiente a un agregado estrecho, M1.

1.1.1. Transacciones (Newcomb, 1885; Fisher, 1911)

Idea: Newcomb formuló la ecuación de cambio y Fisher la convirtió en la formulación de referencia. El dinero se demanda solo para transacciones, y los saldos reales necesarios son una fracción técnica, fijada por las instituciones de pago, del volumen de transacciones.

Supuestos: Fisher parte de la identidad $MV \equiv PT$, donde V es la **velocidad-transacción**, cuántas compras financia en promedio una unidad monetaria, y T el número de transacciones. El dinero solo es unidad de cuenta y medio de cambio, y V y T son estables, la primera por razones institucionales y la segunda por el pleno empleo.

Solución: no hay optimización. Imponiendo el equilibrio del mercado monetario, la demanda agregada de dinero es mecanicista y muy estable:

$$\frac{M^d}{P} = \frac{1}{V_0}\,T_0$$

Estática comparativa: solo se desplaza si cambian las instituciones de pago o las transacciones, y el tipo de interés no la afecta. Todo aumento de la oferta monetaria se traslada íntegramente a los precios, de modo que la política monetaria determina el nivel de precios, pero es neutral sobre las variables reales (**dicotomía clásica**), y la fiscal es ineficaz con la renta.

Conclusiones: es la formulación de referencia de la teoría cuantitativa, pero no hay un modo sencillo de calcular el número de transacciones, y si el dinero cumple otras funciones su demanda podría depender del tipo de interés.

1.1.2. Saldos de caja (escuela de Cambridge, 1917)

Idea: Marshall, Pigou (1917) y el joven Keynes (1923) parten de la decisión individual. Los agentes desean mantener una fracción de su renta en dinero, porque presta servicios útiles, y esa fracción, la k **de Cambridge**, resume todo lo que afecta a la velocidad.

Supuestos: la renta real sustituye a las transacciones, $MV \equiv PY$, y V pasa a ser una velocidad-renta. El dinero es también depósito de valor y reporta utilidad, de modo que su demanda podría depender de la riqueza y del tipo de interés, pero V e Y se tratan como estables.

Solución: con el mercado monetario en equilibrio, y $k_0 = 1/V_0$ constante:

$$\frac{M^d}{P} = k_0\,Y_0, \qquad k_0 \in (0,1)$$

Estática comparativa: la renta, y no el número de transacciones, es el principal determinante, y el tipo de interés queda fuera. Los resultados de política coinciden con los de Fisher.

Conclusiones: admite elementos subjetivos y es la primera teoría genuina de la demanda de dinero, pero no da un papel explícito al tipo de interés.

En ambos enfoques la demanda es proporcional a una variable de escala, con una constante institucional. Esa constancia, sin embargo, es un supuesto y no un resultado. Si k depende del coste de oportunidad, entra el tipo de interés, y si depende de expectativas volátiles, la demanda deja de ser estable, que es lo que Keynes explotará.

1.2. Preferencia por la liquidez (Keynes, 1936)

Idea: en la *Teoría general*, Keynes concibe por primera vez el dinero como un activo financiero. Acepta la ecuación de Cambridge a largo plazo, pero a corto plazo la incertidumbre y las expectativas sobre el tipo de interés hacen que el dinero se demande también como depósito de valor. El interés no es la recompensa a la abstinencia, sino la retribución por renunciar a la liquidez.

1.2.1. Supuestos

Entorno: corto plazo, con incertidumbre sobre la evolución futura del tipo de interés. La riqueza se mantiene en dinero, un activo estéril, o en bonos, cuyo precio guarda una relación inversa con su rentabilidad.

Agentes: cada individuo espera un tipo de interés **normal**, i^e, hacia el que cree que tenderá el tipo, que es convencional, cambiante y heterogéneo entre individuos.

1.2.2. Modelo

Motivos: sin optimización explícita, Keynes distingue tres motivos, cada uno ligado a una función del dinero.

- *Transacción*: para gastos previstos, como medio de cambio. Acepta aquí Cambridge, $M_T^d = f(Y^+)$.
- *Precaución*: para gastos imprevistos. Podría depender del tipo de interés, pero Keynes lo omite, $M_P^d = f(Y^+)$.
- *Especulación*: por la expectativa de rendimiento futuro, como depósito de valor, $M_E^d = f(i^-)$. Este motivo es la aportación novedosa, porque liga la demanda de dinero al tipo de interés y rompe la dicotomía clásica.

Demanda individual: la especulativa es discontinua. Si $i > i^e$, el individuo cree que el tipo caerá y el precio de los bonos subirá, y mantiene toda su riqueza en bonos. Si $i < i^e$, espera que suba y la mantiene toda en dinero, W:

$$\frac{M_E^d}{P} = \begin{cases} 0 & \text{si } i > i^e \\ W & \text{si } i < i^e \end{cases}$$

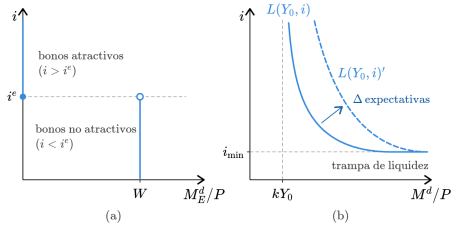
Solución: como los valores críticos, i^e, son heterogéneos, la especulativa agregada es continua, decreciente y convexa. Sumando los tres motivos:

$$\frac{M^d}{P} = L(Y^+, i^-) = kY - h\,i, \qquad k,h > 0$$

- kY recoge la transacción y la precaución, y $-h\,i$ la especulación, núcleo de la preferencia por la liquidez. El tipo nominal es $i = r + \pi^e$, y con $h = 0$ la demanda sería vertical, como en el enfoque neoclásico.

Estática comparativa: la renta eleva la demanda; el tipo de interés la reduce, con una sensibilidad h que Keynes supone elevada, y un cambio en las expectativas sobre el tipo normal desplaza toda la curva, lo que la hace **inestable**.

Trampa de liquidez: Keynes la planteó como posibilidad teórica. Al llegar el tipo a un mínimo, nadie quiere bonos, porque todos esperan que suba, y la demanda se hace infinitamente elástica, $h \to \infty$. En la práctica ese suelo, el ZLB, está en torno a cero.



Demanda de dinero keynesiana: individual discontinua en i^e y agregada decreciente y convexa, horizontal en la trampa de liquidez.

Extensiones:

- *Política económica*: la política monetaria es poco eficaz con la renta por la alta sensibilidad al tipo de interés, el **segundo eslabón débil**, y totalmente ineficaz en la trampa de liquidez. La fiscal es muy eficaz, sobre todo en la trampa, donde no eleva el tipo de interés.

1.2.3. Conclusiones

El abandono de la dicotomía clásica y la trampa de liquidez dan efectividad diferenciada a las políticas fiscal y monetaria. Pero cada individuo guarda toda su riqueza en bonos o en dinero, sin diversificar, y los motivos se postulan sin microfundamentar, tarea de la síntesis neoclásica.

1.3. Síntesis neoclásica

Los motivos keynesianos son el punto de partida de las teorías que buscan microfundamentarlos.

1.3.1. Transacción: inventarios (Baumol, 1952; Tobin, 1956)

Idea: el dinero es un inventario para transacciones del que se dispone gradualmente. Hay una disyuntiva entre el coste de reponerlo y el coste de oportunidad de mantenerlo, de la que resulta una demanda por transacción que depende también del tipo de interés.

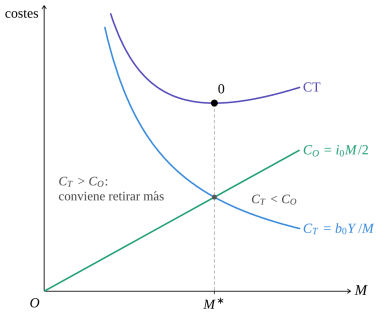
Supuestos: información perfecta y un solo periodo. El individuo recibe su renta real, Y, en una cuenta remunerada al tipo nominal i_0, hace retiradas de cuantía real constante, M, y cada retirada tiene un coste fijo real, $b_0 > 0$.

Planteamiento: elige la retirada que minimiza la suma del coste de transacción, el coste unitario por el número de retiradas, y del coste de oportunidad, el tipo por el saldo medio, $M/2$:

$$\min_M \mathrm{CT} = b_0\frac{Y}{M} + i_0\frac{M}{2}$$

CPO: la retirada óptima iguala el ahorro marginal en costes de transacción al coste marginal de oportunidad, y la convexidad del coste total garantiza el mínimo:

$$M^* = \sqrt{2b_0 Y/i_0} \;\rightarrow\; \frac{M^d}{P} \equiv \frac{M^*}{2} = k(b_0^+, i_0^-)\cdot Y^{1/2}$$



Modelo de inventarios: la retirada óptima, M^*, minimiza el coste total y se sitúa donde ambos costes se igualan.

Estática comparativa: la demanda crece con la renta, pero menos que proporcionalmente, con elasticidad 0,5, lo que refleja **economías de escala** en la gestión de tesorería. Crece con el coste de transacción y cae con el tipo de interés, con elasticidad $-0{,}5$.

Extensiones: con flujos aleatorios, Miller y Orr (1966) proponen una regla de bandas: el agente solo ajusta su cartera, hasta un nivel de retorno, cuando el saldo toca un límite inferior o superior, y la demanda aumenta con la incertidumbre de los flujos. Con recursos ociosos, la política monetaria mueve la renta más que en Keynes, porque la demanda por transacción crece menos que proporcionalmente con ella. Con el tipo de interés fijo y $M^d = kY^{1/2}$, frente a $M^d = kY$ en Keynes, absorber un 1 % más de oferta monetaria exige un 2 % más de renta, y no un 1 %.

Conclusiones: microfundamenta el motivo transacción con elasticidades contrastables, pero supone flujos ciertos y retiradas constantes, y con agentes heterogéneos la suma de raíces cuadradas no es la raíz del agregado.

1.3.2. Precaución: costes de ruptura (Whalen, 1966)

Idea: se guarda efectivo como reserva porque no se sabe cuándo entrará y saldrá dinero. Quedarse sin efectivo tiene un **coste de ruptura** y mantenerlo, un coste de oportunidad.

Planteamiento: el exceso de gasto sobre la renta es aleatorio, de media cero y varianza σ^2. Si supera el saldo, hay que vender bonos y pagar un corretaje, b. Acotando la probabilidad de ruptura por la desigualdad de Chebyshev:

$$\min_M \mathrm{CT} = iM + b\,\frac{\sigma^2}{M^2} \;\rightarrow\; M^* = \left(\frac{2b\sigma^2}{i}\right)^{1/3}$$

Solución: el óptimo iguala el ahorro esperado en corretaje de una unidad más de efectivo al interés perdido. La demanda es $f(b^+, i^-, \sigma^+)$, y la renta entra a través de σ. Según crezca el número o la cuantía de los pagos, la elasticidad-renta está entre 1/3 y 2/3, de nuevo con economías de escala. Si crece el número de pagos, de igual cuantía, σ^2 crece en proporción a Y, y si crece la cuantía de cada pago, lo hace σ:

$$M^* \propto Y^{1/3} \qquad \text{o} \qquad M^* \propto Y^{2/3}$$

Conclusiones: explica que la demanda aumente con la incertidumbre, y por eso puede ser inestable: estabilizar el entorno reduce la liquidez deseada. Pero el efecto de la renta depende de cómo crezcan los pagos, lo que resta capacidad predictiva.

1.3.3. Especulación: carteras mixtas (Tobin, 1958)

Idea: Tobin aplica la selección de cartera de Markowitz. Frente a la disyuntiva keynesiana, los individuos mantienen a la vez dinero y bonos, y la **aversión al riesgo** basta para explicar la preferencia por la liquidez.

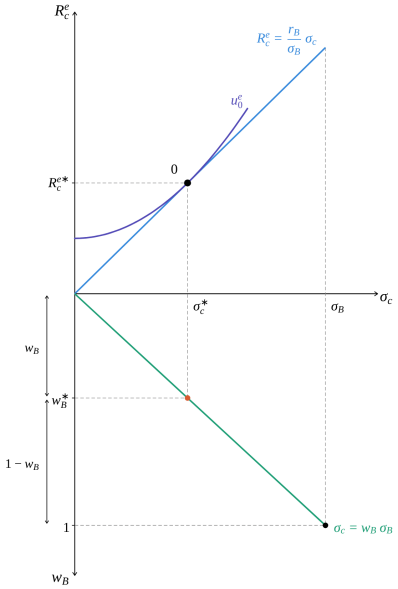
Supuestos: el bono tiene rentabilidad esperada r_B y riesgo $\sigma_B > 0$; el dinero, rentabilidad y riesgo nulos. El agente averso al riesgo tiene curvas de indiferencia crecientes y convexas en el plano riesgo-rentabilidad.

Planteamiento: el agente elige la proporción de bonos, w_B, con $R_c^e = w_B r_B$ y $\sigma_c = w_B \sigma_B$, y maximiza su utilidad esperada sobre la **recta de oportunidades** que resulta:

$$\max_{w_B} u^e = f(R_c^e, \sigma_c) \quad \text{s.a.: } R_c^e = \frac{r_B}{\sigma_B}\,\sigma_c$$

CPO: la pendiente de la curva de indiferencia iguala la de la recta, r_B/σ_B: el agente acepta riesgo hasta que la rentabilidad extra que exige iguala la que ofrece el mercado. La demanda especulativa es $(1-w_B^*)W$, es decir, $f(W^+, \sigma_B^+, r_B^\pm)$. Resulta de la tangencia y, con agentes homogéneos, la agregada hereda sus signos:

$$-\frac{u'_{\sigma_c}}{u'_{R_c^e}} = \frac{r_B}{\sigma_B} \;\rightarrow\; w_B^* \;\rightarrow\; \frac{M^d}{P} = (1-w_B^*)\,W$$



Modelo de carteras de Tobin: tangencia de la curva de indiferencia con la recta de oportunidades y proporción óptima de bonos.

Estática comparativa: la demanda crece con la riqueza y con el riesgo de los bonos. El efecto del tipo es ambiguo, porque el **efecto sustitución** lleva de dinero a bonos, pero el **efecto renta** permite la misma rentabilidad con menos riesgo y eleva el peso del dinero. La demanda solo cae si domina el primero. Así, la política monetaria puede estabilizar la economía: el dinero altera el tipo de interés por el motivo especulación, y este, la inversión.

Conclusiones: explica la tenencia simultánea de dinero y bonos. Pero el dinero no está libre de riesgo, porque la inflación socava su valor. Hay activos casi sin riesgo más rentables, como las letras del Tesoro, que lo dominarían, y solo se aplica a agregados amplios.

En conjunto, la síntesis obtiene en los tres motivos una demanda decreciente en el tipo de interés, en Tobin si domina el efecto sustitución, y potencialmente inestable, pero son demandas individuales que no resuelven la agregación ni integran los motivos.

1.4. Monetarismo

1.4.1. El dinero como bien (Friedman, 1956)

Idea: Friedman reformula la teoría cuantitativa como una teoría de la demanda de dinero. El dinero es un bien de capital más, cuya demanda depende de la renta, de su precio relativo, la rentabilidad de los activos alternativos, y de las preferencias.

Supuestos: las decisiones de consumo y de cartera se toman conjuntamente. El dinero tiene muchos **sustitutivos imperfectos**, financieros y reales, frente a los pocos pero buenos de Keynes, y no se segmenta la demanda por motivos. Como Tobin, parte de una decisión de cartera sobre la riqueza total, humana y no humana, pero postula la función sin resolver el programa:

$$\frac{M^d}{P} = f((Y^P)^+, (W_H/W)^+, r_B^-, r_E^-, \pi^{e-})$$

- Y^P es la renta permanente; W_H/W, la riqueza humana sobre la total, menos líquida; r_B y r_E, las rentabilidades de bonos y acciones, y π^e, la rentabilidad nominal esperada de los activos reales.

Solución: es una ecuación de Cambridge con k variable y la renta permanente como variable de escala:

$$\frac{M^d}{P} = k(r^-, \pi^{e-})\cdot Y^P$$

Estática comparativa: la demanda crece con la renta permanente, que limita los saldos que pueden mantenerse, y con el peso de la riqueza humana, menos líquida y de liquidez más incierta. Cae con la rentabilidad de bonos y acciones, su coste de oportunidad, y con la inflación esperada, que lleva a sustituir dinero por activos reales como el oro o los inmuebles. La demanda nominal, en cambio, sube con el nivel de precios.

Con muchos sustitutivos y la renta permanente, la sensibilidad al tipo de interés es baja, $h \approx 0$, y la demanda es **muy estable**. La inflación esperada liga directamente la demanda de dinero con la de activos y es el núcleo de la transmisión monetarista.

Extensiones:

– *Política económica*: no hay eslabones débiles, porque la inversión es muy sensible al tipo de interés y la demanda de dinero poco. La política monetaria es eficaz a corto plazo por el elevado efecto liquidez, y los cambios del dinero explican los ciclos, pero debe ser reglada y dirigida a la inflación, por los retardos largos y variables y su inefectividad a largo plazo. La fiscal es ineficaz por el efecto expulsión total. Las dos cadenas de transmisión son:

$$\uparrow M^s \to \downarrow i \to \uparrow I \to \uparrow Y$$
$$\uparrow G \to \uparrow Y \to \uparrow M^d \to \uparrow r \to \downarrow I$$

Conclusiones: recupera la estabilidad de la función con un fundamento de elección de cartera, pero postula sus argumentos y se ajusta mejor a agregados amplios.

1.4.2. Hiperinflación (Cagan, 1956)

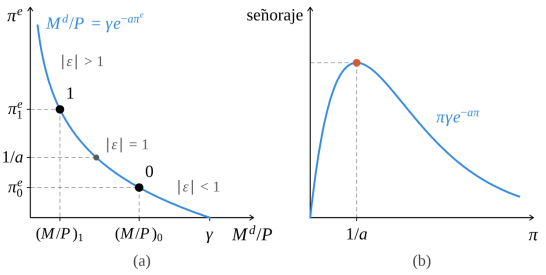
Idea: Cagan, discípulo de Friedman, estima para siete hiperinflaciones europeas la función monetarista reducida a su único argumento relevante, la inflación esperada. Las expectativas son adaptativas, $\pi^e_t - \pi^e_{t-1} = \lambda(\pi_{t-1} - \pi^e_{t-1})$, con $\lambda \in (0,1)$:

$$\frac{M^d}{P} = \gamma \cdot e^{-a\pi^e}$$

– $a > 0$ es la semielasticidad, constante, de modo que la elasticidad, $a\pi^e$, crece con la inflación y es unitaria en $\pi^e = 1/a$. Con $\pi^e = \pi$, el señoraje de estado estacionario es la inflación por los saldos reales:

$$\pi \frac{M^d}{P} = \pi\, \gamma\, e^{-a\pi}$$

Solución: por encima de $1/a$ los saldos reales caen más que proporcionalmente, la **huida del dinero**, y el señoraje disminuye, con un máximo en $\pi = 1/a$. La inflación converge al estado estacionario si $a\lambda < 1$ y se alimenta a sí misma si $a\lambda > 1$.



Modelo de Cagan: demanda de saldos reales decreciente en la inflación esperada y señoraje con forma de curva de Laffer, máximo en $1/a$.

Conclusiones: es la primera estimación econométrica de esa semielasticidad y prueba que la función es estable incluso en hiperinflación, pero la inestabilidad posible procede de las expectativas adaptativas y la función solo vale si los argumentos reales pueden ignorarse.

En conjunto, el monetarismo recupera una demanda estable y poco sensible al tipo de interés, confirmada por Cagan incluso en hiperinflación, pero ni Friedman ni Cagan la derivan de un programa, sino que postulan sus argumentos, y la de Friedman se ajusta mejor a agregados amplios.

Las teorías se ordenan, así, por la sensibilidad de la demanda al tipo de interés, h, y por su estabilidad, que con la sensibilidad de la inversión forman los dos eslabones de la transmisión: con Keynes, una h alta favorece la política fiscal, y con los monetaristas, una h baja y estable, la monetaria. Con $I = I_0 - v\,r$, los dos eslabones débiles keynesianos, v baja y h alta, hacen que una expansión fiscal eleve la renta sin apenas desplazar la inversión y que una monetaria apenas reduzca el tipo y estimule poco la inversión. En el caso monetarista ocurre lo contrario.

2. Desarrollos posteriores

La investigación reciente da a la demanda de dinero una base de agentes que optimizan a lo largo del tiempo en equilibrio general.

2.1. Generaciones solapadas (Samuelson, 1958)

Idea: Allais (1947), Samuelson (1958) y Diamond (1965) desarrollan el modelo de generaciones solapadas, y en su versión sin capital, la de Samuelson, los jóvenes quieren trasladar consumo a la vejez, pero no hay activo en que hacerlo. El dinero fiduciario resuelve el problema si cada generación espera que la siguiente lo acepte, y es depósito de valor e **institución social**.

Supuestos: cada generación vive dos periodos, con dotaciones $y_1 > y_2$, y la población crece a la tasa n. El *stock* nominal de dinero, en manos de los ancianos, crece a la tasa g_M.

CPO: el joven vende parte de su dotación a cambio de dinero, y la ecuación de Euler, con el dinero como único activo, iguala la RMS entre consumo joven y anciano a la rentabilidad real del dinero:

$$\frac{u'_1}{u'_2} = \frac{1}{1+\pi_{t+1}}, \qquad 1+\pi = \frac{1+g_M}{1+n}$$

– Los saldos demandados caen con la impaciencia y, si domina el efecto sustitución, con la inflación esperada, y crecen con la brecha entre dotaciones. En estado estacionario, la inflación es la diferencia entre el crecimiento del dinero y el de la población. Con ρ la tasa de preferencia temporal, las demandas individual y agregada, de los N_t jóvenes, son:

$$m^d_t = f(\pi^{e-}_{t+1}, \rho^-, y^+_1, y^-_2), \qquad \frac{M^d_t}{P_t} = N_t\, m^d_t$$

Equilibrio: puede haber uno no monetario, sin valor del dinero; uno monetario estacionario, y un continuo de equilibrios no estacionarios en que su valor se anula. El dinero tiene valor sin prestar servicios de transacción, y ese valor es una **burbuja racional**.

Extensiones:

– *Política económica*: crear dinero es un impuesto sobre los saldos de los jóvenes. Con $g_M = 0$, la rentabilidad real bruta del dinero es $1+n$ y el dinero sostiene la regla de oro, mientras que con $g_M > 0$ la inflación la aleja de ella, de modo que el dinero no es superneutral. En general, el dinero se aprecia si $g_M < n$ y su valor es constante si $g_M = n$.

Conclusiones: explica por qué se demanda un activo sin valor intrínseco, pero no por qué se demanda cuando hay otros con rendimiento positivo, y el equilibrio monetario es frágil.

2.2. Dinero en la utilidad y sus microfundamentos

Los tres modelos van de menor a mayor exigencia: en el MIU el dinero da utilidad, en el ST ahorra tiempo, y en el CIA sin dinero no hay compra.

2.2.1. MIU (Sidrauski, 1967)

Idea: para saber si la inflación altera la acumulación de capital, Sidrauski introduce los saldos reales en la utilidad de un modelo de crecimiento óptimo, porque proveen servicios de liquidez.

Planteamiento: con utilidad separable y marginales decrecientes, el consumidor elige dinero y bonos sujeto a su restricción presupuestaria, en la que la inflación erosiona los saldos heredados:

$$\max_{\{m_t, b_t\}} \sum_{t=0}^{\infty} \beta^t u(c_t, m_t), \qquad \beta \in (0,1)$$
$$\text{s.a.: } c_t = w_t + (1+r)b_{t-1} - b_t + \frac{1}{1+\pi_t} m_{t-1} - m_t$$

CPO: el cociente de utilidades marginales del dinero y del consumo iguala lo que cuesta, en consumo de hoy, conservar dinero, el interés nominal descontado:

$$\frac{u'_m}{u'_c} = \frac{i}{1+i} \;\to\; \frac{M^d}{P} = f(i^-, c^+)$$

– La demanda es análoga a la keynesiana, con el consumo como variable de escala. Una subida de i exige una utilidad marginal del dinero mayor, es decir, menos saldos reales, y un aumento del consumo eleva la demanda.

Superneutralidad: en el estado estacionario, ni el capital ni la producción dependen del dinero ni de su tasa de crecimiento, que solo determina la inflación y los saldos reales, aunque en la transición puede afectar al capital (Fischer, 1979). El único efecto real es el coste de bienestar de la inflación, que se anula con la **regla de Friedman**, $i = 0$.

Conclusiones: es la modelización dinámica de referencia, pero el dinero en la utilidad es un atajo que no explica por qué se demanda.

2.2.2. Tiempo de compra (McCallum y Goodfriend, 1987)

Idea: el dinero no da utilidad por sí mismo, sino que ahorra tiempo en las transacciones. Ese tiempo es ocio, que sí da utilidad, y así se microfundamenta el MIU.

Planteamiento: la utilidad depende del consumo y del ocio, y el tiempo de compra, $\psi(c_t, m_t)$, crece con el consumo y cae con los saldos reales:

$$l_t = 1 - h_t - \psi(c_t, m_t)$$

CPO: el valor en ocio del tiempo que ahorra una unidad más de dinero, relativo al valor neto del consumo, iguala $i/(1+i)$, la misma condición

del MIU, ahora con contenido. La demanda es $f(i^-, c^+, w^+)$, pues el salario real es el precio del tiempo ahorrado. Con la utilidad indirecta $v(c_t, m_t) \equiv u(c_t, 1 - h_t - \psi(c_t, m_t))$, la condición es:

$$\frac{v_m}{v_c} = \frac{-u_l\,\psi_m}{u_c - u_l\,\psi_c} = \frac{i}{1+i}$$

Estática comparativa: una mejora de la tecnología de pagos reduce la demanda, una lectura de la inestabilidad tras la innovación financiera. Con oferta de trabajo endógena, la inflación altera el reparto entre ocio y trabajo y el dinero deja de ser superneutral.

Extensiones: el modelo da lugar a dos casos polares. Si ψ no depende de m, el dinero no se demanda, y si el consumo exige dinero en proporción fija, resulta la restricción de efectivo por adelantado.

Conclusiones: conecta con los modelos de inventarios e introduce el salario, pero la tecnología de compra es tan *ad hoc* como la utilidad del dinero.

2.2.3. CIA (Clower, 1967; Lucas, 1980)

Idea: el dinero es intrínsecamente inútil y solo sirve para intercambiar bienes, así que su utilidad se sustituye por una restricción. Clower formula la restricción de liquidez, que Lucas integra en equilibrio general: "los bienes compran dinero y el dinero compra bienes, pero los bienes no compran bienes".

Planteamiento: el mercado de bienes abre antes que el de activos, y el consumo no puede exceder los saldos heredados (restricción de Clower, la segunda):

$$\max_{\{c_t, m_t, b_t\}} \sum_{t=0}^{\infty} \beta^t u(c_t) \quad \text{s.a.:} \begin{cases} c_t + m_t + b_t = w_t + (1+r)b_{t-1} + \frac{m_{t-1}}{1+\pi_t} \\ c_t \le \frac{m_{t-1}}{1+\pi_t} \\ b_{-1},\ m_{-1} \text{ dados} \end{cases}, \qquad \beta \in (0,1)$$

CPO: el valor de relajar la restricción es el tipo de interés nominal, de modo que con $i > 0$ se cumple con igualdad. La contrapartida del MIU es $\mu/u'(c) = i/(1+i)$, con μ el multiplicador de la restricción. Con la restricción activa, los saldos de hoy pagan el consumo de mañana:

$$m_t = (1 + \pi_{t+1})\, c_{t+1}$$

Solución: la demanda es $M^d/P = c$, la ecuación cuantitativa con velocidad unitaria y elasticidad-interés nula. El tipo nominal actúa como un impuesto sobre el consumo. En el estado estacionario el dinero es superneutral, salvo que la restricción se aplique a la inversión (Stockman, 1981), y la política óptima es de nuevo la regla de Friedman.

Conclusiones: da una demanda por transacción sin utilidad del dinero, pero la restricción se impone, la elasticidad-interés nula contradice la evidencia y el resultado depende del orden de apertura de los mercados.

En conjunto, MIU, ST y CIA toman como dado el motivo para demandar dinero y derivan $L(i^-, c^+)$, con elasticidad-interés nula en el CIA, la superneutralidad en el estado estacionario y la regla de Friedman como política óptima, con la limitación del agente representativo.

2.3. Otras aportaciones

Búsqueda (Kiyotaki y Wright, 1993): los agentes se cruzan al azar y el trueque exige la doble coincidencia de deseos. Como emparejarse con dinero es más rápido, hay equilibrio monetario aunque nadie valore el dinero *per se*. Complementa la explicación del depósito de valor, pero en su versión original el dinero es indivisible y los precios quedan fijos, sin análisis de la neutralidad. Lagos y Wright (2005), con dinero divisible y mercados descentralizados y centralizados alternos, lo convierten en un marco manejable para la inflación y la regla de Friedman.

Sustitución de monedas: el MIU con dos monedas da una demanda de moneda nacional que crece con la inflación extranjera, π^*, y cae con la nacional. Si la inflación nacional supera con holgura a la exterior, las carteras se inclinan hacia la divisa, que puede llegar a ser también unidad de cuenta (**dolarización**).

En suma, las generaciones solapadas y la búsqueda explican por qué se demanda dinero, pero, salvo la búsqueda desde Lagos y Wright, son poco operativas para la política, mientras que MIU, ST y CIA toman ese motivo como dado y se centran en la política monetaria.

3. Evidencia empírica e implicaciones de política económica

La sección recoge los hechos estilizados, las implicaciones de política y el debate sobre el dinero digital.

3.1. Hechos estilizados

Renta y riqueza: ambas influyen en los saldos demandados, y Meltzer (1963) encontró que la riqueza no humana explicaba la demanda mejor que la renta.

Tipos de interés: la demanda varía en sentido contrario a los tipos, aunque con poca elasticidad.

Trampa de liquidez: la evidencia la relaciona con el estancamiento de Japón desde finales de los noventa y con los tipos nulos de Estados Unidos entre 2009 y 2015.

Velocidad: los coeficientes estimados apenas cambiaron hasta mediados de los setenta, y desde 1974 la función sobrepredijo los saldos (Goldfeld, 1976). La innovación financiera de los ochenta hizo inestables la velocidad y la demanda: en EE. UU., la velocidad de M1 cayó en lugar de subir.

3.2. Implicaciones de política económica

Cuanto más estable y menos sensible al tipo de interés es la demanda de dinero, como en los clásicos y los monetaristas, más potente y predecible es la política monetaria. Cuanto más inestable y sensible, como en Keynes, más protagonismo gana la fiscal, sobre todo en la trampa de liquidez. Con la síntesis neoclásica, ambas son eficaces.

Los modelos de equilibrio general dinámico añaden la superneutralidad en el estado estacionario: la política monetaria solo determina la inflación y el coste de bienestar de los saldos reales, salvo que el dinero afecte a la producción o a la oferta de trabajo.

La inestabilidad de la función llevó a los bancos centrales a abandonar los agregados monetarios como objetivo intermedio.

3.3. Economía sin efectivo y dinero digital

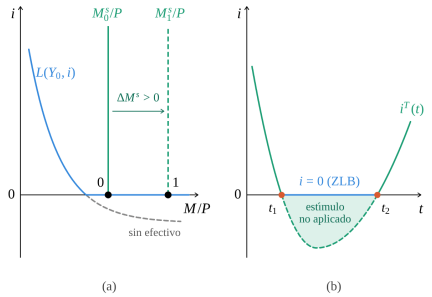
Origen: el debate, cuya formulación de referencia es la de Rogoff (2016), cobra fuerza cuando, tras la crisis financiera global, los tipos oficiales llegan a su límite inferior.

Idea: es dinero cualquier cosa con liquidez y credibilidad bastantes para ser medio de cambio, de modo que el dinero electrónico lo es igual que los billetes. La pregunta es si conviene eliminar el efectivo, y el argumento central es monetario.

Razonamiento: mientras exista efectivo, el tipo nominal no puede bajar sensiblemente de cero, porque los agentes atesorarían billetes antes que pagar por prestar. Con precios rígidos, ese suelo impide el tipo real negativo que vaciaría el mercado en una recesión profunda. Caben cuatro respuestas:

- Política monetaria no convencional, con riesgo para la estabilidad financiera.
- Política fiscal expansiva, que no está disponible para todos los países.
- Reformas estructurales que flexibilicen los precios, costosas y lentas.
- **Eliminación del efectivo**, que suprime el ZLB. Rogoff propone retirar primero los billetes de mayor denominación, lo que además haría aflorar economía sumergida.

Desarrollo gráfico: con efectivo, la demanda de dinero se vuelve horizontal en $i = 0$, porque a tipo nulo el efectivo, que rinde un 0 % nominal, y los bonos son sustitutos perfectos. Una expansión monetaria lleva el mercado del punto 0 al 1 sin reducir el tipo, porque los saldos se mantienen ociosos. Si el tipo objetivo de la regla de Taylor es negativo, el área entre él y cero mide el estímulo no aplicado. Sin efectivo desaparece el tramo horizontal y el tipo oficial podría bajar de cero.



Límite inferior de los tipos: la demanda de dinero se vuelve horizontal en $i = 0$ y el tipo oficial no puede seguir al tipo objetivo negativo.

Monedas digitales de banco central: el debate ha evolucionado hacia las CBDC, como el euro digital, que buscan preservar el acceso al dinero público sin eliminar el efectivo. Su diseño, remuneración y límites de tenencia, determina su efecto sobre los depósitos y la transmisión.

Conclusiones: eliminar el efectivo ataca directamente la causa del ZLB, sin depender de reformas lentas ni de margen fiscal, pero supone perder anonimato y resiliencia en los pagos, excluye a los menos digitalizados y concentra el poder de emisión.

En suma, la evidencia confirma los ingredientes de las teorías, con una función relativamente estable hasta mediados de los setenta y desestabilizada después por la innovación financiera, y el debate sobre el efectivo traslada la pregunta de la eficacia de las políticas al diseño del propio dinero.

1. Fundamentos del diseño de la política monetaria

Los fundamentos del diseño son el concepto de dinero, su oferta y su demanda, y el equilibrio entre ambas, que da lugar al proceso de creación de dinero.

1.1. Concepto de dinero

Definición conceptual: Hicks caracteriza el dinero por las funciones que desempeña y no por su naturaleza. Cumple cuatro:

- *Unidad de cuenta*: denomina bienes y activos y reduce drásticamente el número de precios relativos frente al trueque.
- *Medio de cambio*: se acepta generalmente para liquidar transacciones y supera la doble coincidencia de deseos.
- *Depósito de valor*: transfiere poder adquisitivo del presente al futuro.
- *Estándar de pago diferido*: en él se denominan las deudas.

Definición operativa: interesa a la política económica y se refiere a los **agregados monetarios**, pasivos del sistema monetario en manos del sector tenedor de dinero, ordenados por liquidez decreciente y ampliados históricamente por la innovación financiera. En la escala del Eurosistema:

- La **base monetaria** o dinero de alta potencia, $H = E + R^b$, es el pasivo del banco central: el efectivo en manos del público y las reservas bancarias.
- **M1**, $M_1 = E + D$, suma al efectivo los depósitos a la vista y constituye el dinero en sentido estricto, el que se identifica con la oferta monetaria, M^s. Las reservas forman parte de H, pero no de M_1, porque son un activo de los bancos y no dinero en manos del sector tenedor de dinero.
- **M2** añade los depósitos a plazo de hasta dos años y los disponibles con preaviso de hasta tres meses, el cuasidinero. Así, $M_2 = M_1 + D_2$, y una transferencia de una cuenta de ahorro a una cuenta corriente deja M_2 inalterado, pero eleva M_1.
- **M3** añade instrumentos negociables de las instituciones monetarias. Como la sustitución entre sus componentes apenas lo altera, el Eurosistema lo sigue en su análisis monetario. En $M_3 = M_2 + D_3$, D_3 son las cesiones temporales, las participaciones en fondos del mercado monetario y los valores de deuda con vencimiento hasta dos años.

1.2. Oferta de dinero: el sistema monetario

La oferta depende del sistema monetario, el conjunto de instituciones cuyos pasivos circulan como dinero: el banco central y la banca comercial.

Banco central: tiene la exclusiva de emitir moneda de curso legal, diseña y ejecuta la política monetaria, vigila a las entidades y es prestamista de última instancia. Su pasivo es la base monetaria, y su activo, las reservas internacionales y el crédito interno, al Gobierno y a la banca. Con patrimonio neto nulo, la base crece comprando activos exteriores o dando crédito interno. Su balance simplificado es:

$$\underbrace{R + \text{CI}}_{\text{activo}} = \underbrace{E + R^b}_{H}, \qquad \text{CI} = \text{AFG} + \text{AFSP}$$

Banca comercial: a diferencia de otros intermediarios, crea dinero mediante la **expansión múltiple de los depósitos**. Mantiene reservas legales, fijadas por el coeficiente de caja, y voluntarias, que dependen del tipo de interés y de la incertidumbre. Sus coeficientes suman el encaje, $q = q_L + q_V$, de modo que $R^b = qD$. Su pasivo son los depósitos y la financiación del banco central, y su activo, las reservas y el crédito al público. Las reservas garantizan la estabilidad y la credibilidad del sistema bancario, pues nada obliga a que entradas y salidas de efectivo coincidan. Su balance simplificado es:

$$\underbrace{R^b + \text{Crédito}}_{\text{activo}} = \underbrace{D + \text{AFSP}}_{\text{pasivo}}$$

Sistema monetario: al consolidar ambos balances se cancelan las partidas cruzadas. El pasivo es la oferta monetaria, M_1, y el activo, los activos exteriores netos y el crédito interno, los dos caminos por los que nace el dinero. El balance consolidado es:

$$\underbrace{E + D}_{M_1} = R + \underbrace{\text{AFG} + \text{Crédito}}_{\text{CI}}$$

1.3. Demanda de dinero

Idea: el banco central emite base monetaria, pero la cantidad de dinero que circula la determinan el público, que decide cuánto dinero mantener y en qué forma, y los bancos, que deciden sus reservas.

Enfoque teórico: las dos concepciones clásicas difieren en el papel de cada variable.

- *Keynesiana*, $M^d/P = kY - hi$: suma un componente transacción y otro especulación, decreciente con el tipo de interés, y en ella los bonos son los únicos sustitutivos del dinero. Es muy inestable, porque depende de expectativas convencionales y cambiantes, y dificulta la eficacia de la política monetaria por dos eslabones débiles: la elevada sensibilidad de la demanda de dinero al tipo de interés, con la trampa de liquidez en el límite, y la escasa sensibilidad de la inversión al tipo de interés.
- *Monetarista*: el dinero es un bien más, con muchos sustitutivos imperfectos, financieros y reales, y su demanda es estable y depende sobre todo de la renta permanente. Su forma es $M^d/P = f(Y^{P^+}, r_B^-, r_E^-, \pi^{e^-}, (W_H/W)^+)$, con los bonos, las acciones y los activos reales, cuyo rendimiento es la inflación esperada, como sustitutivos imperfectos.

Enfoque empírico: ambas admiten una forma reducida común, con Y la renta disponible del público, porque la riqueza, más atractiva en teoría, es difícil de estimar, e i el tipo nominal del activo sustitutivo:

$$\frac{M^d}{P} = L(Y^+, i^-), \qquad i = r + \pi^e$$

Las estimaciones para M_1 dieron una elasticidad-renta a largo plazo en torno a 0,7 y una elasticidad-interés reducida y estable, lo que sostuvo el control de los agregados. Su ruptura desde 1974, cuando la función sobrepredijo la demanda (*the case of the missing money*), es el antecedente del abandono de esa estrategia.

Demanda de base monetaria: el público reparte su dinero entre efectivo y depósitos según el **coeficiente de efectivo**, $c_M = E/M^s$, que depende de los costes de transacción, la bancarización y la confianza en los bancos. Los depósitos obligan a los bancos a mantener reservas en proporción q. La suma de efectivo y reservas es la demanda de base, $H^d = E^d + R^{b,d}$, y es frente a ella, y no frente a la demanda de dinero, donde el banco central fija su oferta.

1.4. Creación de dinero

Mercado de reservas: forma parte del interbancario, donde el tipo de interés iguala oferta y demanda de reservas. Es un mercado intervenido, porque el banco central actúa en él para influir en el tipo al que se relacionan todos los demás. La cadena se formaliza con tres modelos de complejidad creciente, en una economía cerrada sin imperfecciones. Sus agentes son el banco central, que emite H, el público, que decide la composición de su demanda de dinero, y los bancos, que fijan sus reservas en una proporción $q \in (0, 1)$ de los depósitos.

Depósito íntegro: toda la base se deposita. Cada banco guarda qH y presta el resto, que vuelve entero a otro banco, y así sucesivamente. Sumando las rondas, el **multiplicador bancario** es $b = 1/q > 1$:

$$D = \sum_{n=0}^{\infty} (1-q)^n H = \frac{1}{q} H$$

Ratio efectivo-depósitos: si el público mantiene efectivo en proporción $c_D = E/D$, el equilibrio del mercado de base, $E^d + R^{b,d} = H$, da $M^s = \frac{1+c_D}{c_D+q} H$. El multiplicador sigue siendo mayor que uno, pero depende también del público. Cuanto más efectivo se guarda fuera del sistema, el dinero debajo del colchón, menor es, y tiende a la unidad cuando todo el dinero es efectivo.

Ratio efectivo-dinero: con el coeficiente c_M, el mismo multiplicador toma su forma habitual, porque c_M se observa en los agregados:

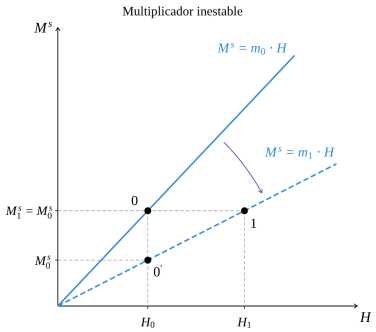
$$M^s = \underbrace{\frac{1}{c_M + q(1 - c_M)}}_{m = f(c_M,\, q^-)} H$$

Control de la oferta monetaria: cada unidad de base genera $m \geq 1$ unidades de dinero. El banco central fija H con operaciones de mercado abierto y q_L con los requerimientos de reservas, pero no controla las reservas voluntarias ni el coeficiente de efectivo. Su control es, por tanto, **indirecto e imperfecto**, tanto mejor cuanto más estable sea m, y el control de la propia base puede impedirlo un tipo de cambio fijo o la financiación del déficit.

Multiplicador estable e inestable: con c_M y q constantes, $M^s = mH$ es un rayo desde el origen más inclinado que la bisectriz, y una expansión de la base eleva el dinero con precisión, baja el tipo de interés y, en el IS-LM, eleva la renta. Una crisis de confianza, en la que el público retira efectivo y los bancos acumulan reservas, hace girar el rayo hacia abajo, de modo que con la misma base el dinero cae, y para mantenerlo hay que ampliar la base.

Así ocurrió tras 2008: los balances de los bancos centrales se multiplicaron mientras los agregados apenas crecían, con $m < 1$ en la Reserva Federal. Por eso el control de la base es un instrumento poco fiable para controlar el dinero.

Conclusiones: los modelos explican el funcionamiento del mercado monetario, base de todos los tipos de interés, y por qué el control del dinero es imperfecto. Episodios como la Gran Recesión o el escándalo del Libor muestran su importancia. Su limitación es que omiten el sector exterior. La intervención en el mercado de divisas altera los activos exteriores del banco central, y la esterilizada compensa su efecto sobre la base con



Multiplicador inestable: un aumento de c_M o de q hace girar el rayo hacia abajo y obliga a ampliar la base de H_0 a H_1.

el crédito interno. La no esterilizada puede sostener el tipo de cambio, pero choca con una política monetaria orientada a objetivos internos, y la eficacia de la esterilizada depende de la sustituibilidad imperfecta entre bonos nacionales y extranjeros.

2. Diseño de la política monetaria

El diseño se articula en cuatro elementos, y su evolución la ordenan dos debates: objetivos y variable operativa, y discrecionalidad y credibilidad. El modelo IS-PC-MR determina la política óptima.

2.1. Elementos del diseño

Definición: el **diseño de la política monetaria** es la concepción de la actuación de la autoridad monetaria sobre la cantidad y el precio del dinero, y tiene cuatro elementos.

Objetivo final: la variable, poco controlable directamente, que es el fin último: estabilidad de precios, estabilidad financiera, pleno empleo o redistribución. El BCE busca la estabilidad de precios, y la Reserva Federal, además, el máximo empleo y tipos a largo plazo moderados.

Objetivo intermedio: una variable controlable con un retardo razonable y precisión alta, con relación estable con el objetivo final, como los agregados, el tipo de cambio o los tipos a medio y largo plazo, variables que hoy se usan solo como indicadores.

Variable operativa: la que el banco central controla en el día a día, la base monetaria o el tipo interbancario a corto plazo: el €STR, sucesor del EONIA desde 2022, en el BCE, y el *fed funds rate* en la Reserva Federal. Hay consenso en que la apropiada es el tipo de interés.

Instrumentos: los convencionales, operaciones de mercado abierto, facilidades permanentes y requerimientos de reservas, y desde 2008 los no convencionales. Entre 1930 y 1980 hubo un cuarto, el método directo, con requerimientos de márgenes y techos al interés de los depósitos.

2.2. Objetivos y variable operativa

2.2.1. Doctrina de reservas (1920-1990)

Idea: arranca de la doctrina de la posición de reservas de la Reserva Federal en los años veinte y culmina en los objetivos de agregados de los setenta y ochenta. Con el esquema de objetivos e instrumentos de Tinbergen (1952), tiene **dos niveles**: la relación entre objetivo final e intermedio, y entre intermedio y variable operativa.

Elementos: el objetivo final es la estabilidad de la inflación promedio, sin cifra clara ni comunicación; el intermedio, los agregados monetarios, por el predominio de la visión monetarista; la variable operativa, la base monetaria, y los instrumentos, los convencionales. El razonamiento monetarista parte de la ecuación dinámica de la teoría cuantitativa con velocidad estable, de la que Friedman deriva una regla de crecimiento constante de la masa monetaria:

$$g_M + g_V = \pi + g_Y, \;\; g_V = 0 \;\rightarrow\; g_M = \pi + g_Y$$

Regla de Poole (1970): en un IS-LM estocástico, el instrumento depende del origen de los *shocks*. En rigor, depende de sus varianzas relativas y de las pendientes de la IS y la LM.

- Si el sector monetario es más inestable, $\sigma_M^2 > \sigma_R^2$, los *shocks* de demanda de dinero dominan y conviene controlar el tipo de interés.
- Si lo es el real, $\sigma_R^2 > \sigma_M^2$, conviene controlar los agregados, lo que en los setenta también apoyaba la evidencia. Los *shocks* del petróleo, sin embargo, fueron sobre todo de oferta, que el IS-LM de Poole no recoge.

Conclusiones: alcanzó su mayor relevancia entre 1979 y 1982, con Volcker, y redujo la inflación a costa del empleo y la producción. Pero el objetivo intermedio se controla mal, por los retardos, el conocimiento imperfecto de la transmisión y un vínculo inestable con el final, y los incentivos cortoplacistas degeneran en inconsistencia dinámica. La estrategia se aplicó en la mayoría de economías avanzadas desde mediados de los setenta.

2.2.2. Seguimiento directo de la inflación (1990-hoy)

Idea: la innovación y la liberalización financiera de los ochenta volvieron inestables la demanda de dinero y su velocidad, y llevaron al **objetivo directo de inflación**, adoptado primero por Nueva Zelanda (1990) y después por Canadá, Reino Unido, Suecia, Australia, el BCE desde 1999 con una estrategia afín, la Reserva Federal (2012) y el Banco de Japón (2013). El objetivo directo es una estrategia de **un nivel** (objetivo final y tipo a corto plazo, sin objetivo intermedio). También lo adoptaron numerosas economías emergentes, como Brasil, Chile, México, Polonia y Sudáfrica.

Objetivo final: la estabilidad de precios a medio plazo, cuantificada sobre un IPC, como objetivo puntual, en torno al 2 %, como banda o como punto con banda de tolerancia, del 1 al 3 % en Canadá y Nueva Zelanda. Debe ser verificable, flexible frente a *shocks* imprevistos y creíble, para anclar expectativas. Hay consenso sobre sus bondades: la estabilidad de precios evita los costes de la inflación y la deflación, mejora la señal de los precios relativos y reduce las primas de riesgo y las distorsiones en los impuestos y la Seguridad Social. El índice es el armonizado de precios de consumo en la UEM y el deflactor del consumo en EE. UU. El BCE cuantificó su objetivo por primera vez en 1998 y lo revisó en 2003, "por debajo, pero cerca, del 2 %", y en 2021, con un 2 % simétrico.

Objetivo intermedio: lo sustituye la **previsión de inflación** a unos dos años: si se desvía del objetivo, se ajusta el tipo (Svensson, 1997a). El BCE la formalizó en dos pilares, el análisis económico y el monetario, con un valor de referencia para M3 del 4,5 %, y la revisión de 2021 los sustituyó por un marco integrado. El análisis monetario era una herencia de la doctrina de reservas que ningún otro banco central mantuvo, y la revisión anual del valor de referencia se abandonó en 2003.

Variable operativa: el tipo de interés a corto plazo, con instrumentos convencionales y no convencionales.

Conclusiones: su transparencia ancla las expectativas y, si es creíble, reduce las ratios de sacrificio, además de ser robusta frente a la innovación financiera. Pero depende de la calidad de la previsión, puede ser procíclica ante *shocks* de oferta y relega la estabilidad financiera, la crítica más repetida tras 2008. Es procíclica porque un *shock* de oferta positivo reduce la inflación y la expansión resultante puede alimentar burbujas, mientras que uno negativo lleva a una contracción que agrava la recesión.

2.2.3. Seguimiento del PIB nominal

Idea: la estabilidad de la senda del PIB nominal, del orden del 4-5 % anual, habría permitido actuar antes en 2008 y reparte los *shocks* de oferta entre precios y producción, pero el PIB se mide con menos frecuencia y más revisiones y es más difícil de evaluar. El objetivo suma un 2 % de inflación y un 2-3 % de crecimiento real, y la noción de senda obliga a compensar los años malos, de modo que en recesión la inflación puede superar el 2 %.

2.2.4. Seguimiento del nivel de precios

Idea: una senda del nivel de precios obliga a compensar la inflación perdida. En el límite inferior eleva las expectativas de inflación y reduce el tipo real cuando el nominal no puede bajar (Eggertsson y Woodford, 2003). La Reserva Federal (2020) y el BCE (2021) incorporaron versiones atenuadas, pero exige un compromiso difícil de sostener, y tras la inflación de 2021-2023 ambos retiraron esos elementos de compensación en sus revisiones de 2025. La Reserva Federal volvió a un objetivo de inflación flexible, y el BCE sustituyó la tolerancia a desviaciones al alza por una respuesta enérgica o persistente ante desviaciones amplias y sostenidas en cualquier dirección.

2.3. Discrecionalidad y credibilidad

Idea: el banco central minimiza una **función de pérdida** sujeta a la estructura de la economía. La cuestión es si lo hace periodo a periodo o con una regla, y cómo hacer creíble el compromiso.

Discrecionalidad: da flexibilidad frente a *shocks* imprevistos, como la pandemia de COVID-19, pero genera **inconsistencia dinámica** (Kydland y Prescott, 1977), porque con un objetivo de *output* superior al natural, hay incentivo a la inflación sorpresa una vez fijadas las expectativas. Con expectativas racionales, el resultado es un **sesgo inflacionista** sin ganancia de *output* (Barro y Gordon, 1983a).

Reglas: atan las manos de la autoridad, desde el crecimiento constante del dinero de Friedman hasta una regla de tipos. La más influyente es la **regla de Taylor** (1993):

$$i_t = \underbrace{\bar{r}_t + \pi^T}_{\tilde{i}_t} + \alpha_\pi \underbrace{(\pi_t - \pi^T)}_{\tilde{\pi}_t} + \alpha_y \underbrace{(y_t - \bar{y}_t)}_{\tilde{y}_t}$$

$$\tilde{i}_t = \alpha_\pi \tilde{\pi}_t + \alpha_y \tilde{y}_t$$

- Si la inflación supera el objetivo, el tipo nominal debe subir más que proporcionalmente, $\alpha_\pi > 1$, para que suba el real (el **principio de Taylor**).
- Las reglas facilitan la comprensión pública, la gestión de expectativas y la delegación en una institución evaluable. Sus costes son la rigidez

y, en la de Taylor, su carácter *backward looking* y la falta de microfundamentos. Hay versiones *forward looking*, con la inflación esperada (Clarida et al., 2000).

Mecanismos de credibilidad: entre la discrecionalidad y la regla rígida hay tres.

- *Reputación* (Barro y Gordon, 1983b): en un juego repetido, con estrategias de castigo y un banco central que valore el futuro, la reputación resuelve la inconsistencia sin regla formal. El castigo consiste en que los agentes elevan sus expectativas de inflación tras una sorpresa.
- *Delegación* (Rogoff, 1985): un banquero central conservador reduce el sesgo a costa de estabilizar peor el *output*. La independencia se correlaciona con menos inflación sin coste en *output* (Alesina y Summers, 1993). Los países de la OCDE han adoptado esta solución.
- *Contratos* (Walsh, 1995; Svensson, 1997b): penalizan la inflación o asignan un objetivo inferior al óptimo social, y alcanzan el resultado bajo compromiso, frente al *second best* de Rogoff. El objetivo de inflación con rendición de cuentas es su versión práctica.

2.4. Política óptima: el modelo IS-PC-MR

Planteamiento: el banco central elige el tipo de interés real que minimiza su pérdida intertemporal, sujeto a la IS dinámica y a la curva de Phillips neokeynesiana:

$$\min_{r_t} \ \mathbb{L}_0 \sum_{t=0}^{\infty} \beta^t \left[(\pi_t - \pi^T)^2 + \omega \tilde{y}_t^2\right], \qquad \beta \in (0,1)$$

$$\text{s.a.:} \begin{cases} \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}\left(r_t - \bar{r}_t\right) \\ \pi_t - \pi^T = \beta \mathbb{E}_t[\pi_{t+1} - \pi^T] + \kappa \tilde{y}_t + \varepsilon_t^{\pi} \end{cases}$$

- ω es el peso de la brecha de producción, $\bar{r}_t$ el tipo real natural, $1/\theta$ la sensibilidad de la brecha al tipo de interés, κ la pendiente de la curva de Phillips, ε_t^{π} un *shock* de costes y β el factor de descuento.

CPO: bajo discrecionalidad, el banco central toma las expectativas como dadas y obtiene la **regla monetaria**, MR. Bajo compromiso, la condición depende también de la brecha pasada:

$$\text{Discrecionalidad:} \quad \tilde{y}_t = -\frac{\kappa}{\omega}(\pi_t - \pi^T)$$

$$\text{Compromiso:} \quad \tilde{y}_t - \tilde{y}_{t-1} = -\frac{\kappa}{\omega}(\pi_t - \pi^T)$$

La MR señala ante cada *shock* el punto de la curva de Phillips que minimiza la pérdida, y la IS lo traduce en un tipo de interés con la forma de Taylor.

CSO: la pérdida es cuadrática y las restricciones lineales, de modo que la CPO es necesaria y suficiente y la política óptima es única.

Solución: del programa se desprenden tres resultados.

1. El instrumento es el tipo de interés real, que el banco central fija a través del nominal y de las expectativas de inflación, y la cantidad de dinero queda determinada por la demanda. La doctrina actual se deriva del modelo.
2. El peso ω define el tipo de banco central: con ω alto es "paloma" y prioriza el *output*, con una MR muy inclinada, mientras que con ω bajo es "halcón" y prioriza la inflación, con una MR plana.
3. Con un objetivo de brecha positivo, $\tilde{y}^T > 0$, la pérdida pasa a ser $(\pi_t - \pi^T)^2 + \omega(\tilde{y}_t - \tilde{y}^T)^2$ y aparece el sesgo inflacionista de Barro-Gordon, que con expectativas racionales solo produce inflación. Además, bajo compromiso el banco central promete responder a los *shocks* de hoy también mañana, la **dependencia de la historia**, que ancla expectativas, reduce el coste de estabilización (Clarida et al., 1999) y fundamenta el *forward guidance*.

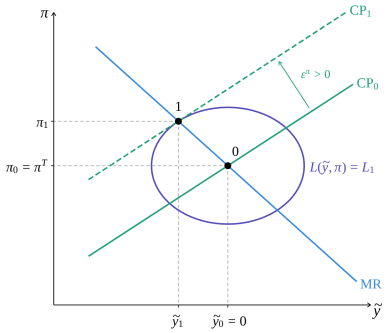
Desarrollo gráfico: en el plano $(\tilde{y}, \pi)$, la curva de Phillips es creciente, con pendiente κ, y sin perturbaciones pasa por el objetivo con brecha nula. Las curvas de isopérdida rodean ese punto, $(0, \pi^T)$, y la pérdida crece al alejarse de él. La MR, decreciente, con pendiente $-\omega/\kappa$, une sus tangencias con las curvas de Phillips. Sin perturbaciones, el equilibrio tiene brecha nula e inflación en el objetivo.

Estática comparativa: un *shock* de costes desplaza la curva de Phillips hacia arriba, y la respuesta óptima es el punto de la MR donde la nueva curva es tangente a una isopérdida, que reparte el ajuste entre más inflación y menos *output* según ω. El objetivo de inflación resuelve esa disyuntiva con bandas y un horizonte de medio plazo.

En suma, el diseño moderno descansa en un objetivo de estabilidad de precios cuantificado, en el tipo a corto plazo como variable operativa, por el control imperfecto del dinero, y en compromisos creíbles que resuelven la inconsistencia dinámica.

3. Instrumentación de la política monetaria

Si el diseño fija qué objetivo alcanzar, el **marco operativo** determina cómo, con instrumentos convencionales, no convencionales desde 2008 y la pandemia, y cualitativos.



Regla monetaria óptima: el banco central elige sobre la curva de Phillips el punto que minimiza su función de pérdida.

3.1. Instrumentos convencionales

3.1.1. Operaciones de mercado abierto

Definición: operaciones de inyección o drenaje de liquidez, generalmente por subasta y contra colateral, que son el principal instrumento mediante la fijación del tipo oficial. Sirven para señalizar la posición del banco central, gestionar la liquidez y controlar los tipos. En el Eurosistema solo acuden las entidades de contrapartida autorizadas, y desde octubre de 2008 las MRO se adjudican a tipo fijo con adjudicación plena.

Instrumentación: abarca cinco tipos.

- *Transacciones revertidas*, las más habituales: liquidez temporal contra colateral que se recompra o revende al vencimiento, como las MRO semanales y las LTRO del Eurosistema.
- *Transacciones en firme*: compras o ventas definitivas, para operaciones estructurales.
- *Certificados de deuda* del banco central, que drenan liquidez.
- *Swaps de tipos de cambio*, que inyectan o drenan moneda nacional sin alterar la posición cambiaria. Combinan la compra al contado de divisa con su venta simultánea a plazo, o a la inversa.
- *Depósitos a plazo fijo*, para absorber excesos puntuales. El BCE los usó semanalmente entre 2010 y 2014 para esterilizar las compras del SMP.

Efectos: alteran la base y, vía multiplicador, la oferta monetaria (*efecto cantidad*), y la cotización y el tipo de los valores comprados (*efecto precio*), porque el rendimiento corriente de un bono es inverso a su precio:

$$r_B = \frac{\text{cup}}{P_B} \cdot 100$$

Conclusiones: son flexibles en magnitud y dirección, pero su efecto depende del multiplicador, exigen mercados de colateral profundos y pierden eficacia en el límite inferior de los tipos, donde el dinero y los activos a corto plazo pasan a ser casi sustitutos perfectos.

3.1.2. Facilidades permanentes

Definición: permiten a las entidades, por su iniciativa, obtener o depositar liquidez a un día en condiciones fijas y generalmente desfavorables. Su objetivo es acotar las fluctuaciones del tipo interbancario a un día y proveer o absorber liquidez de emergencia. La de crédito sigue la regla de Bagehot (1873), según la cual el banco central debe prestar libremente, a tipos penalizadores y contra buen colateral.

Efectos: la **facilidad de crédito**, la marginal del Eurosistema o la ventanilla de descuento de la Reserva Federal, fija un techo a los tipos interbancarios, y la **de depósito**, un suelo, que en la Reserva Federal forman la remuneración de las reservas y la facilidad de *reverse repo*. Entre ambas delimitan el **pasillo** en el que se sitúa el tipo de las operaciones principales. Con liquidez abundante, el tipo se pega al suelo y el pasillo degenera en un **sistema de suelo**, el de la Reserva Federal desde 2008 y el del Eurosistema de hecho desde 2015 y de derecho desde su marco operativo de 2024. Ese marco estrecha el pasillo, sitúa el tipo de las MRO 15 pb por encima del de la facilidad de depósito y hace de este el tipo que guía la orientación de la política, con liquidez abundante suministrada a demanda.

Conclusiones: acotan el tipo a un día sin que el banco central intervenga y son una red de seguridad ante necesidades puntuales. Pero acudir a la de crédito puede estigmatizar a la entidad, y con liquidez abundante el control del tipo pasa a depender de la remuneración de las reservas, cuyo coste crece con el balance.

3.1.3. Coeficiente de caja

Definición: cada banco debe mantener en el banco central una fracción de sus depósitos, $q_L = R_L^b/D$. Estabiliza la demanda de liquidez, porque basta cumplirlo en media, y crea un déficit estructural de liquidez que refuerza el control de los tipos. Ese dinero queda inmovilizado, y el banco central fija el coeficiente y decide si remunera las reservas mínimas.

Efectos: subirlo reduce el multiplicador y la oferta monetaria. El Euro-

sistema lo sitúa en el 1 % desde 2012, sin remunerarlo desde 2023, y la Reserva Federal lo eliminó en 2020.

Conclusiones: es coactivo y preciso y afecta a la vez a todas las entidades sujetas, pero es de acción discontinua, sin remunerar grava a la banca frente a los intermediarios no sujetos y ha perdido popularidad. La competencia regulatoria entre países para atraer fondos explica también esa pérdida.

3.2. Instrumentos no convencionales

Intentan restaurar los canales de transmisión dañados por la crisis financiera, por tres causas: la crisis de confianza, con primas de riesgo crecientes en la periferia; la ruptura de la transmisión, y el límite inferior de los tipos. En ese límite, los agentes absorben el exceso de liquidez como saldos ociosos y los aumentos de la base no reducen el tipo a corto plazo. Todas estas medidas se pensaron como transitorias, con una retirada gradual ligada a la normalización.

Compras masivas de activos: la **expansión cuantitativa** (QE) compra activos escogidos en gran volumen para actuar sobre los tipos a largo plazo y las primas de riesgo. Opera por la transmisión directa, el reequilibrio de carteras y la señalización de tipos bajos prolongados. La inició el Banco de Japón en 2001; la Reserva Federal encadenó QE1 a QE3 desde 2008 y superó los 8,9 billones de dólares de balance con la pandemia; el BCE agrupó sus programas en el APP, con el QE propiamente dicho desde 2015, el PSPP, y añadió el PEPP en 2020, hasta 1,85 billones. Las compras netas cesaron en 2022 y desde 2023 el balance se reduce (*quantitative tightening*). Permiten seguir relajando cuando el tipo oficial ya no puede bajar, pero exponen el balance a pérdidas cuando suben los tipos y acercan la política monetaria a la financiación del sector público. Las pérdidas surgen porque la remuneración de las reservas creadas puede superar el rendimiento de los bonos comprados.

OMA no convencionales: destacan por su plazo o condicionalidad. Las LTRO a tres años (2011-2012) inyectaron más de un billón de euros, pero la liquidez se empleó en sanear balances. De ahí las **TLTRO** (2014-2021), con un tipo más favorable cuanto más crédito a empresas y hogares, y las PELTRO de la pandemia. Sostienen la financiación bancaria a plazos que el mercado ya no ofrece, pero su efecto sobre el crédito depende de la demanda de préstamos y de los balances bancarios. Las LTRO a tres años ofrecían un tipo en torno al 1 %, una amplia gama de colaterales y ninguna estigmatización, y las PELTRO, siete operaciones de 2020, un tipo 25 pb inferior al de las MRO. La financiación prolongada mantiene, además, la dependencia de la banca respecto del banco central.

Políticas innovadoras de tipos: los **tipos negativos**, iniciados por el Riksbank en 2009 y aplicados por el BCE entre 2014 y 2022, muestran que el límite efectivo está bajo cero y amplían el margen de bajada, pero con rendimientos decrecientes y daño a la rentabilidad bancaria, y por debajo de cierto umbral (*reversal rate*) la bajada contrae el crédito. El **sistema dual** (BCE, 2019) exime del tipo negativo parte de las reservas y, con las TLTRO, permite tipos distintos para prestamistas y prestatarios, aunque desdibuja la frontera con la política fiscal. Genera un efecto renta positivo sobre acreedores y deudores a la vez, y lo anticiparon el BNS (2015) y el BoJ (2016). En Suiza los tipos negativos, del −0,75 % entre 2015 y 2022, respondieron a las presiones apreciatorias sobre el franco, activo refugio.

Recomposición del balance: cambia la composición del activo, no necesariamente su tamaño. La **relajación del crédito** sustituye unos activos por otros y la **operación *twist*** de la Reserva Federal (2011-2012) vende bonos a corto y compra a largo para aplanar la ETTI. Actúan sobre plazos concretos sin ampliar necesariamente el balance, pero su alcance lo acotan los activos que el banco central puede vender, y la relajación del crédito genera ganadores y perdedores.

Líneas de *swap*: permutas de divisas entre bancos centrales que proveen dólares a los bancos de otras áreas cuando se cierra su financiación. Las de la Reserva Federal con los principales bancos centrales son permanentes desde 2013, y en marzo de 2020 volvieron a ser decisivas. Abaratan la financiación en dólares fuera de Estados Unidos sin que la Reserva Federal asuma riesgo de crédito ni cambiario, pero solo alcanzan a los bancos centrales con los que ella acuerda. El banco central extranjero entrega su moneda a cambio de dólares al tipo de mercado, se compromete a recomprarla al mismo tipo en un plazo de un día a tres meses, más un interés, y asume el riesgo de crédito al prestarlos a sus bancos.

3.3. Instrumentos cualitativos y transversales

Comunicación: el *forward guidance* informa sobre el curso futuro de la política. Puede ser cualitativo, basado en el calendario o dependiente del estado de la economía, y si es creíble reduce los tipos a largo sin mover el oficial. La orientación **délfica**, que solo anuncia una previsión, puede leerse como señal de debilidad, frente a la **odiseica**, que compromete la actuación futura (Campbell et al., 2012). La Reserva Federal pasó de la orientación cualitativa (2008) a la de calendario (2011) y a la de umbrales (2012), y el BCE la incorporó en julio de 2013. La declaración de Draghi en 2012, dispuesto a hacer "todo lo necesario para preservar el euro", no fue *forward guidance* sobre tipos, sino un compromiso verbal que precedió a las OMT. La comunicación actúa sin coste para el balance y sigue disponible en el límite inferior, pero su eficacia depende de la credibilidad, porque el compromiso es dinámicamente inconsistente. Junto a ella figura la *persuasión moral*, cuyo ejemplo paradigmático es el rescate privado de Long-Term Capital Management que orquestó la Reserva Federal de Nueva York en 1998 sin aportar financiación.

Control del crédito: topes y coeficientes de inversión obligatoria dirigen fondos a sectores prioritarios. Actúan con rapidez y de forma selectiva, pero distorsionan la asignación y la desintermediación erosiona su eficacia, como mostró el *corset* del Banco de Inglaterra (1973-1980).

Rebaja de las calificaciones: el Eurosistema rebajó el umbral del colateral de A- a BBB- en 2008 y admitió los *fallen angels* en 2020, para que una decisión de las agencias no endureciera de forma automática la liquidez. Pero los recortes de valoración solo compensan en parte el mayor riesgo, y suspender el umbral para Grecia (2010) supuso juzgar el riesgo de un soberano, una decisión casi fiscal. Los activos que pierden la calificación exigida siguen siendo admisibles con recortes de valoración mayores, como los *fallen angels* hasta BB.

Compras condicionadas: la disposición anunciada a comprar deuda soberana sin límite *ex ante*, con condiciones, descarta crisis de deuda autocumplidas y reduce las primas sin comprar. Son las **OMT** (2012), para bonos de uno a tres años de países con programa del MEDE, nunca activadas pero eficaces para reducir las primas, y el **TPI** (2022), para países cuya prima se aleje de los fundamentos y cumplan los criterios fiscales de la UE. La condicionalidad limita el riesgo moral, pero puede retrasar la ayuda a quienes más la necesitan, y la compra de deuda de Estados concretos expone al banco central a impugnaciones por financiación monetaria, como las OMT ante el TJUE. En 2015 (asunto *Gauweiler*), el TJUE confirmó que los Tratados autorizan un programa de las características de las OMT.

3.4. Debates recientes

Dinero helicóptero (Friedman, 1969): inyectar liquidez directamente a los consumidores en lugar de comprar bonos. Ningún banco central lo ha aplicado, porque difumina la frontera entre política fiscal y monetaria. Rebautizado como *QE for the people*, se contempla como último recurso frente a la deflación sistémica.

Monedas digitales: eliminar el efectivo (Rogoff, 2016) suprimiría el límite inferior y permitiría tipos negativos, a costa del anonimato y de la exclusión. Las CBDC minoristas, como el euro digital, cuya fase de preparación concluyó en 2025 y cuyo reglamento negocian desde julio de 2026 el Parlamento Europeo y el Consejo, se plantean como complemento del efectivo, con límites de tenencia y sin remuneración para evitar la desintermediación. Por eso rinden un 0 % nominal y mantienen el límite inferior.

En suma, a los instrumentos que operan sobre el tipo a corto plazo se han sumado otros sobre el tamaño y la composición del balance y sobre las expectativas. El límite inferior de los tipos no agota la política monetaria, pero los instrumentos que lo sortean la acercan a la fiscal.

A37. Política monetaria (II). Los mecanismos de transmisión de la política monetaria convencional. Rigideces de precios, rigideces de salarios y fricciones financieras. Los mecanismos de transmisión de la política monetaria no convencional

1. Transmisión de la política monetaria convencional: canales

El **mecanismo de transmisión** es el sistema de canales por los que los impulsos del banco central llegan desde las variables instrumentales hasta los objetivos finales. Tiene tres etapas: el impulso, con operaciones de mercado abierto, facilidades o coeficientes; la transmisión por los canales, empezando por el tipo a corto plazo, y los efectos sobre la inflación y la actividad, que el banco central no puede predecir ni controlar del todo.

1.1. Canal de los tipos de interés

Idea: el tipo a corto plazo, que el banco central guía con sus tipos oficiales, es la variable operativa. Por la ecuación de Fisher, $i = r + \pi^e$, si las rigideces hacen que la inflación esperada se ajuste lentamente, una bajada del tipo nominal reduce el tipo real. Si persiste, mueve también los tipos largos, que son una media de los cortos actuales y esperados, relación que depende de las expectativas recogidas en la ETTI. Una pendiente positiva indica, según la hipótesis pura de las expectativas, que se esperan subidas, aunque con prima por plazo una pendiente moderada es compatible con tipos esperados constantes.

Efectos: el cambio de los tipos oficiales se transmite por tres vías.

- *Sustitución*: abarata el consumo presente. Por la ecuación de Euler, el consumo crece a una tasa aproximada de $(r-\rho)/\theta$, de modo que un menor tipo real lo adelanta.
- *Renta*: eleva la renta disponible de los deudores y reduce la de los acreedores. Prevalece la primera, porque los deudores tienen mayor propensión al consumo.
- *Componentes permanentes*: reduce el coste de uso del capital, $P_K(r+\delta)$, eleva el *stock* deseado y la inversión, y el gasto en duraderos a crédito.

Su intensidad depende de las elasticidades del gasto, del peso de cada componente, de la posición financiera neta y de que los tipos sean fijos o variables. A largo plazo, al ajustarse los precios, el efecto sobre r se disipa. En el modelo IS-PC-MR, estos efectos se resumen en la curva IS.

1.2. Canal del crédito (Bernanke y Gertler, 1995)

Idea: la política monetaria puede aumentar el crédito bancario, al ampliar el pasivo de los bancos o el valor del colateral de los prestatarios. Bernanke y Gertler distinguen el canal en sentido estricto y el canal en sentido amplio, y un modelo sencillo de riesgo moral recoge ambos.

1.2.1. Canal en sentido estricto: el crédito bancario

Una política expansiva amplía los pasivos de los bancos, que pueden conceder más préstamos, y estos elevan la inversión, el consumo a crédito y la renta nominal. El BCE lo ha utilizado con las LTRO, incluidas las de tres años de 2011-2012, y con las TLTRO, vinculadas a la concesión de préstamos a la economía real. Esquemáticamente:

$$\uparrow M^s \to \uparrow \text{pasivos bancarios} \to \uparrow \text{préstamos} \to \uparrow I, C \to \uparrow PY$$

1.2.2. Canal en sentido amplio: el balance de los prestatarios

Idea: el **acelerador financiero** recoge el efecto del precio de los activos y de las imperfecciones del crédito, por costes de agencia y asimetrías informativas. Opera por dos vías:

- Un menor tipo reduce las tasas de descuento, eleva el valor presente de los flujos de caja y de las empresas, y con él el colateral y el crédito. Esquemáticamente:

$$\downarrow i \to \uparrow \text{valor presente} \to \uparrow \text{colaterales} \to \uparrow \text{crédito} \to \uparrow PY$$

- La compra de activos por el banco central, como el APP y el PEPP del BCE, eleva su precio, y por la misma cadena, el crédito.

Asimetría: las pymes lo sufren más que las grandes empresas, que tienen financiación alternativa, y también los sectores con más riesgo, como los innovadores. Donde predomina la tradición anglosajona de mercados de capitales gana peso el efecto riqueza; en la continental, el canal bancario. La Unión de los Mercados de Capitales (2015), relanzada en 2025 como Unión de Ahorros e Inversiones, busca reducir esa dependencia bancaria.

1.2.3. Modelo de Holmström y Tirole (1997)

Idea: Holmström y Tirole formalizan los dos canales con riesgo moral. Si el empresario puede desviar parte del rendimiento, los inversores solo financian lo que el proyecto puede comprometerles, y la inversión depende del patrimonio neto y del coste de los fondos, dos variables que mueve la política monetaria.

Supuestos: un empresario con responsabilidad limitada aporta su patrimonio, PN, e invierte I, que rinde RI si hay éxito. Si se esfuerza, el éxito tiene probabilidad p_H. Si no, $p_L = p_H - \Delta p$ y obtiene un beneficio privado BI. Los inversores, neutrales al riesgo, piden $1+r$, con r fijado por la política monetaria.

Planteamiento: para que se esfuerce, el empresario ha de recibir en caso de éxito al menos $BI/\Delta p$, de modo que a los inversores solo se les puede prometer la **renta comprometible**, $\rho_0 I$. Con $\rho_1 = p_H R$:

$$\max_I \ [\rho_1 - (1+r)]\, I \quad \text{s.a.:} \quad \rho_0\, I \geq (1+r)(I - \text{PN})$$

CPO: con $\rho_1 > 1 + r > \rho_0$, el multiplicador es positivo, la restricción se satura y el empresario invierte cuanto puede financiar. Una unidad de patrimonio invertida rinde más que prestada, y esa diferencia es la **prima de financiación externa**, que aquí adopta la forma de racionamiento.

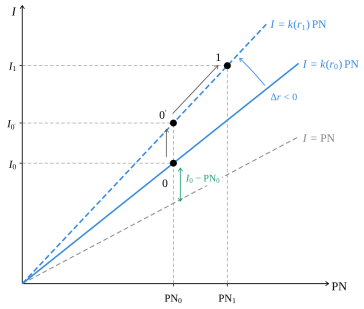
CSO: el objetivo y la restricción son lineales en I, por lo que la condición de Kuhn-Tucker es suficiente.

Solución: la inversión es proporcional al patrimonio:

$$I^* = k(r)\,\text{PN}, \qquad k(r) = \frac{1}{1 - \rho_0/(1+r)} > 1$$

Desarrollo gráfico: en el plano patrimonio neto-inversión, la *recta de autofinanciación*, $I = \text{PN}$, recoge la inversión financiable solo con el patrimonio, y la *capacidad de inversión*, $I = k(r_0)\,\text{PN}$, sale del origen con pendiente $k(r_0) > 1$, tanto mayor cuanto menor es el tipo, porque un tipo mayor reduce la financiación externa respaldada por la renta comprometible. Con el patrimonio PN_0, el empresario invierte $I_0 = k(r_0)\,\text{PN}_0$ (punto 0).

Estática comparativa: una bajada del tipo eleva la inversión por dos vías. Aumenta el **multiplicador de los fondos propios**, $k(r)$, y, con el mismo patrimonio, se invierte más; y revaloriza los activos, de modo que el patrimonio crece y la inversión lo hace en $k(r_1)\,\Delta\text{PN}$, más que el propio patrimonio, que es la amplificación del canal amplio.



Canal del crédito en el modelo de Holmström y Tirole: la bajada del tipo gira la capacidad de inversión (0→0′) y la revalorización del patrimonio la eleva (0′→1).

Extensiones: con bancos que supervisan y deben arriesgar capital propio, las empresas de patrimonio intermedio dependen del capital bancario, el canal estricto. Cualquier endurecimiento, como un *credit crunch*, recae sobre las peor capitalizadas.

Conclusiones: con un único supuesto, el riesgo moral, explica por qué la inversión depende del patrimonio y del capital bancario y por qué los efectos son amplificados y asimétricos, pero es estático, de equilibrio parcial, y toma como dado el efecto sobre el precio de los activos.

1.3. Riqueza, tipo de cambio y q de Tobin

1.3.1. Canal de la riqueza

Idea: la política monetaria altera la riqueza por su volumen o por la valoración de los activos, con el sustrato del ciclo vital (Modigliani y Brumberg, 1954; Ando y Modigliani, 1963) y la renta permanente (Friedman, 1957). Las compras de activos del BCE, APP y PEPP, actúan también por este canal.

- *Efecto riqueza directo*: más liquidez, dados los precios, eleva los saldos reales, la riqueza y el consumo. Es decir, $\uparrow M^s/P \to \uparrow W \to \uparrow C$.
- *Efecto Pigou*: si suben los precios, cae el valor real de los activos nominales y el consumo, en sentido contrario. Es decir, $\uparrow M^s \to \uparrow P \to \downarrow A/P \to \downarrow W \to \downarrow C$.
- *Reajuste de cartera*: el exceso de liquidez se coloca en activos financieros y reales, que se revalorizan y sirven de colateral, lo que eleva crédito y gasto.

1.3.2. Canal del tipo de cambio

Idea: con tipo flexible y plena movilidad del capital, una expansión reduce el tipo de interés, provoca salidas de capital y deprecia la moneda. Por la paridad descubierta, un tipo nacional inferior al exterior exige una apreciación esperada, de modo que la moneda se deprecia hoy por encima de su nivel de largo plazo (**overshooting** de Dornbusch).

Efectos: la depreciación eleva la competitividad-precio y la demanda si se cumple Marshall-Lerner, y encarece los *inputs* importados y eleva la inflación. El efecto neto sobre la cuenta corriente es empírico, porque los menores tipos elevan también las importaciones. En el IS-PC-MR de economía abierta, este canal permite al banco central mover menos el

tipo. Más allá del tipo de cambio, la política monetaria afecta también a la balanza de pagos por los flujos de capital aunque no haya riesgo cambiario, como dentro de la zona del euro, y en el límite su reversión brusca da lugar a *sudden stops*.

1.3.3. Canal de la q de Tobin

Idea: una expansión rebaja los tipos y encarece las acciones, de modo que sube la q de Tobin (A34 §2.2.2). Con $q > 1$ basta emitir pocas acciones para financiar mucha inversión, es decir, $\downarrow i \rightarrow\uparrow V \rightarrow\uparrow q \rightarrow\uparrow I$, con V el valor de mercado de las empresas.

1.4. Canal de las expectativas

Idea: los tipos a medio y largo plazo dependen de los cortos esperados, y el banco central influye en las expectativas señalizando el curso futuro de la política. Iterando la IS dinámica neokeynesiana, la brecha de producción depende de toda la senda esperada de tipos reales:

$$\tilde{y}_t = -\frac{1}{\theta}\sum_{k=0}^{\infty}\mathbb{E}_t\left[r_{t+k} - \bar{r}_{t+k}\right]$$

Por eso anunciar tipos futuros más bajos estimula el gasto actual aunque el tipo oficial no cambie (el *forward guidance*). Sus efectos son difíciles de predecir: una subida de tipos puede leerse como señal de crecimiento robusto o como freno al crecimiento. Estas expectativas llegan a la economía real por las decisiones de inversión y de consumo de bienes duraderos. El canal fue fundamental en la crisis del euro, cuando el *whatever it takes* de Draghi (2012) estabilizó las expectativas y redujo las primas de riesgo.

1.5. Canal de la redistribución

Idea: la visión tradicional supone propensiones al consumo parecidas, pero las fricciones financieras y la distinta liquidez las hacen heterogéneas, de modo que, al repartir ganancias y pérdidas, la política monetaria altera el consumo agregado aunque la renta total no cambie.

Vías: Auclert (2019) distingue tres, que probablemente amplifican el efecto.

- *Heterogeneidad de las rentas*: la expansión eleva más el gasto si responden más las rentas de los hogares con mayor propensión al consumo.
- *Canal de Fisher*: la inflación no anticipada traslada riqueza de los acreedores a los deudores, que gastan más.
- *Exposición a los tipos*: una bajada beneficia a quienes revisan antes sus deudas que sus activos, como con hipotecas a tipo variable.

En suma, con precios rígidos la bajada del tipo reduce el tipo real y estimula el gasto. El crédito lo amplifica, sobre todo en los peor capitalizados, y la riqueza, el tipo de cambio, la q de Tobin, las expectativas y la redistribución lo extienden.

2. Efectos de la política monetaria: de la neutralidad a las rigideces y fricciones financieras

Hasta el consenso actual sobre la neutralidad a largo plazo, las escuelas discutieron si la política monetaria es efectiva y qué efectos reales produce.

2.1. Teoría clásica: la neutralidad del dinero

Teoría cuantitativa: la ecuación de cambio, identidad popularizada por Fisher (1911), iguala el valor de las compras y el de las ventas, con V la velocidad-transacción del dinero y T el número de transacciones, valoradas al precio P de una transacción representativa:

$$MV \equiv PT$$

En la versión de Cambridge (Pigou, 1917), $M^d = kPY$ con $k = 1/V$ constante, lo que en equilibrio, $M^s = M^d = M$, equivale a $MV = PY$. En su lectura causal, con V y la producción estables, los cambios en M se trasladan proporcionalmente a P, de modo que la inflación es un fenómeno monetario.

Efectividad: el dinero es un velo y rige la **dicotomía clásica**: las variables nominales no afectan a las reales y la política monetaria es inefectiva.

Conclusiones: explica la relación a largo plazo entre dinero e inflación, que la evidencia respalda, pero descansa en una velocidad estable, que la demanda especulativa y la innovación financiera cuestionan, y no explica los efectos a corto plazo.

2.2. Keynesianismo, síntesis y monetarismo

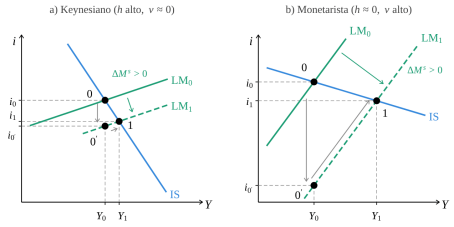
2.2.1. Keynesianismo

Idea: a raíz de la Gran Depresión, Keynes (1936) formula la preferencia por la liquidez. La demanda especulativa de dinero rompe la dicotomía, porque la política monetaria afecta a las variables reales a través del tipo de interés, pero con poca eficacia por dos **eslabones débiles**: una inversión poco sensible al tipo, dominada por los *animal spirits*, $v \approx 0$, y una demanda de dinero muy sensible, h alto, que en el límite da la trampa de liquidez.

Razonamiento: en el IS-LM, una expansión desplaza la LM y tiene tres efectos.

- *Efecto liquidez*: el exceso de dinero se coloca en bonos y el tipo baja, aunque poco.
- *Efecto renta*: el menor tipo eleva la inversión y la renta, aunque poco, por su baja sensibilidad.
- *Efecto expulsión*: la mayor renta eleva la demanda de dinero por transacción y el tipo rebota, lo que modera la subida de la renta.

Desarrollo gráfico: en el plano (Y, i), panel a, la IS es muy inclinada, porque la inversión apenas responde al tipo, y la LM casi horizontal, porque la demanda de dinero es muy sensible a él. Por eso la política monetaria es poco eficaz sobre la renta y la inversión.



Política monetaria expansiva en el IS-LM: efectos liquidez (0→0′) y renta y expulsión (0′→1) en el caso keynesiano (a) y en el monetarista (b).

Conclusiones: da al tipo de interés un papel en la transmisión y la trampa de liquidez anticipa el problema del límite inferior, pero su escasa eficacia descansa en elasticidades que el monetarismo rechaza, y con precios fijos no distingue corto y largo plazo.

2.2.2. Síntesis neoclásica

Idea: con la lectura de Samuelson y Solow (1960), la curva de Phillips se convierte en un **menú** de política, en el que las autoridades pueden comprar un paro más bajo con más inflación.

Razonamiento: con $\pi_t = f(u_t)$ y $f' < 0$, la política monetaria es eficaz en cualquier horizonte. Si mantiene en el tiempo una inflación estable y más alta, el desempleo queda de forma permanente en un nivel más bajo.

Conclusiones: ofrece una relación estimable entre inflación y desempleo, pero la disyuntiva permanente exige ilusión monetaria, y la estanflación de los setenta mostró que la curva se había desplazado.

2.2.3. Monetarismo

Idea: Friedman (1968) y Phelps (1967) introducen las expectativas en la curva de Phillips. La política monetaria tiene efectos reales elevados a corto plazo, pero a largo plazo se cumple la neutralidad.

Razonamiento: la crítica monetarista tiene dos vertientes.

- *Oferta*: la disyuntiva permanente exigiría trabajadores con ilusión monetaria. Con expectativas adaptativas, $\pi^e_t = \pi^e_{t-1} + \lambda(\pi_{t-1} - \pi^e_{t-1})$, la curva de Phillips se amplía con las expectativas:

  $$\pi_t = \pi^e_t - \kappa_u(u_t - \bar{u}), \qquad \kappa_u > 0$$

  A largo plazo, con $\pi_t = \pi^e_t$, es vertical sobre la tasa natural, de modo que ninguna inflación sostenida altera el paro (**superneutralidad**).
- *Demanda*: no hay eslabones débiles. La inversión es muy sensible al tipo, IS elástica, y la demanda de dinero, que depende de la renta permanente y tiene muchos sustitutivos, apenas lo es, LM inelástica. El exceso de dinero se gasta en toda clase de activos y la renta sube mucho a corto plazo: en el IS-LM, los efectos liquidez, renta y expulsión son elevados.

Conclusiones: explica por qué la disyuntiva solo es transitoria, pero descansa en expectativas adaptativas, con errores sistemáticos, y con expectativas racionales la política anticipada no tiene efectos reales ni a corto plazo.

2.2.4. Efectividad de la política monetaria

Para Keynes, la política monetaria es poco eficaz, por los dos eslabones débiles, y para la síntesis, eficaz en cualquier horizonte.

Reglas: los monetaristas la reconocen eficaz a corto plazo, pero quieren atarla a reglas y dedicarla a controlar la inflación, por tres argumentos: la ecuación perdida, que deja sin determinar el reparto entre precios y producción; los retardos largos y variables (Friedman, 1961), que pueden volverla procíclica, y su inefectividad a largo plazo, por las expectativas adaptativas.

2.3. NMC: vuelta a la no intervención

2.3.1. Consideraciones generales

Supuestos: microfundamentos dinámicos de equilibrio general, competencia perfecta con precios y salarios flexibles, **expectativas racionales** e información incompleta: los agentes no distinguen si un cambio de su precio es específico o general (problema de extracción de señales). De ellos se deriva la **oferta de Lucas**:

$$y_t = \bar{y}_t + \zeta(p_t - p^e_t) + \varepsilon^s_t, \qquad \zeta > 0$$

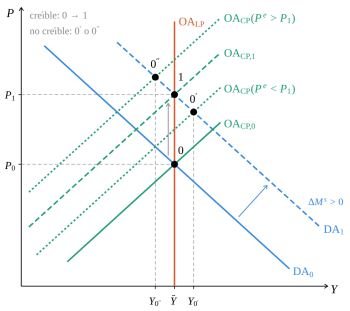
Aquí $\bar{y}_t$ es el *output* natural, ε_t^s un *shock* de oferta de esperanza nula y ζ combina la parte del cambio de precio que se atribuye a precios relativos y la respuesta de la oferta. Los *shocks* monetarios, como la política no anticipada o las perturbaciones de la velocidad, actúan a través de la sorpresa de precios, $p_t - p_t^e$, de modo que solo afectan a la producción en la medida en que son sorpresivos.

La NMC revitaliza, así, el pensamiento clásico con un instrumental mucho más sofisticado.

2.3.2. Sargent y Wallace (1975, 1976)

Razonamiento: si la política sigue una regla conocida, los agentes la incorporan a sus expectativas y solo eleva los precios. Solo la parte no anticipada tiene efectos reales, y transitorios. Un anuncio creíble de expansión desplaza la demanda y la oferta a corto plazo a la vez, y solo suben los precios. Si no es creíble, la sorpresa tiene efectos expansivos o contractivos según los agentes esperen precios menores o mayores.

Desarrollo gráfico: en el plano renta-precios, la oferta a largo plazo es vertical sobre el *output* natural, $\bar{Y}$. La oferta a corto plazo es creciente, una por cada nivel de precios esperado, y la demanda agregada, decreciente, porque unos precios más altos reducen los saldos reales. En el punto 0, con $\bar{Y}$ y P_0, las expectativas se cumplen.



Crítica de Sargent y Wallace: una expansión creíble y anticipada solo eleva los precios (0→1); si no es creíble, tiene efectos reales transitorios (0 o 0).

Aun así, la NMC no recomienda las sorpresas, porque reducen la credibilidad futura y aumentan la incertidumbre.

Conclusiones: separa los efectos de la política anticipada y de la no anticipada, con el respaldo de los primeros contrastes (Barro, 1977), y refuerza el argumento a favor de reglas sencillas y predecibles. Pero exige precios y salarios plenamente flexibles. Con contratos nominales plurianuales (Fischer, 1977), la política anticipada tiene efectos reales, y Mishkin (1982) halla que el dinero anticipado también afecta a la producción.

2.3.3. Inconsistencia dinámica (Kydland y Prescott, 1977)

Idea: una política es **dinámicamente inconsistente** cuando, fijados los contratos, a la autoridad le conviene incumplirla. Los agentes lo anticipan y no es creíble. Si la autoridad persigue un paro inferior al natural, resulta un **sesgo inflacionista** sin ganancia de empleo (Barro y Gordon, 1983a), y la eficacia depende de la credibilidad (de ahí las reglas, la independencia, el banquero conservador y los contratos). El banquero conservador procede de Rogoff (1985) y los contratos de desempeño, de Walsh (1995).

Conclusiones: explica el sesgo sin suponer irracionalidad y fundamenta los objetivos de inflación, pero exige que la autoridad persiga un paro inferior al natural, discutible en un banco central independiente, y es estático.

Para la NMC, la política discrecional solo tiene efectos por sorpresa y, anticipada, da más inflación con el mismo empleo.

2.4. NEK: vuelta al intervencionismo

2.4.1. Rigideces y fricciones financieras

Idea: para la NEK, la no neutralidad no exige sorpresas: aun con expectativas racionales, bastan competencia imperfecta, información imperfecta y contratos duraderos.

Rigidez de precios: con contratos escalonados (Taylor, 1980) el nivel de precios hereda inercia. En Calvo (1983), cada periodo una empresa cambia su precio con probabilidad $1-\xi$, y al agregar se obtiene la **curva de Phillips de la NEK**:

$$\pi_t = \beta\,\mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t, \qquad \kappa > 0,\ \beta \in (0,1)$$

- Los precios duran por término medio $1/(1-\xi)$ periodos. Cuanto mayor es ξ, menor es κ, más plana la curva y más recae el impulso sobre la producción. Con precios flexibles, $\xi \to 0$, κ tiende a infinito, la curva se vuelve vertical y se recupera la neutralidad, y con $\xi \to 1$, κ tiende a cero. Además, la curva es prospectiva, así que la política actúa también por las expectativas.
- Los **costes de menú** (Mankiw, 1985) completan el cuadro: no ajustar el precio es una pérdida privada de segundo orden, pero la suma tiene efectos de primer orden.

Rigidez de salarios: los contratos salariales escalonados explican buena parte de la persistencia de un *shock* monetario (Christiano et al., 2005). Con salarios rígidos *à la* Calvo (Erceg et al., 2000) se pierde la **divina coincidencia**, es decir, el banco central ya no puede estabilizar a la vez la brecha y las inflaciones de precios y salarios. La razón es que un *shock* de productividad exige ajustar el salario real moviendo precios o salarios nominales, ambos costosos.

Rigideces reales: la escasa sensibilidad del precio o salario real deseado a la actividad no rompe por sí sola la neutralidad, pero con fricciones nominales pequeñas genera una rigidez nominal sustancial (Ball y Romer, 1990). En la curva de Phillips, la rigidez real reduce κ para un ξ dado.

Fricciones financieras: la prima de financiación externa crece al caer el patrimonio y la política la mueve en el mismo sentido que el tipo, lo que da el acelerador financiero (Bernanke, Gertler y Gilchrist, 1999). Así, el ajuste es lento y vía cantidades y se refuerzan los canales de la sección 1. Cuando el sistema financiero está dañado, el banco central puede actuar directamente sobre las primas y la oferta de crédito, que es la lógica de la política no convencional.

2.4.2. IS-PC-MR y desinflación

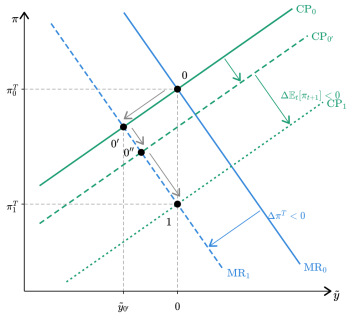
Planteamiento: el modelo se compone de la IS dinámica, la CP de la NEK y la regla de Taylor, que llevada a la IS da la MR, aquí en su versión óptima bajo discrecionalidad:

$$\begin{aligned}
\text{IS:}\quad & \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(r_t - \bar{r}_t)\\
\text{CP:}\quad & \pi_t = \beta\,\mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t\\
\text{Regla:}\quad & \tilde{i}_t = \alpha_\pi\,\tilde{\pi}_t + \alpha_y\,\tilde{y}_t, \qquad \alpha_\pi > 1\\
\text{MR:}\quad & \tilde{y}_t = -\frac{\kappa}{\omega}(\pi_t - \pi^T)
\end{aligned}$$

La IS recoge el canal de los tipos de interés, con $r_t = i_t - \mathbb{E}_t[\pi_{t+1}]$, y en la MR ω es el peso de la brecha en la pérdida del banco central y π^T el objetivo de inflación.

Desarrollo gráfico: en el plano brecha-inflación, la CP es creciente, con pendiente κ. Con β próximo a la unidad, corta la vertical de brecha nula a la altura de $\mathbb{E}_t[\pi_{t+1}]$. La MR es decreciente y pasa por el objetivo con brecha nula, porque ante más inflación el banco central sube el tipo real y contrae la demanda. En el punto 0, la inflación y la esperada coinciden con π_0^T y el tipo real iguala al natural.

Estática comparativa: al reducir el objetivo de π_0^T a π_1^T, la MR baja. Con un anuncio **creíble**, la inflación esperada y la CP bajan a la vez y la economía pasa de 0 a 1 sin brecha, por el canal de las expectativas. Con **credibilidad imperfecta**, la CP sigue en CP_0, el banco central sube el tipo real por encima del natural y la economía pasa a 0 , con brecha negativa. Al revisarse las expectativas, la CP baja y la economía converge a 1. Los puntos de producción perdidos por cada punto de inflación reducida miden la **ratio de sacrificio**.



Reducción del objetivo de inflación en el modelo IS-PC-MR con credibilidad imperfecta. La MR baja (0→0) y la revisión de las expectativas baja la CP hasta el nuevo equilibrio (1).

Extensiones: un banco central que se compromete a responder también en el futuro ancla las expectativas y reduce el coste de estabilizar.

Conclusiones: separa los dos canales que llevan a la inflación, la brecha (vía tipo real) y las expectativas, y explica por qué la credibilidad abarata la desinflación. Pero con la CP prospectiva una desinflación creíble no cuesta producción, en contra de la evidencia. La inercia de la CP híbrida, que añade la inflación pasada, la encarece incluso con credibilidad, y el modelo prescinde del crédito y del límite inferior.

Para la NEK, la política monetaria puede evitar que pérdidas microeconómicas de segundo orden se conviertan en pérdidas macroeconómicas de primer orden y ayudar a estabilizar el *output*, aunque a largo plazo sus efectos se diluyen al ajustarse las expectativas.

En suma, la eficacia depende de la flexibilidad de precios y salarios y de

las expectativas: el menú permanente de la síntesis se reduce al corto plazo con el monetarismo y a las sorpresas con la NMC, y la NEK devuelve a la política anticipada efectos reales transitorios.

3. Evidencia empírica y política no convencional

Además de la teoría, la autoridad monetaria atiende a la evidencia: limitaciones generales y específicas, hechos según el horizonte y la transmisión de la política no convencional.

3.1. Metodología

Los instrumentos son endógenos, porque el banco central los ajusta a la evolución de la economía, de modo que la simple correlación entre tipos o dinero y *output* no prueba la no neutralidad. Se recurre a los **vectores autorregresivos estructurales** (SVAR) y al enfoque narrativo (Romer y Romer, 1989), que identifica los *shocks* a partir de registros históricos.

3.2. Limitaciones generales

Incertidumbre: los bancos centrales no conocen del todo los mecanismos de transmisión. Por mucho que se hayan modelizado, la transmisión y el funcionamiento mismo de la economía solo se conocen en parte, y las decisiones pierden eficacia por ello.

Retardos: son largos y variables. Tras una subida de tipos, los precios tocan su mínimo a los 29 meses de media (Havranek y Rusnak, 2013). El desfase interno incluye el de reconocimiento, por el retraso de la estadística y la duda sobre si la perturbación es transitoria, y el de acción, poco acusado porque la política monetaria es ágil, al contrario que la fiscal. El externo va de la medida a sus efectos. Los retardos pueden volver procíclica una política anticíclica.

Globalización: en una economía pequeña integrada en los mercados de capitales, los tipos dependen de los internacionales y las políticas se vinculan por los efectos desbordamiento, de ahí foros como el G20 o Jackson Hole. Tras 2008, la expansión cuantitativa de la Reserva Federal tuvo efectos desbordamiento sobre las economías emergentes, y Rajan (2014) la criticó como una expansión monetaria competitiva, con beneficios sobre todo internos y costes sobre todo en el exterior.

3.3. Limitaciones específicas

Asimetría: con la "metáfora de la cuerda", la restrictiva es más eficaz que la expansiva (Cover, 1992), porque la expansión no asegura más gasto. En la primera etapa de liquidez ilimitada del BCE, los bancos sanearon balances en lugar de prestar. Retirar dinero recorta de forma directa la capacidad de gasto del sector privado, mientras que inyectarlo no garantiza que aumente, y el resultado depende, entre otros factores, del multiplicador bancario.

Límites institucionales: tampoco la restrictiva está exenta de dificultades, porque suscita presiones sociales. Si los impuestos recaen en las rentas medias y bajas y la deuda está en manos de las altas, subir los tipos beneficia a estas. El crecimiento cae, por la inversión y por la apreciación, y a las pymes les cuesta más financiarse.

Velocidad inestable: la innovación financiera debilitó la relación entre dinero y precios y llevó a abandonar los agregados como objetivo intermedio. La mayoría de los bancos centrales de economías avanzadas pasó a objetivos directos de inflación o a estrategias mixtas.

Aritmética monetarista desagradable (Sargent y Wallace, 1981): con una política fiscal sistemáticamente expansiva y un tipo real superior al crecimiento, la deuda acabará monetizándose y, como los agentes lo anticipan, la autoridad no controla la inflación ni siquiera hoy. Con tipo de cambio fijo y libre circulación de capitales, el **trilema** de Mundell-Fleming suprime la autonomía monetaria.

3.4. Hechos observados según el horizonte

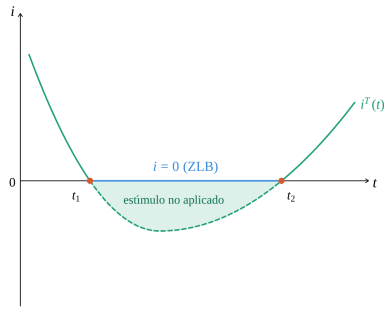
Corto plazo: las fases de expansión monetaria rápida tienden a ir por delante de las de mayor crecimiento del PIB real. Los periodos de inflación más alta coinciden, en general, con un crecimiento real más intenso, y los recortes de tipos se anticipan a las expansiones del *output*. Estas correlaciones no prueban causalidad.

Largo plazo: la correlación entre crecimiento monetario e inflación roza la unidad, 0,92-0,96 según el agregado (McCandless y Weber, 1995), lo que avala el origen monetario de la inflación. La del dinero con el PIB real es nula en la muestra completa, aunque positiva en la OCDE, lo que desmiente la disyuntiva permanente de la síntesis y respalda la dicotomía clásica.

3.5. Transmisión de la política no convencional

3.5.1. El límite inferior efectivo

Idea: la Reserva Federal llevó su tipo al límite inferior a finales de 2008, y el BCE lo alcanzó entre 2012 y 2014. Entonces el canal de los tipos se bloquea aunque la regla prescriba recortes. El límite es un **límite inferior efectivo** algo bajo cero, porque por debajo de cierto tipo negativo se prefiere el efectivo, que rinde un 0 % nominal. Si cae la inflación esperada, el tipo real sube y puede haber una espiral deflacionista. Las operaciones de mercado abierto ordinarias pierden además eficacia, porque con tipos nulos el dinero y las letras del Tesoro son casi sustitutos perfectos, y en 2008-2012 se sumó un canal del crédito dañado.



Límite inferior de los tipos: entre t_1 y t_2 el tipo de la regla de Taylor es negativo, pero el oficial no baja de cero; el área es el estímulo no aplicado.

Prima por plazo: las medidas no convencionales actúan sobre los demás determinantes de un tipo largo, los tipos cortos esperados y la prima por plazo:

$$i_t^{(n)} \approx \frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}_t[i_{t+k}] + \rho_t^{(n)}$$

3.5.2. Canales de las medidas no convencionales

Compras de activos: operan por el reequilibrio de carteras y la duración, que reduce la prima por plazo y desplaza a los vendedores hacia otros activos, siempre que no sean sustitutos perfectos, porque si lo fueran, sin otras fricciones, las compras solo actuarían por las expectativas (teorema de Wallace); por la señalización de tipos bajos prolongados; por la liquidez en episodios de tensión, como en marzo de 2020, cuando el banco central actúa como comprador de último recurso, y por el tipo de cambio, con efectos desbordamiento sobre otras economías. Las compras refuerzan además la riqueza y el colateral.

Forward guidance: comprometerse a mantener el tipo en el límite más tiempo del que prescribe la regla (Eggertsson y Woodford, 2003) rebaja los tipos esperados. Exige credibilidad, porque el compromiso es dinámicamente inconsistente, y sus efectos observados fueron muy inferiores a los de los modelos, el *forward guidance puzzle* (Del Negro et al., 2023).

Tipos negativos: se transmiten como un recorte, pero estrechan el margen bancario. Por debajo del **tipo de reversión** reducen el capital de los bancos y contraen el crédito (Brunnermeier y Koby, 2018). Ese umbral es más alto cuanto menores son la cartera de renta fija y el capital de los bancos, y sube con el tiempo, por lo que conviene agotar antes los recortes. El BCE introdujo en 2019 la remuneración de las reservas en dos tramos para aliviar ese coste.

Financiación condicionada: las TLTRO actúan sobre el crédito en sentido estricto. Prestan a los bancos a varios años a un tipo tanto menor cuanto más crédito conceden a empresas no financieras y hogares, salvo para la compra de vivienda, y en la TLTRO III ese tipo llegó a situarse 50 pb por debajo del de la facilidad de depósito. Condicionan así la liquidez a que llegue a la economía real, algo que no aseguraron las LTRO a tres años.

3.5.3. Evidencia sobre su eficacia

El efecto es sustancial sobre los activos y moderado sobre la actividad. Las compras de la Reserva Federal de 2008-2010 redujeron la prima por plazo a diez años entre 30 y 100 pb, y la caída de los tipos largos se debió sobre todo a menores primas de riesgo y no a menores tipos esperados (Gagnon et al., 2011), aunque el efecto depende de los activos comprados (Krishnamurthy y Vissing-Jorgensen, 2011). Los anuncios del APP rebajaron el rendimiento soberano a diez años entre 30 y 50 pb, y en torno al doble en Italia y España (Altavilla et al., 2015), y las medidas del BCE elevaron el crecimiento del PIB en unos 0,6 puntos anuales en 2015-2020, y sin ellas la inflación habría sido unos 0,4 puntos más baja (Altavilla et al., 2021). Combinar instrumentos resultó más eficiente que usar uno solo. Krishnamurthy y Vissing-Jorgensen hallan apoyo para los canales de señalización, de demanda de activos seguros y de inflación, pero poco para el de duración.

En suma, el límite inferior no agota la política monetaria, pero hace su transmisión más incierta, porque depende de la credibilidad, de la segmentación de los mercados y de la salud de los bancos. Por eso la estrategia del BCE de 2021 exigió, cerca de ese límite, medidas especialmente enérgicas o persistentes, y la evaluación de 2025 lo reformuló de forma simétrica, ante desviaciones amplias y duraderas en cualquier sentido.

1. Efectos vía gasto público no productivo y gravamen

El análisis parte del modelo IS-LM, que sirve de antecedente, y pasa después a las perspectivas de la nueva macroeconomía clásica y de la nueva economía keynesiana.

1.1. Antecedentes: el modelo IS-LM

1.1.1. Efectos de la política fiscal

Idea: a corto plazo, con precios fijos, la renta la determina la demanda agregada, y una expansión fiscal desplaza la IS a la derecha con dos efectos.

Efecto multiplicador: el mayor gasto, o los menores impuestos, elevan la renta a través del multiplicador. Con la función de consumo keynesiana, $C = C_0 + c(Y - T_0)$:

- El multiplicador del gasto, a tipo de interés constante, es $1/(1-c) > 1$, y el de los impuestos, $-c/(1-c)$, menor en valor absoluto porque una rebaja solo eleva la demanda en la parte de la renta que se consume. Ambos dependen de la teoría del consumo, y la renta permanente y el ciclo vital los reducen.
- El **multiplicador del presupuesto equilibrado**, suma de ambos, es la unidad (Haavelmo, 1945).
- En el IS-LM, cuando el tipo de interés se ajusta, el multiplicador es menor, por el efecto expulsión:

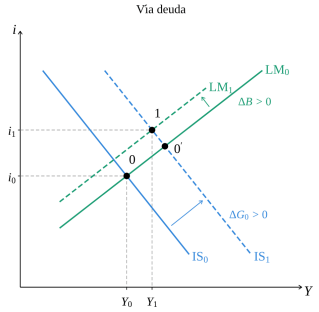
$$\frac{\partial Y}{\partial G_0} = \frac{1}{1 - c + vk/h}$$

Efecto expulsión: la mayor renta eleva la demanda de dinero por transacción, los agentes venden bonos, su precio cae y sube el tipo de interés, lo que expulsa gasto privado. Es decir, $\downarrow B^d \rightarrow \downarrow P_B \rightarrow \uparrow i$.

Otros efectos: en economía abierta, el desbordamiento hacia las importaciones reduce el multiplicador. Según la financiación:

- *Impuestos*: reducen el multiplicador, pero no lo anulan, porque el del presupuesto equilibrado es positivo.
- *Deuda*: añade el **efecto cartera**. La mayor oferta de bonos eleva la rentabilidad exigida y desplaza la LM a la izquierda, con más expulsión.
- *Emisión de dinero*: añade el **efecto liquidez**, que desplaza la LM a la derecha y evita la expulsión, aunque para los monetaristas se diluye a largo plazo.

Desarrollo gráfico: en el plano (Y, i), la expansión desplaza la IS de IS_0 a IS_1 y eleva la renta y el tipo. Con deuda, la mayor oferta de bonos desplaza además la LM a la izquierda y el tipo sube aún más, con más expulsión.



Financiación vía deuda: la expansión fiscal expulsa parcialmente la inversión.

1.1.2. Debate

La importancia de cada efecto depende de las pendientes de las curvas, es decir, de cuánto responden al tipo de interés la inversión, v, y la demanda de dinero, h. En el multiplicador del IS-LM, k es la sensibilidad de la demanda de dinero a la renta.

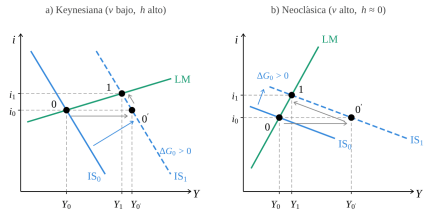
Perspectiva keynesiana: la inversión, gobernada por los instintos animales, apenas responde al tipo, IS rígida, y el motivo especulación hace muy elástica la LM, hasta la trampa de liquidez. La política fiscal es muy efectiva: multiplicador elevado, expulsión reducida y multiplicador del gasto superior a la unidad.

Perspectiva neoclásica: el gasto privado reacciona con fuerza al tipo real, IS elástica, y la demanda de dinero, con muchos sustitutivos, es rígida, $h \approx 0$. La política fiscal es poco efectiva: multiplicador reducido, expulsión total o casi total y multiplicador del gasto inferior a la unidad.

Desarrollo gráfico: a tipo constante, el multiplicador llevaría la economía del punto 0 al 0 , pero la subida del tipo la devuelve al punto 1. En la perspectiva keynesiana, la subida apenas reduce la inversión y Y_1 queda cerca de $Y_{0'}$, y en la neoclásica expulsa casi tanta inversión como gasto añade y Y_1 apenas supera a Y_0.

1.2. Perspectiva NMC: ciclo real con sector público

Idea: la lectura neoclásica del IS-LM se formaliza hoy con un modelo del ciclo real ampliado con sector público, en la línea de Baxter y King (1993).



Expansión fiscal en el IS-LM según las pendientes: la expulsión, de $Y_{0'}$ a Y_1, es pequeña en la perspectiva keynesiana (a) y casi total en la neoclásica (b).

1.2.1. Supuestos

Entorno: economía cerrada con *shocks* tecnológicos y fiscales, horizonte infinito, competencia perfecta, mercados completos y plena flexibilidad de precios, salarios y tipo de interés.

Agentes: muchos agentes idénticos con expectativas racionales y vida infinita, es decir, con altruismo intergeneracional perfecto. El sector público realiza gasto no productivo, recauda impuestos de suma fija y financia su déficit con deuda, de modo que los impuestos no distorsionan la elección consumo-ocio, aunque la afectan por el efecto riqueza.

Tecnología: Cobb-Douglas con progreso técnico neutral en el sentido de Hicks, rendimientos constantes a escala y depreciación δ. La tecnología crece a un ritmo exógeno, g_A, y sus desviaciones de la tendencia siguen un AR(1):

$$\tilde{a}_t = \rho_A \tilde{a}_{t-1} + \phi_t, \qquad \rho_A \in (-1, 1)$$

1.2.2. Modelo

Hogares: maximizan su utilidad intertemporal en consumo y ocio sujetos a su restricción presupuestaria:

$$\max_{\{c_t, l_t, k_{t+1}, b_{t+1}\}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, 1 - l_t), \qquad \beta \in (0,1)$$

$$\text{s.a.:} \begin{cases} c_t + k_{t+1} - (1-\delta)k_t + b_{t+1} - b_t = y_t - T_t + r_t b_t \\ k_0,\ b_0 \text{ dados} \end{cases}$$

Las CPO dan la condición ocio-consumo, $\text{RMS}_{c,1-l} = w_t$, y las ecuaciones de Euler de los bonos y del capital, cuya ausencia de arbitraje exige que rindan lo mismo:

$$u'(c_t) = \beta \mathbb{E}_t[u'(c_{t+1})(1 + r_{t+1})]$$

Empresas: con una Cobb-Douglas, igualan la productividad marginal de cada factor a su precio, con beneficio nulo:

$$f'_{l,t} = (1-\alpha)\frac{y_t}{l_t} = w_t, \qquad f'_{k,t} = \alpha \frac{y_t}{k_t} = r_t + \delta$$

y, en términos netos, $f'_{k,t} - \delta = r_t$, la rentabilidad del capital de la ecuación de Euler de los hogares.

Sector público: el déficit primario más los intereses de la deuda se financian con emisión neta de bonos, y los *shocks* de gasto siguen un AR(1), cuya persistencia, ρ_g, distingue los transitorios de los permanentes:

$$g_t - T_t = b_{t+1} - (1 + r_t) b_t$$

Equilibrio: los mercados de trabajo y de capital se vacían en todos los periodos, de modo que en cada momento la inversión iguala al ahorro.

Estática comparativa: si el gasto aumenta por sorpresa, los agentes saben que tendrán que pagarlo con más impuestos y su renta permanente cae. El *output* no crece por la demanda agregada, sino por el **canal riqueza**: los individuos, más pobres, ofrecen más trabajo. Las empresas no piden más horas, porque el gasto no altera la productividad marginal del trabajo.

Dinámica: depende de la permanencia del *shock*.

- *Transitorio*: el efecto riqueza es pequeño, la producción y las horas suben poco y caen el consumo, por el efecto riqueza, y la inversión, por el efecto expulsión, de modo que el multiplicador está entre cero y uno.
- *Permanente*: el efecto riqueza es mayor y la producción sube más. A ello se suma el efecto ampliación de Baxter y King (1993). Con más empleo, el capital gana productividad marginal, sube el tipo real y se estimula la inversión, y el multiplicador puede superar la unidad.

Extensiones: el modelo deja cuatro.

- *Equivalencia ricardiana* (Barro, 1974): los bonos no son riqueza neta y, dada la senda de gasto, financiarlo con impuestos o deuda es irrelevante, porque el desahorro público se convierte en ahorro privado. Con todo, la política fiscal sigue teniendo efectos reales, y las

limitaciones de la equivalencia, como las restricciones de liquidez, la ilusión fiscal, el altruismo imperfecto, los impuestos distorsionantes y el gasto productivo, motivan el resto del análisis.
– *Efectos no keynesianos* (Giavazzi y Pagano, 1990): con agentes ricardianos, una consolidación creíble y permanente puede ser expansiva, porque anticipa menos impuestos distorsionantes y reduce las primas de riesgo. El efecto atracción resultante sobre la inversión privada puede compensar la caída de la demanda pública.
– *Suavización impositiva* (Barro, 1979): con impuestos distorsionantes la senda de impuestos sí importa. Como el coste de la distorsión crece con el tipo, lo óptimo es mantener constante el tipo esperado y dejar que la deuda absorba los *shocks*.
– *Aritmética monetarista desagradable* (Sargent y Wallace, 1981): si la autoridad fiscal no ajusta sus déficits, su restricción intertemporal se impone a la monetaria.

1.2.3. Conclusiones

Es la base metodológica de los modelos posteriores, y su lección es que el gasto no productivo solo eleva el *output* a costa del consumo privado, y que financiarlo con impuestos distorsionantes reduce la producción potencial. Pero depende de la calibración y, sobre todo, es walrasiano: no hay rigideces y los ciclos son respuestas óptimas a perturbaciones reales.

1.3. Perspectiva NEK: IS-PC-MR con gasto público

Idea: la lectura keynesiana del IS-LM, que subraya las rigideces, se formaliza hoy en el modelo canónico neokeynesiano ampliado con gasto público.

1.3.1. Supuestos

Competencia monopolística, precios rígidos a la Calvo y mercado de trabajo competitivo, de modo que la política monetaria no es neutral a corto plazo. Los agentes tienen expectativas racionales, el banco central sigue una regla y los hogares son ricardianos. El trabajo es el único factor variable, sin acumulación de capital.

1.3.2. Modelo

Planteamiento: el modelo se resume en tres ecuaciones log-linealizadas, la IS dinámica, la curva de Phillips neokeynesiana y la regla de Taylor:

$$\text{IS: } \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(r_t - \bar{r}_t) + \varepsilon_t^g$$

$$\text{CP: } \pi_t = \beta \mathbb{E}_t[\pi_{t+1}] + \kappa \tilde{y}_t$$

$$\text{Regla: } i_t = \bar{i}_t + \alpha_\pi \tilde{\pi}_t + \alpha_y \tilde{y}_t + \varepsilon_t^m$$

– En la IS, con $r_t = i_t - \mathbb{E}_t[\pi_{t+1}]$, el tipo natural, $\bar{r}_t$, aumenta con la impaciencia, con el gasto público y con el crecimiento esperado de la productividad.
– En la regla, $\bar{i}_t = \bar{r}_t + \pi^T$ es el tipo nominal neutral y $\alpha_\pi > 1$ (principio de Taylor).

Equilibrio: sin *shocks*, la brecha es nula, la inflación está en el objetivo y el principio de Taylor garantiza que el equilibrio es único.

Desarrollo gráfico: en los planos brecha-desviación del tipo real respecto del natural, arriba, y brecha-inflación, abajo, la IS es decreciente, porque un tipo real más alto contrae la demanda, la CP crece con pendiente κ y la MR, que resulta de llevar la regla a la IS, es decreciente. En el punto 0, la brecha es nula, la inflación está en el objetivo y el tipo real iguala al natural.

Estática comparativa: con rigideces, el gasto público se transmite por dos canales.

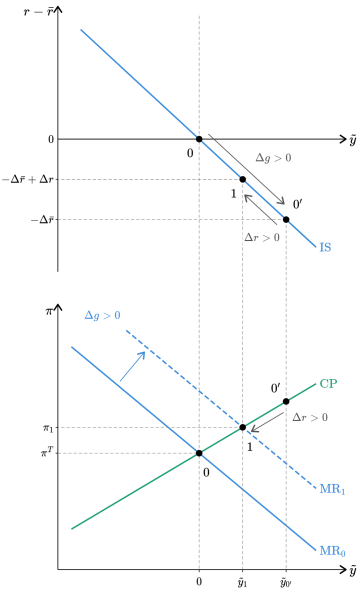
– *Canal de la riqueza*: como en el ciclo real, los hogares se sienten más pobres, consumen menos y trabajan más, lo que eleva el *output* natural.
– *Canal del* mark-up: las empresas no ajustan sus precios y atienden la demanda produciendo más, para lo que contratan más trabajo. El margen se estrecha y el salario real sube, lo que eleva el *output* efectivo, aunque el consumo de los hogares ricardianos sigue la ecuación de Euler y solo aumenta el de los no ricardianos.

El efecto neto sobre la brecha es ambiguo y depende de la respuesta monetaria. Si domina el *mark-up*, el gasto eleva el tipo natural y, a tipo real constante, la brecha aumentaría, pero el banco central, siguiendo la regla de Taylor, eleva el tipo real y contrarresta parte del impulso, con un multiplicador inferior a la unidad. La economía se sitúa así en el punto 1, con una brecha menor que en 0 y una inflación por encima del objetivo.

Dinámica: un *shock* transitorio opera sobre todo por el *mark-up*, con un multiplicador de impacto mayor que el del permanente. Uno permanente eleva más el *output* natural y su multiplicador se aproxima al neoclásico.

Extensiones: tres precisan cuándo es efectiva la política fiscal.

– *Restricciones de liquidez* (Galí, López-Salido y Vallés, 2007): los agentes no ricardianos consumen su renta corriente, no sufren el efecto riqueza y el *mark-up* eleva su consumo. Si son bastantes y el gasto se financia en parte con deuda, el multiplicador supera la unidad.



Expansión fiscal en el IS-PC-MR: a tipo real constante la economía pasaría de 0 a 0 , y con la regla de Taylor se sitúa en 1.

– *Política monetaria* (Woodford, 2011): con tipo real constante, el multiplicador es exactamente uno; con un objetivo estricto de inflación o la regla de Taylor, inferior a uno; en el límite inferior de los tipos, la mayor inflación esperada reduce el tipo real y es superior a uno, siempre que el gasto no se prolongue más allá de ese periodo. Si un mayor ahorro precautorio, como tras la crisis de 2008, hace negativo el tipo natural, la pérdida de *output* es grande y la política fiscal, especialmente efectiva.
– *Uniones monetarias* (Galí y Monacelli, 2008): como la política monetaria común no responde a la inflación de un país, el multiplicador es mayor, aunque inferior a la unidad si el país financia el gasto con sus impuestos (Farhi y Werning, 2016). La política fiscal óptima de cada país debe corregir las distorsiones que la política monetaria común no atiende.

1.3.3. Conclusiones

Recupera la efectividad de la política fiscal a corto plazo sin renunciar a los microfundamentos, y muestra que depende de las restricciones de liquidez y de los regímenes monetario y cambiario. Pero solo considera gasto no productivo financiado con suma fija y no permite analizar el crecimiento.

1.4. Evidencia empírica

Consumo privado: es la predicción que separa ambos modelos. Con SVAR, el consumo aumenta tras un *shock* de gasto (Blanchard y Perotti, 2002). Con *shocks* de gasto militar fechados al anunciarse, cae, como predice el ciclo real, y Ramey (2011) atribuye la diferencia a que los del SVAR llegan tarde, cuando los agentes ya conocían el aumento del gasto.

Hogares no ricardianos: entre el 20 % y el 40 % de las devoluciones del impuesto sobre la renta de 2001 en EE. UU. se gastó en no duraderos en el trimestre en que se recibieron (Johnson et al., 2006), contra la renta permanente. Las devoluciones se conocían de antemano, lo que apoya la existencia de hogares restringidos por liquidez.

Multiplicador: con datos de EE. UU. desde 1889, queda por debajo de la unidad con desempleo alto o bajo (Ramey y Zubairy, 2018). El relativo entre estados de una unión monetaria ronda 1,5 (Nakamura y Steinsson, 2014). El primer resultado concuerda con la perspectiva neoclásica y el segundo, en una unión cuya política monetaria no responde a un solo estado, con la neokeynesiana.

En suma, el gasto no productivo reduce el consumo privado por el efecto riqueza y, si es transitorio, también la inversión y el ahorro nacional. Su efecto sobre el *output* depende de las rigideces, las restricciones de liquidez y la política monetaria.

2. Efectos vía diseño del sistema de pensiones

El diseño del sistema de pensiones es la principal vía por la que el sector público altera el ahorro de los hogares y, con él, la acumulación de capital.

2.1. Sistemas de pensiones

Idea: los sistemas públicos responden a la función redistributiva y a la corrección de la miopía del ahorro individual, y pesan mucho en las economías avanzadas, en España en torno al 12 % del PIB. Se financian con cotizaciones, que gravan el trabajo y, con reparto, reducen el ahorro.

2.1.1. Capitalización (*fully funded*)

Definición: cada trabajador acumula sus cotizaciones en una cuenta individual y cobra en la jubilación el capital y su rendimiento, $(1+r)$, sin redistribución intergeneracional. Es de aportación definida, con conversión del saldo en una renta vitalicia, y a menudo con un pilar solidario. La reforma chilena de 1981 lo recuperó como sistema nacional obligatorio de gestión privada, y lo aplican también otros países latinoamericanos y Australia.

Conclusiones: es robusto al envejecimiento, puede elevar el ahorro nacional, profundiza los mercados de capitales y distorsiona menos el trabajo. Pero traslada al trabajador los riesgos financiero y de longevidad, penaliza las lagunas de cotización y la transición exige que una generación pague dos veces.

2.1.2. Reparto (*pay as you go*)

Definición: lo que cotizan hoy los trabajadores paga las pensiones de los jubilados de hoy, con redistribución intergeneracional. Su rendimiento implícito es el crecimiento de la masa salarial, $(1+n)(1+g_y)$, por el número de cotizantes y su renta (Samuelson, 1958; Aaron, 1966). Es de prestación definida, con edad, cotización y revalorización fijadas por ley. El reparto puro se generalizó tras la II Guerra Mundial y rige en la mayoría de países europeos y en la Seguridad Social de EE. UU. El Estado garantiza el pago, y los riesgos demográfico y político recaen sobre las generaciones futuras.

Conclusiones: paga desde el primer día, reparte los riesgos entre el Estado y los cotizantes y es inmune a las crisis financieras. Pero su rendimiento cae con el envejecimiento y la ralentización del crecimiento, reduce el ahorro nacional y la desconexión entre cotización y prestación actúa como un impuesto sobre el trabajo. Ese menor rendimiento obliga a subir cotizaciones, bajar prestaciones o retrasar la jubilación.

2.1.3. Sistemas mixtos

Definición: combinan un pilar de reparto con pilares de capitalización, según la arquitectura de tres pilares del Banco Mundial (1994), y rinden una media ponderada de ambos. En las **cuentas nocionales** (NDC), el reparto calcula la pensión como si fuera de capitalización, con un rendimiento ligado a la masa salarial. Así ocurre en Suecia, y también combinan pilares Países Bajos y Dinamarca.

Conclusiones: diversifican los riesgos y las cuentas nocionales recuperan el vínculo entre cotización y prestación sin coste de transición, pero la coordinación es compleja y la transición conlleva un doble pago atenuado.

2.2. Generaciones solapadas (Diamond, 1965)

Supuestos: Allais (1947), Samuelson (1958) y Diamond (1965) plantean dos periodos con certidumbre: los jóvenes trabajan, consumen y ahorran, y los ancianos consumen lo ahorrado. La población crece a la tasa n y la renta laboral, a g_y. Cada sistema se introduce como una cotización τ en la juventud y una pensión en la vejez.

Planteamiento: el joven elige su ahorro con utilidad logarítmica:

$$\max_{s_t} \ln c_t^j + \beta \ln c_{t+1}^a \quad \text{s.a.:} \begin{cases} c_t^j = (1-\tau)\, y_t - s_t \\ c_{t+1}^a = (1+r)\, s_t + p_{t+1} \end{cases}$$

CPO: sustituyendo las restricciones en la función objetivo y derivando respecto a s_t, se obtiene la ecuación de Euler:

$$\frac{1}{c_t^j} = \frac{\beta(1+r)}{c_{t+1}^a}$$

CSO: la segunda derivada respecto a s_t, $-1/(c_t^j)^2 - \beta(1+r)^2/(c_{t+1}^a)^2$, es negativa, luego la CPO es suficiente.

Solución: la ecuación de Euler, común a todos los casos, da un ahorro privado que cae con la cotización y con el valor actual de la pensión:

$$s_t^* = \frac{\beta}{1+\beta}(1-\tau)\, y_t - \frac{p_{t+1}}{(1+\beta)(1+r)}$$

- *Sin pensiones*: el ahorro, $s_t^* = \frac{\beta}{1+\beta}\, y_t$, es solo privado, depende de la impaciencia y de la renta, y podría ser insuficiente.
- *Capitalización*: con $p_{t+1} = (1+r)\tau y_t$, el ahorro privado, $s_t^* = \frac{\beta}{1+\beta}\, y_t - \tau y_t$, cae exactamente en la cotización, de modo que el **ahorro nacional** y el capital no varían. Al agente solo le importa el rendimiento, que es el mismo.
- *Reparto*: con $p_{t+1} = (1+n)(1+g_y)\tau y_t$, la cotización se transfiere a los jubilados y no se invierte, de modo que el ahorro nacional cae siempre, y el privado cae más que la cotización si el rendimiento del reparto supera al tipo de interés. El ahorro privado es:

$$s_t^* = \frac{\beta}{1+\beta}\, y_t - \tau y_t \left[1 + \frac{(1+n)(1+g_y) - (1+r)}{(1+\beta)(1+r)}\right]$$

Bienestar: si $(1+n)(1+g_y) > 1+r$ se prefiere el reparto, y si es menor, la capitalización. La caída del ahorro bajo reparto no implica pérdida de bienestar, porque en una economía **dinámicamente ineficiente**, con $r < n + g_y$, el reparto mejora a todas las generaciones. Pero con cohortes menos numerosas tras el *baby boom* y una renta que crece menos desde los setenta, su rendimiento ha caído.

Conclusiones: aísla con claridad el efecto de cada sistema sobre el ahorro nacional y explica por qué el envejecimiento hace insostenibles sistemas que fueron óptimos. Pero omite la incertidumbre sobre la longevidad y los rendimientos, el efecto de la cotización sobre el trabajo y los costes de transición.

2.3. Evidencia empírica

Alemania: su reparto bismarckiano ha respondido con reformas paramétricas: el pilar voluntario Riester (2001), el factor de sostenibilidad (2004), que liga la revalorización a la ratio entre cotizantes y pensionistas, y el retraso de la jubilación a 67 años (2007). La tasa de sustitución ha caído, aunque en 2025 se prorrogó hasta 2031 un suelo del 48 %.

Chile: la capitalización profundizó el mercado de capitales, pero su efecto sobre el ahorro nacional es discutido por el coste fiscal de la transición, y las bajas tasas de sustitución obligaron a un pilar solidario (2008) y a una cotización adicional del empleador con un componente de reparto (2025). Esas bajas tasas se deben a las lagunas de cotización, la informalidad y las comisiones.

Suecia: la reforma aprobada en 1998 y aplicada desde 1999 destina 16 de sus 18,5 puntos de cotización a cuentas nocionales y 2,5 a capitalización, con un mecanismo de equilibrio automático que se activó en 2010. El mecanismo reduce la revalorización cuando los pasivos del pilar de reparto superan a sus activos, de modo que traslada a los cotizantes parte del riesgo a cambio de garantizar la solvencia.

España: un reparto de prestación definida con pilares complementarios poco desarrollados, presionado por una fecundidad de 1,10 hijos por mujer en 2024, el efecto sustitución de pensiones nuevas más altas y una tasa de dependencia que seguiría aumentando hasta el 73,2 % en 2076, según las proyecciones del INE (2026). Además, los tipos próximos a cero de la década de 2010 recortaron la rentabilidad del fondo de reserva y de los planes complementarios. Las reformas de 2011, con la edad legal a 67 años y un periodo de cómputo más largo, y de 2013, con el índice de revalorización y el factor de sostenibilidad, después derogados, actuaron sobre el gasto. Las de 2021-2023 actúan sobre el ingreso, con el mecanismo de equidad intergeneracional, el destope gradual de las bases máximas y la cuota de solidaridad, y recuperan la revalorización con el IPC, con evaluación periódica de la AIReF. En términos del modelo, la baja fecundidad reduce n y el rendimiento implícito del reparto, y un r bajo hace el reparto relativamente más atractivo, pero obliga a ahorrar más para complementar la pensión pública.

En suma, la capitalización deja inalterado el ahorro nacional y el reparto lo reduce. El envejecimiento ha mermado el rendimiento del reparto y las reformas convergen hacia sistemas mixtos.

3. Efectos vía gasto público productivo

El último canal es el gasto que entra en la función de producción, cuyos efectos se estudian con modelos de crecimiento exógeno y endógeno.

3.1. Gasto público productivo

Definición: el **gasto público productivo** entra en la función de producción y eleva la productividad privada: la educación, que eleva el capital humano, y las infraestructuras y otros bienes públicos, que el mercado no proveería lo bastante por sus externalidades.

Disyuntiva: aumenta la producción potencial, pero se financia con impuestos distorsionantes, que reducen la renta disponible, la inversión y el consumo privados.

3.2. Crecimiento exógeno: Ramsey-Cass-Koopmans

3.2.1. Supuestos

Un agente representativo con utilidad CES y aversión al riesgo θ ofrece una unidad de trabajo. El sector público grava el consumo, τ_c, la renta laboral, τ_w, y la del capital, τ_r, y dedica lo recaudado a gasto productivo con presupuesto equilibrado. La tecnología es $y = A(g)k^\alpha$, donde el gasto eleva la productividad total, $A'(g) > 0$, sin ser acumulable.

3.2.2. Modelo

Hogares: con su riqueza inicial, b_0, dada, maximizan su utilidad descontada sujetos a una restricción en la que los impuestos gravan cada uso y fuente de renta:

$$\max_c \int_0^\infty e^{-(\rho-n)t} \frac{c^{1-\theta}-1}{1-\theta}\, dt, \qquad \theta > 0$$
$$\text{s.a.:}\ (1+\tau_c)c + \dot{b} = (1-\tau_w)w + [(1-\tau_r)r - n]\, b$$

Del hamiltoniano se obtiene la ecuación de Euler:

$$g_c = \frac{1}{\theta}\left[(1-\tau_r)\, r - \rho\right]$$

- El impuesto sobre el capital introduce una **cuña** entre el rendimiento social del ahorro y el privado, de modo que el hogar ahorra menos.

– Los impuestos sobre el consumo y el trabajo no aparecen: gravan por igual el consumo de hoy y el de mañana, y una oferta de trabajo inelástica. Reducen la renta, pero no distorsionan la decisión intertemporal, y por eso la imposición óptima recomienda gravar el consumo antes que el ahorro.
– El gasto público eleva el rendimiento del capital y el ahorro.

Empresas: el gasto productivo es una externalidad que no internalizan, de modo que el equilibrio competitivo no es eficiente por sí solo. La demanda de capital es $A(g)f'(k) = r + \delta$, y con rendimientos constantes el trabajo percibe el residuo, $w = A(g)\left[f(k) - kf'(k)\right]$.

Sector público: con presupuesto equilibrado, financiar el gasto con τ_c o τ_w eleva el capital por el efecto productividad, y con τ_r el efecto es ambiguo. La composición de la recaudación importa tanto como su nivel. Además, el gasto se autofinancia en parte, porque eleva las bases imponibles. La restricción del sector público es:

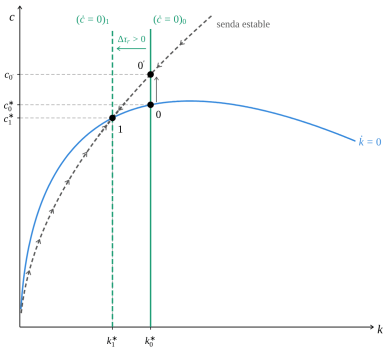
$$g = \tau_w w + \tau_r r b + \tau_c c$$

Equilibrio: con $b = k$ y la restricción pública en la de los hogares, $c + g + \dot{k} + (n+\delta)k = A(g)f(k)$. En estado estacionario, el rendimiento neto del capital iguala la impaciencia y el consumo es el del *locus* $\dot{k} = 0$:

$$(1-\tau_r)\left[A(g)f'(k^*) - \delta\right] = \rho \qquad c^* = A(g)f(k^*) - (n+\delta)k^* - g$$
$$g_c = \dot{k}/k = 0$$

Desarrollo gráfico: en el plano k-c, a la izquierda de k^* el consumo crece y a su derecha decrece. El equilibrio es un punto de silla, de modo que, dado k_0, una única senda converge a (k^*, c^*).

Estática comparativa: los impuestos sobre el consumo y el trabajo no alteran el capital ni el consumo de estado estacionario. El impuesto sobre el capital los reduce, y el gasto productivo eleva el capital y la renta. Una subida no anticipada y permanente de τ_r, compensada con otros impuestos para no alterar g, desplaza a la izquierda el *locus* $\dot{c} = 0$: el consumo salta a la nueva senda estable y la economía converge a menos capital y consumo.



Subida del impuesto sobre el capital en el modelo de Ramsey-Cass-Koopmans: el consumo salta a la senda estable (0→0) y converge a menos capital y consumo (1).

Disyuntiva: el tamaño óptimo del sector público iguala la productividad marginal del gasto al coste marginal de la distorsión necesaria para financiarlo. La política fiscal tiene **efectos nivel**, pero no **efectos tasa**, porque el crecimiento a largo plazo sigue siendo el del progreso técnico exógeno.

Extensiones: con trabajo elástico, τ_w y τ_c también distorsionan. El tipo óptimo sobre el capital a largo plazo es nulo (Chamley, 1986; Judd, 1985), aunque puede ser positivo si la elasticidad de sustitución intertemporal es inferior a la unidad (Straub y Werning, 2020). Con capital público acumulable (Arrow y Kurz, 1970), la inversión pública afecta transitoriamente al crecimiento. El tipo nulo de Chamley y Judd se debe a que gravar el capital equivale a gravar el consumo futuro a un tipo creciente con el horizonte, y se debilita con horizontes finitos, restricciones de liquidez o rentas puras del capital.

Conclusiones: ordena los impuestos por su coste de acumulación, con el del capital como el más dañino.

3.3. Crecimiento endógeno: el modelo de Barro (1990)

Idea: el gasto productivo eleva la productividad del capital privado y su financiación detrae renta, de modo que existe un tamaño del sector público que maximiza el crecimiento. Se presenta con ahorro exógeno, como extensión de Solow-Swan, de modo que la figura impositiva es irrelevante, porque la distorsión del rendimiento privado del capital solo aparece con hogares optimizadores.

Supuestos: el gasto entra como *input*, con rendimientos constantes en k y g conjuntamente, de modo que si g crece con k no hay rendimientos decrecientes y el modelo es AK:

$$y = Ak^\alpha g^{1-\alpha}, \qquad A > 0,\ \alpha \in (0,1)$$

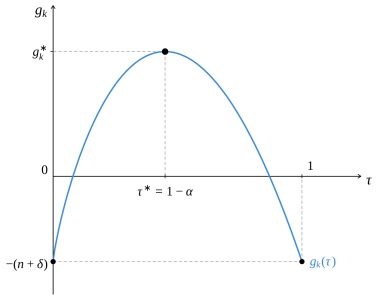
Planteamiento: con sector público, $Y = C + I + G$, y el gasto, $G = \tau Y$, se financia con un impuesto proporcional. En niveles y per cápita:

$$\text{Renta:} \quad Y = AK^\alpha L^{1-\alpha}(G/L)^{1-\alpha} \Rightarrow y = A^{1/\alpha}\tau^{\frac{1-\alpha}{\alpha}}k, \qquad A > 0,\ \alpha \in (0,1)$$
$$\text{Inversión:} \quad \dot{K} = s(1-\tau)Y - \delta K \Rightarrow \dot{k} = s(1-\tau)y - (n+\delta)k$$
$$\text{Consumo:} \quad C = (1-s)(1-\tau)Y \Rightarrow c = (1-s)(1-\tau)y$$

Equilibrio: sin dinámica de transición, la renta, el capital y el consumo crecen a la misma tasa constante, $g_k(\tau) = s(1-\tau)A^{1/\alpha}\tau^{(1-\alpha)/\alpha} - (n+\delta)$, máxima en $\tau^* = 1-\alpha$.

– El tipo aparece dos veces: como $(1-\tau)$, el **efecto distorsión**, que reduce la renta disponible y el ahorro, y como $\tau^{(1-\alpha)/\alpha}$, el **efecto productividad** del gasto.
– El crecimiento es máximo cuando el peso del ingreso público iguala la participación del gasto productivo en la tecnología.

Desarrollo gráfico: la tasa de crecimiento tiene forma de U invertida, porque con $\tau = 0$ no hay gasto y nada se produce, y con $\tau = 1$ no queda renta que ahorrar. A la izquierda del máximo domina el efecto productividad y a la derecha, la distorsión. En ambos extremos, el capital por trabajador varía a la tasa $-(n+\delta) < 0$.



Modelo de Barro (1990): la tasa de crecimiento es máxima cuando el tipo impositivo iguala la elasticidad del *output* al gasto productivo.

Estática comparativa: a diferencia del modelo exógeno, la política fiscal tiene **efectos tasa**.

Extensiones: en la versión original, con hogares a la Ramsey, el crecimiento es creciente con g/y si se financia con suma fija, aunque el tamaño que maximiza el bienestar sigue siendo $1-\alpha$. Con el impuesto proporcional, el tipo que maximiza el crecimiento es el mismo. Con oferta de trabajo endógena (Baxter y King, 1993), los aumentos transitorios de la inversión pública expulsan inversión privada y apenas expanden el *output*, pero uno permanente lleva a más producción, inversión, consumo y empleo, porque el capital público eleva la productividad del trabajo y del capital privado.

Conclusiones: fue de las primeras aportaciones sobre el tamaño del sector público en crecimiento endógeno e ilustra de forma intuitiva el debate sobre ese tamaño. Pero fija la oferta de trabajo, trata el gasto como flujo y no distingue partidas de gasto ni figuras impositivas.

3.4. Evidencia empírica

Inversión pública: Aschauer (1989) estima para EE. UU. (1949-1985) una elasticidad del *output* al capital público en torno a 0,4, con un fuerte efecto de las infraestructuras básicas sobre la productividad. La literatura posterior encuentra un efecto positivo pero mucho menor, y el metaanálisis de Bom y Ligthart (2014) la sitúa en torno a 0,1. Para la OCDE, la inversión pública eleva el *output* a corto y largo plazo, con efecto atracción sobre la privada, sobre todo con holgura y si es eficiente (Abiad et al., 2016).

Consumo público: en 98 países entre 1960 y 1985, su peso en el PIB se relaciona negativamente con el crecimiento, y la inversión pública no es significativa (Barro, 1991), acorde con la distinción de su modelo entre gasto productivo e improductivo.

Imposición: en un panel de 22 países de la OCDE, los impuestos distorsionantes reducen el crecimiento y el gasto productivo lo eleva (Kneller et al., 1999). Con 21 países de la OCDE, Arnold et al. (2011) establecen una jerarquía de daño creciente: los impuestos recurrentes sobre la propiedad inmobiliaria, el consumo, la renta personal y, los más dañinos, sociedades. En España corresponden al IBI, IVA, IRPF e impuesto sobre sociedades.

En suma, el gasto productivo eleva el nivel de renta o su tasa de crecimiento, pero su financiación distorsionante fija un tamaño óptimo del sector público, y la composición del gasto y de los impuestos importa tanto como su nivel.

A39. Déficit público. Conceptos. Financiación y sus consecuencias macroeconómicas. Dominación monetaria y dominación fiscal. La dinámica de la deuda pública y su sostenibilidad

1. Fundamentos del análisis del déficit y la deuda

Antes de estudiar la financiación y la dominancia, se fijan los conceptos esenciales y se repasa el debate sobre la efectividad de la política fiscal y de su financiación.

1.1. Conceptos esenciales

Saldo presupuestario: el saldo observado recoge la recaudación y el gasto primario, ambos netos del ciclo y de las medidas *one-off*, el servicio de la deuda, el componente cíclico y las propias medidas *one-off*, φ_t:

$$\mathrm{SP}_t = T_t - (G1_t + iB_{t-1}) + \varepsilon(y_t - \bar{y}_t) + \varphi_t$$

- $\varepsilon > 0$ es la semielasticidad del saldo a la brecha de producción: cuántos puntos de PIB mejora el saldo por cada punto de brecha positiva, vía estabilizadores automáticos. La Comisión Europea la estima en torno a 0,5-0,6 para España y la media de la UE. Los estabilizadores actúan por la mayor recaudación y el menor gasto en prestaciones.

De esta expresión se derivan tres saldos:

- El **saldo primario**, $\mathrm{SP}_t + iB_{t-1} = T_t - G1_t + \varepsilon(y_t - \bar{y}_t) + \varphi_t$, excluye los intereses de la deuda y se acerca más a lo que controla el Gobierno, aunque aún incluye el ciclo.
- El **saldo primario cíclicamente ajustado** corrige además el componente cíclico.
- El **saldo primario estructural**, $T_t - G1_t$, corrige también las medidas *one-off* y es el relevante para evaluar la orientación de la política fiscal.

Señoraje e impuesto inflacionario: el señoraje, el ingreso real que el Estado obtiene al emitir dinero, no es idéntico al impuesto inflacionario, el gravamen implícito que la inflación impone a los tenedores de saldos monetarios.

Dominancia monetaria y fiscal (Sargent y Wallace, 1981): con **dominancia monetaria**, la conducta habitual, el banco central fija la senda del dinero o de los tipos y el Tesoro ajusta sus saldos para que la deuda sea sostenible. Con **dominancia fiscal**, el Tesoro fija la senda de déficits y el banco central debe generar el señoraje necesario para evitar una deuda explosiva, con lo que pierde el control de los precios.

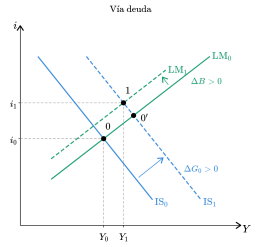
Políticas activas y pasivas (Leeper, 1991): una autoridad es activa si fija su instrumento con independencia de la restricción presupuestaria intertemporal del sector público, y pasiva si lo ajusta para garantizar su cumplimiento. La dominancia monetaria es una política monetaria activa con una fiscal pasiva, y la fiscal, la combinación inversa.

1.2. Efectividad de la política fiscal

1.2.1. Antecedentes: el modelo IS-LM

En el IS-LM con precios fijos, una política fiscal expansiva desplaza la IS a la derecha, con un efecto multiplicador y un efecto expulsión cuyo peso depende de la sensibilidad al tipo de interés de la inversión, v, y de la demanda de dinero, h. La forma de financiación añade efectos:

- *Vía impuestos*: el efecto contractivo no anula el expansivo, porque el gasto actúa directamente y los impuestos solo a través de la propensión marginal a consumir, resultado conocido como *teorema del multiplicador del presupuesto equilibrado* (Haavelmo, 1945). La IS se desplaza a la derecha, pero menos que sin financiación impositiva, y para los monetaristas esta financiación hace aún más ineficaz la política fiscal.
- *Vía deuda*: el keynesianismo añade el *efecto reajuste de cartera*. Más bonos abaratan los ya emitidos, elevan la rentabilidad exigida y desplazan la LM a la izquierda. El efecto cartera es tanto mayor cuanto menor es la sustituibilidad entre bonos y dinero. El monetarismo añade un efecto riqueza directo, que desplaza la IS a la derecha, y otro indirecto, que refuerza la expulsión. Con sus parámetros, la expulsión es total, y la política fiscal puede llegar incluso a contraer la actividad.



Financiación del déficit vía deuda en el IS-LM.

- *Vía monetización*: añade el *efecto liquidez* o efecto Keynes. La LM se desplaza a la derecha, el tipo sube menos y, a corto plazo, la renta crece más que con cualquier otra financiación. Como hay más dinero para atender la mayor demanda de saldos, el efecto expulsión es menor.

Para los primeros keynesianos, la política monetaria sirve para financiar el gasto, pero no para estimular la demanda, por los dos eslabones débiles. Para los monetaristas, solo es eficaz a corto plazo, porque sus efectos se diluyen por la teoría cuantitativa. Además, el efecto liquidez va acompañado de los efectos riqueza y del efecto cartera.

1.2.2. Perspectiva NMC: el modelo del ciclo real

Tras la crítica de Lucas, lo relevante es la respuesta de los agentes a la senda esperada de impuestos, déficits y deuda. Se derivan tres resultados:

- **Equivalencia ricardiana** (Barro, 1974): con mercados de capitales perfectos, expectativas racionales, altruismo intergeneracional perfecto e impuestos de suma fija, los bonos no son riqueza neta y, dado el gasto, elegir entre impuestos y deuda es irrelevante. En economía abierta, el mayor ahorro privado financia el déficit y no hay déficits gemelos. Sus supuestos fallan con restricciones de liquidez, ilusión fiscal, altruismo imperfecto, impuestos distorsionantes o gasto productivo. Con restricciones de liquidez, el sector público se financia más barato que el privado y la deuda sí es riqueza neta, y la ilusión fiscal y el altruismo imperfecto trasladan la carga a generaciones posteriores.
- **Efectos no keynesianos** (Giavazzi y Pagano, 1990): un ajuste creíble y duradero del gasto puede estimular la actividad, porque anticipa menos impuestos, reduce la incertidumbre y las primas de riesgo y atrae inversión privada.
- **Suavización impositiva** (Barro, 1979): con impuestos distorsionantes de coste convexo, lo óptimo es mantener constante el tipo esperado y usar el déficit para amortiguar los *shocks* transitorios de gasto. El déficit es el resultado de una política óptima, y el tipo impositivo y la ratio de deuda siguen un paseo aleatorio. El Gobierno se endeuda en las recesiones y amortiza en las expansiones, y cada *shock* no anticipado altera permanentemente el nivel de la deuda.

1.2.3. Perspectiva NEK: el modelo IS-PC-MR

Las rigideces nominales y las fricciones financieras devuelven a la política fiscal su función estabilizadora, y su efecto depende del entorno:

- Con hogares *hand-to-mouth*, las transferencias tienen efectos reales y la equivalencia ricardiana falla en proporción a su peso. Estos hogares consumen su renta corriente, de modo que también las rebajas impositivas tienen efectos reales.
- Con objetivo de inflación, el banco central contrarresta el impulso y el multiplicador es inferior a la unidad.
- Con tipo de cambio fijo o en una unión monetaria, la política monetaria no reacciona y el multiplicador es mayor.
- En el límite inferior de los tipos, el ZLB, el tipo real cae y el multiplicador puede superar la unidad.

En suma, la financiación no es neutral salvo bajo equivalencia ricardiana estricta, y el efecto de una medida depende del régimen monetario. De ahí las dos preguntas siguientes: si el déficit puede sostenerse y qué relación guardan las dos autoridades.

2. Financiación del déficit y su sostenibilidad

Hoy el debate no es qué financiación es más expansiva, sino la sostenibilidad. Un **déficit sostenible** procede de una política que puede prolongarse indefinidamente con una ratio deuda-PIB nominal constante. Uno **insostenible**, con $r > g_Y$, es estructural y permanente y provoca un efecto bola de nieve. Por eso la ratio deuda-PIB pesa mucho en la valoración del riesgo-país. La sostenibilidad remite a la solvencia, no a la liquidez. La solvencia se refiere al sujeto, que el valor actual de sus superávits primarios futuros cubra su deuda, y la sostenibilidad, a la política, de modo que se puede ser solvente sin que la política sea sostenible, al menos durante un tiempo, pero no al revés.

2.1. Financiación vía deuda: dinámica y sostenibilidad

Origen: la lógica de sostenibilidad de la restricción presupuestaria intertemporal del Gobierno, RPIG, se remonta a Domar (1944), y Blanchard (1990) la sistematizó en los indicadores fiscales de los organismos internacionales.

2.1.1. Supuestos

Economía cerrada, en la que el déficit solo se financia con deuda o base monetaria. Los agentes privados absorben la deuda al tipo de mercado y se cumple la ecuación de Fisher, $1 + i_t = (1 + r_t)(1 + \pi_t)$. El Gobierno respeta su restricción en cada periodo, y tipos, inflación y crecimiento son constantes en el análisis gráfico.

2.1.2. Modelo

Planteamiento: el modelo parte de la restricción presupuestaria del Gobierno y la expresa en proporción del PIB:

- *Restricción presupuestaria del Gobierno*: el gasto primario más los intereses de la deuda heredada, menos la recaudación, se financia con nueva deuda o con base monetaria:

$$G1_t + i_t B_{t-1} - T_t = (B_t - B_{t-1}) + (H_t - H_{t-1})$$

- *Ecuación dinámica de la deuda*: se procede en tres pasos. Se divide la

restricción entre el PIB nominal y las variables en minúscula pasan a ser ratios sobre el PIB. Con la ecuación de Fisher y la descomposición del crecimiento nominal en inflación y crecimiento real, la deuda heredada cuesta $(1+r_t)/(1+g_{Y,t})$, con r_t el tipo real *ex post*. Por último, la financiación monetaria se agrupa en el señoraje sobre el PIB, σ_t. Restando b_{t-1} a ambos lados:

$$\Delta b_t = d1_t + \frac{r_t - g_{Y,t}}{1+g_{Y,t}} b_{t-1} - \sigma_t$$

La dinámica de la deuda depende, así, de tres factores:

- El **déficit primario**, $d1_t = g1_t - \tau_t$.
- La **relación entre tipo real y crecimiento**: si $g_Y > r$, la deuda pierde peso. Si $r > g_Y$, los intereses pesan cada vez más y hace falta más deuda para pagarlos, un círculo vicioso o **efecto bola de nieve**, estilo Ponzi.
- El **señoraje**, que alivia, al menos transitoriamente, el recurso a la deuda.

Equilibrio: si el Estado no puede emitir dinero, $\sigma_t = 0$, el déficit es sostenible si la ratio es estable. La **condición de sostenibilidad** es:

$$-d1_t = \frac{r_t - g_{Y,t}}{1+g_{Y,t}} b_{t-1}$$

El superávit primario debe igualar el coste efectivo de la deuda viva corregido por el crecimiento, aproximadamente, $(r - g_Y)\, b$. La intuición es que el crecimiento "paga" parte de la deuda como razón del PIB, y la condición es útil porque el saldo primario es lo que el Gobierno controla. El supuesto $\sigma_t = 0$ corresponde a un Estado que ha cedido la política monetaria a una autoridad independiente.

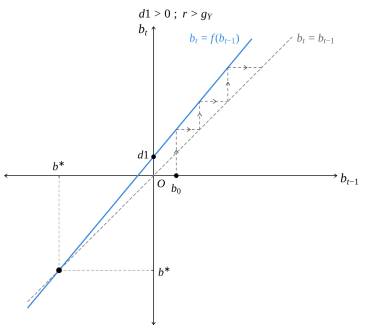
Estática comparativa: las soluciones ante una deuda insostenible se ordenan según la variable de la EDD sobre la que inciden: la monetización eleva σ, a costa de inflación; la consolidación fiscal reduce $d1$; el *default*, la quita o la reprogramación reducen b; las políticas de crecimiento elevan g_Y, y la represión financiera contiene r. Esta canaliza el ahorro privado hacia la deuda pública con topes a los tipos de los depósitos, la nacionalización de fondos de pensiones o límites a la inversión exterior.

Dinámica: fuera del equilibrio, la EDD es una ecuación en diferencias de primer orden. En el plano b_{t-1}-b_t es una recta con ordenada en el origen $d1_t - \sigma_t$ y pendiente $(1+r)/(1+g_Y)$. Cuanto mayor es el tipo real respecto al crecimiento, más crece la deuda. La bisectriz recoge los puntos de ratio estable, y su corte con la recta da la ratio estacionaria:

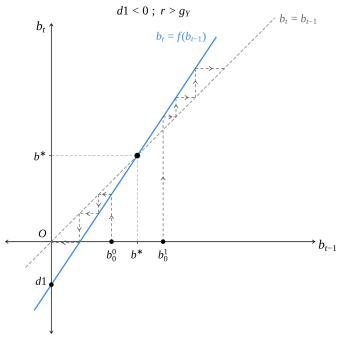
$$b^* = \frac{d1\,(1+g_Y)}{g_Y - r}$$

Caben cuatro casos, con r, g_Y y π constantes:

- *Déficit primario y* $r < g_Y$: la ratio converge a b^* desde cualquier punto de partida, porque la pendiente de la recta es menor que uno. El déficit es sostenible.
- *Superávit primario y* $r < g_Y$: converge a un valor negativo y el sector público acumula activos. El superávit es sostenible y basta con mantener la situación.
- *Déficit primario y* $r > g_Y$: la deuda sigue una senda explosiva (efecto bola de nieve), sea cual sea el punto de partida, porque la pendiente es mayor que uno, y será necesaria una consolidación fiscal.
- *Superávit primario y* $r > g_Y$: el resultado es ambiguo. El umbral b^*, positivo porque $d1 < 0$ y $g_Y < r$, separa las ratios iniciales que se reducen de las que crecen sin límite, y el superávit debe ajustarse al nivel inicial de deuda.



Déficit primario y $r > g_Y$: la ratio deuda/PIB sigue una senda explosiva, el efecto bola de nieve.



Superávit primario y $r > g_Y$: el umbral b^* separa las ratios que se reducen de las que crecen sin límite.

Extensiones: dos relajan los supuestos.

– *Incidencia del sector privado*: la emisión neta de deuda no puede superar el excedente financiero privado, $S^p - I^p$. Un déficit puede ser insostenible aun con $r < g_Y$ si el sector privado no absorbe la deuda, y el ajuste llega por la prima de riesgo o por el recurso al exterior. En proporción del PIB, la restricción es:

$$\frac{B_t - B_{t-1}}{P_t Y_t} \leq \frac{S_t^p - I_t^p}{P_t Y_t}$$

– *Función de reacción fiscal* (Bohn, 1998, 2007): se estima la respuesta del saldo primario a la deuda heredada, con μ_t los determinantes exógenos, como el ciclo o el gasto transitorio:

$$-d1_t = \mu_t + \eta\, b_{t-1} + \varepsilon_t$$

Que $\eta > 0$ basta para cumplir la RPIG, con independencia de r y g_Y, porque el Gobierno responde a más deuda con mayores superávits. Con deudas elevadas, sin embargo, la respuesta decrece (*fatiga fiscal*).

2.1.3. Conclusiones

Aspectos positivos: subraya la importancia de destinar los fondos a gasto corriente o a inversión, por su distinto efecto sobre el crecimiento, y de obtener un tipo competitivo generando credibilidad.

Limitaciones: ignora el nivel de la ratio, pues mantener el 100 % exige mucho más esfuerzo que el 20 %, y toma r y g_Y como dados, cuando la propia deuda eleva la prima de riesgo y expulsa capital privado. Cuanto mayor es la ratio, mayor suele ser el tipo que exigen los mercados, de modo que deuda y tipo se retroalimentan.

2.2. Financiación vía monetización

2.2.1. Señoraje e impuesto inflacionario

Señoraje: es el valor real del aumento de la cantidad de dinero, es decir, el ritmo de emisión por los saldos reales sobre los que recae:

$$\sigma_t = \frac{\Delta M_t}{P_t} \;\rightarrow\; \sigma_t = g_{M,t}\frac{M_t}{P_t}$$

Impuesto inflacionario: la inflación reduce el valor real de los saldos monetarios, y su medida admite dos criterios:

– **Friedman** (1971): con producción y velocidad constantes, la teoría cuantitativa da $g_M = \pi$, y el impuesto, $T^\pi = \pi\, M/P$, con la inflación como tipo y los saldos reales como base, coincide con el señoraje.
– **Marty** (1967): si la producción crece, solo hay impuesto cuando el dinero crece más que el PIB. La parte $g_Y M/P$ es **señoraje no inflacionario**, y $T_t^\pi = \sigma_t - g_{Y,t} M/P$. Con elasticidad renta unitaria de la demanda de dinero, $g_{M,t} = \pi_t + g_{Y,t}$, de modo que el impuesto inflacionario propiamente dicho es $T_t^\pi = \pi_t M/P$.

2.2.2. Señoraje óptimo (Cagan, 1956)

Idea: Cagan estudió siete hiperinflaciones en seis países europeos y se preguntó si la inflación reduce la demanda de dinero, vía expectativas, y se perpetúa a sí misma.

Supuestos: el mercado monetario está siempre en equilibrio, las variables reales apenas varían, las expectativas son adaptativas, $\dot\pi_t^e = \lambda(\pi_t - \pi_t^e)$, y la demanda de saldos reales es $m_t - p_t = c - a\pi_t^e$, con semielasticidad constante $-a$. En ella, $c = \kappa_0 + \kappa_1 y_t - a r_t$ recoge la demanda por factores reales distintos de la inflación esperada, y la relación es exponencial, $M_t/P_t = e^{c - a\pi_t^e}$, de modo que cada punto de inflación esperada reduce los saldos reales en torno a un a %.

Planteamiento: en estado estacionario, la inflación esperada iguala el crecimiento monetario y el Gobierno elige g_M para maximizar el señoraje:

$$\max_{g_M}\; \sigma = g_M\, e^{c - a g_M}$$

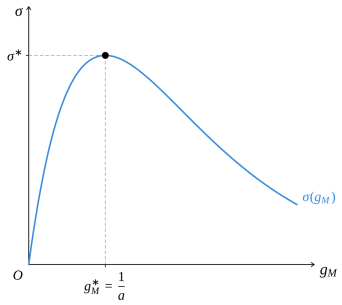
CPO: subir g_M eleva el tipo del impuesto pero reduce su base, y el óptimo iguala ambos efectos en el margen.

Solución: la CSO confirma un máximo, y la tasa óptima y el señoraje máximo son:

$$g_M^* = \frac{1}{a} \;\rightarrow\; \sigma^* = \frac{e^{c-1}}{a}$$

Desarrollo gráfico: la **curva de Laffer del señoraje** alcanza su máximo donde la elasticidad de la base respecto al tipo es -1.

Estática comparativa: el señoraje máximo crece con c y cae con a. La tasa óptima no depende de c, porque una mayor demanda de saldos solo amplía la base sobre la que recae el impuesto.



Curva de Laffer del señoraje, con máximo en $g_M^* = 1/a$.

Dinámica: si $a\lambda < 1$, el equilibrio es estable y la hiperinflación solo nace de un crecimiento monetario excesivo. Si $a\lambda > 1$, es inestable y la hiperinflación es una profecía autocumplida. Un salto de la inflación revisa al alza la expectativa, reduce la demanda de saldos reales y genera más inflación. Cagan halló $a\lambda < 1$ en la mayoría de los episodios, explicados por el crecimiento monetario, pero algo superior a la unidad en Rusia (1921-1924) y Alemania (1922-1923), lo que apunta a un componente autoalimentado de las expectativas.

Aspectos positivos: la financiación monetaria está acotada, y las hiperinflaciones se explican por un Gobierno que busca más señoraje del que permite la demanda de dinero.

Limitaciones: el máximo no es el óptimo social, por los costes de la inflación, y sustituir dinero por deuda no evita necesariamente la inflación, como muestra la aritmética monetarista desagradable.

2.2.3. Monetización a largo plazo (Blinder y Solow, 1973)

Idea: incorporan la restricción presupuestaria del Gobierno al IS-LM y cuestionan que monetizar sea más expansivo que emitir bonos: $PG + iB = PT + \dot{B} + \dot{M}$, con $T = \tau Y$. En ella, iB son los intereses de la deuda, $\dot{B}$ la emisión de deuda y $\dot{M}$ los ingresos de señoraje.

Razonamiento: a largo plazo el presupuesto se equilibra y los *stocks* dejan de variar, $\dot{B} = \dot{M} = 0$. Con dinero, $dY/dG = 1/\tau$, porque la renta crece hasta que la recaudación adicional cubre el gasto. Con bonos, hay que cubrir también los intereses de la deuda acumulada, y $dY/dG = [1 + i\,(dB/dG)]/\tau > 1/\tau$.

Con monetización, el tipo sube menos y el efecto liquidez expande la renta antes. A largo plazo, sin embargo, la deuda es más expansiva, porque obliga a crear más activos hasta que el presupuesto se equilibra, y esa mayor riqueza impulsa el consumo. Frente a la tesis monetarista de expulsión total, los efectos riqueza y de cartera tienen un sesgo expansivo, siempre que el sistema sea estable: con un efecto riqueza débil, los intereses crecen más que la recaudación y la deuda se vuelve explosiva.

Aspectos positivos: incorpora los efectos riqueza y la dinámica de la deuda.

Limitaciones: hereda los precios fijos y la falta de expectativas del IS-LM. Por ello ignora la respuesta de los agentes que subraya la equivalencia ricardiana.

2.3. Financiación en una economía abierta

El Gobierno puede financiarse también en el exterior, y los efectos de la monetización dependen del régimen cambiario. La restricción incorpora el tipo de cambio.

Deuda externa: la restricción incorpora la deuda en divisas, valorada al tipo de cambio. Influyen la prima de riesgo, que puede volver explosiva la senda, y la depreciación, que eleva el valor de esa deuda (*efecto hoja de balance*). El FMI desarrolló el *enfoque de la hoja de balance* tras las crisis emergentes de finales de los noventa para detectar descalces de plazos, de divisas y de estructura del capital. Dividiendo por el PIB nominal se obtiene una EDD para la deuda externa análoga a la interna, en la que el coste efectivo incorpora la depreciación observada. El endeudamiento externo puede ser beneficioso a corto plazo, pero debe seguir una senda no explosiva.

Monetización con tipo de cambio fijo: si no esteriliza, la base monetaria crece con el crédito interno, $\uparrow H =\uparrow \mathrm{CI} + R$. Para defender la paridad, el banco central esteriliza vendiendo divisas por el mismo importe, $\overline{H} =\uparrow \mathrm{CI} + \downarrow R$, y cada operación consume reservas y deteriora los fundamentos. Como los agentes lo prevén, un ataque especulativo se lleva las reservas restantes y el banco central abandona la paridad, como en las crisis de primera generación (Krugman, 1979a). La raíz es que, con reservas finitas, financiar el déficit y defender la paridad dejan de ser compatibles antes o después.

Monetización con tipo de cambio flexible: no hay que esterilizar, pero la monetización baja el tipo de interés, provoca salida de capitales, deprecia la moneda y puede generar inflación importada. Es decir, $\downarrow i \rightarrow \uparrow E$, con E en unidades de moneda nacional por divisa.

En suma, la sostenibilidad depende de $r - g_Y$ y de la respuesta del saldo primario a la deuda, y con deuda externa también de la prima de riesgo y del tipo de cambio. La monetización solo alivia transitoriamente, acotada por el señoraje máximo y con inflación. Queda por ver quién decide.

3. Dominancia monetaria y dominancia fiscal

La restricción presupuestaria intertemporal vincula a ambas autoridades: si una es activa, la otra debe ser pasiva. El modelo de Leeper (1991) ofrece el marco en el que se ordenan las teorías.

3.1. El modelo de Leeper (1991)

Idea: formaliza la distinción de Sargent y Wallace en un modelo dinámico con expectativas racionales en el que cada autoridad sigue una regla. Solo la combinación de una activa y otra pasiva da un equilibrio determinado.

3.1.1. Supuestos

Mundo neoclásico con precios flexibles, mercados perfectos e incertidumbre sobre *shocks* fiscales y monetarios. Los consumidores tienen expectativas racionales y una utilidad con dinero, MIU, y la economía es de dotación, con renta exógena. Así, el modelo aísla la determinación de la inflación y la deuda.

3.1.2. Modelo

Planteamiento: los hogares eligen las sendas de consumo y saldos reales:

$$\max_{\mathbf{c},\mathbf{m}} \sum_{t=0}^{\infty} \beta^t u(c_t, m_t) \quad \text{s.a.:} \begin{cases} c_t + (b_t - b_{t-1}) + \left(m_t - \frac{m_{t-1}}{1+\pi_t}\right) = y_t - T_t + r_t b_{t-1} \\ b_{-1},\ m_{-1} \text{ dados} \end{cases},$$

CPO: la demanda de dinero iguala el coste de una unidad de dinero hoy a su utilidad como liquidez más su valor real mañana, y la ecuación de Euler iguala la relación marginal de sustitución intertemporal al rendimiento bruto de los bonos.

Solución: combinándolas, la relación marginal de sustitución entre saldos y consumo iguala el coste de oportunidad del dinero, con $1 + i_t = (1 + r_{t+1})(1 + \pi_{t+1})$:

$$\frac{u'_{m_t}}{u'_{c_t}} = \frac{i_t}{1 + i_t}$$

La demanda de dinero cae, así, con el tipo nominal y con la inflación esperada. Por este canal, el del señoraje, la política fiscal puede acabar influyendo en los precios.

Sector público: el déficit primario más los intereses reales se financian con bonos y señoraje, y cada autoridad sigue una regla, con *shocks* AR(1):

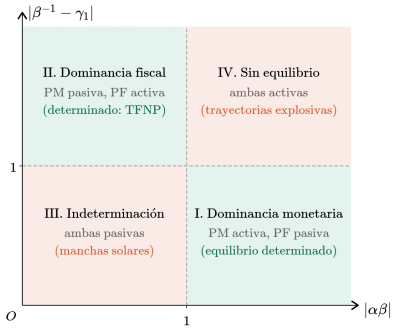
$$i_t = \alpha_0 + \alpha\pi_t + \varepsilon_t^m$$
$$T_t = \gamma_0 + \gamma_1 b_{t-1} + \varepsilon_t^f$$

La regla monetaria responde a la inflación con α, y la fiscal, a la deuda heredada con γ_1.

Equilibrio: linealizado, la solución depende de dos parámetros:

- La autoridad monetaria es **activa** si $|\alpha\beta| > 1$, es decir, si sube el tipo real ante la inflación (principio de Taylor), y pasiva en caso contrario.
- La autoridad fiscal es **pasiva** si $|\beta^{-1} - \gamma_1| < 1$, es decir, si los impuestos responden a la deuda más de lo que cuesta su servicio, $\gamma_1 > \beta^{-1} - 1 \approx r$. La deuda se estabiliza entonces, y la autoridad es activa en caso contrario.

Regiones: en la I, **dominancia monetaria**, que corresponde al caso convencional, la política monetaria determina la inflación y la restricción se cumple vía impuestos. En la II, **dominancia fiscal**, el nivel de precios lo determina la restricción presupuestaria del Gobierno. Ambas dan un equilibrio único. Con las dos pasivas, región III, hay indeterminación y equilibrios de manchas solares, mientras que con las dos activas, región IV, no hay equilibrio y la deuda explota. En el plano $|\alpha\beta|$-$|\beta^{-1} - \gamma_1|$, las rectas $|\alpha\beta| = 1$ y $|\beta^{-1} - \gamma_1| = 1$ delimitan las cuatro regiones. La estimación de los parámetros de las reglas permite saber qué régimen rige, y la coordinación no es opcional, sino un requisito de la determinación del equilibrio.



Modelo de Leeper: solo las regiones I y II, una autoridad activa y otra pasiva, dan un equilibrio determinado.

Dinámica: bajo dominancia monetaria, una rebaja de impuestos financiada con deuda la absorben los impuestos futuros y la inflación no varía. Bajo dominancia fiscal, la deuda solo se absorbe si su valor real cae y la inflación sube. Incluso un *shock* monetario contractivo la eleva, porque encarece los intereses. Si las reglas cambian de régimen, la expectativa de dominancia fiscal futura afecta ya a la inflación.

Aspectos positivos: demuestra que la estabilidad de precios depende de la combinación de reglas monetarias y fiscales, con implicaciones para el diseño de la UEM, y fundamenta la teoría fiscal del nivel de precios.

Limitaciones: reglas lineales, sin sector exterior y con un agente representativo ricardiano. Por ello no captura las restricciones de liquidez que reforzarían la dominancia fiscal.

3.2. Dominancia fiscal

3.2.1. Aritmética monetarista desagradable (Sargent y Wallace, 1981)

Idea: una política fiscal sistemáticamente expansiva impide al banco central controlar la inflación, hasta el punto de que financiar el déficit con deuda llega a ser más inflacionista que monetizarlo.

Razonamiento: la secuencia tiene cuatro pasos.

1. Si el banco central no monetiza, el déficit se financia con deuda.
2. Como $r > g_Y$, la deuda crece como una bola de nieve, y el Estado compite por los fondos prestables. La deuda acumulada ensancha además la diferencia entre r y g_Y.
3. Cuando la deuda se vuelve insostenible, el Gobierno solo puede pagar los intereses creando dinero.
4. Como los agentes lo prevén, la demanda de saldos cae y la inflación sube ya hoy.

Paradójicamente, una política monetaria contractiva hoy puede generar más inflación mañana si la autoridad fiscal no coopera, porque exige más deuda y, dado el límite a su absorción, más dinero en el futuro. La inflación es, así, un fenómeno más fiscal que monetario, y cuanto más se retrase la monetización, peor, porque habrá más intereses que monetizar. Este resultado fundamenta la disciplina fiscal como complemento de la independencia del banco central, pero descansa en supuestos fuertes: $r > g_Y$, un límite exógeno a la deuda y una autoridad fiscal que no reacciona a la monetaria. Si alguno de ellos falla, la paradoja deja de estar garantizada.

3.2.2. Teoría fiscal del nivel de precios

Idea: la TFNP (Leeper, 1991; Sims, 1994; Woodford, 1995; Cochrane, 2001) generaliza esa intuición. Bajo dominancia fiscal, la RPIG deja de ser una restricción para convertirse en una condición de equilibrio, sin necesidad de señoraje. El valor real de la deuda heredada debe igualar el valor actual de los superávits esperados:

$$\frac{B_{t-1}}{P_t} = \mathbb{E}_t \sum_{s=0}^{\infty} \beta^s \left(T_{t+s} - G_{t+s}\right)$$

Si la autoridad fiscal fija los superávits con independencia de la deuda y la monetaria no reacciona, un aumento del gasto sin ajuste fiscal futuro se traslada a P_t, que sube hasta restaurar la igualdad. Por ello, la independencia del banco central es necesaria pero no suficiente para la estabilidad de precios, una tensión central en el debate sobre la arquitectura de la UEM. La TFNP integra, además, la política fiscal en la determinación del nivel de precios y da el marco con el que se analiza la madurez de la deuda. Su supuesto central es, con todo, controvertido, y los dos regímenes pueden ser observacionalmente equivalentes.

3.3. Dominancia monetaria

3.3.1. Aritmética fiscal desagradable (King, 1995)

Idea: es el argumento simétrico al de Sargent y Wallace, en el contexto de los bancos centrales independientes con objetivos de inflación de los noventa: un banco central activo y creíble disciplina a la autoridad fiscal, porque la RPIG se convierte en una restricción efectiva sobre el presupuesto. Si el Gobierno gasta más sin subir impuestos, no puede esperar que la inflación licúe la deuda. Los mercados anticipan un ajuste y, si no llega, la prima de riesgo se dispara. La carga del ajuste recae así sobre el presupuesto, con costes políticos y sociales, como en la crisis de deuda soberana de la periferia del euro, en 2010-2014. Fundamenta las reglas fiscales que acompañan a los objetivos de inflación.

El resultado presupone que la autoridad fiscal cede. Si no, el conflicto puede acabar en crisis de deuda o en una dominancia fiscal *de facto*. Esta se manifestaría como una presión sobre el banco central para que acomode la política fiscal.

3.3.2. Debate sobre las reglas fiscales

Definición: una **regla fiscal** es una restricción permanente sobre la política fiscal expresada mediante un indicador sintético de los resultados presupuestarios. Ancla las expectativas sobre la deuda y construye espacio fiscal. Hay cuatro tipos:

- De **saldo**, como la regla chilena de superávit estructural. Limitan el déficit, es decir, exigen el equilibrio presupuestario.
- De **deuda**.
- De **gasto**, como la suiza o la española, $g_G \leq \bar{g}_Y$. La española limita el crecimiento del gasto computable a la tasa de referencia de crecimiento del PIB a medio plazo.
- De **ingresos**.

Las complementan instituciones fiscales independientes como la AIReF (2013) y marcos como el PEC (1997) y el Pacto Presupuestario Europeo (2012), diseñados para hacer coherente una política monetaria activa, la del BCE, con políticas fiscales pasivas en los Estados miembros. La generalización de las reglas acompañó a la independencia de los bancos centrales en los noventa.

Aspectos positivos: limitan la tendencia de la discrecionalidad a acumular déficits y refuerzan la credibilidad, lo que permite a la política fiscal ser pasiva. El banco central puede centrarse así en la estabilidad de precios.

Limitaciones: se critican por procíclicas. La reforma del PEC (2024) combina un ancla de sostenibilidad, una ratio de deuda en senda descendente verosímil, con una senda de gasto neto como instrumento, aplicada mediante planes fiscales nacionales a medio plazo.

3.4. Otros desarrollos y evidencia

3.4.1. Teorema de irrelevancia (Wallace, 1981)

Idea: es el Modigliani-Miller de la banca central. Con mercados completos, sin fricciones y con agentes ricardianos, la composición del balance del banco central es irrelevante, porque los agentes neutralizan las operaciones de mercado abierto ajustando su cartera. Eggertsson y Woodford (2003) lo extienden al ZLB, donde las compras de activos son irrelevantes si no alteran las expectativas sobre la senda futura de tipos. La potencia de la política monetaria descansa, así, en el incumplimiento de estas condiciones, y la evidencia sobre las compras de activos sugiere que el balance sí importa. El teorema trata esas operaciones como cambios en la composición de los pasivos del sector público consolidado, que no alteran la senda de déficits ni la deuda pública neta.

3.4.2. Madurez de la deuda (Cochrane, 2001)

Idea: bajo dominancia fiscal, un *shock* fiscal reduce el valor real de los pasivos públicos, y la madurez decide cómo. Con deuda a corto, el ajuste es un salto inmediato del nivel de precios y una subida de tipos encarece enseguida la deuda. Con deuda a largo, cae hoy el precio de los bonos y el ajuste se reparte en el tiempo. La gestión de la madurez es, así, un instrumento de política. Las compras de activos, al cambiar deuda a largo por reservas a corto, acortan la madurez de los pasivos consolidados. Conecta la gestión de la deuda con la política monetaria, aunque comparte las críticas a la TFNP. Con deuda a largo, además, el presupuesto queda parcialmente aislado de las decisiones del banco central.

3.4.3. Evidencia empírica

La relación entre déficit observado e inflación no es concluyente. Lo relevante es el **déficit estructural persistente**, que acaba acompañado de más crecimiento monetario e inflación, es decir, de dominancia fiscal *de facto*. Esa asociación es coherente con la aritmética monetarista desagradable.

En la OCDE tras Bretton Woods, la monetización fue más frecuente donde el banco central era menos independiente. El "divorcio" entre el Tesoro y la Banca d'Italia (1981) precedió a la desinflación, y Maastricht prohibió la financiación monetaria, hoy en el artículo 123 del TFUE. Cuanto mayor es la independencia del banco central, mayor es la probabilidad de dominancia monetaria. El divorcio italiano suprimió la obligación de la Banca d'Italia de adquirir los títulos no colocados en subasta.

En suma, Leeper ordena las teorías: la aritmética monetarista y la TFNP describen la dominancia fiscal, y la aritmética fiscal y las reglas, la monetaria. Una mayor independencia se asocia con menos inflación (Alesina y Summers, 1993), aunque un déficit estructural persistente puede imponer la dominancia fiscal.

A40. Efectividad e interrelación de las políticas monetaria y fiscal en las principales economías desarrolladas desde la Gran Recesión de 2008. El valor de los multiplicadores fiscales

1. Efectividad de las políticas monetaria y fiscal

Antes de estudiar su interrelación, la efectividad de cada política se mide por separado a través de su multiplicador, primero el monetario y después el fiscal.

1.1. Multiplicador monetario

1.1.1. Concepto

Enfoque operativo: en sentido estricto, el **multiplicador monetario** relaciona la base monetaria, H, con la oferta monetaria, y depende de los coeficientes de efectivo, $c_D = E/D$, y de reservas, q:

$$M^s = \frac{1+c_D}{c_D+q}H \quad \rightarrow \quad m = f(c_D^-, q^-)$$

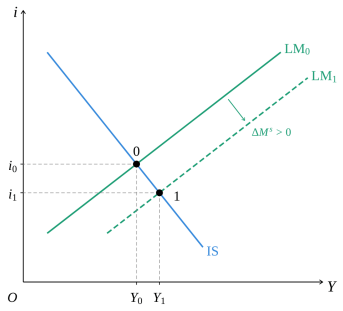
La autoridad controla, así, la oferta monetaria de forma imperfecta, porque el multiplicador depende del público y de los bancos, es procíclico y cae bruscamente en las crisis financieras, como la de 2008.

Enfoque teórico: en este tema interesa el **multiplicador de la política monetaria**, la variación del *output* de equilibrio ante una variación unitaria de la oferta real de dinero. En el IS-LM de economía cerrada:

$$M_M \equiv \frac{\partial Y}{\partial (M^s/P)} = \frac{v/h}{1-c(1-\tau)+v\frac{k}{h}} > 0$$

– c es la propensión marginal al consumo, τ el tipo impositivo, v la sensibilidad de la inversión al tipo de interés, y k y h las de la demanda de dinero a la renta y al tipo.

La transmisión es indirecta: más dinero baja el tipo, eleva la inversión y la renta. La política monetaria es más efectiva cuanto más sensible es la inversión al tipo, con una IS más plana, y cuanto menos lo es la demanda de dinero, con una LM más rígida. En la **trampa de liquidez**, $h \to \infty$, el multiplicador es nulo; en el caso clásico, $h = 0$, alcanza su máximo, $M_M = 1/k$, la velocidad de circulación.



Expansión monetaria en el IS-LM: la LM se desplaza, baja el tipo y sube la renta de Y_0 a Y_1.

1.1.2. Perspectivas teóricas

Visión keynesiana: la política monetaria es poco efectiva, porque la demanda de dinero es muy sensible al tipo, por el motivo especulación, y la inversión, dominada por los *animal spirits*, responde poco. Es como una cuerda, que sirve para tirar pero no para empujar, y por eso los keynesianos priorizan la política fiscal.

Visión de la teoría del ciclo real: el multiplicador es nulo incluso a corto plazo. Con precios flexibles rige la dicotomía clásica, el dinero es neutral y la correlación entre dinero y *output* se interpreta como causalidad inversa (King y Plosser, 1984), porque el dinero interno responde a la actividad.

Visión NEK: las rigideces nominales, como los costes de menú o la fijación escalonada de precios a la Calvo, dan un multiplicador positivo a corto plazo, porque el tipo nominal mueve el tipo real, aunque a largo plazo se recupera la neutralidad. Esta visión fundamenta el consenso previo a 2008: la política monetaria como instrumento principal, con reglas de tipo de interés, y la fiscal limitada a los estabilizadores automáticos.

Crecimiento: la cuestión es la **superneutralidad**. Sidrauski (1967) la demuestra con dinero en la función de utilidad, pero Tobin (1965) obtiene un efecto real positivo, porque la inflación desplaza la cartera hacia capital. La evidencia matiza ambos, pues la inflación elevada perjudica el crecimiento, con efectos no lineales a partir de ciertos umbrales.

1.2. Multiplicador fiscal

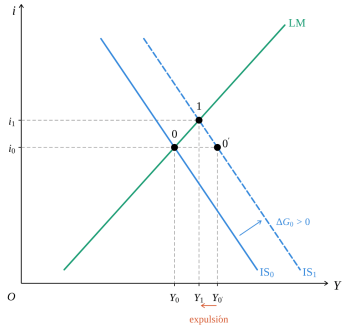
1.2.1. Concepto

Definición: el **multiplicador fiscal** mide la variación del *output* de equilibrio ante una variación unitaria del gasto, los impuestos o las transferencias. Se distingue el multiplicador de impacto, el acumulado en un horizonte y el de valor presente. La evidencia se refiere, salvo indicación, al acumulado del gasto.

Lado del gasto: en el IS-LM, el numerador recoge el impulso inicial y el mecanismo multiplicador keynesiano, y el denominador, las filtraciones: el ahorro, los impuestos y el **efecto expulsión** monetario, vk/h:

$$M_G \equiv \frac{\partial Y}{\partial G_0} = \frac{1}{1-c(1-\tau)+v\frac{k}{h}} > 0$$

– El multiplicador es mayor cuanto mayor es c y menores son τ, v y la pendiente de la LM, k/h.
– En la trampa de liquidez desaparece la expulsión y se alcanza el máximo keynesiano, $1/[1-c(1-\tau)]$; en el caso clásico la expulsión es total y $M_G = 0$.



Expansión fiscal en el IS-LM: la diferencia entre $Y_{0'}$ e Y_1 mide el efecto expulsión.

Lado del ingreso: una subida del tipo impositivo reduce la renta disponible, el consumo y la renta, $M_\tau = -cY/[1-c(1-\tau)+vk/h] < 0$. Con un impuesto de suma fija, el multiplicador, M_T, es menor en valor absoluto que el del gasto, porque los impuestos operan a través del consumo y una parte, $1-c$, se filtra al ahorro. De esa asimetría nace el **multiplicador del presupuesto equilibrado** (Haavelmo, 1945): en el modelo keynesiano simple, un gasto financiado con impuestos de suma fija tiene un multiplicador igual a la unidad.

1.2.2. Perspectivas teóricas

Visión keynesiana: el multiplicador es superior a la unidad y la política fiscal es el instrumento principal, gracias a la capacidad ociosa, al consumo dependiente de la renta corriente y a un efecto expulsión limitado.

Visión de la teoría del ciclo real: el multiplicador es inferior a la unidad o nulo, por dos mecanismos:

– La **equivalencia ricardiana** (Barro, 1974): con horizonte infinito, expectativas racionales, mercados completos e impuestos de suma fija, los hogares ahorran ante los impuestos futuros y financiar con deuda o con impuestos es equivalente.
– El **canal neoclásico del gasto**: un aumento permanente de G genera un efecto riqueza negativo que eleva la oferta de trabajo. El *output* sube menos que el gasto y cae el consumo. A largo plazo, el multiplicador puede superar la unidad por la acumulación de capital.

Visión NEK: el multiplicador es **estado-dependiente** (reducido con una regla de Taylor activa y elevado en el límite inferior de los tipos, el ZLB). Los hogares no ricardianos con restricciones de liquidez, o *rule-of-thumb*, lo elevan.

Crecimiento: a largo plazo importan la composición y la deuda. La inversión pública productiva eleva la producción potencial (Aschauer, 1989). Las *consolidaciones fiscales expansivas* (Giavazzi y Pagano, 1990) operarían vía expectativas, pero la evidencia del FMI las encuentra, en general, contractivas a corto plazo. Reinhart y Rogoff (2010) asociaron una deuda superior al 90 % del PIB con menos crecimiento, resultado que cuestionaron Herndon et al. (2013).

En suma, la magnitud de los multiplicadores es condicional: depende del régimen cambiario, la movilidad del capital, la posición cíclica, las expectativas y el marco institucional. En condiciones normales la política monetaria es el instrumento principal, pero en el ZLB la fiscal recupera protagonismo. La pregunta relevante es la combinación óptima de ambas, que Mundell (1962) formuló como **principio de asignación** y Leeper (1991) como coordinación activa-pasiva.

2. Interrelación de las políticas monetaria y fiscal

Ambas políticas se interrelacionan en dos planos: a corto plazo, por la respuesta de la política monetaria al impulso fiscal, crítica en el límite inferior de los tipos; y en el plano intertemporal, por la restricción presupuestaria del sector público, que da lugar a los regímenes de dominancia.

2.1. Corto plazo: IS-PC-MR y ZLB

Origen: es el modelo canónico de tres ecuaciones de la NEK, en la versión con límite inferior de los tipos de Eggertsson y Woodford (2003).

Idea: la efectividad de la política fiscal depende del régimen monetario: es baja cuando el banco central es activo y contrarresta el impulso, y alta cuando está restringido por el límite inferior.

Planteamiento: una IS dinámica, una curva de Phillips neokeynesiana y una regla de tipo de interés:

$$\begin{aligned} \text{IS:}\ & \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(i_t - \mathbb{E}_t[\pi_{t+1}] - \bar{r}_t) + g_t \\ \text{CP:}\ & \pi_t = \beta \mathbb{E}_t[\pi_{t+1}] + \kappa \tilde{y}_t \\ \text{Regla:}\ & i_t = \bar{i}_t + \alpha_\pi \tilde{\pi}_t + \alpha_y \tilde{y}_t \end{aligned}$$

– $\tilde{y}_t$ es la brecha de producción, $\bar{r}_t$ el tipo real natural, g_t el impulso fiscal, κ la pendiente de la curva de Phillips, $\bar{i}_t = \bar{r}_t + \pi^T$ el tipo nominal neutral y $\alpha_\pi > 1$ el principio de Taylor. La política fiscal entra como un desplazamiento de la IS.

Condiciones normales: un impulso fiscal eleva la brecha y la inflación, el banco central sube el tipo nominal más que proporcionalmente a la inflación y el tipo real contrarresta parte del impulso, de modo que el multiplicador es inferior a la unidad.

Límite inferior de los tipos: si un *shock* negativo de demanda lleva el tipo natural por debajo de $-\pi^T$, la regla prescribe un tipo nominal negativo, pero $i_t \geq 0$, porque los agentes pueden mantener efectivo. El tipo real queda por encima del natural, la brecha es negativa, la inflación cae y la menor inflación esperada eleva aún más el tipo real, en una **espiral deflacionista**. La interrelación se invierte:

1. La política monetaria convencional pierde eficacia y solo actúa sobre las expectativas, con *forward guidance*, el compromiso de mantener el tipo en cero más allá de lo que prescribiría la regla, o sobre los tipos a largo y las primas, con compras de activos.
2. El multiplicador fiscal es elevado: el tipo nominal no sube y la mayor inflación esperada baja el tipo real, de modo que la expulsión se convierte en atracción (*crowding in*). Christiano, Eichenbaum y Rebelo (2011) obtienen multiplicadores de unos 3,7 en su modelo simple y 2,3 en el de escala media. En general, superan la unidad siempre que el estímulo se limite al periodo en el ZLB.
3. La coordinación se vuelve necesaria: una expansión fiscal con el compromiso monetario de no reaccionar a la inflación es la vía más efectiva para salir del ZLB, lo que difumina la frontera entre ambas políticas.

Aspectos positivos: explica por qué la política fiscal volvió a estabilizar cuando la Gran Recesión y la pandemia situaron a las principales economías en el límite inferior durante más de una década.

Limitaciones: supone $i_t \geq 0$, cuando el límite efectivo está algo por debajo de cero, y trata la política fiscal como exógena, sin restricción intertemporal, por lo que no permite analizar la sostenibilidad ni la dominancia.

2.2. Intertemporal: modelo de Leeper (1991)

Idea: Leeper formaliza la distinción de Sargent y Wallace (1981) en un modelo dinámico con expectativas racionales. La restricción presupuestaria intertemporal del sector público vincula a ambas autoridades.

Planteamiento: los consumidores tienen expectativas racionales y dinero en la función de utilidad, y el sector público financia el déficit con deuda o con base monetaria. La autoridad monetaria fija el tipo según la inflación corriente, y la fiscal, los impuestos según la deuda real heredada, ambas con *shocks*:

$$\begin{aligned} i_t &= \alpha_0 + \alpha \pi_t + \varepsilon_t^m \\ T_t &= \gamma_0 + \gamma_1 b_{t-1} + \varepsilon_t^f \end{aligned}$$

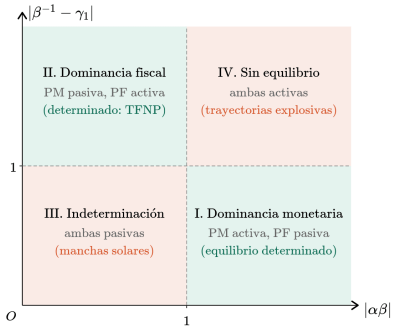
– La regla monetaria es una regla de Taylor sin término de brecha, de modo que $|\alpha\beta| > 1$ equivale aproximadamente al principio de Taylor, con β próximo a la unidad.

Una autoridad es **activa** si fija su instrumento con independencia de la restricción presupuestaria y **pasiva** si lo ajusta para garantizar su cumplimiento.

Equilibrio: de las cuatro regiones, solo dos dan un equilibrio determinado:

– *Dominancia monetaria*: política monetaria activa, $|\alpha\beta| > 1$, y fiscal pasiva, $|\beta^{-1} - \gamma_1| < 1$. El banco central determina la inflación y los impuestos estabilizan la deuda.
– *Dominancia fiscal*: política fiscal activa y monetaria pasiva. El nivel de precios se ajusta para que se cumpla la restricción presupuestaria.

Si ambas son activas, la deuda es explosiva; si ambas son pasivas, el equilibrio está indeterminado. La coordinación no es, por tanto, opcional, sino un requisito de la determinación del equilibrio.



Modelo de Leeper (1991).

Aspectos positivos: la estabilidad de precios depende de la combinación de reglas, no solo del banco central, y la taxonomía sustenta la teoría fiscal del nivel de precios.

Limitaciones: reglas lineales, economía cerrada y un agente representativo ricardiano.

2.3. Teorías de la dominancia

Cada teoría explica una región del modelo de Leeper:

– *Dominancia fiscal*: en la *aritmética monetarista desagradable* (Sargent y Wallace, 1981), si la autoridad fiscal no ajusta sus déficits, una política monetaria restrictiva hoy exige más deuda y más monetización mañana. Los agentes lo anticipan y la inflación surge ya, como fenómeno fiscal. La *teoría fiscal del nivel de precios* (Leeper, 1991; Sims, 1994; Woodford, 1995) lo generaliza, pues los precios se ajustan para que el valor real de la deuda nominal iguale el valor actual de los superávits esperados.
– *Dominancia monetaria*: en la *aritmética fiscal desagradable* (King, 1995), un banco central creíble fija tipos altos que comprometen la financiación del Gobierno y fuerzan el ajuste presupuestario. Las reglas fiscales, como el PEC y su reforma de 2024, son su corolario institucional.
– *Otros desarrollos*: por el *teorema de irrelevancia* (Wallace, 1981), sin fricciones las operaciones de mercado abierto no afectan a los déficits, la deuda neta ni los precios, y Eggertsson y Woodford (2003) lo extienden al ZLB. La potencia monetaria descansa en que fallen esos supuestos. Según Cochrane (2001), bajo dominancia fiscal el ajuste recae sobre los precios corrientes si la deuda es a corto, y sobre la inflación futura si es a largo.

Ambos desarrollos razonan con el balance consolidado del banco central y el Tesoro. En suma, la interacción opera en los dos planos y tiene implicaciones directas para la UEM, donde la independencia del BCE convive con veintiuna políticas fiscales nacionales.

3. Evidencia desde la Gran Recesión: multiplicadores e interrelación

Queda por contrastar la teoría con los datos: las técnicas de medición, los multiplicadores fiscal y monetario y la interrelación desde 2008.

3.1. Técnicas de medición

VAR: los SVAR estiman los multiplicadores con funciones de impulso-respuesta. Blanchard y Perotti (2002) identifican los *shocks* fiscales con los retardos institucionales, suponiendo que el gasto no reacciona en el mismo periodo al *output*. Ramey (2011) muestra que las previsiones y los *shocks* narrativos anticipan los *shocks* VAR, que están, así, mal fechados. En política monetaria, Christiano, Eichenbaum y Evans (1999) ordenan el tipo de interés detrás del *output* y los precios, al suponer que reacciona a ellos en el mismo periodo, pero no a la inversa. Los resultados son sensibles a la identificación, los retardos y la muestra.

Narrativas: identifican cambios exógenos de política en registros históricos, como actas, discursos o documentos legislativos, para aislar las decisiones motivadas por objetivos de largo plazo, como la consolidación o la lucha contra la inflación, de las respuestas al ciclo. Romer y Romer (2010) estiman un multiplicador impositivo en torno a 3 en EE. UU. (1945-2007): cada dólar de subida de impuestos reduce el PIB en tres a los dos años y medio. En política monetaria, el mismo enfoque usa los registros y las previsiones internas de la Fed. Evitan la endogeneidad de los VAR, pero son subjetivas y difíciles de replicar.

Experimentos naturales: explotan variaciones exógenas institucionales o geográficas, como el gasto militar regional en EE. UU. En el ámbito monetario, Jordà, Schularick y Taylor (2015) explotan el trilema de Mundell-Fleming en un panel de 17 países (1870-2013). Con tipo fijo y libre movilidad de capital, cada país importa la política monetaria del ancla, lo que permite estudiar su efecto sobre el crédito hipotecario y la vivienda. Su identificación causal es robusta, pero su validez externa, limitada.

3.2. Evidencia sobre el multiplicador fiscal

Enfoque narrativo (Ramey y Shapiro, 1998; Ramey, 2011): con los episodios de gasto militar en EE. UU., fechados cuando se anuncian, el multiplicador del gasto se sitúa entre 0,6 y 1,2. El consumo privado y los salarios reales caen, como predice el canal neoclásico.

SVAR (Blanchard y Perotti, 2002): con datos de EE. UU. (1960-1997), $M_G \in [0{,}84;\ 1{,}29]$ y $M_T \in [-0{,}78;\ -0{,}7]$. A diferencia del enfoque narrativo, el gasto eleva el consumo privado, pero reduce la inversión, y lo difícil de conciliar con la teoría keynesiana es que también las subidas de impuestos la reduzcan.

Estado de la economía: con VAR de transición suave, Auerbach y Gorodnichenko (2012) obtienen 2,2-2,5 en recesión y no más de 0,6 en expansión, por la capacidad ociosa y las restricciones de liquidez. Ramey y Zubairy (2018), con proyecciones locales y datos de 1889-2015, lo cuestionan: el multiplicador queda en 0,6-0,7 con desempleo alto o bajo. Atribuyen la discrepancia a suponer recesiones de 20 trimestres cuando duran 3,3 de media. En el ZLB, algunas especificaciones dan hasta 1,5.

Metarregresiones (Gechert y Will, 2012): en 89 estudios, el multiplicador depende sobre todo del contexto y del método. La inversión pública da los más altos, y las transferencias y los impuestos, los más bajos. El FMI, con VAR de cambio de régimen, encuentra también multiplicadores mayores en recesión y propone calibrar el multiplicador de cada país en dos pasos, según su apertura, rigidez laboral, estabilizadores, régimen cambiario, deuda y posición cíclica.

Panel internacional (Ilzetzki et al., 2013): en 44 países, el multiplicador es:

- mayor en los países de renta alta que en los países en desarrollo;
- relativamente grande con tipo predeterminado y nulo con tipo flexible, como predice Mundell-Fleming;
- menor en las economías más abiertas, por la filtración hacia las importaciones, y
- negativo a largo plazo con deuda superior al 60 % del PIB.

Enfoque regional (Nakamura y Steinsson, 2014): con el gasto de defensa por estados, el *multiplicador relativo de economía abierta* ronda 1,5. No es el agregado, porque la respuesta monetaria y fiscal federal se cancela entre regiones, pero informa sobre el estímulo dentro de una unión monetaria.

Contexto: por instrumentos, en siete modelos DSGE el multiplicador del gasto, de 1,0-1,5 el primer año, supera al de transferencias, 0,3-0,8, y al de impuestos, 0,2-0,5 (Coenen et al., 2012). Blanchard y Leigh (2013) mostraron que el FMI usó multiplicadores de unos 0,5 cuando los reales estaban entre 0,9 y 1,7, y la austeridad en Grecia, Portugal, Irlanda y España contrajo el PIB más de lo previsto.

Persistencia: el efecto sigue una U invertida, con un máximo entre el primer y el segundo año y una duración de unos cinco. Es mayor con histéresis (DeLong y Summers, 2012) y con acomodación monetaria.

3.3. Evidencia sobre el multiplicador monetario

SVAR (Bernanke y Mihov, 1998): para medir la política hay que identificarla en el mercado de reservas, con indicadores distintos según el periodo. Un *shock* expansivo eleva el *output* con un máximo a los doce o dieciocho meses, y los precios responden más tarde y de forma persistente.

Medidas no convencionales: el QE1 de la Fed redujo la prima por plazo a diez años entre 30 y 100 puntos básicos (Gagnon et al., 2011), con efectos que dependen de los activos comprados: Krishnamurthy y Vissing-Jorgensen (2011) hallan evidencia de los canales de señalización, de demanda de activos seguros y de inflación, y de un canal hipotecario en el QE1, y relativizan el de duración. Los anuncios del APP del BCE (2014-2015) redujeron los tipos soberanos a diez años en 30-50 pb (Altavilla et al., 2015).

Efectividad real: la transmisión a la economía real fue más débil de lo esperado. Según el Banco de Inglaterra (Joyce et al., 2011), su primer QE, de 200.000 M£, elevó el PIB en torno a un 1,5-2 % en su máximo. La asimetría entre el impacto financiero, elevado, y el real, moderado, se explica por las restricciones crediticias de la banca y el desapalancamiento privado.

***Forward guidance* y tipos negativos**: el compromiso de la Fed de mantener tipos bajos hasta 2014 buscaba reducir los tipos esperados y, con ellos, los de largo plazo. Su efectividad depende de la credibilidad, un problema de inconsistencia dinámica (Kydland y Prescott, 1977). El BCE mantuvo la facilidad de depósito en negativo entre 2014 y 2022, con un mínimo de -0,50 %. Su experiencia y las del Riksbank, el Banco Nacional Suizo y el BoJ muestran que el límite inferior efectivo está por debajo de cero, aunque con rendimientos decrecientes y el riesgo de un *reversal rate* (Brunnermeier y Koby, 2018).

3.4. Evidencia sobre la interrelación

Antes de 2008, la monetización del déficit fue más frecuente donde los bancos centrales eran menos independientes, como Italia, que donde lo eran más, como EE. UU. o Alemania.

3.4.1. La Gran Recesión (2009-2015)

EE. UU.: la ARRA (2009) ascendía a 787.000 M$ y la Fed mantuvo los tipos en el 0-0,25 % de 2008 a 2015, con tres QE y *forward guidance*. Desde 2011, la Budget Control Act y el *sequester* giraron hacia la restricción, y el déficit federal bajó del 9,8 % del PIB en 2009 al 2,8 % en 2014, una consolidación en el ZLB, donde el modelo predice multiplicadores más altos.

Reino Unido: el presupuesto de emergencia de junio de 2010 abrió una consolidación plurianual, mientras el Banco de Inglaterra, con el tipo en el 0,5 % desde 2009, ampliaba sus compras de deuda de 200.000 a 375.000 M£ entre 2011 y 2012. Como en EE. UU., la restricción fiscal coincidió con una política monetaria que solo podía actuar comprando activos.

3.4.2. Deuda soberana europea y Japón

Zona del euro (2010-2012): la consolidación agresiva bajo dominancia monetaria, con multiplicadores elevados, generó una espiral recesión-austeridad. El *whatever it takes* de Draghi (2012) y las OMT dieron a la política monetaria un papel cuasifiscal, al reducir las primas sin monetización directa.

Japón (Abenomics, 2013-2020): las tres flechas de Abe, el caso más claro de interacción deliberada, combinaron QQE, estímulo fiscal y reformas. Los resultados fueron mixtos: el 2 % solo se alcanzó en 2022, por un *shock* energético externo, y el BoJ cerró la etapa en marzo de 2024, al abandonar los tipos negativos.

3.4.3. La pandemia (2020-2021)

La coordinación alcanzó niveles sin precedentes en tiempos de paz: estímulos del 10-20 % del PIB y compras récord de deuda, más de 1,7 BN€ en el PEPP del BCE y unos 3,5 BN$ en la Fed. La línea entre ambas políticas se difuminó, y en la zona del euro las compras llegaron a calificarse de financiación monetaria, una lectura que el BCE rechazó (Schnabel, 2020).

Inflación de origen fiscal: la inflación de 2021-2023 reavivó el debate sobre la dominancia fiscal. En clave de la teoría fiscal del nivel de precios, Bianchi, Faccini y Melosi (2023) distinguen la deuda respaldada por superávits futuros de la que no lo está. Los *shocks* fiscales no financiados sostuvieron la recuperación, pero causaron un aumento persistente de la inflación, y su modelo predice fuera de muestra el efecto del ARPA de 2021. La lectura rival atribuye la mayor parte del repunte a *shocks* de precios, como la energía, los alimentos o la escasez de bienes, y un efecto menor, aunque más persistente, a la tensión del mercado de trabajo (Bernanke y Blanchard, 2025).

3.4.4. El endurecimiento (2022-2026)

Subidas de tipos: la Fed subió del 0-0,25 % al 5,25-5,50 % en 16 meses, y el BCE llevó la facilidad de depósito de -0,50 % al 4 % en 14, con déficits aún elevados (una autoridad monetaria activa sin una fiscal todavía pasiva). Como predice la aritmética fiscal desagradable de King (1995), el servicio de la deuda de EE. UU. pasó de unos 350.000 M$ en 2021 a unos 900.000 M$ en 2024, la prima italiana repuntó hasta unos 250 pb en otoño de 2022 y creció la presión por un ajuste fiscal futuro.

Reino Unido (2022): es el ejemplo más nítido de un fallo de credibilidad fiscal. El minipresupuesto de septiembre, con rebajas de impuestos financiadas con deuda, elevó unos 130 pb en tres sesiones el rendimiento de los *gilts* a 30 años, y su caída de precio obligó a los fondos LDI de los planes de pensiones a vender, lo que realimentó la caída. El Banco de Inglaterra compró 19.300 M£ de *gilts*, con indemnización del Tesoro, y aun así subió el tipo 75 pb, hasta el 3 %. Separó las compras temporales, para la estabilidad financiera, de los tipos, para la inflación.

En la zona del euro, el TPI (julio de 2022) buscó evitar la fragmentación sin desencadenar una crisis de deuda. El endurecimiento solo es sostenible si la autoridad fiscal acaba siendo pasiva, la tensión central del modelo de Leeper. Tras las bajadas de 2024-2025, la Fed y el BCE volvieron a subir los tipos en 2026.

Pérdidas de los bancos centrales: al remunerar las reservas por encima del rendimiento de sus carteras, incurrieron en pérdidas, que llegan al Tesoro: el BCE perdió 7944 M€ en 2024 y 1254 M€ en 2025, que compensará con beneficios futuros; la Fed suspendió sus transferencias en 2022 y registra un activo diferido, de 243.481 M$ a finales de 2025, y el Tesoro británico indemniza al APF del Banco de Inglaterra. Se debate si remunerar todas las reservas o solo un tramo, y si el endurecimiento cuantitativo, que materializa pérdidas al vender bonos, es también una decisión fiscal. El BCE dejó de remunerar las reservas mínimas en 2023.

1. Concepto, costes y efectos sobre el bienestar

El análisis empieza por definir y medir la inflación y sigue con sus costes, entre ellos los que impone su persistencia, y sus beneficios, de cuya comparación resulta la inflación óptima.

1.1. Concepto y medición

Definición: la **inflación** es el aumento sostenido del nivel general de precios. El carácter sostenido la distingue de una subida aislada, que altera el nivel de precios y solo transitoriamente su tasa de variación. Como fenómeno dinámico, exige considerar las expectativas.

Medición: hay tres medidas principales.

- El **IPC** es un índice de Laspeyres de una cesta representativa del consumo, nacional e importado, y es la referencia de los objetivos de inflación y de la indiciación de salarios y pensiones. El BCE usa el IPCA y la Reserva Federal, el deflactor del consumo privado.
- El **deflactor del PIB** es un índice de Paasche de lo producido en el país.
- La **inflación subyacente** excluye la energía y los alimentos no elaborados. Aproxima la tendencia de los precios sobre la que actúa la política monetaria.

Distinciones: por intensidad, es moderada, de un dígito, galopante, de dos o tres, o hiperinflación, por encima del 50 % mensual; por previsibilidad, esperada o no esperada, con costes distintos; por su causa, de demanda, de costes, estructural, inercial o monetaria. La relevante para la política es la sostenida y esperada, porque la no esperada no puede mantenerse con una inflación estable, pues los agentes acaban anticipándola.

1.2. Los costes de la inflación

Idea: los costes dependen de la magnitud, la volatilidad y la anticipación. Los de la inflación esperada subsisten aunque se anticipe, y los de la no esperada proceden de la sorpresa.

1.2.1. Costes microeconómicos

Conducta de los agentes: en una expansión, la inflación puede adelantar el consumo y la inversión y alimentar burbujas, porque los agentes anticipan sus compras al esperar precios más altos. Además, la *ilusión monetaria* lleva a confundir magnitudes nominales y reales, de modo que los agentes, al ignorar muchos efectos de la inflación, no optimizan adecuadamente.

Mercado de bienes: salvo que sea pura, la inflación hace que los precios informen peor de la escasez relativa, con costes de búsqueda e incertidumbre, y obliga a revisar precios, con **costes de menú**.

Mercado de trabajo y de factores: reduce el valor real de los salarios no ajustados y de la riqueza nominal, y, si se concentra en la energía o las materias primas, altera las cadenas de valor. Esa pérdida de poder de compra puede, además, enfriar el consumo.

Mercados financieros: afecta a las tres funciones del dinero. Como medio de pago se acepta peor, como unidad de cuenta pierde fiabilidad y como depósito de valor pierde poder de compra. Surgen así los **costes de suela de zapatos**, porque, al reducir el efectivo, los agentes acuden a otros activos y multiplican las conversiones en saldos líquidos, con sus costes de transacción. Su coste de bienestar (Bailey, 1956) es el área entre la demanda de saldos reales y el eje del tipo, de $i = 0$ al vigente, menos la transferencia $i\,m(i)$. Con la demanda de Cagan crece aproximadamente con el cuadrado del tipo nominal, ronda el 1 % de la renta al pasar de una inflación del 10 % a cero (Lucas, 2000) y desaparece con la regla de Friedman. Además, la inflación no anticipada redistribuye arbitrariamente de acreedores a deudores, y la alta y volátil añade primas de riesgo a los tipos nominales, que encarecen la financiación en términos reales.

1.2.2. Costes macroeconómicos

Tipo de interés real: con inflación esperada, $i = r + \pi^e$, y a largo plazo la inflación se traslada íntegramente a los tipos nominales, el **efecto Fisher**. La no esperada reduce el tipo real *ex post* y favorece a los deudores, mientras que a largo plazo el tipo real queda determinado por factores reales.

Sector exterior: con tipo fijo o traslación incompleta, una inflación superior a la de los socios resta competitividad-precio y deteriora la balanza comercial si se cumple Marshall-Lerner.

Crecimiento: la incertidumbre sobre los precios relativos y la rentabilidad acorta los horizontes y reduce la inversión. La evidencia asocia inflaciones elevadas y volátiles con menor crecimiento, aunque el efecto solo es apreciable por encima de las tasas moderadas.

1.2.3. Costes desde la política económica

Política monetaria: una inflación alta y volátil desancla las expectativas, con lo que la curva de Phillips de corto plazo se desplaza y la política pierde capacidad de estabilizar. Además, volver a la estabilidad cuesta producción y empleo, la *ratio de sacrificio*, que mide los puntos de producción perdidos por cada punto de inflación reducida.

Persistencia: la inflación tiende a permanecer cerca de su valor anterior, como la inercia física, y el coste de desinflar depende de sus fuentes (Fuhrer, 2006).

- La **persistencia intrínseca** procede de la propia inflación y hace las desinflaciones más largas y costosas y la respuesta a los *shocks* más persistente.
- La **persistencia heredada** procede de determinantes con dinámica propia, como el desempleo, la brecha o el coste marginal.

La distinción importa porque la heredada desaparece con sus causas, mientras que la intrínseca obliga a desinflar de forma gradual y costosa. Qué parte es intrínseca es el punto en que discrepan las teorías, y el criterio con el que la evidencia las juzga.

Política fiscal: dos distorsiones.

- El **efecto Tanzi-Olivera**: como se recauda con demora, el valor real de lo recaudado cae por debajo del gasto y el déficit se amplía.
- La **progresividad en frío**: con impuestos progresivos no indiciados, las subidas nominales llevan a tramos superiores, aunque el coste se debe al diseño del impuesto, no a la inflación.

En suma, la inflación esperada cuesta suela de zapatos, menú y distorsiones fiscales, pequeños a tasas moderadas; la no esperada, redistribución arbitraria, incertidumbre y pérdida de información de los precios. Por eso una inflación moderada y estable cuesta menos que una baja pero volátil. Ambos tipos de costes justifican que la estabilidad de precios sea el objetivo central de la política monetaria. La solución a muchos de estos costes pasa por la indiciación de salarios, tramos impositivos y deuda.

1.3. Beneficios e inflación óptima

Microeconómicos: con tasas bajas, la conducta **casi racional** de quien no incorpora la inflación a sus decisiones permite un desempleo sostenible menor que con inflación nula (Akerlof et al., 2000). El beneficio se agota, sin embargo, a tasas más altas, cuando ignorar la inflación sale caro y los agentes la incorporan a sus contratos. Con salarios nominales rígidos a la baja, la inflación reduce los salarios reales sin recortar los nominales y "engrasa" el mercado de trabajo, como sostenía Tobin. Además, grava indirectamente la economía sumergida, que estrecha las bases imponibles y se paga sobre todo en efectivo, de modo que la inflación alcanza una base que los impuestos ordinarios no tocan. Esto justifica una inflación óptima positiva, aunque la ganancia es pequeña.

Macroeconómicos: a corto plazo, si $\pi^e > i$, el tipo real es negativo y estimula la inversión. Por el **efecto Mundell-Tobin**, la inflación esperada desplaza la cartera hacia activos reales, eleva la acumulación de capital y reduce el tipo real. Por este canal puede tener efectos reales permanentes en un modelo con dinero, aunque pequeños. Y, por el mismo efecto que perjudica a los acreedores, reduce el valor real de las deudas de hogares, empresas y Estado, lo que facilita salir de las crisis de sobreendeudamiento y es la contrapartida de la deflación de deuda.

Política económica: una inflación positiva aleja los tipos nominales del límite inferior y da margen para bajarlos en las recesiones. Además, el *señoraje*, el ingreso real que el Estado obtiene por la emisión de moneda, permite financiar gasto sin subir impuestos y completa con los impuestos y la deuda sus vías de financiación.

Inflación óptima: la **regla de Friedman** (1969) sitúa el óptimo en un tipo nominal nulo, es decir, por la relación de Fisher, en una deflación igual al tipo real:

$$i = r + \pi^e = 0 \quad \Longleftrightarrow \quad \pi^e = -r$$

Solo así el coste de mantener dinero iguala su coste de producción, que es nulo. El consenso se aparta de ella por tres razones: el sesgo al alza de los índices de precios, la rigidez salarial a la baja y el margen frente al límite inferior de los tipos. De ahí procede el objetivo del 2 %, una inflación reducida, controlada y anticipable.

En suma, la inflación no es neutral y su análisis fija el objetivo de la política monetaria, aunque los costes a tasas moderadas parecen reducidos (Lucas, 2000) y el nivel óptimo, del 2 %, el 3 % o el 4 %, sigue en debate.

2. Las causas de la inflación: escuelas y curvas de Phillips

Las causas se ordenan en cinco tipos: de demanda, de costes, estructural o de puja distributiva, inercial o de expectativas, y monetaria. Las escuelas difieren en cuál destacan y en cómo la fundamentan.

2.1. Teoría clásica y teoría keynesiana

2.1.1. La teoría cuantitativa del dinero

Origen: nace en la Escuela de Salamanca, con Martín de Azpilcueta (1556), que atribuyó la subida de precios a la plata americana, antes que Jean Bodin (1568). Su formulación moderna se debe a Fisher (1911) y a la escuela de Cambridge, que dieron forma algebraica a la relación causal de Hume (1752).

Idea: a largo plazo, precios y dinero se mueven en la misma proporción, de modo que la inflación es siempre un fenómeno monetario.

Ecuación de cambio (Newcomb, 1885; Fisher, 1911): es una identidad, $M \cdot V_T = P \cdot T$, que Fisher convierte en relación causal al suponer

estables la velocidad-transacción, fijada por la tecnología de pagos, y las transacciones, fijadas por factores reales: $P = (V_0/T_0)\,M$. El dinero es un velo: se cumplen la **neutralidad monetaria** y la **dicotomía clásica**.

Ecuación de Cambridge (Marshall, Pigou): $M = k \cdot P \cdot Y$, con $k = 1/V_Y$. La conclusión es la misma, pero se lee como una demanda de dinero, y k, una decisión de los agentes, puede depender del tipo de interés y de las expectativas, de lo que partirá Keynes. En ambas, M es la cantidad de dinero, P el nivel general de precios, T el volumen de transacciones e Y la renta real.

De cualquiera de las dos versiones se sigue que, con dinero exógeno, sus cambios se trasladan proporcionalmente a los precios a largo plazo. Esos cambios solo afectan al nivel de precios, que es la neutralidad, y las variables reales se determinan con independencia de las monetarias, que es la dicotomía.

Fija la conclusión de largo plazo que nadie ha desmentido: la inflación sostenida es un fenómeno monetario. Pero, al comparar equilibrios estáticos, explica el nivel de precios, o un salto aislado, mejor que su subida sostenida, que exigiría una expansión monetaria mantenida que la propia neutralidad deja sin justificación.

2.1.2. La teoría keynesiana: demanda y costes

Idea: la Gran Depresión puso en duda el paradigma neoclásico. Con rigideces de precios y salarios, la inflación no es solo monetaria. Más dinero puede tener efectos reales, porque con salarios nominales rígidos abarata el salario real, y la inflación puede nacer de la demanda o de los costes con independencia del dinero. Distingue dos orígenes:

- La **inflación de demanda**, formalizada con la *brecha inflacionista* por Keynes (1940) y Smithies (1942), aparece en el pleno empleo o cerca de él, cuando la oferta no puede seguir a la demanda. Los consumidores compiten entonces por los productos y presionan al alza los precios.
- La **inflación de costes** aparece aun con desempleo elevado, por la fijación de *mark-ups* con poder de mercado, como cuando suben los precios del petróleo y, con ellos, los costes de todas las empresas que usan el combustible.

Sigue siendo útil para el diagnóstico, porque una misma inflación puede requerir respuestas distintas según su origen, pero también compara equilibrios estáticos y no explica la dinámica. Una subida continuada de los precios exigiría que el exceso de demanda, o la escasez de oferta, durase también.

2.2. Síntesis neoclásica y monetarismo

2.2.1. Síntesis: la curva de Phillips

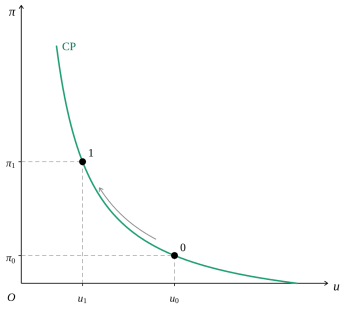
Origen: Phillips (1958) halló una relación inversa entre el crecimiento de los salarios nominales y el desempleo en el Reino Unido, Lipsey (1960) la fundamentó y Samuelson y Solow (1960) la convirtieron en una relación entre inflación y desempleo.

Idea: la inflación nace de la presión de la demanda sobre el mercado de trabajo. La relación es estable y ofrece un menú a la política económica. Por la ley de Okun:

$$\pi_t = -\kappa_u(u_t - \bar{u}) \qquad \Longleftrightarrow \qquad \pi_t = \kappa\,(y_t - \bar{y}), \qquad \kappa_u > 0,\ \kappa > 0$$

con $\kappa_u, \kappa > 0$, de modo que la inflación aparece cuando la producción supera a su potencial, se anula cuando lo iguala y se vuelve negativa cuando queda por debajo.

Desarrollo gráfico: en el plano desempleo-inflación la curva es decreciente, y una expansión monetaria o fiscal lleva la economía a lo largo de ella del punto 0 al 1, con $u_1 < u_0$ y $\pi_1 > \pi_0$.



Curva de Phillips original: una expansión de la demanda lleva la economía de 0 a 1.

Es una teoría de demanda: una expansión sostenida mantiene la producción por encima del potencial a cambio de una inflación más alta pero estable. Se apoyaba, sin embargo, en que los trabajadores no incorporaban la inflación a sus demandas, y la estanflación de los setenta mostró que la curva se había desplazado.

2.2.2. Monetarismo: la curva ampliada

Origen: Friedman recuperó la teoría cuantitativa como teoría de la demanda de dinero (1956) e introdujo las expectativas en la curva de Phillips (1968), a la vez que Phelps (1967).

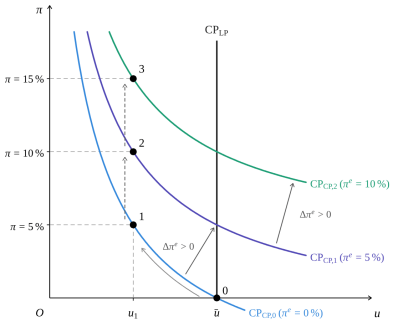
Reformulación cuantitativa: la demanda de dinero depende de la renta permanente y es estable. En tasas, $\pi = g_M + g_V - g_Y$. Con velocidad estable, la inflación es el exceso del crecimiento monetario sobre el real, y es "siempre y en todas partes un fenómeno monetario" (Friedman, 1963). Las causas de demanda o de costes solo la sostienen si el dinero las acomoda. Friedman lee la ecuación de Cambridge como una demanda de dinero, $M/P = k(\cdot)\,Y$, que depende también de los rendimientos de los activos alternativos y es más estable que la función de consumo keynesiana.

Curva ampliada por las expectativas: si los trabajadores negocian en términos reales, la curva se desplaza con la inflación esperada, con $\bar{u}$ la tasa natural de paro:

$$\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u}), \qquad \kappa_u > 0$$

- El **componente inercial**: con expectativas adaptativas, la inflación esperada es la pasada y se autoalimenta.
- El **componente de demanda**: solo la producción por encima del potencial presiona los precios. La curva original es el caso $\pi_t^e = 0$.
- La **aceleración**: mantener el desempleo por debajo de la tasa natural exige una inflación creciente, sostenida por un crecimiento monetario cada vez mayor, porque cada periodo la esperada alcanza a la observada.

Dinámica: a corto plazo, cada curva corresponde a una inflación esperada. A largo plazo, la curva es vertical en $\bar{u}$. Si la autoridad fija $u_1 < \bar{u}$, la inflación pasa al 5 %. Al revisarse las expectativas, la curva sube y hace falta un 10 %, y luego un 15 %. Cuando deja de acelerar, la economía vuelve a $\bar{u}$ con la inflación alcanzada. Las curvas de corto plazo son decrecientes porque, dada la inflación esperada, solo una inflación superior a ella lleva el desempleo por debajo de la tasa natural.



Hipótesis aceleracionista: mantener $u_1 < \bar{u}$ exige una inflación creciente (puntos 1, 2 y 3).

Es el punto de partida de toda la teoría moderna de la inflación, pero las expectativas adaptativas suponen errores sistemáticos y la tasa natural queda sin explicar, lo que aportará el modelo de Layard y Nickell.

2.3. Nueva macroeconomía clásica

2.3.1. Información imperfecta (Lucas, 1972)

Idea: con expectativas racionales, los agentes usan toda la información y sus errores no siguen un patrón sistemático. Si aun así existe un *trade-off* a corto plazo, es porque su información sobre los precios agregados es incompleta, un problema de extracción de señales.

Razonamiento: los productores no distinguen si una subida de su precio es un *shock* específico, ante el que convendría producir más, o general, ante el que no. Atribuyen al componente específico una fracción, el **coeficiente de extracción de señales**:

$$\vartheta = \frac{\sigma_i^2}{\sigma_i^2 + \sigma_p^2}$$

donde σ_i^2 es la varianza de los precios relativos y σ_p^2 la del nivel general, y cuanto mayor es la primera, más verosímil es que el *shock* sea específico.

Con inflación muy variable, ϑ es cercana a cero y los productores no ajustan la producción, mientras que con precios estables las sorpresas tienen efectos reales transitorios. La oferta agregada es $y_t = \bar{y}_t + \zeta\,(p_t - p_t^e) + \varepsilon_t^s$: solo las sorpresas tienen efectos reales, la política sistemática es inefectiva (Sargent y Wallace, 1975) y la inflación anticipada no compra actividad. Con $\zeta > 0$ creciente en ϑ, el efecto de las sorpresas es menor cuanto más variable es la inflación, y una autoridad que intente sorprender de forma sistemática deja de sorprender.

Aspectos positivos: sitúa las expectativas y la credibilidad en el centro, con una implicación contrastable entre países.

Limitaciones: la confusión es poco plausible con índices mensuales, no explica la persistencia y Mishkin (1982) halla efectos reales de la política anticipada.

2.3.2. Inconsistencia dinámica (Barro y Gordon, 1983a)

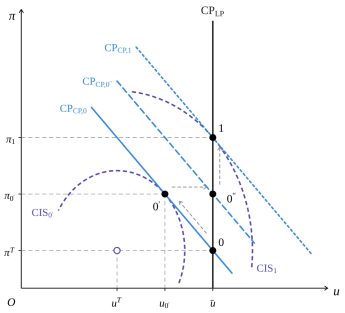
Idea: Kydland y Prescott (1977) mostraron que una política anunciada es **dinámicamente inconsistente** si, una vez que los agentes han actuado, a la autoridad le conviene incumplirla. Barro y Gordon (1983a) lo aplican a la inflación. La falta de credibilidad genera por sí sola un **sesgo inflacionista**, sin ganancia de empleo.

Supuestos: los agentes tienen expectativas racionales y fijan sobre ellas sus contratos. El Gobierno minimiza una pérdida social, con un objetivo de inflación, π^T, y otro de desempleo inferior a la tasa natural, $u^T < \bar{u}$. Las expectativas son $\pi_t^e = \mathbb{E}_{t-1}[\pi_t]$, y las curvas de indiferencia social son elipses en torno a (u^T, π^T), tanto mejores cuanto más cerca de ese objetivo.

Planteamiento: en el punto 0, la autoridad anuncia que controlará la inflación y los agentes firman sus contratos. Con salarios fijados, le conviene sorprender con más inflación para bajar el desempleo, punto 0 . Los agentes revisan sus expectativas, la curva de corto plazo se desplaza, punto 0 , y, tras varios periodos, se llega al punto 1, un equilibrio de Nash en el que la curva de corto plazo es tangente a una curva de indiferencia social.

Equilibrio: como los agentes prevén la secuencia, reclaman desde el principio el salario del punto 1: mismo desempleo, el natural, y más inflación. El anuncio inicial no es un equilibrio perfecto en subjuegos. La inflación puede nacer, así, de que el público no se fíe de los anuncios de política económica.

Desarrollo gráfico: en el plano desempleo-inflación, la curva de largo plazo es vertical en $\bar{u}$ y las de corto plazo, una para cada inflación esperada, tienen pendiente negativa. En el punto 1, la $\text{CP}_{\text{CP},1}$ es tangente a la CIS_1 sobre la de largo plazo, con el desempleo en $\bar{u}$ y una inflación π_1 superior al objetivo π^T del punto 0.



Barro y Gordon: del compromiso (0) al equilibrio discrecional (1), con sesgo inflacionista y sin ganancia de empleo.

Estática comparativa: el sesgo es mayor cuanto más baja es la meta de desempleo respecto a la natural, cuanto más peso da el Gobierno al desempleo y cuanto más reduce el paro una sorpresa dada. En todos los casos el desempleo de equilibrio sigue siendo la tasa natural, y los parámetros solo alteran la inflación.

Extensiones: las soluciones al sesgo son la independencia del banco central, las reglas transparentes, el **banquero conservador** (Rogoff, 1985) y los contratos de desempeño (Walsh, 1995). El banquero conservador tiene más aversión a la inflación que la sociedad, y los contratos vinculan la remuneración del gobernador al cumplimiento del objetivo de inflación.

Aspectos positivos: fundamenta el diseño institucional moderno de la política monetaria.

Limitaciones: supone $u^T < \bar{u}$ y que se conocen las preferencias del Gobierno. Con incertidumbre surgen la reputación y la señalización (Backus y Driffill, 1985).

2.4. Nueva economía keynesiana

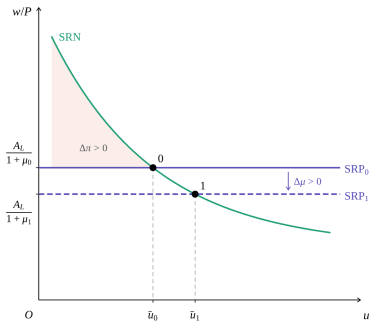
2.4.1. Puja distributiva (Layard y Nickell, 1985)

Idea: la inflación resulta de una **puja distributiva**. Los trabajadores negocian un salario real que crece con su poder y las empresas fijan un margen sobre el coste. Cuando las aspiraciones son incompatibles, porque el desempleo es bajo, la puja se resuelve en inflación. La tasa a la que son compatibles es la **NAIRU**.

Razonamiento: se obtiene la misma curva ampliada, pero $\bar{u}$ depende de parámetros estructurales: crece con el margen y con las prestaciones, el salario mínimo o las cotizaciones, y cae con la productividad. El salario negociado, $w/P^e = f(u, z)$, decrece con el desempleo, y el pagado, $A_L/(1+\mu)$, es horizontal. Por debajo de la NAIRU, el salario negociado supera al pagado y la inflación crece, mientras que por encima el negociado queda por debajo y la inflación se reduce. La NAIRU iguala ambos salarios, con A_L la productividad del trabajo y μ el margen:

$$f(\bar{u}, z) = \frac{A_L}{1+\mu}$$

Solo las políticas que actúan sobre la competencia en los mercados de bienes y de trabajo la reducen de forma permanente.



Puja distributiva: un mayor margen ($\Delta\mu > 0$) eleva la NAIRU de $\bar{u}_0$ a $\bar{u}_1$.

Explica inflaciones distintas con la misma política monetaria y fundamenta las políticas de rentas y las reformas estructurales, pero la NAIRU no es observable y el modelo no explica la dinámica del ajuste.

2.4.2. Fijación escalonada de precios (Calvo, 1983)

Origen: Calvo propuso la fijación escalonada como alternativa a los contratos de Fischer (1977) y Taylor (1980), y es hoy el estándar de los modelos DSGE. En él la revisión de precios es aleatoria y no depende del tiempo transcurrido desde la última, a diferencia de los contratos de duración fija, y es analíticamente más manejable.

Idea: las empresas tienen poder de mercado, pero cada periodo solo una fracción $1-\xi$ revisa su precio, de forma aleatoria, y lo fija con un margen sobre los costes marginales presentes y esperados. Agregando y loglinealizando, la **curva de Phillips neokeynesiana** es:

$$\pi_t = \beta \mathbb{E}_t[\pi_{t+1}] + \kappa\, \tilde{y}_t, \qquad \kappa > 0,\ \beta \in (0,1)$$

donde $\tilde{y}_t$ es la brecha de producción. En la calibración típica (Galí, 2008), $\xi = 2/3$, con precios fijos durante unos tres trimestres, y $\beta = 0{,}99$.

Con κ creciente en la frecuencia de ajuste, la inflación de hoy depende de la esperada para mañana y del exceso de demanda actual. Es puramente prospectiva: la inflación resume la senda esperada de brechas futuras, no hay inercia y la desinflación creíble no cuesta. Es la oferta de los modelos de los bancos centrales y hace de la credibilidad y la comunicación los instrumentos centrales, pero no reproduce la persistencia observada.

2.4.3. Curvas de Phillips híbridas

Críticas: la CP-NEK predice un ajuste inmediato, cuando la inflación cae con retraso tras una contracción (Mankiw, 2001), no explica la persistencia y las empresas apenas se informan sobre la inflación (Coibion et al., 2018). Por ello se considera que la inflación pasada predice mejor la futura que la esperada.

Desarrollos: recuperan la inercia del triángulo de Gordon.

- **Gordon** (1982): su modelo del triángulo formaliza la persistencia con tres vértices, la inercia, la demanda y la oferta, más un componente estocástico. Los determinantes son la inflación de los n periodos anteriores, ponderada por a_i, que recoge la persistencia intrínseca, como la hipótesis aceleracionista; el diferencial entre el desempleo y la NAIRU, $\bar{u}$, ponderado por b, y los factores de oferta, s_t, ponderados por c:
$$\pi_t = \sum_{i=1}^{n} a_i \pi_{t-i} - b(u_t - \bar{u}) + c \cdot s_t + \varepsilon_t$$
- **Galí y Gertler** (1999): reglas de ajuste sencillas, un retardo de la inflación y el coste marginal real, aproximado por la participación del trabajo en la renta, en lugar de la brecha. Esta especificación es la más usada.
- **Christiano, Eichenbaum y Evans** (2005): la indiciación a la inflación pasada de quienes no optimizan genera persistencia endógena.
- **Mankiw y Reis** (2002): rigidez de la información, no de los precios. Como informarse es costoso, cada periodo solo una parte de las empresas actualiza su información y el resto fija sus precios con información atrasada, lo que genera un ajuste gradual.

La curva híbrida es $\pi_t = \gamma_b \pi_{t-1} + \gamma_f \mathbb{E}_t[\pi_{t+1}] + \kappa \tilde{y}_t$. La desinflación tiene un coste, la ratio de sacrificio, mayor cuanto más pesa la inercia y menos credibilidad hay. El triángulo explica la estanflación y sirve para estimar la NAIRU y la ratio de sacrificio. Su límite es que la inercia descansa en retardos sin fundamento microeconómico, en la indiciación o en rigideces de información postuladas, expuestas a la crítica de Lucas. Además,

el peso estimado de la inercia frente a las expectativas varía con la especificación y con la medida del coste marginal.

En suma, las teorías convergen en $\pi_t = \pi_t^e - \kappa_u(u_t - \bar{u})$, coinciden a largo plazo con la teoría cuantitativa y difieren en cómo se forman las expectativas, de qué depende la tasa natural y de dónde vienen los efectos reales. En la NMC, la inflación anticipada no tiene efectos reales: Lucas explica los efectos de las sorpresas, y Barro y Gordon, la inflación por falta de credibilidad, que resuelven los bancos centrales independientes y los objetivos explícitos, cuyo éxito explica la baja inflación de 1990 a 2020. Ambas, con precios flexibles, predicen desinflaciones creíbles sin coste. Ninguna explica bien la persistencia sin supuestos *ad hoc*, ni la inflación importada y de oferta de los episodios recientes, y todas tratan la economía como cerrada.

3. Hiperinflación y deflación

En ambos extremos el dinero deja de cumplir sus funciones. En la hiperinflación pierde la de depósito de valor y acaba perdiendo la de medio de pago, y en la deflación su valor crece y los agentes lo atesoran. Ambos son fenómenos monetarios en su mecánica y fiscales o financieros en su origen.

3.1. La hiperinflación

3.1.1. Concepto y origen

Definición: según Cagan (1956), la **hiperinflación** supera el 50 % mensual, lo que multiplica los precios por más de cien en un año. Es rara y se asocia a guerras y colapsos del Estado: Alemania en 1923, Hungría en 1946, Zimbabue en 2008 o Venezuela entre 2017 y 2020.

Origen: es fiscal. Un Estado sin impuestos ni deuda recurre al señoraje. La inflación reduce la demanda de dinero, lo que obliga a emitir más para el mismo ingreso real, y el proceso se acelera con las expectativas. La hiperinflación es la forma extrema de la inflación monetaria, con la particularidad de que la aceleran las expectativas y la demanda de dinero, y no solo la oferta.

3.1.2. El modelo de Cagan (1956)

Supuestos: las variables reales son constantes, el Gobierno fija g_M para financiar el déficit y las expectativas son adaptativas.

Planteamiento: la demanda de saldos reales depende solo de la inflación esperada, que se revisa en proporción al error cometido:

$$m_t - p_t = c - a\pi_t^e$$
$$\dot{\pi}^e = \lambda\,(\pi - \pi^e)$$

Aquí $c > 0$ recoge la demanda de dinero por factores reales, $a > 0$ su sensibilidad a la inflación esperada y $\lambda > 0$ la velocidad de aprendizaje.

Equilibrio: la demanda de dinero da el nivel de precios, y su versión en tasas, $g_M - \pi = -a\,\dot{\pi}^e$, con la HEA, la inflación observada en función de la esperada:

$$p_t = m_t - c + a\pi_t^e$$
$$\pi = \frac{g_M - a\lambda\,\pi^e}{1 - a\lambda} \qquad \pi^e = \pi = g_M$$

En el estado estacionario, la inflación observada, la esperada y el crecimiento monetario coinciden.

Desarrollo gráfico: en el plano inflación esperada-inflación observada, la recta $\pi(\pi^e)$ tiene pendiente $-a\lambda/(1-a\lambda)$, negativa si $a\lambda < 1$ y positiva y mayor que uno si $a\lambda > 1$. Corta a la bisectriz, $\pi = \pi^e$, en el equilibrio, el punto 0, con $\pi^e = \pi = g_M$.

Estática comparativa: el nivel de precios es menor cuanto mayor es c y crece con la cantidad de dinero, y la inflación de equilibrio, con g_M. El modelo fue un poderoso argumento a favor de controlar los agregados monetarios. Cuando la inflación se acelera, los precios suben más cuanto mayores son a y λ, que intensifican la retroalimentación.

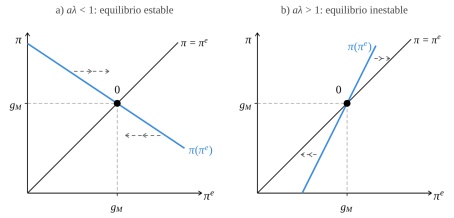
Dinámica: la inflación esperada evoluciona según $\dot{\pi}^e = \lambda(g_M - \pi^e)/(1 - a\lambda)$. Si $a\lambda < 1$, el equilibrio es estable y la hiperinflación solo se explica por una emisión excesiva y creciente. Si $a\lambda > 1$, un salto de la inflación eleva la expectativa, reduce la demanda de saldos y el exceso de dinero genera aún más inflación, de modo que la hiperinflación es una **profecía autocumplida**, sin que el dinero crezca de forma continua.

Con expectativas racionales, son las expectativas sobre el nivel de precios futuro las que lo determinan, y un Gobierno sin credibilidad perpetúa la inflación también por el lado de la demanda de dinero.

Señoraje: es $\sigma = g_M \cdot (M/P)$, el tipo por la base, y, en estado estacionario, $\sigma = g_M\, e^{c - a g_M}$. Aumentar g_M eleva el tipo pero reduce la base, porque los agentes huyen del dinero. La condición de primer orden da el máximo:

$$e^{c - a g_M}(1 - a g_M) = 0 \quad \Longrightarrow \quad g_M^* = \frac{1}{a}, \qquad \sigma^* = \frac{e^{c-1}}{a}$$

que crece con c y decrece con a. El máximo es la cima de una curva de Laffer, y las hiperinflaciones aparecen cuando el Gobierno quiere más señoraje del que permite la demanda de dinero.



Estabilidad en el modelo de Cagan: converge si $a\lambda < 1$ (a) y se autoalimenta si $a\lambda > 1$ (b).

El modelo formaliza por primera vez la hiperinflación y trata la inflación como un impuesto, pero supone una economía cerrada, y el máximo de señoraje no es el óptimo social. Con tipo de cambio fijo, monetizar el déficit haría perder reservas y acabaría forzando el abandono de la paridad. Tampoco endeudarse en lugar de emitir dinero evita la inflación, según la "aritmética monetarista desagradable" de Sargent y Wallace (1981).

3.1.3. Estabilización (Sargent, 1982)

Idea: las hiperinflaciones de Austria, Hungría, Polonia y Alemania terminaron de golpe, con un coste menor del que predecía la ratio de sacrificio, gracias a un **cambio de régimen** creíble que combinó:

1. Un ajuste fiscal que eliminó el déficit, la causa de la inflación. Como esta procedía de las expectativas de financiación monetaria del déficit, solo un cambio institucional visible podía cambiarlas.
2. Un banco central independiente que no podía financiar al Tesoro, lo que hacía creíble el ajuste.
3. Un ancla nominal, a menudo una nueva moneda o un tipo de cambio fijo, que daba a los agentes una referencia para fijar precios.

Los programas heterodoxos de los ochenta añadieron controles de precios y salarios, con éxito solo con ajuste fiscal. La credibilidad, no la contracción gradual, es el instrumento, aunque la conclusión se traslada peor a inflaciones moderadas, en las que los contratos y la indiciación hacen la desinflación más gradual y costosa.

3.2. La deflación

Definición: la **deflación** es la caída sostenida del nivel general de precios. Es benigna si procede de *shocks* de oferta positivos y recesiva si procede de *shocks* de demanda negativos. No es simétrica a la inflación por tres asimetrías: las deudas nominales, la rigidez salarial a la baja y el límite inferior de los tipos. La benigna acompaña a una mayor productividad o a importaciones más baratas, y la recesiva, a una crisis financiera o a una contracción del crédito.

Costes: Fisher (1933) explicó la Gran Depresión por la **deflación de deuda**: la caída de precios eleva la deuda real, los deudores venden activos, los precios caen más y, cuanto más pagan, más deben. La rigidez salarial eleva el salario real y el desempleo. Y en la **trampa de liquidez** (Keynes, 1936), con $i = 0$, la deflación esperada eleva el tipo real, $r = i - \pi^e$, cuando el banco central ya no puede bajar los tipos. La deflación recesiva es el escenario en que la política monetaria convencional es inoperante, y por eso los bancos centrales la temen más que a una inflación moderada. Japón desde los noventa es el caso de referencia, y la "japonización", el riesgo de una deflación prolongada con tipos en cero y crecimiento débil. La deflación de deuda es, además, precursora del acelerador financiero, aunque su efecto agregado depende de que los deudores gasten más que los acreedores. La experiencia reciente muestra, con todo, que el límite inferior efectivo está algo por debajo de cero y que las medidas no convencionales conservan cierta capacidad de actuación.

Política económica: la política monetaria recurre al *forward guidance*, a la expansión cuantitativa, a los tipos negativos y a los objetivos de nivel de precios o de PIB nominal. La fiscal recupera eficacia, porque con tipos en cero no hay expulsión. La experiencia de Japón, de la zona del euro en 2014-2016 y de EE. UU. tras 2008 muestra que las medidas no convencionales evitan la espiral, pero no siempre devuelven la inflación al objetivo, y que lo eficaz ha sido combinarlas con la política fiscal. La deflación es el argumento más fuerte a favor de la coordinación. Los objetivos de nivel de precios o de PIB nominal sirven porque comprometen al banco central a compensar la deflación pasada con inflación futura.

En suma, la hiperinflación muestra que la inflación es, en última instancia, fiscal y de expectativas, y la deflación, que la caída de precios es más peligrosa que su subida. Ambas justifican un objetivo positivo y creíble.

1. Concepto y antecedentes

El punto de partida es la definición de ciclo y los conceptos con que lo analizan las teorías modernas. Sobre esa base se repasan las primeras explicaciones y los modelos del acelerador, que dieron su primera formalización.

1.1. Concepto de ciclo económico

Definición: los **ciclos económicos** son fluctuaciones recurrentes pero no periódicas de la actividad agregada de un país, de duración superior a un año (Burns y Mitchell, 1946). El análisis moderno los mide como desviaciones respecto de la tendencia secular y estudia tres conceptos:

1. El **impulso**: la perturbación, exógena o endógena, que causa la desviación.
2. La **propagación**: los mecanismos de amplificación, difusión y persistencia, como la suavización del consumo, los retardos en comprender los *shocks*, la sustitución intertemporal de ocio y trabajo o los costes de ajuste.
3. Los **comovimientos**: la correlación, ρ, entre el componente cíclico de cada variable y el del PIB. Son procíclicas si $\rho > 0$, contracíclicas si $\rho < 0$ y acíclicas si $\rho \approx 0$.

Clasificación: por el signo de ρ, las variables son procíclicas, como el consumo, la inversión, el empleo o los beneficios; contracíclicas, como el desempleo, por la ley de Okun, o el déficit público, por los estabilizadores, y acíclicas, como el gasto público en consumo o los tipos reales. Por el desfase, son coincidentes, como la producción industrial; adelantadas, útiles para predecir, como la inversión residencial, el PMI, la pendiente de la curva de tipos o la bolsa, y retrasadas, como la inflación, los tipos a corto o los costes laborales unitarios. El desfase se mide con la correlación $\rho(x_t, y_{t+j})$, máxima con $j = 0$ en las coincidentes, con $j > 0$ en las adelantadas y con $j < 0$ en las retrasadas.

1.2. Primeras aportaciones

Hasta la década de 1890 no se analizó el ciclo de forma sistemática. La excepción es Juglar (1862), que lo atribuyó a fuerzas internas del sistema y distinguió prosperidad, crisis y liquidación. Las aportaciones posteriores se clasifican, siguiendo a Haberler (1937), por la naturaleza del impulso.

1.2.1. Teorías no monetarias

- **Fenómenos naturales** (Jevons, 1878; Moore, 1914): las manchas solares, que en Jevons siguen un ciclo de unos once años, o las lluvias afectan a las cosechas y, por ellas, a la actividad. Estas teorías introducen la idea de un impulso exógeno regular.
- **Sobreinversión no monetaria** (Tugan-Baranowsky, 1894; Spiethoff, 1902; Cassel, 1918b): el capital prestable acumulado en la depresión desata un auge de bienes de capital, desproporcionado frente al de consumo, hasta que se agotan los fondos y llega la crisis.
- **Infraconsumo** (Hobson, 1909): la desigualdad lleva a ahorrar más de lo que puede invertirse rentablemente, y el exceso de capacidad acaba en depresión, cuyo remedio es redistributivo.
- **Principio de aceleración** (Aftalion, 1909; Clark, 1917): la inversión depende de la variación del consumo, no de su nivel, de modo que pequeñas fluctuaciones del consumo se amplifican en la inversión. Los retardos en la construcción del capital hacen que la respuesta llegue tarde y se sobrepase.
- **Teorías psicológicas** (Marshall, 1923; Pigou, 1927): los errores de optimismo del auge engendran los de pesimismo de la crisis.
- **Teoría keynesiana** (Keynes, 1936): el origen está en la eficiencia marginal del capital, volátil por los *espíritus animales*, y el multiplicador traslada la caída de la inversión a la renta.
- **Innovación** (Schumpeter, 1939): las innovaciones llegan en racimos, financiadas con crédito, y su imitación genera el auge. La recesión absorbe la innovación y liquida lo obsoleto, la *destrucción creativa*. Schumpeter superpone los ciclos de Kitchin, Juglar y Kondratiev. Al generalizarse la innovación, los beneficios extraordinarios se erosionan, el crédito se devuelve y los precios caen.

1.2.2. Teorías monetarias

- **Puramente monetaria** (Hawtrey, 1919): la expansión del crédito baja el tipo y estimula las existencias de los comerciantes, hasta que la pérdida de reservas obliga a restringirlo, y la contracción sigue el camino inverso.
- **Sobreinversión monetaria** (Mises, 1912; Hayek, 1929): el crédito sitúa el tipo de mercado por debajo del natural (el que igualaría el ahorro y la inversión voluntarios), alarga la estructura productiva sin ahorro real y financia el auge con ahorro forzoso. La crisis liquida la malinversión.
- **Monetarista** (Friedman y Schwartz, 1963): con velocidad estable, la oferta monetaria se transmite al gasto nominal y, con retardos largos y variables, a la producción.

1.3. Modelos del acelerador y ciclos endógenos

Con la *Teoría general* aparecen el multiplicador y, recuperado, el acelerador. Combinados forman una ecuación en diferencias de la que se deriva el ciclo.

1.3.1. Modelo del oscilador (Samuelson, 1939)

Idea: un impulso al gasto autónomo eleva la renta más que el propio impulso, por el multiplicador. Ese aumento induce nueva inversión, por el acelerador, y así sucesivamente.

Planteamiento: el consumo depende de la renta pasada, $C_t = C_0 + cY_{t-1}$, y la inversión inducida, de la variación del consumo, $I_t = I_0 + v\,(C_t - C_{t-1})$, con v el acelerador. Es positiva mientras el consumo crece y negativa cuando cae, y eso introduce la dinámica. A diferencia de la inversión keynesiana estándar, no responde al tipo de interés.

Equilibrio: con $Y_t = C_t + I_t$ resulta la *ecuación fundamental del ciclo*, de segundo orden:

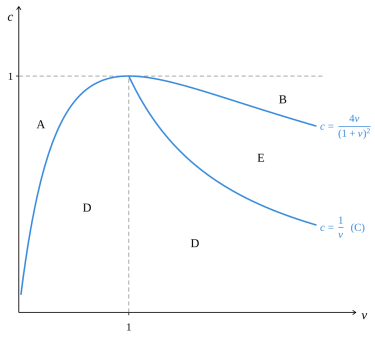
$$Y_t - c(1+v)Y_{t-1} + vcY_{t-2} = I_0 + C_0$$

Estática comparativa: el nivel estacionario es $Y^* = (C_0 + I_0)/(1-c)$, y un aumento permanente del gasto autónomo lo eleva en el multiplicador keynesiano, $1/(1-c)$, que no depende del acelerador. El acelerador solo decide cómo se recorre el camino.

Dinámica: depende de las raíces de la ecuación característica, $r^2 - c(1+v)\,r + vc = 0$:

- Con **raíces reales** no hay ciclos, y la trayectoria converge o diverge según su valor absoluto sea o no menor que la unidad en ambas raíces.
- Con **raíces complejas** la trayectoria es cíclica, y el ciclo es amortiguado, constante o explosivo según su módulo, $\sqrt{vc}$, sea menor, igual o mayor que la unidad.

En el plano v-c, la frontera $c = 4v/(1+v)^2$, que recoge los pares con discriminante nulo, separa las raíces reales, por encima, de las complejas, por debajo, y la hipérbola $c = 1/v$ separa las trayectorias convergentes de las explosivas. Delimitan cinco regiones: convergencia y divergencia sin ciclos, A y B; ciclos constantes, C, sobre la hipérbola; amortiguados, D, y explosivos, E. Con raíces complejas, la desviación respecto de Y^* sigue $A\,R^t\cos(\varphi t + \psi)$, de periodo $2\pi/\varphi$ y con una amplitud que el módulo, $R = \sqrt{vc}$, multiplica cada periodo.



Modelo de Samuelson: regiones del plano acelerador-propensión a consumir.

El modelo es la primera formalización del ciclo, con la interacción multiplicador-acelerador como propagación. Pero con $c \approx 0{,}8$ y un acelerador superior a $1/c = 1{,}25$, $vc > 1$, de modo que predice ciclos explosivos, o una divergencia monótona si v supera 2,62, y no los ciclos persistentes y acotados que se observan.

1.3.2. Modelo del oscilador revisado (Hicks, 1950)

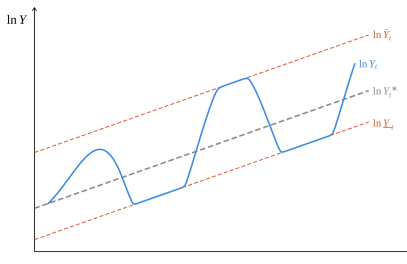
Idea: Hicks hace explosivo el modelo de Samuelson, pero lo contiene entre un techo y un suelo que crecen con la tendencia, de modo que el ciclo se superpone al crecimiento.

Planteamiento: la inversión autónoma crece a una tasa exógena, $I_{0,t} = I_0(1+\bar{g}_Y)^t$. El **suelo** procede de que la inversión bruta inducida no puede ser negativa. El capital solo cae al ritmo de la depreciación, mientras la inversión autónoma sigue creciendo, de modo que el suelo, $\underline{Y}_t$, crece con la tendencia. El **techo** es la producción de pleno empleo, $\bar{Y}_t = \bar{Y}_0(1+\bar{g}_Y)^t$. La inversión suma el componente autónomo y el inducido por el acelerador:

$$I_t = I_{0,t} + v\,(Y_{t-1} - Y_{t-2})$$

Desarrollo gráfico: en el plano t-$\ln Y$, donde toda magnitud que crece a tasa constante es una recta, el techo, $\ln \bar{Y}_t$, la tendencia, $\ln Y_t^*$, y el suelo, $\ln \underline{Y}_t$, crecen a la tasa $\bar{g}_Y$, y la senda del PIB oscila entre el techo y el suelo en torno a la tendencia.

Dinámica: en la expansión, multiplicador y acelerador llevan el PIB al techo. Allí el crecimiento se limita al de los factores, la inversión inducida cae y el PIB desciende. La caída se detiene en el suelo, donde la desinversión ya no puede acelerarse y la inversión autónoma reactiva el acelerador. La renta vuelve a crecer y se inicia un nuevo ciclo.



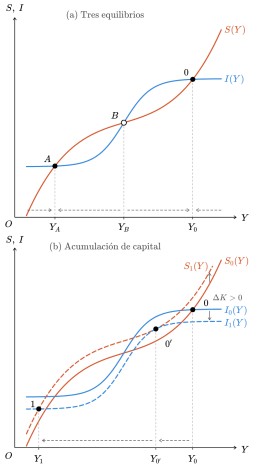
Modelo de Hicks: el ciclo oscila entre el techo de pleno empleo y el suelo de la inversión autónoma.

Introduce límites realistas e integra ciclo y crecimiento, pero descansa en un acelerador rígido y mecánico, y el techo y el suelo son restricciones impuestas.

1.3.3. Ciclos endógenos (Kaldor, 1940)

Idea: es la primera teoría keynesiana del ciclo plenamente endógeno, sin acelerador. Toma de Keynes el ahorro y la inversión como determinantes de la renta y de Kalecki la dependencia de la inversión respecto del capital acumulado. Como el ahorro y la inversión no son lineales en la renta, hay tres equilibrios:

- Con **actividad baja**, el exceso de capacidad deprime la propensión a invertir y la resistencia a recortar el consumo eleva la de ahorrar: es estable.
- Con **actividad intermedia**, la propensión a invertir supera a la de ahorrar: es inestable. Cualquier desviación se amplifica hacia uno de los extremos.
- Con **actividad alta**, los límites de oferta deprimen la inversión y la saturación del consumo eleva el ahorro: es estable.



Modelo del ciclo de Kaldor (1940).

La acumulación o desacumulación de capital desplaza ambas funciones y desestabiliza los extremos, de modo que el sistema oscila por sí solo. Anticipa la dinámica no lineal, pero sus funciones se postulan y genera un ciclo regular, cuando los observados son recurrentes pero no periódicos.

En suma, las teorías tradicionales legaron la distinción entre impulso y propagación y su primera formalización, pero solo generan ciclos acotados con supuestos *ad hoc* y reglas mecánicas, sin optimización ni expectativas.

2. Teorías modernas del ciclo

Tras la crítica de Lucas (1976), las teorías del ciclo se microfundamentan. Las **teorías del ciclo de equilibrio**, de la NMC, distinguen un impulso nominal, Lucas, o real, el ciclo real; las **teorías del ciclo de desequilibrio**, de la NEK, lo conciben como variaciones de los excesos de oferta y demanda; la nueva síntesis neoclásica integra ambas.

2.1. Ciclos nominales (Lucas, 1972; 1973; 1975)

Idea: el impulso es una perturbación monetaria y la transmisión, un problema de extracción de señales, y la persistencia no queda bien explicada. Se apoya en la parábola de las islas de Phelps (1969) y en la sustitución intertemporal de Lucas y Rapping (1969).

2.1.1. Supuestos

Mundo neoclásico, con precios flexibles, pero con información imperfecta: los agentes conocen el precio de su isla, p_{it}, pero no el nivel general, p_t, y no distinguen un *shock* específico, ante el que conviene producir más, de uno general, ante el que no. Tienen expectativas racionales, $p_t^e \equiv \mathbb{E}_{t-1}[p_t]$, y el dinero tiene un componente previsible y una sorpresa, ε_t:

$$m_t = \mathbb{E}_{t-1}[m_t \mid \Omega_{t-1}] + \varepsilon_t, \qquad \varepsilon_t \sim N(0, \sigma_m^2)$$

2.1.2. Modelo

Oferta agregada: los productores atribuyen al *shock* específico una probabilidad igual al peso de la varianza de los precios relativos, $\vartheta = \sigma_i^2/(\sigma_i^2 + \sigma_p^2)$. Con precios muy volátiles, $\vartheta \to 0$ y apenas reaccionan, mientras que con precios estables $\vartheta \to 1$. La oferta de cada isla es $y_{it} = \bar{y}_t + \zeta\,(p_{it} - p_t^e)$, con $\zeta \equiv \vartheta\eta > 0$ y η la capacidad de ajustar la producción. Al agregar, el componente específico promedia a cero y se añaden los *shocks* de oferta, ε_t^s.

Demanda agregada: procede de la teoría cuantitativa, con v_t un *shock* de velocidad, de esperanza nula, que recoge los *shocks* exógenos de demanda, y decrece con el nivel de precios, porque un precio mayor reduce los saldos reales. El modelo queda así:

$$\text{OA: } y_t = \bar{y}_t + \zeta\,(p_t - p_t^e) + \varepsilon_t^s$$
$$\text{DA: } y_t = m_t - p_t + v_t$$

Salvo por los *shocks* de oferta, la producción no se desvía de su tendencia si $\vartheta = 0$, como en una hiperinflación, si $\eta = 0$ o si los cambios de precios están plenamente anticipados.

Equilibrio: igualando, con expectativas racionales, solo los componentes no anticipados mueven la brecha:

$$\tilde{y}_t = \frac{\zeta}{1+\zeta}(\varepsilon_t + v_t - \varepsilon_t^s) + \varepsilon_t^s$$

porque la sorpresa de precios depende únicamente de ellos, $p_t - \mathbb{E}_{t-1}p_t = \frac{1}{1+\zeta}(\varepsilon_t + v_t - \varepsilon_t^s)$.

Estática comparativa: la producción responde a las sorpresas monetarias y de demanda con $\zeta/(1+\zeta)$ y a las de oferta con $1/(1+\zeta)$. En el plano renta-precios, la oferta de largo plazo es vertical en la producción tendencial y la de corto plazo, creciente. Un cambio monetario anticipado desplaza a la vez la demanda y la oferta de corto plazo, se traslada íntegramente a los precios y es neutral. Su parte no anticipada eleva transitoriamente la producción, hasta que los agentes infieren el verdadero nivel de precios y revisan sus expectativas. Los *shocks* de oferta afectan a la capacidad productiva y pueden no ser transitorios. El ciclo es, así, una sucesión de equilibrios.

2.1.3. Conclusiones

Aspectos positivos: es el primer modelo microfundamentado del ciclo. Obtiene efectos reales del dinero con precios flexibles, compatibles con la neutralidad a largo plazo, y predice que el *trade-off* producción-inflación es peor donde la demanda es más volátil, lo que Lucas (1973) respaldó con datos de dieciocho países. Barro (1977) halló que solo el dinero no anticipado afecta al desempleo.

Limitaciones: no explica la persistencia, por falta de un mecanismo de propagación; la confusión informativa es poco plausible con estadísticas rápidas y baratas; no explica los *shocks* reales, como las crisis del petróleo, y requiere una oferta de trabajo más elástica que la observada.

2.2. Ciclos reales (Kydland y Prescott, 1982)

Idea: Kydland y Prescott (1982) y Long y Plosser (1983) explican el ciclo por la conducta optimizadora ante perturbaciones tecnológicas estocásticas y autocorrelacionadas. El impulso es tecnológico; la transmisión, la sustitución intertemporal de consumo y ocio y la acumulación de capital, y la persistencia, la autocorrelación de los *shocks* y el *time to build*.

2.2.1. Supuestos

Economía cerrada con información perfecta, horizonte infinito, competencia perfecta y precios flexibles. Agentes idénticos con expectativas racionales y utilidad logarítmica en consumo y ocio, sin sector público. Tecnología Cobb-Douglas con progreso técnico aumentador del trabajo y depreciación completa.

2.2.2. Modelo

Hogares: el programa del hogar es:

$$\max_{c_t, l_t, k_{t+1}} \mathbb{E}_0 \sum_{t=0}^{\infty} \beta^t u(c_t, 1-l_t) \quad \text{s.a.: } c_t + k_{t+1} - k_t = w_t l_t + r_t k_t, \qquad \beta \in (0,1)$$

La concavidad estricta de la utilidad y de la tecnología, con la condición de transversalidad, hace las CPO necesarias y suficientes. La relación marginal entre ocio y consumo iguala el salario, y la ecuación de Euler estocástica iguala la utilidad marginal del consumo presente con la esperada de invertir:

$$u'(c_t) = \beta\,\mathbb{E}_t[u'(c_{t+1})\,(1 + r_{t+1})]$$

Empresas: maximizan su beneficio, $y_t - w_t l_t - (r_t + \delta)k_t$, y remuneran los factores por su productividad marginal, el trabajo al salario y el capital a su coste de uso. Las CPO son $f'(l_t) = w_t$ y $f'(k_t) = r_t + \delta$.

Tecnología: crece a una tasa g_A con un componente aleatorio AR(1), $a_t = \rho_A a_{t-1} + \phi_t$, y la producción es:

$$y_t = k_t^{\alpha}(A_t l_t)^{1-\alpha}, \qquad \ln A_t = \ln \bar{A} + g_A\, t + a_t$$

Equilibrio: el hogar ahorra una fracción constante, $k_{t+1} = \alpha\beta\, y_t$, y trabaja horas constantes, porque los efectos de la tecnología sobre el salario y el tipo de interés se cancelan. El consumo y el ahorro sí fluctúan. Con b el peso del ocio en la utilidad, $u(c_t, 1-l_t) = \ln c_t + b\ln(1-l_t)$, los valores solución son:

$$k^*_{t+1} = \alpha\beta\, y_t, \qquad c^*_t = (1-\alpha\beta)\, y_t, \qquad l^*_t = \bar{l}(\alpha, \beta, b)$$

Estática comparativa: un *shock* tecnológico positivo desplaza a la derecha la oferta agregada a largo y a corto plazo y, en menor medida, la demanda, porque eleva la renta esperada y la rentabilidad de la inversión. La economía pasa a más producción y un nivel de precios algo menor. Como el equilibrio está siempre sobre la oferta de largo plazo, no se abre brecha respecto del potencial, y la fluctuación del *output* es la del propio potencial.

Dinámica: la desviación del *output* respecto de su tendencia determinista sigue un AR(2):

$$\tilde{y}_t = (\alpha + \rho_A)\, \tilde{y}_{t-1} - \alpha\rho_A\, \tilde{y}_{t-2} + (1-\alpha)\, \phi_t$$

que resulta de sustituir el AR(1) de la tecnología en $\tilde{y}_t = \alpha\, \tilde{y}_{t-1} + (1-\alpha)\, a_t$. Sus raíces, α y ρ_A, son reales y menores que la unidad, de modo que el modelo no genera oscilaciones propias, y la recurrencia del ciclo se debe a la llegada continua de perturbaciones.

El ciclo dura más cuanto mayor es ρ_A. Un *shock* transitorio expande producción, consumo, inversión y salarios. Consumo e inversión varían en la misma proporción que la producción, porque la tasa de ahorro, $\alpha\beta$, es constante. El tipo real sube primero y luego cae. Uno permanente desplaza la propia tendencia. Nelson y Plosser (1982) muestran que el PIB se describe mejor con raíz unitaria, y si los *shocks* permanentes son frecuentes explican a la vez el ciclo y el crecimiento.

Bienestar: las fluctuaciones son la respuesta óptima a las perturbaciones, el propio óptimo de Pareto fluctúa, y toda intervención reduce el bienestar.

Extensiones: las indivisibilidades del trabajo elevan la elasticidad de la oferta de trabajo. La depreciación incompleta hace más sensibles la inversión y el empleo, y el sector público rompe el vínculo entre producción y salario real.

2.2.3. Conclusiones

Aspectos positivos: con modelos sencillos se reproducía gran parte del ciclo de EE. UU. (1950-1979). Los *shocks* tecnológicos explicarían en torno al 70 % de la volatilidad del PIB (Kydland y Prescott, 1991), y las versiones calibradas recogen la suavización del consumo y la volatilidad de la inversión. El ciclo real supuso una revolución metodológica que adoptó también la NEK.

Limitaciones: es difícil atribuir las recesiones a regresiones tecnológicas; el modelo básico carece de propagación; el empleo apenas fluctúa y el salario y la productividad son demasiado procíclicos, con una elasticidad de la oferta de trabajo irreal, y el dinero es neutral, contra la evidencia de los VAR.

2.3. Desequilibrio y fricciones financieras

La NEK parte de fallos de mercado, como el poder de mercado o la información asimétrica, que generan rigideces, de modo que los mercados no se vacían, el ajuste es lento y por cantidades y caben fallos de coordinación.

2.3.1. Racionamiento del crédito (Stiglitz y Weiss, 1981)

Idea: con información asimétrica, los bancos racionan el crédito en lugar de subir el tipo. El impulso son perturbaciones de demanda, reales o monetarias; la transmisión, la rigidez del tipo, y la persistencia, el ajuste escalonado de precios.

Supuestos: bancos competitivos e inversores sin alternativa a la banca, ambos neutrales al riesgo. Los proyectos solo difieren en su probabilidad de éxito, p_j, con el mismo rendimiento esperado, $\bar{X} = p_j X_j$, y solo el inversor conoce su riesgo. El banco presta B y el inversor devuelve $(1+i)B$ con responsabilidad limitada, una asimetría en los pagos que incentiva el riesgo excesivo.

Inversores: con responsabilidad limitada, piden crédito si su beneficio esperado, $\bar{X} - p_j(1+i)B$, no es negativo, y ese beneficio crece con el riesgo. Al subir el tipo baja el umbral de probabilidad de éxito y los proyectos seguros se retiran, la **selección adversa**.

Bancos: eligen el tipo i^* que maximiza su rendimiento esperado por unidad prestada, $(1+i)\,\bar{p}(i)$, con $\bar{p}(i)$ la probabilidad media de éxito de quienes piden crédito. Subir el tipo tiene un efecto recaudación, positivo, y un **efecto deterioro de cartera**, negativo, que domina por encima de i^*. En el tipo óptimo se cumple:

$$\bar{p}(i^*) + (1+i^*)\,\bar{p}'(i^*) = 0$$

donde el primer sumando es el efecto recaudación y el segundo, el de deterioro de cartera.

Equilibrio: si a i^* la demanda supera a la oferta, hay **racionamiento** (exceso de demanda y ajuste vía cantidades). El banco prefiere racionar a subir el tipo más allá de i^*, inferior al walrasiano, i^W, y el racionamiento es $Z = L^d(i^*) - L^s(i^*)$.

Estática comparativa: una caída de los depósitos aumenta el racionamiento y contrae el crédito, la inversión y el PIB. Una expansión monetaria lo reduce. Los efectos son asimétricos, porque recaen sobre las empresas más endeudadas y las pequeñas.

Explica el canal del crédito y anticipa el acelerador financiero, pero las garantías atenúan la asimetría y el modelo es de equilibrio parcial.

2.3.2. Acelerador financiero (Bernanke y Gertler, 1989)

Idea: si el coste de financiarse depende del patrimonio de los prestatarios, y este cae en las recesiones, la fricción amplifica y propaga el ciclo. La **prima de financiación externa** cae con la parte de la compra de capital financiada con recursos propios:

$$\rho_{\mathrm{fe},t} = \rho_{\mathrm{fe}}\left(\frac{\mathrm{PN}_t}{q_t K_t}\right), \qquad \rho'_{\mathrm{fe}}(\cdot) < 0$$

donde PN_t es el patrimonio neto, q_t el precio del capital y K_t el capital que se adquiere, de modo que $q_t K_t - \mathrm{PN}_t$ es la financiación externa necesaria. Como el prestamista no observa sin coste el resultado de los proyectos, la financiación externa es más cara que la interna.

Razonamiento: una perturbación negativa opera por dos canales. Por el **patrimonio neto** (Bernanke y Gertler, 1989), caen beneficios y precios de los activos, sube la prima y la inversión cae más de lo que justifica el *shock*. Por el **colateral** (Kiyotaki y Moore, 1997), el mismo activo produce y garantiza la deuda, de modo que su abaratamiento endurece el límite de endeudamiento y fuerza ventas. Como el patrimonio se recompone despacio, el efecto persiste. Bernanke, Gertler y Gilchrist lo bautizaron como *acelerador financiero* (1996) y lo integraron en un DSGE neokeynesiano (1999). No es exclusivo de la NEK, porque también surge con precios flexibles.

La crisis de 2008 confirmó su relevancia, aunque olvida el capital de los intermediarios y capta mal la dinámica no lineal de las crisis.

2.3.3. Ciclo financiero

Inestabilidad financiera (Minsky, 1986): la fragilidad nace del propio auge, que desplaza la financiación de posiciones cubiertas a especulativas y Ponzi.

Definición: el **ciclo financiero** es el conjunto de interacciones, mutuamente reforzadas, entre la percepción del valor y del riesgo, la actitud ante el riesgo y las restricciones de financiación, que generan auges seguidos de contracciones (Borio, 2014). Se mide con el componente de medio plazo del crédito privado sobre el PIB y de los precios de los inmuebles. Sus hechos estilizados son tres:

- Dura mucho más que el ciclo económico, de uno a ocho años: unos 16 años de media en siete economías avanzadas desde los sesenta, y casi 20 en los ciclos con máximo posterior a 1998.
- Las crisis financieras de origen interno se produjeron en su máximo o cerca de él. Tras las crisis bancarias graves, el producto cae en promedio más de un 9 %, el desempleo sube unos 7 puntos en algo más de cuatro años y el precio real de la vivienda cae un 35 % en unos seis años (Reinhart y Rogoff, 2009).
- Su amplitud y duración han crecido desde mediados de los ochenta.

Interacción: las recesiones que coinciden con su fase contractiva son mucho más profundas, con caídas del PIB en torno a un 50 % mayores (Drehmann et al., 2012). Por eso estabilizar el ciclo económico no basta, y la política macroprudencial usa el colchón de capital anticíclico, ligado a la desviación del crédito sobre el PIB. Como el ciclo financiero es más largo, una economía puede crecer con inflación baja y estable mientras acumula desequilibrios que solo afloran en la crisis.

2.3.4. Otras aportaciones

Los **contratos escalonados** (Fischer, 1977; Taylor, 1980) hacen que las perturbaciones nominales persistan más allá de los contratos, y los **fallos de coordinación** (Blanchard y Kiyotaki, 1987; Cooper y John, 1988) generan equilibrios subóptimos e incluso múltiples. Con negociaciones no sincronizadas, cada empresa evita alejar sus salarios de los ya pactados por las demás, y por eso las perturbaciones duran más que los contratos.

2.4. Nueva síntesis: IS-PC-MR

Idea: el modelo canónico de Clarida, Galí y Gertler (1999) es un DSGE de tres ecuaciones que acepta las expectativas racionales y la microfundamentación de la NMC, pero incorpora las rigideces de la NEK. El ciclo resulta a la vez de *shocks* reales y de la no neutralidad del dinero a corto plazo.

Supuestos: competencia monopolística y precios a la Calvo en el mercado de bienes, con un mercado de trabajo competitivo. Hay cuatro agentes con expectativas racionales: hogares aversos al riesgo, empresas,

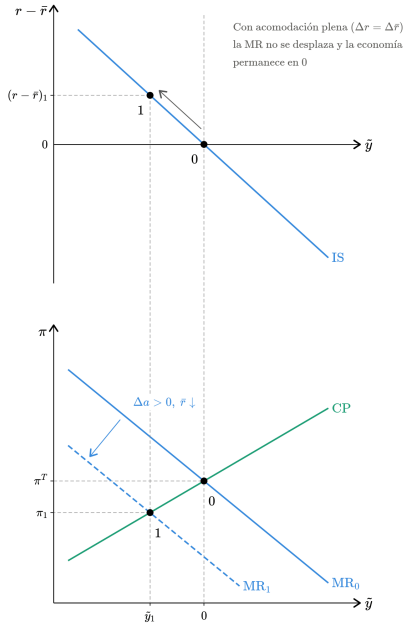
sector público y banco central. Por simplicidad, se prescinde del capital y, con él, de la inversión.

Planteamiento: una IS dinámica, la curva de Phillips de la NEK y la regla de Taylor:

$$\text{IS: } \tilde{y}_t = \mathbb{E}_t[\tilde{y}_{t+1}] - \frac{1}{\theta}(r_t - \bar{r}_t)$$
$$\text{CP: } \pi_t = \beta\,\mathbb{E}_t[\pi_{t+1}] + \kappa\,\tilde{y}_t$$
$$\text{Regla: } i_t = \bar{i}_t + \alpha_\pi\,\tilde{\pi}_t + \alpha_y\,\tilde{y}_t, \qquad \alpha_\pi > 1$$

- La IS procede de la ecuación de Euler: la demanda crece con la brecha esperada, porque los hogares alisan su consumo, y cae con el tipo real, $r_t = i_t - \mathbb{E}_t[\pi_{t+1}]$. La CP procede de Calvo: la inflación crece con la esperada y con la brecha. Aquí θ es la aversión relativa al riesgo, inversa de la elasticidad de sustitución intertemporal, y β el factor de descuento.
- Por la concepción wickselliana, $\bar{r}_t$ es el tipo natural, $\bar{r}_t = \rho + \theta\,\mathbb{E}_t[\Delta\bar{y}_{t+1}]$, que crece con la impaciencia, ρ, y con el crecimiento esperado del *output* natural. Si el tipo real lo supera, la brecha se vuelve negativa.
- Con la regla, el tipo nominal sube más que la inflación, el *principio de Taylor*, que estabiliza la economía y garantiza la unicidad del equilibrio. Aquí $\bar{i}_t = \bar{r}_t + \pi^T$ es el tipo neutral, π^T el objetivo de inflación y $\tilde{\pi}_t = \pi_t - \pi^T$ la desviación respecto de él. Sustituida en la IS, da la curva MR: ante más inflación, menos *output*.

Estática comparativa: ante un *shock* de costes, el banco central acepta transitoriamente una brecha negativa para devolver la inflación al objetivo, y ante una expansión fiscal sube el tipo, de modo que el multiplicador depende de la acomodación monetaria. Un *shock* tecnológico positivo transitorio reduce el tipo natural. Si la regla traslada esa caída al tipo neutral, no se abre brecha ni hay conflicto entre objetivos. Si no la traslada, el tipo real queda por encima del natural, la demanda crece menos que el *output* natural, la MR se desplaza a la izquierda y se abren brechas negativas de producción e inflación.



IS-PC-MR ante un *shock* tecnológico positivo no acomodado: la MR se desplaza y la economía pasa de 0 a 1, con brechas negativas.

A diferencia del ciclo real, la fluctuación depende de la respuesta del banco central, no solo del impulso real. El modelo es la referencia de los bancos centrales, pero no explica la persistencia de la inflación, prescinde del sector financiero y le cuesta afrontar grandes perturbaciones o el límite inferior de los tipos.

En suma, las cuatro teorías comparten la microfundamentación y se distinguen por el impulso y la propagación que privilegian. Cuál describe mejor el ciclo es una cuestión empírica.

3. Análisis empírico del ciclo

Medir el ciclo exige separar la tendencia del componente cíclico. Con esa medida se obtienen los hechos estilizados que toda teoría debe reproducir.

3.1. Medición de los ciclos

Descomposición: la serie se descompone en tendencia, ciclo, estacionalidad e irregular, de forma aditiva o multiplicativa, que con logaritmos se vuelve aditiva. Primero se elimina la estacionalidad, el patrón que se repite cada año, con medias móviles o modelos ARIMA. La forma aditiva, $Y_t = T_t + S_t + C_t + \varepsilon_t$, conviene cuando la magnitud de las fluctuaciones no varía en el tiempo, y la multiplicativa, $Y_t = T_t \cdot S_t \cdot C_t \cdot \varepsilon_t$, cuando son proporcionales al nivel de la serie, como en la mayoría de los agregados.

Brecha de producción: es la desviación porcentual del PIB respecto del potencial, $\tilde{y}_t = y_t - \tau_t$. El potencial es, en lectura estadística, la tendencia suave de la serie y, en lectura económica, el *output* compatible con una inflación estable o con el uso pleno de los recursos. Con $y_t = \ln Y_t$ y $\tau_t = \ln \bar{Y}_t$, la brecha es:

$$\tilde{y}_t = y_t - \tau_t \approx \frac{Y_t - \bar{Y}_t}{\bar{Y}_t}$$

Métodos univariantes: la tendencia lineal supone un potencial que crece a tasa constante, lo que la raíz unitaria pone en duda. El **filtro de Hodrick-Prescott** (1997) elige una tendencia estocástica que equilibra ajuste y suavidad:

$$\min_{\{\tau_t\}} \sum_{t=1}^{T}(y_t - \tau_t)^2 + \lambda \sum_{t=2}^{T-1}[(\tau_{t+1} - \tau_t) - (\tau_t - \tau_{t-1})]^2$$

donde λ penaliza los cambios en la pendiente de la tendencia.

- La CPO da un sistema lineal en el que la tendencia es un promedio ponderado de la serie, más suave cuanto mayor es λ, y la convexidad garantiza el mínimo. La solución es $\tau = (I + \lambda K'K)^{-1}\mathbf{y}$, con K la matriz de segundas diferencias.
- Si $\lambda \to \infty$, la tendencia es lineal y, si $\lambda = 0$, coincide con la serie y no hay ciclo. Se usan $\lambda = 1600$ para datos trimestrales, 6,25 para anuales (Ravn y Uhlig, 2002) y 129 600 para mensuales.

Métodos multivariantes: la función de producción del FMI, la Comisión o la OCDE estima el potencial con el capital, el trabajo ajustado por la NAIRU y la productividad tendencial. Como el funcionamiento de la economía no se conoce con certeza, no hay un método único, y la brecha se revisa a menudo.

Usos: la brecha sitúa la posición cíclica, con infrautilización si es negativa y sobreutilización si es positiva. Guía la política monetaria, porque una brecha positiva anticipa presiones inflacionistas, y permite medir el tono fiscal, porque el saldo estructural se obtiene descontando el componente cíclico.

3.2. Hechos estilizados

Estimados los componentes cíclicos de los principales agregados, los hechos estilizados se resumen en tres:

- La **volatilidad** se mide con la cuasidesviación típica del componente cíclico, relativa a la del PIB. En muchas economías de la OCDE es mayor en la inversión, el consumo duradero y el comercio, similar en el empleo y menor en el capital, la productividad, el salario real, el consumo no duradero y el gasto público.
- En los **comovimientos**, además de las regularidades ya vistas, la productividad del trabajo es débilmente procíclica, el salario real es débilmente procíclico o acíclico, según la estimación, y la inflación es procíclica cuando dominan los *shocks* de demanda y contracíclica cuando dominan los de oferta.
- La **persistencia** se mide con el coeficiente de autocorrelación de primer orden del componente cíclico y es elevada en el PIB.

3.3. Las teorías frente a los hechos

Frente a los hechos, las teorías se complementan más que se excluyen:

- **Lucas** explica los efectos reales del dinero no anticipado, pero, sin propagación, su brecha es ruido blanco, cuando la autocorrelación observada es elevada.
- El **ciclo real**, calibrado, reproduce la ordenación de volatilidades, la prociclicidad del consumo y la inversión y la persistencia del *output*, heredada de ρ_A, pero falla en el mercado de trabajo y en los efectos del dinero.
- La **NEK** explica la no neutralidad del dinero y sus efectos asimétricos, pero no genera por sí sola los comovimientos agregados.
- La **nueva síntesis** recoge los efectos de los *shocks* tecnológicos y monetarios y la prociclicidad de la inflación ante la demanda, pero no la persistencia de la inflación sin indiciación.

Ninguna reproduce todos los hechos. De ahí los DSGE de tamaño medio (Smets y Wouters, 2007), que combinan *shocks* reales y nominales, rigideces de precios y salarios, hábitos de consumo y costes de ajuste, y se estiman sobre esos hechos. Con todo, la medición es inseparable de la teoría, porque el filtro y la definición de potencial condicionan los propios hechos estilizados.

A43. Crecimiento económico (I). Acumulación de capital y progreso técnico exógeno. El modelo de Solow. El modelo de Solow aumentado con acumulación de capital humano. El modelo de Ramsey-Cass-Koopmans

1. Modelo de Harrod-Domar (1939, 1946)

Origen: Roy Harrod (1939) y Evsey Domar (1946) propusieron dos modelos de raíz keynesiana con las mismas implicaciones. Su valor es sobre todo conceptual, porque plantean el problema de la estabilidad del crecimiento al que responde Solow-Swan.

Idea: la inversión tiene una naturaleza dual: crea demanda vía multiplicador y capacidad vía acelerador. Solo una tasa de crecimiento, la garantizada, hace que ambas crezcan al mismo ritmo. Nada asegura que la economía se sitúe en ella, y si se desvía, se aleja. El crecimiento transita así al filo de la navaja.

1.1. Supuestos

Entorno: mundo keynesiano en tiempo discreto, con una economía cerrada, desarrollada y sin sector público, $Y = C + I$. Los precios relativos no ajustan las proporciones de factores ni el ahorro a la inversión, y la inversión depende de las expectativas de los empresarios.

Agentes: las familias ahorran una fracción constante y exógena de la renta, $S = sY$. Las empresas ajustan su capital al deseado según el acelerador, $K^* = vY$, donde v es la ratio capital-output fijada por la tecnología.

Tecnología: Leontief, sin sustituibilidad entre factores, sin depreciación ni progreso técnico, y con una fuerza de trabajo que crece a la tasa exógena n:

$$Y = \text{mín}(K/v, L/u)$$

1.2. Modelo

Planteamiento: no hay programa de optimización. El modelo compara tres tasas de crecimiento:

- *Tasa garantizada* (Harrod): de la igualdad entre ahorro e inversión, $sY_t = K_{t+1} - K_t = v(Y_{t+1} - Y_t)$, se obtiene la tasa que mantiene el equilibrio del mercado de bienes en cada periodo:

$$g_w = \frac{Y_{t+1} - Y_t}{Y_t} = \frac{s}{v}$$

  que aumenta con la propensión al ahorro y disminuye con la intensidad de capital de la tecnología.
- *Tasa natural*: es la máxima compatible con el pleno empleo, $g_n = n$ (con progreso técnico, $n + g_A$).
- *Tasa efectiva*: g_Y, la que resulta de las decisiones de inversión de los empresarios, guiadas por los *animal spirits*.
- *Condición de Domar*: el aumento de la inversión amplía la demanda vía multiplicador, $\Delta I/s$, y su nivel añade una capacidad I/v. Igualando ambos incrementos, la inversión ha de crecer a $\Delta I/I = s/v$, la tasa garantizada.

Dinámica: para Harrod el equilibrio $g_Y = g_w$ es inestable, y la economía discurre por el **filo de la navaja**. Identifica dos desviaciones:

- *Exceso de ahorro* ($g_Y < g_w$, $S > I$): la producción no encuentra comprador. Los empresarios rebajan sus ventas esperadas, el acelerador reduce el capital deseado, invierten menos y la demanda se contrae aún más, en una espiral implosiva. Así explicaban Harrod y Domar la década de 1930.
- *Exceso de inversión* ($g_Y > g_w$, $S < I$): el proceso opera al revés y el crecimiento se vuelve explosivo.

Largo plazo: nada garantiza que $g_w = g_n$, porque s, v y n son parámetros independientes. Si $g_w > g_n$, el capital crece más deprisa que el trabajo y aparece capacidad ociosa permanente; si $g_w < g_n$, desempleo creciente y permanente.

Extensiones: la inestabilidad justifica la intervención pública, con una política fiscal anticíclica. Si además s/v queda por debajo del crecimiento objetivo, g_Y^T, la brecha de ahorro (*savings gap*) entre s y $v\,g_Y^T$ solo se cierra con ahorro público o ayuda exterior.

Salidas al filo de la navaja: o una función de producción continua que permita ajustar v, la vía de Solow-Swan, o una tasa de ahorro endógena, la de la escuela de Cambridge (Kaldor, 1956; Pasinetti, 1962). En esta, el ahorro depende de la distribución de la renta y los precios mueven la participación de los beneficios hasta que $s/v = n$.

Aspectos positivos: con Harrod-Domar arranca la teoría moderna del crecimiento. Lleva el multiplicador-acelerador a un marco dinámico y plantea si el crecimiento equilibrado es estable.

Limitaciones: los coeficientes fijos generan una inestabilidad que la evidencia no muestra.

2. Modelo de Solow-Swan (1956)

Origen: Solow (1956) y, de forma independiente, Swan (1956).

Idea: sustituir la Leontief por una función neoclásica con sustituibilidad entre capital y trabajo. La ratio capital-output se ajusta y desaparece el filo de la navaja. Por los rendimientos decrecientes del capital, solo el progreso técnico sostiene el crecimiento per cápita a largo plazo.

2.1. Supuestos

Entorno: información y competencia perfectas, tiempo continuo, economía cerrada, desarrollada y sin sector público.

Agentes: las familias son agentes tipo Robinson Crusoe (productores, consumidores y propietarios de factores y tecnología) y ahorran e invierten una fracción constante y exógena de la renta, $S = I = sY$, sin que el ahorro se derive de ninguna decisión.

Tecnología: Cobb-Douglas neoclásica, con A Hicks-neutral y constante, α y $1 - \alpha$ las participaciones de capital y trabajo, depreciación δ y población creciente a n:

$$Y = AK^\alpha L^{1-\alpha}, \quad \alpha \in (0,1)$$

- *Rendimientos constantes a escala*: $f(\lambda K, \lambda L) = \lambda f(K, L)$.
- *Productividades marginales positivas y decrecientes*, con cruzadas positivas.
- *Condiciones de Inada*: $\lim_{K\to 0} f'_K = \infty$ y $\lim_{K\to\infty} f'_K = 0$.

2.2. Modelo sin progreso técnico

Planteamiento: sin optimización, porque el ahorro es exógeno. Sin sector exterior ni público, $Y = C + I$, y cada ecuación de comportamiento pasa de niveles a per cápita y a tasas de crecimiento:

$$\begin{aligned}
&\text{Renta:} && Y = AK^\alpha L^{1-\alpha} \;\Rightarrow\; y = Ak^\alpha && \Rightarrow\; g_y = \alpha g_k, \qquad A > 0,\ \alpha \in (\\
&\text{Inversión:} && \dot{K} = sY - \delta K \;\Rightarrow\; \dot{k} = sAk^\alpha - (n+\delta)k && \Rightarrow\; g_k = sAk^{\alpha-1} - (n+\delta)\\
&\text{Consumo:} && C = (1-s)Y \;\Rightarrow\; c = (1-s)Ak^\alpha && \Rightarrow\; g_c = g_y
\end{aligned}$$

Como la renta per cápita depende solo del capital per cápita, entender el capital es entender el crecimiento.

Ecuación fundamental: la variación del capital per cápita, $\dot{k}$, es la diferencia entre el ahorro bruto per cápita y la inversión de reposición, que mantiene k ante la depreciación y los nuevos trabajadores. Mientras el ahorro bruto supere la reposición, $\dot{k} > 0$, la economía acumula capital y crece.

Tasas de crecimiento: la renta crece a una tasa proporcional a la del capital y el consumo, al ritmo de la renta. La economía crece mientras el ahorro bruto por unidad de capital supere la tasa de reposición. La tasa depende positivamente de s y A, negativamente de n y δ, y ambiguamente de α, cuyo efecto es positivo solo si $k > 1$, porque $\partial g_k/\partial\alpha = sAk^{\alpha-1}\ln k$.

Equilibrio: cuando $\dot{k} = 0$, el ahorro iguala la reposición y no quedan recursos para aumentar k. Los niveles y las tasas de crecimiento de estado estacionario son:

$$k^* = \left(\frac{sA}{n+\delta}\right)^{1/(1-\alpha)} \qquad y^* = A^{1/(1-\alpha)}\left(\frac{s}{n+\delta}\right)^{\alpha/(1-\alpha)} \qquad c^* = (1-s)y^*$$

$$g_y^* = g_k^* = g_c^* = 0 \qquad g_Y^* = g_K^* = n$$

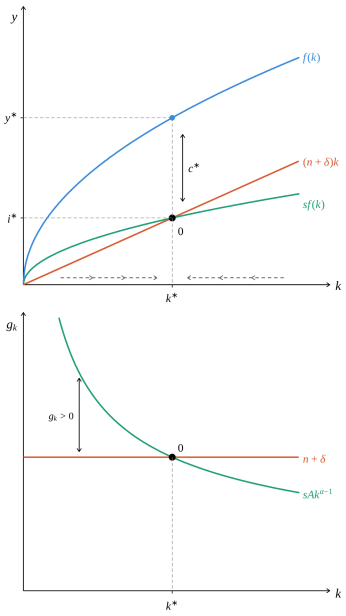
Sin progreso técnico, las variables per cápita no crecen y las agregadas crecen al ritmo de la población. Los precios de los factores, iguales a sus productividades marginales, son constantes, $w^* = (1-\alpha)Ak^{*\alpha}$ y $r^* = \alpha Ak^{*(\alpha-1)} - \delta$. El estado estacionario tiene cuatro propiedades:

- *Existencia*: por las dos condiciones de Inada, que hacen que ahorro y reposición se corten en un $k^* > 0$. La del origen hace que el ahorro supere a la reposición para k pequeño, y la del infinito, que quede por debajo para k grande.
- *Unicidad*: el ahorro por trabajador es cóncavo y la reposición, lineal.
- *Estabilidad*: la economía gravita hacia k^* desde cualquier punto, la diferencia fundamental con Harrod-Domar.
- *Ausencia de crecimiento endógeno*: por Inada en el infinito, la acumulación no sostiene por sí sola el crecimiento per cápita.

Desarrollo gráfico: en el plano k-y, el cruce de la curva de ahorro, $sf(k)$, con la recta de reposición, $(n+\delta)k$, determina k^* en el punto 0. A su izquierda k crece y a su derecha cae. En el panel inferior, g_k es la distancia entre $sAk^{\alpha-1}$ y $n + \delta$.

Estática comparativa: la alteración permanente de un parámetro tiene un **efecto nivel** permanente y un **efecto tasa** solo transitorio.

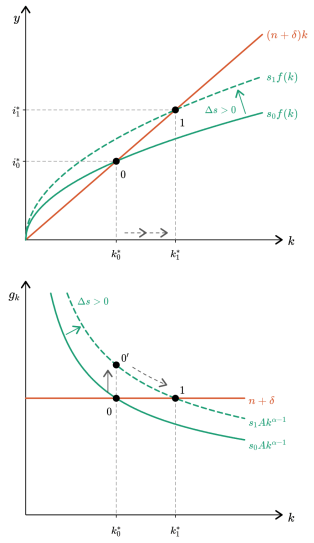
- *Aumento del ahorro*: el capital estacionario pasa de k_0^* a k_1^*. Durante la transición la economía crece más deprisa, y luego vuelve a la tasa inicial. El consumo cae a corto plazo, porque la inversión absorbe una fracción mayor de la producción, y solo sube a largo si se partía de un capital inferior al de la regla de oro. Ahorrar más, como se recomienda a menudo a los países pobres, eleva su nivel de renta, no su crecimiento a largo plazo.
- *Caída de n*: la recta de reposición pivota hacia abajo, porque hay que equipar a menos trabajadores nuevos, y k^* aumenta con el mismo patrón.
- *Mejora tecnológica*: un aumento de una vez de A desplaza la curva de ahorro con el mismo patrón. Solo un crecimiento continuo de A tiene efecto tasa permanente.



Modelo de Solow-Swan: el ahorro por trabajador iguala la inversión de reposición en k^*.

– *Destrucción de capital*: no altera ninguna curva. La economía vuelve a k^* creciendo transitoriamente más deprisa, como Alemania y Japón tras 1945.

Desarrollo gráfico: con un aumento del ahorro, en el panel superior la curva de ahorro pivota hacia arriba, de $s_0 f(k)$ a $s_1 f(k)$, y corta la recta de reposición en un capital mayor ($0 \to 1$). En el inferior, $s_0 A k^{\alpha-1}$ sube a $s_1 A k^{\alpha-1}$, de modo que en k_0^* la tasa g_k salta de cero a un valor positivo ($0 \to 0'$) y decrece hasta anularse en 1.



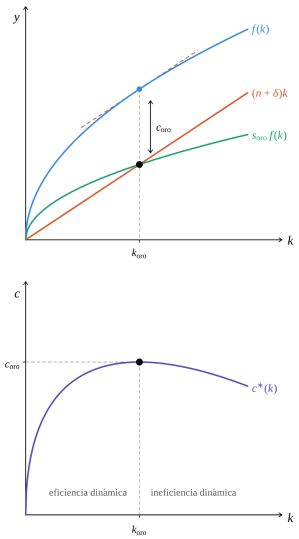
Aumento de la tasa de ahorro en Solow-Swan: k^* pasa de k_0^* a k_1^* ($0 \to 1$) y g_k salta en k_0^* ($0 \to 0'$) y se anula de nuevo en 1.

Regla de oro (Phelps, 1961): es la tasa de ahorro que maximiza el consumo de estado estacionario, $c^* = f(k^*) - (n+\delta)k^*$:

$$f'(k_{\text{oro}}) = n + \delta \quad \Rightarrow \quad s_{\text{oro}} = \alpha$$

La tasa de ahorro oro coincide con la participación del capital: se ahorra toda la renta del capital y se consume toda la del trabajo, con $k_{\text{oro}} = (\alpha A/(n+\delta))^{1/(1-\alpha)}$. Gráficamente, el consumo estacionario es la distancia vertical entre $f(k)$ y la recta de reposición. Crece mientras $f'(k) > n+\delta$ y alcanza su cima en k_{oro}, donde ambas pendientes coinciden.

– *Exceso de ahorro* ($k^* > k_{\text{oro}}$, $r^* < n$): reducir s hasta s_{oro} produce un salto inmediato del consumo y una convergencia a c_{oro}, superior al consumo inicial en todos los periodos. La economía es **dinámicamente ineficiente**, porque acumula capital que no le sirve para consumir más. Esto podría corregirse penalizando fiscalmente el ahorro financiero privado.



Regla de oro: el consumo estacionario es máximo en k_{oro}.

– *Insuficiencia de ahorro* ($k^* < k_{\text{oro}}$): subir s exige un sacrificio a corto plazo, aunque el consumo aumenta a largo plazo. Ninguna trayectoria factible eleva el consumo en todos los periodos, de modo que la economía es dinámicamente eficiente, y ahorrar más depende de la impaciencia de los consumidores.

Con ahorro exógeno, nada garantiza que se ahorre s_{oro}.

Extensiones: tres corolarios.

– *Ineficiencia dinámica*: como $r^* = f'(k^*) - \delta$, el exceso de ahorro equivale a $r^* < n = g_Y^*$. Abel, Mankiw, Summers y Zeckhauser (1989) comparan la renta bruta del capital con la inversión bruta. En las economías avanzadas la primera supera a la segunda, lo que indica eficiencia dinámica, aunque los tipos inferiores al crecimiento tras 2008 reabrieron el debate.
– *Paradoja de Lucas* (1990): con la Cobb-Douglas, si un país tiene diez veces la renta per cápita de otro y $\alpha = 1/3$, su capital per cápita es mil veces mayor y su rendimiento cien veces menor. El capital debería fluir masivamente a los pobres, y no lo hace. Sin riesgos ni fricciones, la explicación exige que las diferencias de renta se deban a la tecnología o al capital humano, o que la elasticidad de sustitución supere la unidad.
– *Elasticidad de sustitución*: la Cobb-Douglas impone $\sigma = 1$. Con una CES y $\sigma > 1$, se incumple Inada en el infinito y el crecimiento per cápita puede sostenerse sin progreso técnico. Con $\sigma < 1$, el estado estacionario puede no existir con ahorro bajo. Además, con $\sigma \neq 1$ las participaciones factoriales dejan de ser constantes.

2.3. Modelo con progreso técnico

Planteamiento: para un crecimiento sostenido, se añade un **progreso técnico exógeno** a la tasa g_A que, como el maná, no consume recursos. Puede ser neutral en el sentido de Hicks, $AF(K,L)$; aumentador del trabajo o neutral en el sentido de Harrod, $F(K,AL)$, o aumentador del capital o neutral en el sentido de Solow, $F(AK,L)$.

Teorema de Uzawa (1961): solo el progreso aumentador del trabajo es compatible con un crecimiento equilibrado, con tasas y participaciones factoriales constantes. Cualquier otra forma haría variar la relación capital-producto sin límite, aunque con la Cobb-Douglas las tres son indistinguibles. Por unidad de trabajo efectivo, $\tilde{x} = X/(AL)$:

$$\text{Renta:} \quad Y = K^\alpha (AL)^{1-\alpha} \;\Rightarrow\; \tilde{y} = \tilde{k}^\alpha \qquad \Rightarrow\; g_{\tilde{y}} = \alpha g_{\tilde{k}}, \qquad \alpha \in (0,1$$

$$\text{Inversión:} \quad \dot{K} = sY - \delta K \quad \Rightarrow\; \dot{\tilde{k}} = s\tilde{k}^\alpha - (g_A + n + \delta)\tilde{k}$$

Ahora hay que equipar también a las nuevas unidades de eficiencia, y $g_y = g_A + g_{\tilde{y}}$.

Equilibrio: en el estado estacionario, el capital y la renta por unidad de trabajo efectivo son constantes:

$$\tilde{k}^* = \left(\frac{s}{g_A + n + \delta}\right)^{1/(1-\alpha)} \qquad \tilde{y}^* = \tilde{k}^{*\alpha}$$

$$g_y^* = g_k^* = g_c^* = g_A \qquad g_Y^* = g_K^* = g_A + n$$

Las variables per cápita crecen a la tasa del progreso técnico y las agregadas, a esa tasa más la población.

El crecimiento equilibrado reproduce los hechos estilizados de Kaldor: relación capital-producto, tipo de interés y participaciones constantes, y salario real creciente a g_A.

Estática comparativa: g_A es el único parámetro con efecto tasa permanente. La acumulación explica la transición, no el estado estacionario. Esa conclusión es el mayor logro del modelo, coherente con el residuo de la contabilidad del crecimiento, y su mayor limitación, porque deja sin explicar la variable que lo explica todo.

2.4. Dinámica de transición y convergencia

Velocidad de convergencia: log-linealizando la ecuación fundamental, cada año se cierra una fracción constante λ de la distancia al estado estacionario, y la mitad se recorre en $t_{1/2} = \ln 2/\lambda$:

$$\lambda = (1-\alpha)(g_A + n + \delta)$$

que procede de $\dot{\ln \tilde{k}} \approx -\lambda(\ln \tilde{k} - \ln \tilde{k}^*)$, con solución:

$$\ln \tilde{k}(t) - \ln \tilde{k}^* = e^{-\lambda t}\left(\ln \tilde{k}(0) - \ln \tilde{k}^*\right)$$

Es mayor cuanto menor es α, porque los rendimientos decrecientes son más intensos. Con $\alpha \approx 1/3$ y $g_A + n + \delta \approx 0{,}06$, $\lambda \approx 4$ % ($t_{1/2} \approx 17$ años), el doble de la observada, en torno al 2 %.

Convergencia entre países: tres formulaciones.

- *β-convergencia absoluta*: los pobres crecen más que los ricos. Se estima con $g_y = \beta_0 - \beta \ln y_0 + \varepsilon$, existe si $\beta > 0$ y exige el mismo estado estacionario. La evidencia de 1960-1990 no la apoya en muestras grandes de países, aunque sí la de las últimas décadas, y eso al principio se leyó como evidencia contraria al modelo neoclásico. Sí se observa entre regiones de un mismo país, a tasas de alrededor del 2 %, como entre los estados de EE. UU. (Barro y Sala-i-Martin, 1991), con resultados similares para las regiones europeas y las prefecturas japonesas.
- *β-convergencia condicional*: controlando por las diferencias en parámetros, $g_y = \beta_0 - \beta \ln y_0 + \gamma z + \varepsilon$, sí aparece una correlación parcial negativa. Basta con que un parámetro no coincida para que cada país se dirija a su propio estado estacionario y deje de estar asegurado que los pobres crezcan más. Barro y Sala-i-Martin (1992a) obtienen en 98 países una tasa de alrededor del 2 % ($t_{1/2} \approx 35$ años).
- *σ-convergencia*: la dispersión de la renta per cápita disminuye. La β-convergencia absoluta es condición necesaria, pero no suficiente.

2.5. Extensiones

2.5.1. Modelo de Solow aumentado con capital humano (Mankiw, Romer y Weil, 1992)

Idea: Mankiw, Romer y Weil (MRW) atribuyen los fallos cuantitativos del modelo a una participación del capital de un tercio. Si el capital incluye el humano, se acerca a dos tercios, los rendimientos decrecientes se suavizan y se corrigen la convergencia demasiado rápida y los diferenciales de renta demasiado pequeños.

Supuestos: el capital humano, H, entra como tercer factor. Se ahorran fracciones exógenas de la renta en cada capital, s_k y s_h, ambos se deprecian a δ, y $\alpha + \beta < 1$, de modo que se mantiene Inada y no hay crecimiento endógeno:

$$Y = K^\alpha H^\beta (AL)^{1-\alpha-\beta}, \qquad \alpha, \beta > 0,\ \alpha + \beta < 1$$

Equilibrio: con $Y = C + I_K + I_H$, dos leyes de acumulación, $\dot{\tilde{k}} = s_k \tilde{y} - (g_A + n + \delta)\tilde{k}$ y $\dot{\tilde{h}} = s_h \tilde{y} - (g_A + n + \delta)\tilde{h}$, cuyos *loci* se cortan en un estado estacionario estable. Los niveles y las tasas de estado estacionario son:

$$\tilde{k}^* = \left(\frac{s_k^{1-\beta} s_h^{\beta}}{g_A + n + \delta}\right)^{1/(1-\alpha-\beta)} \qquad \tilde{h}^* = \left(\frac{s_k^{\alpha} s_h^{1-\alpha}}{g_A + n + \delta}\right)^{1/(1-\alpha-\beta)}, \qquad \alpha, \beta > 0,\ \alpha + \beta < 1$$

$$g_y^* = 0 \qquad g_y^* = g_A$$

Las tasas de crecimiento son, pues, las del modelo canónico.

Renta estacionaria: en logaritmos, depende de las dos tasas de inversión y de la tasa de reposición:

$$\ln y^* = \ln A_0 + g_A t + \frac{\alpha \ln s_k + \beta \ln s_h - (\alpha+\beta)\ln(g_A + n + \delta)}{1-\alpha-\beta}$$

Estática comparativa: efecto nivel, no tasa. Las elasticidades de y^* respecto a s_k y s_h son $\alpha/(1-\alpha-\beta)$ y $\beta/(1-\alpha-\beta)$. La primera es mayor que en el canónico, $\alpha/(1-\alpha)$, porque más ahorro eleva la renta y, con ella, el capital humano. La de $g_A + n + \delta$, $-(\alpha+\beta)/(1-\alpha-\beta)$, es mayor en valor absoluto.

Dinámica: log-linealizando, el crecimiento entre dos fechas depende de la renta inicial y de los determinantes del estado estacionario, que es la regresión de convergencia condicional que estima la literatura. La velocidad cae con la participación conjunta de ambos capitales:

$$\lambda = (1-\alpha-\beta)(g_A + n + \delta)$$

con la que la renta se aproxima a su estado estacionario según:

$$\ln \tilde{y}_t - \ln \tilde{y}_0 = (1 - e^{-\lambda t})\left[\ln \tilde{y}^* - \ln \tilde{y}_0\right]$$

Aspectos positivos: MRW obtienen $\alpha \approx \beta \approx 1/3$. La velocidad cae del 4 % al 2 %, el R^2 pasa de unos 0,6 a unos 0,8 y los diferenciales predichos se acercan a los observados.

Limitaciones: aproximan s_h con la tasa de escolarización secundaria, multiplicada por la proporción de la población en edad de trabajar que está en edad escolar, una medida criticada por imprecisa, que no captura la calidad de la educación, la experiencia ni el aprendizaje en el puesto.

2.5.2. Otras extensiones

Economía abierta (Barro et al., 1995): con movilidad perfecta del capital la convergencia sería instantánea, lo que la evidencia desmiente. Si solo el capital físico se financia fuera y el humano se acumula con ahorro interno, la velocidad es mayor que en la economía cerrada, pero finita y próxima al 2 %.

Tierra y recursos naturales: con tierra fija, recursos agotables a la tasa b y g_A el crecimiento de la PTF, $g_y^* = [g_A - (\gamma+\eta)n - \eta b]/(1-\alpha)$. El crecimiento per cápita solo es positivo si el progreso técnico compensa el freno de la tierra y los recursos. Así responde el enfoque neoclásico a la predicción maltusiana y al *Informe Meadows* (1972).

2.6. Conclusiones

Aspectos positivos: el modelo es la referencia de toda la literatura posterior. Con sustitución entre factores el estado estacionario es estable, y solo el avance tecnológico sostiene el crecimiento de la renta por habitante.

Limitaciones: el progreso técnico es exógeno, y el propio Solow (1957) atribuyó al residuo el 87,5 % del crecimiento de la producción por hora en EE. UU. entre 1909 y 1949. Además, con $\alpha \approx 1/3$ la velocidad de convergencia y los diferenciales de renta son irrealistas, el modelo es unisectorial, el ahorro es exógeno y predice diferenciales de rentabilidad muy superiores a los observados.

3. Modelo de Ramsey-Cass-Koopmans (1928, 1965)

Origen: Ramsey (1928) planteó el problema del ahorro óptimo. Cass (1965) y Koopmans (1965) lo formalizaron como modelo de crecimiento con optimización intertemporal.

Idea: endogenizar la tasa de ahorro a partir de la decisión intertemporal de las familias. El ahorro óptimo puede crecer, decrecer o ser constante en la transición, pero las conclusiones cualitativas de Solow se mantienen.

3.1. Supuestos

Entorno: los supuestos neoclásicos de Solow, con horizonte infinito. La diferencia está en las familias, que deciden en cada momento cuánto ahorrar.

Agentes:

- *Familias*: hogares idénticos, cuyo tamaño crece a la tasa n, $L(t) = e^{nt}$. Operan bajo altruismo intergeneracional puro, como dinastías, lo que justifica el horizonte infinito. También puede interpretarse como el de agentes con una probabilidad de muerte constante, de tipo Poisson. Poseen activos a que rinden r y alquilan su trabajo al salario w.
- *Preferencias*: utilidad CRRA, con $\theta > 0$ el coeficiente de aversión relativa al riesgo y $1/\theta$ la **elasticidad de sustitución intertemporal**. Con $\theta = 1$ la utilidad es logarítmica. Con θ bajo, los hogares toleran más variabilidad del consumo, y con θ alto desean alisarlo fuertemente. La tasa de descuento cumple $\rho > n > 0$, lo que garantiza que la utilidad no diverge:

$$u(c) = \frac{c^{1-\theta} - 1}{1-\theta}, \qquad \theta > 0$$

- *Empresas*: alquilan el capital a su coste de uso, $r + \delta$, y contratan trabajo al salario w.

Tecnología: la Cobb-Douglas de Solow-Swan.

3.2. Modelo

Hogares: maximizan la utilidad descontada sujetos a la restricción dinámica, a sus activos iniciales y a la condición de no-Ponzi:

$$\max_{\{c\}} \int_0^\infty e^{-(\rho-n)t}\, u(c)\, dt \quad \text{s.a.:} \begin{cases} c + \dot{a} = w + (r-n)a \\ a_0 \text{ dado} \\ \lim_{t\to\infty} \lambda(t)\, a(t) \geq 0 \end{cases}$$

El factor $e^{-(\rho-n)t}$ descuenta a la tasa ρ y pondera por el tamaño del hogar, y la condición de no-Ponzi, escrita con el multiplicador $(\lambda(t) = \lambda(0)\, e^{-[R(t)-nt]}$, con $R(t) \equiv \int_0^t r\, dv)$, impide acabar endeudado en el límite. En el hamiltoniano en valor presente u origen, $\mathcal{H} = e^{-(\rho-n)t} u(c) + \lambda[(r-n)a + w - c]$, el multiplicador dinámico λ es el precio sombra de los activos (lo que vale en $t = 0$ una unidad más de ellos).

Por el principio del máximo, las CPO son $e^{-(\rho-n)t}c^{-\theta}=\lambda$, $(r-n)\lambda=-\dot\lambda$ y la transversalidad, $\lim_{t\to\infty}\lambda(t)\,a(t)=0$. Como el hamiltoniano es estrictamente cóncavo en c y lineal en a, son necesarias y suficientes. Tomando logaritmos en la primera, derivando respecto al tiempo y sustituyendo la segunda se obtiene la **ecuación de Keynes-Ramsey**:

$$g_c=\frac{1}{\theta}(r-\rho)$$

El consumo crece cuando la rentabilidad del ahorro supera a la impaciencia, $r>\rho$, porque los agentes ahorran hoy para consumir más mañana, y decrece en caso contrario. En ambos casos, el cambio es más pronunciado cuanto mayor es la elasticidad de sustitución intertemporal, lo que implica también una senda de consumo más volátil ante *shocks*.

Empresas: maximizan su beneficio, $AK^{\alpha}L^{1-\alpha}-(r+\delta)K-wL$, con precios dados, y pagan a cada factor su productividad marginal, $r=f'(k)-\delta$ y $w=f(k)-kf'(k)$. Con rendimientos constantes, el beneficio es nulo.

Equilibrio: en una economía cerrada sin sector público el único activo es el capital, de modo que $a=k$. El capital se acumula según la restricción de recursos, $c+\dot k+(n+\delta)k=f(k)$, donde, a diferencia de Solow, c no es una fracción fija de $f(k)$, sino una variable de control. Imponiendo $\dot c=0$, el tipo de interés iguala la tasa de descuento, $r=\rho$, y rige la **regla de oro modificada**. Como en el estado estacionario el ahorro repone exactamente el capital, $s^*=(n+\delta)k^*/f(k^*)$, con la Cobb-Douglas los niveles y las tasas son:

$$f'(k^*)=\rho+\delta\quad\Rightarrow\quad k^*=\left(\frac{\alpha A}{\rho+\delta}\right)^{1/(1-\alpha)}\qquad s^*=\alpha\,\frac{n+\delta}{\rho+\delta}$$
$$g_c^*=g_k^*=0$$

Como $\rho>n$, la productividad marginal exigida es mayor que en la regla de oro y, por rendimientos decrecientes, $k^*<k_{\text{oro}}$ y $s^*<\alpha$, porque los agentes impacientes no aceptan los sacrificios que exige la regla de oro. La tasa de ahorro estacionaria decrece con la impaciencia y, sin progreso técnico, capital y consumo per cápita no crecen.

Ahorro en la transición: con la Cobb-Douglas, es monótono y su sentido depende de s^* frente a $1/\theta$. En una economía pobre, el tipo de interés alto incentiva el ahorro por efecto sustitución, pero los agentes, que se saben más ricos en el futuro, desean adelantar consumo por efecto renta. Cuanto mayor es $1/\theta$, más pesa el primero.

- $s^*<1/\theta$: domina la sustitución y el ahorro decrece hacia s^*.
- $s^*>1/\theta$: domina el efecto renta y el ahorro crece hacia s^*.
- $s^*=1/\theta$: el ahorro es constante y el modelo reproduce exactamente a Solow.

Desarrollo gráfico: en el plano k-c:

1. *Locus* $\dot c=0$: vertical en k^*. A su izquierda el capital es escaso, el tipo de interés supera a la impaciencia y el consumo crece; a su derecha, al contrario.
2. *Locus* $\dot k=0$: la curva cóncava $c=f(k)-(n+\delta)k$, con su máximo en k_{oro}. Por encima de ella el consumo es demasiado alto y el capital disminuye, y por debajo aumenta. Corta el eje horizontal en el capital máximo sostenible, $\bar k$, donde $f(\bar k)=(n+\delta)\bar k$.
3. *Equilibrios*: el origen es inestable; el capital máximo sostenible con consumo nulo es estable, pero indeseable; y (k^*,c^*), el punto 0, es un **punto de silla**.

Hay una única senda estable. Las trayectorias por encima agotan el capital en un tiempo finito y las de abajo llevan a un consumo nulo e incumplen la transversalidad. Así, el consumo es una **variable de salto**, que se sitúa de inmediato sobre la senda estable, y el capital, una variable predeterminada. La economía converge desde la región de crecimiento ($k<k^*$, $c<c^*$), en la que el tipo de interés cae gradualmente hasta ρ y el crecimiento del consumo, hasta cero, o desde la de desacumulación ($k>k^*$, $c>c^*$).

Con θ elevado, la senda estable discurre muy cerca del *locus* $\dot k=0$, porque los hogares prefieren un consumo alto desde el principio. Con θ reducido es más empinada, porque recortan el consumo inicial para acumular deprisa. Por eso la velocidad de convergencia es mayor cuanto menor es θ. Con $\alpha\approx1/3$ supera con mucho el 2 % observado, y solo un capital amplio, $\alpha\approx2/3$, reconcilia el modelo con la evidencia.

Estática comparativa: tres perturbaciones no anticipadas.

- *Caída de* ρ: el *locus* $\dot c=0$ se desplaza a la derecha. El consumo salta hacia abajo a la nueva senda estable y después capital y consumo crecen hasta valores mayores (el patrón de Solow con $k<k_{\text{oro}}$).
- *Mejora tecnológica permanente*: el *locus* $\dot c=0$ se desplaza a la derecha, porque $f'(k)$ aumenta, y el $\dot k=0$, hacia arriba, porque hay más producción para cada k. El consumo salta, hacia arriba o hacia abajo, y a largo plazo k^* y c^* son mayores.
- *Mejora tecnológica transitoria*: los agentes la anticipan desde el primer momento y el consumo salta a una trayectoria intermedia que deja la economía sobre la senda original justo cuando el *shock* se revierte, un resultado que Solow no puede generar.

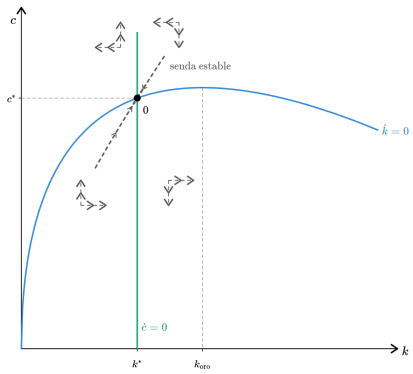
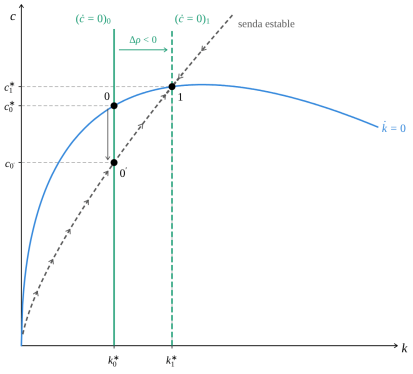


Diagrama de fases de Ramsey-Cass-Koopmans y senda estable hacia el punto de silla.



Caída no anticipada de ρ: el consumo salta a la nueva senda estable ($0\to0'$) y converge a 1.

Bienestar: se cumple el primer teorema del bienestar. El equilibrio descentralizado es Pareto-eficiente y coincide con el del planificador, porque las ecuaciones de Euler son iguales. No cabe ineficiencia dinámica. Con impuestos distorsionantes, externalidades o competencia imperfecta, la equivalencia se rompe.

Extensiones:

- *Teorema de la autopista* (Dorfman et al., 1958; Cass, 1966): con horizonte lejano, la senda óptima permanece casi siempre cerca del estado estacionario.
- *Progreso técnico*: con progreso aumentador del trabajo, la transversalidad exige $\rho>n+(1-\theta)g_A$ y el capital estacionario cumple $f'(\tilde k^*)=\rho+\delta+\theta g_A$: los agentes, que se saben más ricos, descuentan más el futuro. El consumo per cápita crece a g_A, como en Solow.
- *Sector público*: el gasto financiado con impuestos de suma fija desplaza hacia abajo el *locus* $\dot k=0$ y expulsa uno a uno al consumo privado, sin alterar k^*. Por la equivalencia ricardiana, lo mismo ocurre con deuda. Un impuesto sobre la renta del capital, en cambio, reduce k^*, con efecto nivel y no tasa.
- *Oferta de trabajo endógena*: con ocio en la utilidad, se añade la condición intratemporal $-u_l/u_c=w$, y el modelo es el núcleo del ciclo real.
- *Política económica*: las políticas que alteran los parámetros tienen efecto nivel y no tasa, y sin fallos de mercado no hay motivo para intervenir.

Aspectos positivos: microfundamenta el ahorro de Solow, prueba la eficiencia del equilibrio descentralizado e incorpora la impaciencia y la suavización del consumo, que determinan cuánto se ahorra y a qué ritmo se converge.

Limitaciones: la tecnología sigue siendo exógena, predice una convergencia condicional similar a la de Solow y el agente representativo con horizonte infinito excluye la heterogeneidad y los efectos intergeneracionales.

1. Función de producción no neoclásica

Los supuestos neoclásicos (rendimientos constantes a escala, rendimientos marginales decrecientes y condiciones de Inada) impiden el crecimiento endógeno: cada unidad de capital aporta cada vez menos, hasta que la inversión apenas repone la depreciación. La primera solución abandona alguna de esas propiedades y mantiene el resto del andamiaje de Solow, con ahorro exógeno salvo en Barro (1990).

1.1. Modelo AK (Rebelo, 1991)

Origen: Rebelo (1991). Su microfundamentación mediante el *aprendizaje por la práctica* procede de Arrow (1962) y Romer (1986).

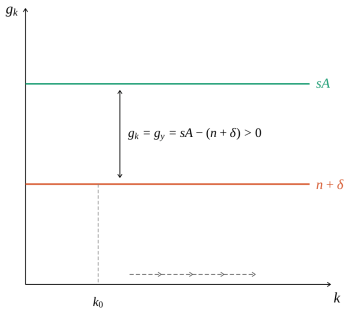
Idea: los supuestos son los de Solow-Swan salvo la función de producción, lineal en el capital *en sentido amplio*: físico, humano, conocimientos, infraestructuras y todo factor acumulable. Así, $Y = BK^{\alpha}H^{1-\alpha}$ equivale a un AK si una unidad de producto se convierte indistintamente en cualquiera de los dos capitales y estos se acumulan en la proporción que iguala sus rendimientos netos, $K/H = \alpha/(1-\alpha)$ con igual depreciación.

Supuestos: $Y = AK$ conserva solo los rendimientos constantes a escala. Pierde los rendimientos marginales decrecientes, porque el producto marginal es siempre A, y las condiciones de Inada. El entorno es el habitual: tecnología constante, ahorro exógeno s, economía cerrada sin sector público, población creciente a n y depreciación δ.

Planteamiento: como en Solow-Swan, con $Y = C + I$:

$$\begin{aligned}
&\text{Renta:} && Y = AK && \Rightarrow\ y = Ak && \Rightarrow\ g_y = g_k\\
&\text{Inversión:} && \dot{K} = sY - \delta K && \Rightarrow\ \dot{k} = sAk - (n+\delta)k && \Rightarrow\ g_k = sA - (n+\delta)\\
&\text{Consumo:} && C = (1-s)Y && \Rightarrow\ c = (1-s)Ak && \Rightarrow\ g_c = g_y
\end{aligned}$$

Desarrollo gráfico: en el plano k-g_k, la curva de ahorro y la de depreciación, divididas por k, son las rectas horizontales sA y $n+\delta$. Si $sA > n+\delta$, no se cruzan, y la distancia entre ellas, g_k, es la misma cualquiera que sea k_0: la economía crece indefinidamente a esa tasa común.



Modelo AK: con rendimientos constantes del capital, el crecimiento es positivo, permanente y sin convergencia.

Dinámica: las diferencias con Solow-Swan son profundas.

1. *Crecimiento sostenido*: la renta per cápita crece sin progreso técnico exógeno, el rasgo que define el crecimiento endógeno.
2. *Política económica*: las políticas que elevan el ahorro o la inversión afectan a la tasa de crecimiento, no solo al nivel de renta.
3. *Sin dinámica de transición*: la economía crece desde el principio a un ritmo invariable.
4. *Sin convergencia*, ni absoluta ni condicional: dos economías con parámetros idénticos crecen al mismo ritmo y mantienen para siempre su distancia relativa.
5. *Efectos permanentes de las recesiones*: sin un estado estacionario que actúe como atractor, no hay reversión a la tendencia.
6. *Sin ineficiencia dinámica*: como $s < 1$, $r = A - \delta > sA - \delta = g_Y$.

Aspectos positivos: es la formulación más sencilla del crecimiento endógeno e ilustra su mecanismo esencial: impedir que la productividad marginal del capital tienda a cero. El AK es, además, el caso $\alpha + \eta = 1$ del modelo de externalidades.

Limitaciones: no predice la convergencia condicional que apoya la evidencia, y supone una sustituibilidad perfecta, poco realista, entre capital físico y humano.

1.2. Modelo Sobelow (Jones y Manuelli, 1990)

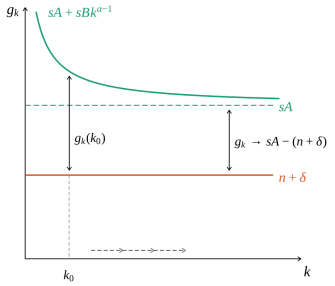
Idea: el modelo de Jones y Manuelli (1990), bautizado *Sobelow* por Sala-i-Martin porque combina las funciones de Solow y de Rebelo, aísla la propiedad sin la cual no hay crecimiento endógeno. Eliminar los rendimientos decrecientes basta, pero no es imprescindible. Lo indispensable es que **falle la condición de Inada en el infinito**, es decir, que la productividad marginal del capital no baje de cierta constante positiva.

Supuestos: la función suma un componente lineal y otro neoclásico, $Y = AK + BK^{\alpha}L^{1-\alpha}$, con rendimientos constantes a escala y decrecientes en el capital, pero $\lim_{K\to\infty} f'_K = A \neq 0$. Con las ecuaciones de comportamiento del AK, $y = Ak + Bk^{\alpha}$ y:

$$\dot{k} = s(Ak + Bk^{\alpha}) - (n+\delta)k \quad \Rightarrow \quad g_k = sA + sBk^{\alpha-1} - (n+\delta)$$

Dinámica: la curva de ahorro por unidad de capital es decreciente y tiene como asíntota la recta sA. Decrece porque el componente neoclásico tiene rendimientos decrecientes y el lineal no, y si $sA < n+\delta$ corta a la recta de depreciación en k^*.

1. $sA > n+\delta$: a corto plazo crece más deprisa que un AK, porque el término neoclásico suma mientras el capital es escaso. Después ese término se desvanece y g_k converge a $sA-(n+\delta) > 0$, con transición hacia una tasa positiva.
2. $sA < n+\delta$: el componente lineal no basta y, como en Solow, la renta per cápita se estanca en el estado estacionario.



Modelo Sobelow: el ahorro por unidad de capital no corta a la depreciación ampliada y el crecimiento es sostenido.

Aspectos positivos: combina convergencia y transición, herencia neoclásica, con crecimiento endógeno a largo plazo, herencia AK, y se ajusta mejor a la evidencia que el AK puro.

Limitaciones: no distingue tipos de capital y postula sin microfundamentar la cota inferior de la productividad marginal.

1.3. Externalidades (Romer, 1986; Lucas, 1988)

Idea: a partir del *learning by doing* de Arrow (1962), la tercera vía deja intacta la tecnología de cada empresa y añade efectos externos. Invertir enseña a producir, y lo aprendido, no rival y difícil de apropiar, se filtra al resto de la economía (*knowledge spillovers*): la inversión de cada empresa eleva la producción de las demás.

Supuestos: la empresa produce con $Y = AK^{\alpha}L^{1-\alpha}E^{\eta}$, con $\eta > 0$. Percibe rendimientos constantes en sus factores privados y toma E como dado, pero en el agregado la externalidad eleva los rendimientos del capital.

- *Visión de Romer* (1986): la externalidad procede del capital agregado, $E = K$, y la tasa de crecimiento depende de la población. Según este **efecto escala**, el crecimiento debería aumentar con la población, algo que los datos no confirman. En forma intensiva, $y = Ak^{\alpha+\eta}L^{\eta}$.
- *Visión de Lucas* (1988): procede del capital per cápita, $E = k$, y el efecto escala desaparece. Es una estilización de manual, porque en Lucas la externalidad procede del capital humano medio.

Planteamiento: con la visión de Lucas, $y = Ak^{\alpha+\eta}$ y:

$$g_k = sAk^{\alpha+\eta-1} - (n+\delta)$$

Desarrollo gráfico: en el plano k-g_k, la curva de ahorro por unidad de capital es decreciente, horizontal o creciente según $\alpha+\eta$ sea menor, igual o mayor que uno, y la distancia a la recta $n+\delta$ es g_k.

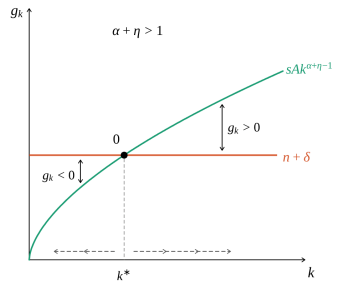
Dinámica: todo depende del tamaño de la externalidad, $\alpha+\eta$.

1. *Pequeña* ($\alpha+\eta < 1$): el ahorro por unidad de capital es decreciente y hay un estado estacionario estable sin crecimiento per cápita. La externalidad aleja el estado estacionario, pero el modelo es cualitativamente equivalente a Solow.
2. *Media* ($\alpha+\eta = 1$): compensa exactamente los rendimientos decrecientes y el modelo es idéntico al AK, con crecimiento si $sA > n+\delta$.
3. *Grande* ($\alpha+\eta > 1$): hay rendimientos crecientes y el estado estacionario es inestable. Por encima, la senda es explosiva, y por debajo, implosiva. Su interés empírico es limitado.

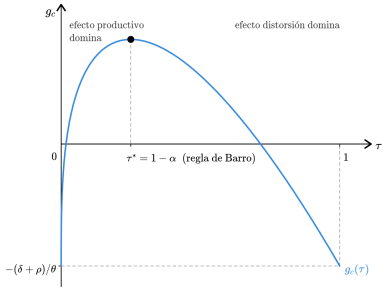
Extensiones: las externalidades dan margen a la intervención para elevar el crecimiento con *subvenciones a la inversión*, porque las empresas no capturan toda la externalidad e invierten menos de lo socialmente óptimo, y con la *promoción de los efectos externos*: *clusters*, centros de investigación y relaciones universidad-empresa. Las subvenciones, junto con las desgravaciones, aumentan el capital en sentido amplio.

Aspectos positivos: endogenizan el progreso técnico con un mecanismo plausible, el desbordamiento del conocimiento, sin abandonar la competencia perfecta.

Limitaciones: la evidencia de *learning by doing* es sectorial, como el 20 % de aprendizaje en semiconductores de Irwin y Klenow (1994), y



Externalidades con $\alpha+\eta>1$: el estado estacionario k^* es inestable.



Modelo de Barro (1990): crecimiento en U invertida respecto al tipo impositivo, con máximo en $\tau^*=1-\alpha$.

no alcanza la escala necesaria para el crecimiento agregado. Además, el progreso técnico es un subproducto no buscado de la inversión, cuando buena parte de la innovación procede de una I+D deliberada.

1.4. Gasto público productivo (Barro, 1990)

Idea: Barro hace de los servicios públicos (infraestructuras, seguridad jurídica, orden público), financiados con impuestos, el factor que sostiene la productividad del capital privado. El sector público es a la vez factor productivo y distorsión, y de ese doble papel resulta un tamaño óptimo del Estado.

Supuestos:

- *Entorno*: familias a la Ramsey, con utilidad CRRA de parámetro θ y descuento ρ, población constante y depreciación δ. El ahorro es aquí endógeno.
- *Agentes*: el sector público provee servicios por trabajador g, no rivales ni congestionables, financiados con un impuesto proporcional sobre la renta y presupuesto equilibrado, $g=\tau y$.
- *Tecnología*: rendimientos constantes en capital y servicios públicos y decrecientes en el capital privado. Sustituyendo $g=\tau y$, resulta un AK cuya constante depende de la política fiscal, creciente en τ:

$$y=Ak^{\alpha}g^{1-\alpha}=\tilde{A}(\tau)\,k,\qquad A>0,\ \alpha\in(0,1)$$

Planteamiento: las familias toman τ y g como dados y maximizan su utilidad intertemporal sujetas a $c+\dot{k}+\delta k=(1-\tau)Ak^{\alpha}g^{1-\alpha}$. El rendimiento neto del capital privado es $r=(1-\tau)\,\alpha\tilde{A}(\tau)-\delta$, con:

$$\tilde{A}(\tau)\equiv A^{1/\alpha}\tau^{(1-\alpha)/\alpha}$$

CPO: la ecuación de Keynes-Ramsey con el rendimiento neto del capital privado:

$$g_c=\frac{1}{\theta}\left[(1-\tau)\,\alpha\tilde{A}(\tau)-\delta-\rho\right]$$

La condición iguala el rendimiento neto del capital privado, descontados el impuesto y la depreciación, a la tasa de descuento más el crecimiento del consumo ponderado por θ. El consumo crece mientras la última unidad de capital, tras impuestos, rinde más de lo que la familia exige por posponer su consumo, y tanto más deprisa cuanto menor es θ.

CSO: el hamiltoniano es estrictamente cóncavo en c y cóncavo en k dado g, y con la transversalidad la CPO es suficiente. El crecimiento es positivo si $(1-\tau)\,\alpha\tilde{A}-\delta>\rho$ y la utilidad está acotada si $\rho>(1-\theta)g_c$.

Equilibrio: con tecnología lineal no hay transición, y consumo, capital y renta crecen desde el principio a g_c. El impuesto tiene dos efectos contrapuestos: reduce el rendimiento privado, efecto *distorsión* $(1-\tau)$, y financia un factor productivo, efecto *productivo* $(\tilde{A}(\tau))$. Maximizando $(1-\tau)\tau^{(1-\alpha)/\alpha}$ se obtiene la **regla de Barro**:

$$\tau^*=1-\alpha$$

$$g_c=g_k=g_y=\frac{1}{\theta}\left[\alpha^2\tilde{A}(1-\alpha)-\delta-\rho\right]$$

El tipo óptimo iguala la elasticidad del producto respecto al gasto público, de modo que el sector público debe tener el tamaño que corresponde a su contribución productiva. Ese tipo maximiza también la utilidad, porque, sin transición, crecimiento y consumo inicial dependen de τ solo a través del producto neto de impuestos por unidad de capital.

Desarrollo gráfico: en el plano τ-g_c, la tasa de crecimiento es una U invertida con máximo en $\tau^*=1-\alpha$. En $\tau=0$ no hay servicios públicos ni producción, y en $\tau=1$ no hay rendimiento privado. A la izquierda del máximo domina el efecto productivo y a la derecha, el de distorsión. En ambos extremos se anula el rendimiento del capital privado después de impuestos y la tasa alcanza su mínimo, $-(\delta+\rho)/\theta<0$.

Estática comparativa: un aumento de A o una reducción de ρ o δ elevan el crecimiento para todo τ, sin alterar τ^*, que solo depende de la tecnología.

Extensiones: el gasto en consumo público no eleva $\tilde{A}$ y solo distorsiona, lo que ofrece una lectura de la evidencia de Barro (1991): el consumo público reduce el crecimiento y la inversión pública no. La composición del gasto importa más que su volumen. Los impuestos de suma fija eliminan la distorsión, aunque el tamaño óptimo sigue siendo $g/y=1-\alpha$, que responde a la condición de eficiencia productiva, $\partial y/\partial g=1$. Con impuestos proporcionales el equilibrio es ineficiente, porque las familias no internalizan que su inversión, al elevar la recaudación, financia servicios públicos para todos. Si los servicios son congestionables, es decir, si el que recibe cada empresa depende de g/k, el tamaño óptimo es el mismo, pero el impuesto sobre la renta pasa a ser el instrumento eficiente, al internalizar la congestión.

Aspectos positivos: es el modelo de referencia sobre tamaño del sector público y crecimiento, y fundamenta la distinción entre gasto productivo e improductivo.

Limitaciones: la no rivalidad es un supuesto fuerte, porque la congestión es la regla en las infraestructuras. El resultado depende de que el gasto sea un flujo y no un *stock*, y la evidencia sobre tamaño del Estado y crecimiento es débil.

En suma, basta con que el rendimiento del capital amplio no caiga a cero para que el ahorro, y con él la política económica, determine la tasa de crecimiento. Pero el ahorro sigue siendo exógeno, salvo en Barro, el efecto escala de Romer (1986) no tiene respaldo y ninguno explica cómo se genera el conocimiento.

2. Capital humano: Uzawa-Lucas (1965; 1988)

Idea: el modelo de Uzawa (1965), popularizado por Lucas (1988), distingue explícitamente entre capital físico y humano y endogeniza el ahorro con hogares a la Ramsey.

2.1. Supuestos

Entorno: la población es constante y se normaliza a uno.

Agentes: los hogares maximizan una utilidad CRRA con dos controles, el consumo c y la fracción u, y dos estados, K y H.

Tecnología:

- *Factores*: capital físico, K, y humano, H, ambos rivales y acumulados con tecnologías distintas. Una fracción $u\in(0,1)$ del capital humano produce bienes y el resto, $1-u$, se dedica a la educación. Con la noción de capital humano de la escuela de Chicago, el trabajo pasa a ser un factor acumulable.
- *Sector de bienes finales*: produce con una Cobb-Douglas, $Y=AK^{\alpha}(uH)^{1-\alpha}$, y el capital físico se acumula con la producción no consumida, $c+\dot{K}+\delta K=Y$.
- *Sector educativo*: con productividad B, podría emplear ambos capitales, con una participación α_H del físico. Como la educación usa mucho más intensivamente el capital humano, $\alpha_H<\alpha$, y Uzawa y Lucas se quedan con el caso extremo en que solo emplea ese factor, $\alpha_H=0$:

$$\dot{H}=B(1-u)H-\delta H$$

  El supuesto convierte la educación en un **AK encubierto**, lineal en el factor acumulable, y de ahí nace el crecimiento endógeno.

2.2. Modelo

Hogares: eligen el consumo y la fracción u del capital humano que dedican a producir:

$$\max_{\{c,u\}}\int_0^{\infty}e^{-\rho t}\frac{c^{1-\theta}-1}{1-\theta}\,dt\quad\text{s.a.:}\begin{cases}c+\dot{K}+\delta K=AK^{\alpha}(uH)^{1-\alpha},\qquad A>0,\ \alpha\in(0,\\ \dot{H}=B(1-u)H-\delta H\\ K_0,\ H_0\ \text{dados}\end{cases}$$

El hamiltoniano en valor corriente tiene dos multiplicadores, λ y μ, y, como uH impide su concavidad conjunta, se supone la condición de suficiencia de Arrow. Las cuatro CPO, más las dos de transversalidad, dan tres resultados:

- *Keynes-Ramsey*: $g_c=(1/\theta)(r-\rho)$, con $r=\alpha AK^{\alpha-1}(uH)^{1-\alpha}-\delta$.

- *Asignación del tiempo*: el valor de una unidad de capital humano en la producción de bienes iguala su valor en la educación, $\lambda(1-\alpha)AK^{\alpha}(uH)^{-\alpha}=\mu B$.
- *Euler del capital humano*: $\mu(B-\delta)=\rho\mu-\dot{\mu}$, de modo que su precio sombra decrece a la tasa $B-\delta-\rho$, y con $\alpha_H=0$ evoluciona con independencia de la tecnología de bienes, lo que aísla el crecimiento de A y α.

Empresas: maximizan su beneficio, $AK^{\alpha}(uH)^{1-\alpha}-(r+\delta)K-w_H\,uH$, con precios dados y w_H la remuneración del capital humano empleado, y pagan a cada capital su productividad marginal, de donde sale el r anterior. Con rendimientos constantes, el beneficio es nulo. Al capital humano empleado le corresponde $w_H=(1-\alpha)AK^{\alpha}(uH)^{-\alpha}$.

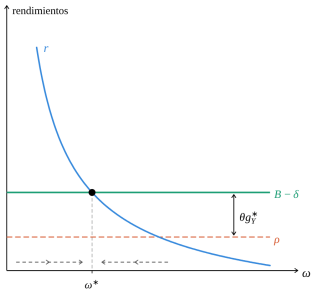
Equilibrio: en la senda de crecimiento equilibrado, K/H y u son constantes, y la condición de **arbitraje entre capitales** iguala el rendimiento del capital físico al del humano. Con Keynes-Ramsey y la acumulación del capital humano, $g_H^*=B(1-u^*)-\delta$, se obtienen la fracción de tiempo dedicada a producir y la tasa común de crecimiento:

$$r=B-\delta \qquad u^*=1-\frac{B-\delta-\rho+\theta\delta}{\theta B}$$

$$g_c^*=g_K^*=g_H^*=g_Y^*=\frac{1}{\theta}(B-\delta-\rho)$$

Para crecer hace falta $B>\delta+\rho$, es decir, la educación debe cubrir el olvido de conocimientos y la impaciencia. Solo B eleva el crecimiento a largo plazo, y la educación es el motor. La lógica es la del AK, pero sin suponer capitales sustitutos perfectos. Con utilidad logarítmica, $u^*=\rho/B$. Sin externalidades, la solución descentralizada coincide con la del planificador, como establece el primer teorema fundamental de la economía del bienestar.

Desarrollo gráfico: con $\omega\equiv K/H$ en el eje horizontal, el rendimiento del capital físico es decreciente en ω y el del humano, $B-\delta$, una horizontal que no decrece (la recta AK que sostiene el crecimiento). La tasa de descuento, ρ, es otra horizontal. En ω^* se igualan ambos rendimientos, y la distancia entre $B-\delta$ y ρ es θg_Y^*.



Modelo de Uzawa-Lucas: en ω^* se igualan los rendimientos de ambos capitales.

Estática comparativa: el crecimiento aumenta con la productividad educativa, B, y disminuye con la impaciencia, ρ, la tasa de olvido, δ, y, si es positivo, con el deseo de alisar el consumo, θ. Una mayor impaciencia lleva a producir más y, si $\rho>(\theta-1)\delta$, una mayor productividad educativa lleva a estudiar más. Los parámetros de la tecnología de bienes, A y α, solo afectan al nivel de la senda y a ω^*.

Extensiones: Lucas (1988) añade una externalidad del capital humano medio, H_a^{ζ}: la productividad de un trabajador crece con la cualificación de quienes le rodean, y el mercado dedica menos tiempo a la educación que el planificador. Rebelo (1991) muestra que el resultado se mantiene si la educación usa capital físico, con rendimientos constantes en los factores acumulables. Lucas interpreta el modelo como explicación de las diferencias persistentes de crecimiento entre países y de las migraciones de trabajadores cualificados hacia los países con más capital humano. En política, toda mejora permanente de B eleva permanentemente el crecimiento, y con la externalidad de Lucas la subvención a la educación acerca u a su óptimo.

2.3. Dinámica

Punto de silla: en el plano de la ratio de capitales, ω, y la ratio consumo-capital, $\chi\equiv c/K$, hay un único estado estacionario, un punto de silla, y ω es predeterminada y χ, con u, es de salto. Con $\theta>\alpha$, el caso relevante:

1. $\omega<\omega^*$, tras un conflicto que arrasa las infraestructuras pero no el conocimiento: el capital físico rinde mucho, se produce más, $u>u^*$, y la economía crece transitoriamente por encima de su tasa estacionaria. Así se interpreta la rápida reconstrucción de Alemania y Japón tras la Segunda Guerra Mundial.
2. $\omega>\omega^*$, tras una fuga de cerebros: el capital físico rinde poco y los agentes estudian más, $u<u^*$.

A diferencia del AK hay transición, pero la convergencia es hacia una tasa común y no hacia un nivel, de modo que economías iguales con dotaciones distintas mantienen su diferencia de renta. La economía se sitúa de inmediato sobre la senda estable y converge por ella, mientras K, H y c crecen en el límite a g_Y^*.

Aspectos positivos: distingue capitales, hace de la asignación del tiempo entre producir y aprender la decisión que fija el crecimiento, tiene transición y su solución descentralizada es óptima. Explica por qué economías con la misma tecnología crecen a ritmos distintos si difieren en su sistema educativo o en sus preferencias.

Limitaciones: no predice convergencia, descansa en una educación con rendimientos constantes que la evidencia no confirma y no separa el capital humano que produce del que innova.

3. Innovación tecnológica: I+D con competencia imperfecta

Modelizar la innovación como actividad deliberada exige abandonar la competencia perfecta. Por el teorema de Euler, retribuir los factores rivales por su productividad marginal agota el producto y no deja rentas para la I+D. Además, una idea tiene costes fijos elevados y un coste de reproducción casi nulo, y tiende a producirse con poder de mercado. Los modelos de innovación introducen, por tanto, competencia imperfecta y empresas que invierten en progreso técnico a cambio de rentas de monopolio protegidas por patentes. La innovación es *horizontal* en Romer (1990) y *vertical* en Aghion y Howitt (1992).

3.1. Aumento de variedades (Romer, 1990)

Idea: es el primer modelo en el que la innovación resulta de una inversión deliberada en I+D, y consiste en aumentar el número N de variedades de bienes de capital intermedios.

3.1.1. Supuestos

Hay tres clases de agentes: productores de bienes finales en competencia perfecta, inventores con una patente perpetua que les da el monopolio de su variedad y consumidores a la Ramsey. El análisis es de equilibrio general, con horizonte infinito y población constante.

3.1.2. Modelo

Productores finales: maximizan su beneficio, $Y-wL-\sum_j p_j x_j$, con precios dados, y contratan trabajo y cada intermedio hasta que su productividad marginal iguala su precio, de donde sale la demanda isoelástica de cada intermedio, $x_j^d=L\,(\alpha A/p_j)^{1/(1-\alpha)}$, creciente en A y L y decreciente en p_j. Producen con productividad A, empleo L y x_j unidades de cada intermedio:

$$Y=A\,L^{1-\alpha}\sum_{j=1}^{N}x_j^{\alpha}$$

Inventores: pagan un coste fijo F por inventar y deciden por inducción hacia atrás.

- *Precio*: maximizan $(p_j-1)\,x_j^d(p_j)$, con coste marginal unitario, y ante una demanda isoelástica, de elasticidad $1/(1-\alpha)>1$, fijan un **margen** constante sobre el coste marginal, $p_j=1/\alpha>1$, que maximiza su beneficio. Todas las variedades venden lo mismo y dejan el mismo beneficio.
- *Participación*: con patente perpetua, inventar vale Π/r, y por libre entrada se crean variedades hasta que $\Pi/r=F$. Mientras $\Pi/r>F$, entran inventores, y la mayor demanda de fondos eleva el tipo de interés hasta restaurar la igualdad.

Consumidores: siguen la ecuación de Keynes-Ramsey, $g_c=(1/\theta)(r-\rho)$.

Equilibrio: con $p_j=1/\alpha$, la cantidad de cada intermedio es común y la producción resulta proporcional al número de variedades. Combinando Keynes-Ramsey con la libre entrada se obtiene la tasa común a todas las variables:

$$x_j=A^{1/(1-\alpha)}\alpha^{2/(1-\alpha)}L \qquad Y=A^{1/(1-\alpha)}\alpha^{2\alpha/(1-\alpha)}L\cdot N$$

$$g_Y^*=g_y^*=g_N^*=g_c^*=f(A^+,L^+,\theta^-,F^-,\rho^-,\alpha)$$

La expansión horizontal de variedades evita que caiga la productividad marginal del capital y cumple el papel de la linealidad en el AK.

Estática comparativa: la tasa de crecimiento aumenta con A y con el tamaño de la economía, L, y disminuye con F, con ρ y, si es positiva, con θ.

Dinámica: no hay transición, y todas las variables avanzan desde el inicio al mismo ritmo, como en el AK.

Extensiones: en la formulación original, inventar exige trabajo dedicado a la investigación, $L=L_Y+L_A$, y las variedades nacen de la **función de producción de ideas**, lineal en N porque cada investigador se apoya en todas las ideas existentes (*standing on the shoulders of giants*):

$$\dot{N}=\nu L_A N\;\Rightarrow\; g_N=\nu L_A$$

El trabajo se asigna por arbitraje salarial: el salario en investigación, el valor de los diseños que produce un investigador, iguala al de producción.

El salario en investigación es $w_A = \nu N P_N$, con $P_N = \Pi/r$, y el de producción, $w_Y = (1-\alpha)Y/L_Y$. Con utilidad logarítmica:

$$L_A^* = \frac{L - \Lambda\rho/\nu}{1+\Lambda} \qquad \Lambda \equiv \frac{1}{\alpha}$$
$$g_Y^* = \nu L_A^* = \frac{\nu L - \Lambda\rho}{1+\Lambda}$$

El crecimiento es proporcional al trabajo en investigación y aumenta con L (la forma explícita del efecto escala). Ambas versiones comparten la estructura de tres sectores y la producción proporcional a N.

3.1.3. Análisis normativo

El poder de mercado encarece los intermedios, los productores finales usan pocos y se investiga y crece por debajo del óptimo. De menor a mayor eficacia:

1. *Solución de mercado*: el menor crecimiento.
2. *Subvención a la I+D*: acelera la aparición de variedades, pero deja intacta la distorsión de las cantidades.
3. *Subvención a la compra de intermedios*: baja el precio neto al competitivo y, con coste fijo, reproduce la asignación del planificador. Con sector de I+D subsiste la externalidad del conocimiento sobre la investigación futura.

3.1.4. Conclusiones

Aspectos positivos: trata la innovación como inversión deliberada guiada por incentivos e integra en equilibrio general la competencia imperfecta, las patentes y la no rivalidad de las ideas.

Limitaciones: predice un efecto escala que los datos refutan, y no tiene convergencia ni transición.

3.2. Enfoque schumpeteriano

Origen: Aghion y Howitt (1992), que compartieron con Mokyr el Premio Nobel de Economía de 2025, formalizaron, junto con Grossman y Helpman (1991), la **destrucción creativa** de Schumpeter.

Idea: la innovación es vertical. La calidad mejora escalón a escalón en una *escalera de calidad*, y cada innovación vuelve obsoleta la anterior y destruye las rentas del incumbente. Crea valor y destruye valor, y del balance depende que se innove más o menos de lo óptimo.

3.2.1. Supuestos

Economía cerrada, población constante L y agentes neutrales al riesgo que descuentan al tipo r. Los productores del bien final compiten perfectamente, el del intermedio es monopolista mientras posee la mejor calidad, y los investigadores compiten por la siguiente innovación. Con $Y = A_t x^\alpha$, cada innovación multiplica la calidad, $A_{t+1} = \gamma A_t$, con $\gamma > 1$. Las innovaciones llegan según un proceso de Poisson de intensidad $\lambda_I L_A$, y el trabajo se reparte entre producir el intermedio e investigar, $L = x + L_A$. El tiempo hasta la siguiente innovación es aleatorio, con media $1/(\lambda_I L_A)$.

3.2.2. Modelo

Monopolista del intermedio: el problema se resuelve en dos pasos. En el primero, el monopolista de la generación t elige la cantidad de intermedio que maximiza su beneficio dado el salario, con la productividad marginal del sector final como demanda inversa:

$$\max_x \ \Pi_t = p_t(x)\,x - w_t x, \qquad p_t(x) = \alpha A_t x^{\alpha-1}$$

De su CPO, $\alpha^2 A_t x^{\alpha-1} = w_t$, la cantidad y el beneficio decrecen con el salario en unidades de productividad, $\tilde{w}_t \equiv w_t/A_t$. El beneficio es estrictamente cóncavo en x.

Investigadores: después, deciden cuánto trabajo dedicar a investigar comparando su valor esperado con el salario. El **arbitraje** iguala el salario al valor esperado de investigar, con las rentas descontadas al tipo de interés más la tasa de llegada de la innovación que las destruirá. La condición es de libre entrada y tiene solución única en el estado estacionario:

$$w_t = \lambda_I V_{t+1}, \qquad V_{t+1} = \frac{\Pi_{t+1}}{r + \lambda_I L_{A,t+1}}$$

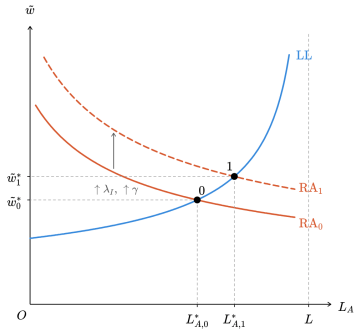
Equilibrio: en el estado estacionario, L_A y $\tilde{w}$ son constantes. La curva de arbitraje, RA, es decreciente, porque más investigación acorta la vida esperada de cada monopolio y reduce el salario que la investigación puede pagar. La del mercado de trabajo, LL, $L = \tilde{x}(\tilde{w}) + L_A$, es creciente, porque un salario mayor libera trabajo para investigar. Su corte fija el esfuerzo investigador, L_A^*, y la tasa media de crecimiento, frecuencia por tamaño de las innovaciones:

$$\text{RA: } \tilde{w} = \frac{\lambda_I \gamma \tilde{\Pi}(\tilde{w})}{r + \lambda_I L_A} \qquad \text{LL: } L = \tilde{x}(\tilde{w}) + L_A$$
$$g_Y^* = \lambda_I L_A^* \ln\gamma$$

El crecimiento es estocástico, a saltos de tamaño $\ln\gamma$ en instantes aleatorios, con una tasa media constante.

Estática comparativa: el crecimiento aumenta con λ_I, γ y L, efecto escala, y disminuye con r y con la competencia de los rivales, que reduce el valor de innovar. Un aumento de λ_I o de γ desplaza hacia arriba la RA y eleva L_A^*, uno de L desplaza la LL hacia la derecha y libera trabajo para investigar, y uno de r desplaza la RA hacia abajo.

1. *Educación superior*: eleva λ_I y la capacidad de absorción.
2. *Propiedad intelectual*: prolonga la vida esperada del monopolio y eleva V.
3. *Competencia*: la evidencia para el Reino Unido muestra una U invertida, en la que la competencia moderada estimula la innovación para escapar de ella, aunque otras muestras no siempre lo confirman.



Modelo de Aghion y Howitt (1992).

Extensiones: tres externalidades hacen ambiguo si se innova de más o de menos, a diferencia de Romer:

- *Desbordamiento intertemporal*: cada innovación eleva la base de las siguientes; subinversión.
- *Business stealing*: el innovador no internaliza las rentas que destruye; sobreinversión.
- *Apropiabilidad incompleta*: no captura todo el excedente del consumidor; subinversión.

Grossman y Helpman (1991) obtienen resultados análogos con escaleras de calidad en un continuo de bienes. Subvencionar la I+D solo es deseable si dominan las dos de subinversión. La política de competencia debe buscar el tramo intermedio de la U, y la protección de patentes tiene un óptimo interior.

3.2.3. Conclusiones

Aspectos positivos: describe la innovación como se observa, con mejoras de calidad que desplazan productos, rotación empresarial y obsolescencia, y da predicciones contrastables sobre estructura de mercado e innovación.

Limitaciones: su normativa es ambigua y, en las versiones básicas, comparte el efecto escala y la falta de convergencia.

3.3. Crítica del efecto escala (Jones, 1995)

Idea: Charles Jones (1995) observa que en Romer y en Aghion-Howitt duplicar los investigadores duplica la tasa de crecimiento, pero los investigadores se han multiplicado desde 1950 y el crecimiento de la productividad no ha aumentado.

Razonamiento: Jones supone rendimientos decrecientes en la producción de ideas, que son **cada vez más difíciles de encontrar**:

$$\dot{N} = \nu L_A N^\phi, \ \ \phi < 1 \ \Rightarrow \ g_y = \frac{n}{1-\phi}$$

El crecimiento a largo plazo depende solo del crecimiento de la población, exógeno, y la fracción de recursos dedicada a investigar solo afecta al nivel. El modelo es **semiendógeno**: la tasa se determina dentro del modelo, pero no responde a la política, como en Solow.

Extensiones: los modelos schumpeterianos de segunda generación eliminan el efecto escala suponiendo que las variedades crecen con la población y diluyen la investigación por producto, y preservan así una tasa sensible a la política. El debate decide si la I+D eleva la tasa de crecimiento o solo el nivel de renta.

Aspectos positivos: concilia los modelos de I+D con un crecimiento estable pese al mayor esfuerzo investigador y conserva la acumulación de ideas no rivales. La hipótesis de ideas más difíciles de encontrar tiene apoyo empírico.

Limitaciones: el crecimiento vuelve a depender de una variable exógena, la población, y si esta deja de crecer, el crecimiento per cápita tiende a desaparecer.

A45. Evidencia empírica sobre el crecimiento económico y la distribución de la renta entre los factores de producción. Contabilidad del crecimiento, con especial referencia a la productividad total de los factores. Convergencia económica internacional

1. Predicciones teóricas y contabilidad del crecimiento

El modelo de Solow-Swan fija las hipótesis que contrasta el análisis empírico, y la contabilidad del crecimiento mide la aportación de los factores y el residuo que el modelo atribuye al progreso técnico, la PTF.

1.1. Predicciones del modelo de Solow-Swan

Punto de partida: el modelo canónico de Solow-Swan (1956), sin imperfecciones, con una Cobb-Douglas con progreso técnico aumentador del trabajo, $Y = K^{\alpha}(AL)^{1-\alpha}$, y una tasa de ahorro exógena y constante. Con $\tilde{k} = K/AL$, la ecuación fundamental es:

$$\dot{\tilde{k}} = s\tilde{k}^{\alpha} - (g_A + n + \delta)\tilde{k}$$

En el estado estacionario, $\dot{\tilde{k}} = 0$, la ratio capital-producto es constante, las variables en unidades de trabajo efectivo no crecen y las per cápita crecen al ritmo del progreso técnico:

$$\frac{K}{Y} = \frac{s}{n + g_A + \delta} \qquad g_{\tilde{k}}^{*} = g_{\tilde{y}}^{*} = 0$$

$$g_k^{*} = g_y^{*} = g_c^{*} = g_A$$

De ahí se siguen cinco predicciones, que son las hipótesis que contrasta el análisis empírico:

1. *Acumulación de capital*: determina el nivel de renta per cápita, pero no su crecimiento a largo plazo, que procede solo del progreso técnico exógeno.
2. *Automatismo*: toda economía alcanza su estado estacionario sin intervención pública, aunque nada garantiza que sea el de la regla de oro.
3. *Alcance de la política económica*: tiene efectos de nivel, pero sobre la tasa de crecimiento solo transitorios, durante la transición entre estados estacionarios.
4. *Exogeneidad del progreso técnico*: con rendimientos constantes y competencia perfecta, por el teorema de Euler, pagar a cada factor su productividad marginal agota la producción y no queda nada para costear el progreso técnico, que empíricamente se mide como residuo.
5. *Convergencia condicional*: cada país se aproxima a su propio estado estacionario. El modelo no predice que los pobres crezcan siempre más que los ricos, es decir, la *convergencia absoluta*.

1.2. Contabilidad del crecimiento

Definición: la **contabilidad del crecimiento** mide qué parte del aumento del PIB procede de una mayor dotación de factores y qué parte del progreso técnico. Admite un enfoque primal, con cantidades, y otro dual, con precios.

1.2.1. Enfoque primal

Versión con capital físico (Solow, 1957): con $Y = Af(K, L)$ y rendimientos constantes, A es la **productividad total de los factores** (PTF), neutral en el sentido de Hicks. Tomando logaritmos y derivando respecto al tiempo, el residuo se obtiene como diferencia:

$$g_{\text{PTF}} = g_Y - \alpha g_K - (1-\alpha) g_L$$

Bajo competencia perfecta, α y $1-\alpha$ son las participaciones de las rentas del capital y del trabajo, observables en la contabilidad nacional (en las economías avanzadas, α ronda un tercio). Formalmente, $\alpha = Af'_K K/Y$ y $1-\alpha = Af'_L L/Y$, y con una Cobb-Douglas g_{PTF} es $1-\alpha$ veces la tasa del progreso técnico aumentador del trabajo.

Como es un residuo, la PTF es "una especie de medida de nuestra ignorancia" (Abramovitz, 1956). Recoge el progreso técnico, pero también factores organizativos, institucionales, culturales y errores de medición. Con datos de EE. UU. de 1909 a 1949, Solow (1957) atribuyó al residuo el 87,5 % del crecimiento del producto por hora trabajada.

Versión con capital humano (Jorgenson y Griliches, 1967): sustituye el trabajo bruto por el capital humano, $H = hL$, que ajusta el trabajo por su calidad. Al imputar a los factores la mejora de la calidad del trabajo, reduce la contribución estimada de la PTF. La participación del capital humano se aproxima por el peso de la masa salarial, y el residuo pasa a ser:

$$g_{\text{PTF}} = g_Y - \alpha g_K - (1-\alpha) g_H$$

1.2.2. Enfoque dual

Idea: parte de la identidad de la renta con rendimientos constantes, $Y \equiv rK + wH$, y expresa el residuo con las tasas de crecimiento de los precios de los factores, el rendimiento del capital, g_r, y el salario, g_w:

$$g_{\text{PTF}} = \alpha g_r + (1-\alpha) g_w$$

Intuición: con rendimientos decrecientes, acumular un factor abarata su remuneración. Si solo se acumulara capital, r caería y w subiría, y si solo se acumulara capital humano, ocurriría lo contrario. Por eso, una subida duradera y simultánea de ambos precios delata progreso técnico. Su ventaja práctica es que los precios de los factores se miden mejor que el *stock* de capital.

Discrepancias: con buena medición y rendimientos constantes, ambos enfoques coincidirían. En el "milagro" de los tigres asiáticos difieren radicalmente, y la causa se ha buscado en la medición del capital y en los impuestos sobre el capital.

1.2.3. Problemas de medición

Capital: no se observa. Se construye por el **inventario permanente**, $K_t = (1-\delta)K_{t-1} + I_t$, y depende de la depreciación supuesta y del *stock* de partida. La mejora de calidad de los bienes de capital exige deflactores hedónicos. Su heterogeneidad complica la agregación, problema que alimentó la controversia de Cambridge (Joan Robinson). En la práctica se usan el índice de volumen de servicios de capital, ponderado con el coste de arrendamiento, y el índice capital-riqueza, ponderado con el de adquisición. Además, el servicio del capital varía con la utilización, y la PTF medida es procíclica aunque la tecnología no lo sea. El peso del *stock* de partida se pierde con el tiempo, y deflactar o no con índices hedónicos, porque un ordenador de hoy no es el de 1990, cambia sustancialmente la contribución del capital.

Trabajo: las horas son la medida idónea, aunque mezclan participación laboral, paro y jornada. El capital humano se aproxima por los años de escolarización, que ignoran la calidad y la experiencia, o por la masa salarial, que confunde calidad con poder de negociación. La calidad de la educación, además, difiere entre países.

Participaciones: exigen repartir la renta mixta de los autónomos y suponer competencia perfecta. Cada decisión desplaza crecimiento entre factores y residuo, y por eso las estimaciones de la PTF varían tanto.

1.2.4. Contabilidad del desarrollo

Idea: Robert Hall y Charles Jones (1999), con la síntesis de Francesco Caselli (2005), aplican la misma lógica a las diferencias de renta entre países en un momento dado: qué parte de la brecha se debe a tener más capital físico y humano por trabajador y qué parte a usarlos con más productividad. La contabilidad del crecimiento, en cambio, descompone la variación de la renta de un país en el tiempo.

Razonamiento: con $Y = K^{\alpha}(AH)^{1-\alpha}$ y h obtenido de los años de escolarización con rendimientos mincerianos:

$$y = \left(\frac{K}{Y}\right)^{\alpha/(1-\alpha)} h\, A$$

Se usa la ratio capital-producto porque en el estado estacionario es proporcional a la tasa de inversión y no depende de A, y así no se atribuye al capital lo que es consecuencia de la productividad. En 127 países, la ratio de renta por trabajador entre los cinco más ricos y los cinco más pobres, 31,7, se descompone en un factor 1,8 de capital físico, 2,2 de capital humano y **8,3 de productividad**. Caselli resume el reparto con la medida de éxito de los factores, donde y_{KH} es la renta que se observaría si todos los países tuvieran la misma A:

$$\text{success} = \frac{\text{var}(\ln y_{KH})}{\text{var}(\ln y)} \approx 0{,}4$$

Los factores explican, por tanto, el 40 % de la dispersión de la renta y la PTF, el 60 %.

Conclusiones: desplaza la pregunta del crecimiento hacia la productividad y sus determinantes últimos, que son las instituciones, la tecnología y la asignación de recursos entre empresas. Pero la PTF sigue siendo un residuo, y el reparto es sensible a la elasticidad de sustitución y a la medición del capital humano.

1.2.5. Enfoque econométrico

Idea: la contabilidad impone las participaciones observadas como elasticidades, lo que exige competencia perfecta y rendimientos constantes. El enfoque econométrico, que se remonta a Cobb y Douglas (1928) y a Arrow, Chenery, Minhas y Solow (1961), estima la función de producción directamente, de modo que las elasticidades y los rendimientos a escala son resultados y no supuestos, y el progreso técnico se identifica como tendencia:

$$\ln Y_t = \ln A_0 + g_{\text{PTF}} t + \alpha \ln K_t + \alpha_L \ln L_t + \varepsilon_t$$

Razonamiento: permite contrastar $\alpha + \alpha_L = 1$. Una elasticidad del capital superior a su participación revela rendimientos crecientes, externalidades o poder de mercado (Hall, 1988), porque con un margen sobre el coste marginal las participaciones infravaloran las elasticidades. Con datos de empresa es hoy la vía principal para medir la mala asignación de recursos (*misallocation*, Hsieh y Klenow, 2009), que explica una parte sustancial de las diferencias de PTF entre países.

Limitaciones: la endogeneidad de los factores, que las empresas eligen conociendo su productividad y que se corrige con estimadores de función de control; la colinealidad del capital con la tendencia, y la sensibilidad a la forma funcional: Cobb-Douglas, CES o translog.

1.2.6. Conclusiones

Aspectos positivos: es sencilla, transparente y poco exigente en información, da descomposiciones comparables y acota dónde buscar las fuentes del crecimiento.

Limitaciones: hereda los problemas de medición, omite posibles factores (tierra, recursos naturales, capital público u organizativo) e identifica las fuentes *próximas*, no las causas *fundamentales*: constatar que un país acumuló capital no explica por qué lo hizo. Esta limitación motiva los estudios de determinantes profundos.

En suma, la PTF se obtiene como residuo y hereda los errores de medición. La contabilidad del desarrollo le atribuye la mayor parte de la brecha entre países, pero ninguna de estas vías identifica las causas fundamentales.

2. Herramientas: convergencia y distribución

El análisis empírico se completa con dos herramientas: el análisis de convergencia y el de la distribución de la renta.

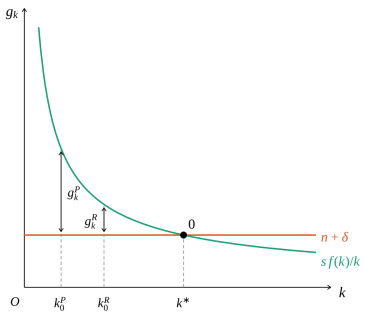
2.1. Análisis de convergencia

Origen: nació para enfrentar a los modelos neoclásicos, que con rendimientos decrecientes predicen convergencia condicional, con los endógenos tipo AK, que con rendimientos constantes del capital ampliado no predicen ninguna.

Desarrollo gráfico: en el plano k-g_k, la curva de ahorro por unidad de capital, $sf(k)/k$, es decreciente y la recta $n+\delta$, horizontal. Su corte es el estado estacionario y la distancia entre ambas, la tasa de crecimiento. Con tasas de ahorro distintas hay una curva por país, y la tasa de crecimiento del capital per cápita es:

$$g_k = \frac{sf(k)}{k} - (n+\delta)$$

1. *Parámetros iguales* (s, n, δ, α y el nivel tecnológico): mismas curvas y mismo estado estacionario. Cuanto menor es el capital inicial, mayor el crecimiento, y hay **β-convergencia absoluta**. Así ocurrió tras la Segunda Guerra Mundial con Alemania y Japón, que diferían de los aliados sobre todo en su capital destruido y crecieron sistemáticamente más rápido.



Convergencia absoluta: con los mismos parámetros, la economía pobre crece más deprisa.

2. *Parámetros distintos*: basta que difiera uno, como la tasa de ahorro, para que los estados estacionarios se separen y el rico, lejos del suyo, pueda crecer más que el pobre. La hipótesis relevante es la **convergencia condicional**: una correlación parcial negativa entre crecimiento y renta inicial, controlando por el estado estacionario.

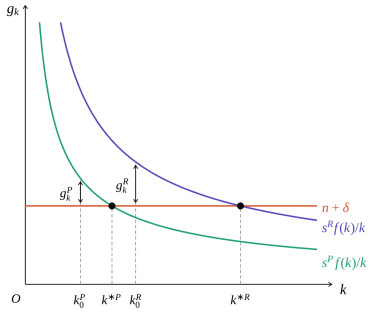
β-convergencia: relación inversa entre renta inicial y crecimiento posterior, con Z^j las variables que aproximan el estado estacionario:

$$g_y^j = \beta_0 - \beta \ln y_0^j + \gamma Z^j + \varepsilon^j$$

Hay convergencia si $\beta > 0$, y menor que $1/T$ para evitar adelantamientos sistemáticos. Sin controles se contrasta la absoluta y con controles, la condicional. La versión no lineal, $\beta = (1-e^{-\lambda T})/T$, da directamente la velocidad, λ, que mide el ritmo al que una economía cierra la brecha con su estado estacionario.

σ-convergencia: la dispersión de la renta per cápita disminuye, $\sigma_t > \sigma_{t+T}$. La β-convergencia absoluta es necesaria, pero no suficiente, porque un pobre puede adelantar al rico y seguir creciendo más, y la dispersión aumentar.

γ-convergencia: mide la movilidad en la clasificación de rentas con el índice de concordancia de Kendall. Si los países cambian sus posiciones relativas a lo largo del tiempo, la jerarquía internacional de rentas no es fija.



Convergencia condicional: cada economía converge a su propio estado estacionario.

Velocidad de convergencia: en Solow-Swan, la aproximación log-lineal en torno al estado estacionario da la velocidad y el plazo en que se cierra la mitad de la brecha:

$$\lambda \approx (1-\alpha)(g_A + n + \delta) \qquad t_{1/2} = \frac{\ln 2}{\lambda}$$

Con $\alpha \approx 1/3$, el modelo básico da un 4 % anual, muy por encima del 2 % que miden los estudios (Barro y Sala-i-Martin, 1992a). Con capital humano (Mankiw, Romer y Weil, 1992), la participación conjunta del capital físico y humano ronda dos tercios y la velocidad baja al 2 %. A ese ritmo, cerrar la mitad de la distancia lleva unos 35 años, y la convergencia es un proceso lento.

Problemas econométricos:

- *Heterogeneidad no observada*: los determinantes omitidos sesgan β. Con panel y efectos fijos la velocidad supera sensiblemente el 2 % (Islam, 1995).
- *Endogeneidad*: la inversión o la escolarización responden al crecimiento y exigen variables instrumentales.
- *Falacia de Galton* (Quah, 1993): un β positivo es compatible con una dispersión constante o creciente, por regresión a la media. Hay que completar con la σ-convergencia y la dinámica de toda la distribución.

2.2. Distribución personal de la renta

El mismo crecimiento agregado es compatible con evoluciones distributivas muy distintas. Destacan cinco instrumentos:

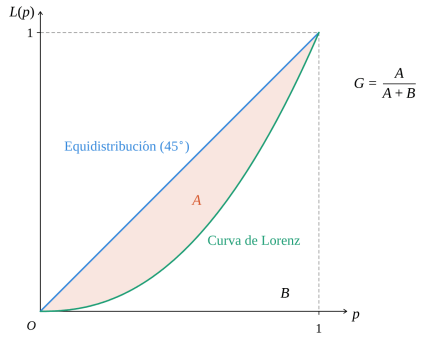
1. *Curva de Lorenz* (1905): proporción acumulada de la renta que percibe cada proporción acumulada de la población, ordenada de menor a mayor renta. La diagonal de 45 grados representa la equidistribución perfecta, y cuanto más se aleja de ella la curva, más desigualdad. Solo da una ordenación parcial: si dos curvas se cruzan, no ordena.
2. *Índice de Gini* (1912): cociente entre el área entre la curva y la diagonal y el área bajo la diagonal, $G = A/(A+B)$, entre 0 y 1. Es el más usado, pero dos distribuciones distintas pueden dar el mismo Gini, y es más sensible al centro que a las colas. Vale 0 con equidistribución y 1 si un individuo concentra toda la renta.
3. *Índice de Theil* (1967): basado en la entropía. Su ventaja es la **descomponibilidad aditiva** en desigualdad dentro de los grupos y entre grupos. Con H individuos, y^h la renta de cada uno e $\bar{y}$ la media, $T = \frac{1}{H}\sum_{h=1}^{H} \frac{y^h}{\bar{y}} \ln \frac{y^h}{\bar{y}}$, y la descomposición, p. ej. entre regiones o niveles educativos, sirve para analizar las fuentes de la desigualdad.
4. *Índice de Atkinson* (1970): normativo, con un parámetro de aversión a la desigualdad, ε, y $A_\varepsilon = 1 - y_e/\bar{y}$ mide la fracción de la renta que la sociedad sacrificaría a cambio de la equidistribución. Aquí y_e es la renta "equivalente igualmente distribuida", la que, repartida por igual, daría el mismo bienestar social que la distribución observada.
5. *Cuentas nacionales distributivas* (Piketty et al., 2018): asignan el 100 % de la renta nacional a los individuos, frente a las encuestas, que infrarrepresentan las rentas altas. Garantizan la coherencia entre agregados y datos distributivos, permiten analizar la distribución antes y después de impuestos y transferencias, y son la base del *World Inequality Report*.

Desarrollo gráfico: en el plano p-$L(p)$, la curva de Lorenz es convexa y queda por debajo de la diagonal de equidistribución. El Gini es el área A entre ambas sobre el área total bajo la diagonal, $A+B$. La curva es convexa porque la población se ordena de menor a mayor renta.

En suma, la β-convergencia debe completarse con la σ-convergencia, y las medidas de desigualdad van de la ordenación parcial de Lorenz al juicio normativo de Atkinson.

3. Evidencia empírica

La evidencia se ordena en cinco ámbitos: el crecimiento y sus fuentes, la convergencia, la distribución personal, la distribución factorial y la transformación estructural.



Curva de Lorenz y coeficiente de Gini.

3.1. Evidencia sobre el crecimiento

3.1.1. Hechos estilizados

Kaldor (1961) resumió la evidencia en seis regularidades, entendidas como promedios históricos y no como constantes válidas en todo momento:

1. *Renta per cápita*: crece a una tasa aproximadamente constante, en torno al 2 % anual en EE. UU.
2. *Capital per cápita*: también crece a una tasa aproximadamente constante.
3. *Ratio capital-producto*: aproximadamente constante, lo que se deriva de las dos primeras si ambas tasas coinciden.
4. *Tasa de retorno del capital*: aproximadamente constante.
5. *Diferencias entre países*: el crecimiento de la productividad difiere de forma importante y persistente, del 2 % al 5 %.
6. *Participaciones factoriales*: aproximadamente constantes.

Todas salvo la quinta son coherentes con el modelo neoclásico con progreso técnico exógeno y Cobb-Douglas, y contribuyeron a fijarlo como canónico. La quinta solo encaja como transición. La sexta es hoy la más discutida, porque la participación del trabajo ha caído desde los años ochenta, y con más intensidad desde comienzos de la década de 2000.

Jones y Romer (2010) proponen seis "nuevos hechos de Kaldor": la extensión del mercado; la aceleración, durante milenios, del crecimiento de la población y de la renta per cápita; la mayor variabilidad de las tasas lejos de la frontera; el papel de la PTF, que explica más de la mitad de las diferencias de renta; el fuerte aumento del capital humano por trabajador, y un premio salarial de la cualificación que no ha caído. El último sugiere una demanda de cualificación sesgada por la tecnología. Los tres primeros motivan los modelos de crecimiento basados en ideas, y el cuarto, la contabilidad del desarrollo. La extensión del mercado refleja la globalización y la urbanización.

3.1.2. PTF frente a acumulación de factores

Dominan los factores: Denison (1985), en EE. UU. entre 1929 y 1982; Maddison (1991), en torno a dos tercios en seis países avanzados entre 1913 y 1987, y Mankiw, Romer y Weil (1992), con un R^2 de 0,6 que sube a 0,8 con capital humano.

Jorgenson y Griliches (1967): el residuo es en gran parte mala medición. Corrigiendo la utilización y la calidad del capital y la composición del trabajo, según educación, edad y sexo, la contribución de la PTF en EE. UU. (1945-1965) baja de casi la mitad a un 3 %. La réplica de Denison y el debate posterior fijaron la posición intermedia que hoy se acepta, según la cual una parte sustancial del residuo es medición, pero no toda.

Tigres asiáticos: Alwyn Young (1995) mostró que el crecimiento de Hong Kong, Singapur, Corea y Taiwán entre 1966 y 1990 se debía sobre todo a la acumulación de factores, con una PTF casi nula en Singapur. Krugman (1994) lo comparó con el crecimiento soviético y predijo su agotamiento. El enfoque dual dio una PTF muy superior, que sugería que las series de capital de Young sobrestimaban la acumulación, pero Young (1998) sostiene que, corregidos los errores, la diferencia es menor. El episodio ilustra por qué los dos enfoques son complementarios y la pregunta "acumulación o PTF" sigue abierta.

Domina la PTF: Solow (1957), Hall y Jones (1999) y Caselli (2005). La discrepancia de Hall y Jones con Mankiw, Romer y Weil procede de la medición del capital humano. Mankiw, Romer y Weil usan la escolarización secundaria, un flujo de inversión en capital humano, con un efecto muy pronunciado sobre la renta. Hall y Jones convierten los años de escolarización, un *stock*, con la ecuación de Mincer, y obtienen diferencias mucho menores. Caselli halla además una parte de los factores mayor dentro de la OCDE, en torno al 60 %, que fuera de ella, algo más de un tercio. En las economías europeas que se acercaron a la frontera tras la Segunda Guerra Mundial, la contribución de la PTF fue mayor que en EE. UU., reflejo de un intenso proceso de modernización y absorción de tecnología.

3.1.3. Determinantes profundos

Primeras regresiones: condicionando a la renta inicial, la escolarización y la inversión se correlacionan positivamente con el crecimiento, y el consumo público, la inflación, las distorsiones del precio de la inversión, el crecimiento demográfico y la inestabilidad política, negativamente. El consumo público no debe confundirse con la inversión pública, y la inestabilidad política se aproxima por revoluciones, golpes de Estado y asesinatos políticos.

Límites extremos (Levine y Renelt, 1992): una variable solo es robusta si su coeficiente conserva signo y significatividad con cualquier combinación de controles. Solo sobreviven la renta inicial y la tasa de inversión. El criterio procede de Leamer, y casi ninguna de las más de cincuenta variables propuestas por la literatura es robusta, por lo que no cabe depositar demasiada confianza en las regresiones de crecimiento.

Dos millones de regresiones (Sala-i-Martin, 1997): con la distribución completa de los coeficientes, una veintena de casi sesenta variables resultan robustas, entre ellas geográficas, institucionales y de apertura.

3.1.4. Enfoque institucional

Idea: Acemoglu, Johnson y Robinson (2001) afrontan la endogeneidad, porque los países ricos pueden permitirse buenas instituciones, instrumentando la protección frente a la expropiación, R_i, con la mortalidad de los colonos europeos:

$$\ln y_i = \beta_0 + \beta_1 R_i + \mathbf{X}_i'\gamma + \varepsilon_i$$

Donde la mortalidad de los siglos XVII-XIX era alta, los europeos no se asentaron y crearon instituciones **extractivas**, diseñadas para expoliar recursos, mientras que donde era baja replicaron instituciones **inclusivas**, que protegían la propiedad. Unas y otras persistieron tras la independencia. Las instituciones tienen un efecto causal de primer orden sobre la renta per cápita. La "reversión de la fortuna" (2002) lo refuerza, pues las zonas colonizadas más ricas en 1500 son hoy relativamente más pobres, lo que es incompatible con efectos geográficos invariables y apunta a las instituciones como causa fundamental del desarrollo.

3.1.5. Desaceleración de la productividad

Origen: la *paradoja de la productividad* de Solow (1987) y el debate del *estancamiento secular* (Summers, 2014; Gordon, 2016). En palabras de Solow, "la era del ordenador se ve en todas partes menos en las estadísticas de productividad".

Hechos: la PTF de EE. UU. creció en torno al 2 % entre 1920 y 1970, cayó a un tercio de ese ritmo entre 1973 y 1995, repuntó con las TIC entre 1995 y 2004 y, desde mediados de la década de 2000, ronda el 0,5 %, con tasas próximas a cero en buena parte de Europa. A esta escala, el primer hecho de Kaldor es solo una aproximación.

Explicaciones:

- *Pesimista* (Gordon): las grandes invenciones de la segunda revolución industrial, como la electricidad, el motor de combustión y el saneamiento, tuvieron efectos únicos que las TIC no igualan, y la demografía, la educación, la desigualdad y la deuda frenan el crecimiento tendencial.
- *Medición*: la producción digital está infravalorada, pero no explica la mayor parte de la desaceleración, que es común a países con distinta intensidad digital.
- *Difusión y asignación*: las empresas de la frontera siguen mejorando y el resto se estanca, lo que apunta a fallos de difusión, a la caída del dinamismo empresarial y a la concentración. Además, las ideas son cada vez más difíciles de encontrar (Bloom, Jones, Van Reenen y Webb, 2020), y mantener el crecimiento de la productividad exige un esfuerzo investigador cada vez mayor, lo que favorece los modelos semiendógenos.

La desaceleración es el hecho estilizado que la teoría del crecimiento debe explicar hoy, y sus lecturas tienen implicaciones opuestas. Si es tecnológica y permanente, las políticas de oferta tienen poco recorrido, mientras que, si es de difusión, la I+D, la formación y la competencia pueden revertirla. Las tres hipótesis no se excluyen, además, y descansan en la PTF, un residuo.

3.2. Evidencia sobre convergencia

1. *Sin convergencia mundial*: en la *Penn World Table* (Summers y Heston, 1991) no hay convergencia β ni σ, lo que se leyó como argumento a favor del crecimiento endógeno. Tomado el mundo en su conjunto, los países pobres no crecen sistemáticamente más que los ricos.
2. *Contrarrevolución neoclásica* (1990-1992): el modelo siempre predijo convergencia condicional. Contrastar la absoluta equivale a ignorar que cada país tiene sus propios parámetros y su propio estado estacionario.
   - *Mankiw, Romer y Weil* (1992): sin controles, β es positivo (0,34) y significativo entre los 22 países de la OCDE, por su semejanza estructural, pero en 98 países (1960-1985) es incluso ligeramente negativo. Con escolarización, inversión y crecimiento demográfico

como controles, es positivo y significativo en las tres muestras, las dos anteriores y una intermedia de 75 países, y la convergencia condicional avanza a un 2 % anual, la velocidad coherente con el modelo de Solow con capital humano.
   - *Barro y Sala-i-Martin* (1991-1992): las regiones de un mismo país tienen parámetros muy parecidos, como el ahorro, la demografía, la tecnología, el gobierno y las instituciones, y permiten contrastar la convergencia absoluta casi en condiciones de laboratorio. En 73 regiones europeas, los estados de EE. UU., los *Länder*, las prefecturas japonesas y las comunidades autónomas, todas las muestras dan β-convergencia absoluta a un ritmo cercano al 2 % anual, unos 35 años para reducir la brecha a la mitad. Tanta uniformidad ha hecho de ese 2 % casi un hecho estilizado.
3. *Sesgo de selección y clubes*: la convergencia de Baumol (1986) entre los países hoy ricos era un artefacto de selección, porque la muestra incluía solo a los que acabaron convergiendo. Con una muestra definida *ex ante* por la renta de 1870, desaparece (De Long, 1988). Quah (1993, 1996) muestra que la distribución mundial no se comprime, sino que tiende a los *picos gemelos*, un club de países ricos y otro de pobres con un centro que se vacía. Hay, por tanto, **clubes de convergencia**, porque el estado estacionario depende también de las condiciones iniciales, como el capital humano, algo coherente con las trampas de pobreza.
4. *Desde 2000*: los países en desarrollo crecen sistemáticamente más que los ricos, y la convergencia absoluta, ausente en 1960-1990, tiende a aparecer desde 1990 y es clara desde 2000, a la vez que convergen las instituciones y políticas (Kremer et al., 2022). La no convergencia anterior fue un fenómeno del periodo y no una propiedad del crecimiento.
5. *China*: el mayor episodio de convergencia de la historia. Desde la reforma y apertura de 1978 creció cerca del 10 % anual durante más de tres décadas, con el ingreso en la OMC (2001) como punto de inflexión, y en torno al 7 % en la década de 2010. Su PIB en paridad de poder adquisitivo supera al de EE. UU. desde mediados de esa década, pero su PIB per cápita sigue en torno a un tercio del estadounidense, y a la sexta parte a tipos de cambio de mercado. La convergencia, aunque muy intensa, dista por tanto de haberse completado.

3.3. Desigualdad y crecimiento

Curva de Kuznets (1955): la desigualdad describe una U invertida con el desarrollo. Aumenta cuando la población pasa de una agricultura de baja productividad y renta homogénea a una industria de alta productividad y renta dispersa, y disminuye cuando la mayoría está ya en el sector moderno y operan la educación y la redistribución. La apoyó la evidencia de corte transversal de los años sesenta y setenta, pero las series largas la desmienten. Piketty (2014) atribuye la caída de 1914-1970 a las guerras, la inflación y la fiscalidad, y la desigualdad crece desde 1980. Milanovic (2016) propone *ondas de Kuznets* sucesivas, que responden a las revoluciones tecnológicas y a la globalización.

Curva del elefante (Lakner y Milanovic, 2016): entre 1988 y 2008 ganaron mucho las clases medias emergentes y, algo menos, el 1 % superior, y se estancaron las clases medias y bajas de los países ricos. La desigualdad entre países, ponderada por población, cae por la convergencia de China e India, la interna sube y la global baja ligeramente desde 2000, la primera reducción desde la Revolución Industrial.

De la desigualdad al crecimiento: efecto negativo por economía política (Alesina y Rodrik, 1994; Persson y Tabellini, 1994) y restricciones de crédito (Galor y Zeira, 1993); Barro (2000), negativo en los pobres y positivo en los ricos; otros estudios muestran que la desigualdad acorta los episodios de crecimiento. La evidencia es frágil, pero el consenso es que la desigualdad elevada no favorece el crecimiento. El canal político opera porque una mayor desigualdad eleva la demanda de redistribución distorsionante, y el crediticio, porque impide a los pobres invertir en capital humano. La fragilidad procede de los problemas de medición de la desigualdad y de identificación.

Estado de la desigualdad: según el *Informe sobre la Desigualdad Mundial 2022* (Chancel, Piketty y otros), confirmado en lo esencial por la edición de 2026, el 10 % superior concentra el 52 % del ingreso mundial y el 76 % de la riqueza. La mitad inferior obtiene menos del 15 % del ingreso nacional, salvo en Europa, y apenas el 2 % de la riqueza. Oriente Medio y el Norte de África es la región más desigual, y los niveles se aproximan a los de comienzos del siglo XX. Además, la riqueza privada ha crecido y la pública ha caído durante cuarenta años, tendencia que agravó la pandemia. Durante ella, los gobiernos se endeudaron con el sector privado por magnitudes del orden del 10-20 % del PIB.

Piketty: la ratio riqueza-renta ha aumentado sin que cayera el rendimiento del capital, lo que eleva la participación del capital, producto de ambos (primera ley), y explica el aumento de la desigualdad desde el final de Bretton Woods. A largo plazo, esa ratio tiende al cociente entre ahorro neto y crecimiento (segunda ley), de modo que un menor crecimiento la eleva. Si $r > g_Y$ de forma persistente, la riqueza heredada crece más que la renta y se concentra, alimentada por el valor inmobiliario, los bienes posicionales y la ventaja fiscal del capital sobre el trabajo.

3.4. Participación del trabajo

Idea: la distribución factorial, distinta de la personal, mide qué fracción de la renta nacional remunera al trabajo y cuál al capital. Su estabilidad, en torno a dos tercios para el trabajo, es el sexto hecho de Kaldor, que reproduce la Cobb-Douglas y que la contabilidad del crecimiento usa como ponderación. Desde los años ochenta, sin embargo, la participación del trabajo ha caído en torno a cinco puntos en la OCDE.

Explicaciones: el abaratamiento del capital, sobre todo TIC, con elasticidad de sustitución superior a uno (Karabarbounis y Neiman, 2014); la deslocalización de las fases intensivas en trabajo; las *empresas superestrella* de márgenes altos (Autor y otros, 2020); el menor poder de negociación, y la medición de la renta mixta, la vivienda y la capitalización de la propiedad intelectual, que antes se trataba como gasto intermedio. El FMI y la OIT documentan la caída con datos comparables desde 1990, con la excepción parcial de algunos países europeos.

La caída conecta crecimiento y desigualdad personal. Como la renta del capital está más concentrada, un desplazamiento de la distribución factorial hacia el capital eleva la desigualdad personal, como en el mecanismo de Piketty. Además, obliga a revisar la elasticidad de sustitución unitaria y la competencia perfecta. Las explicaciones son complementarias, pero la evidencia aún no permite ponderarlas.

3.5. Transformación estructural

Idea: el marco unisectorial no captura la reasignación entre sectores (Kuznets, 1966; Chenery y Syrquin, 1975), que opera por la *demanda*, con preferencias no homotéticas, y por la *oferta*, con progreso técnico desigual que, con demandas poco sustituibles, desplaza el empleo a los sectores estancados (**enfermedad de los costes** de Baumol, 1967).

Patrones: la agricultura pierde peso (ley de Engel); las manufacturas describen una U invertida, cuyo máximo se alcanza hoy con menos renta y menos empleo que en los primeros industrializados, la *desindustrialización prematura* (Rodrik, 2016); los servicios ganan peso, y el sector público crece con la renta (ley de Wagner). La ley de Engel refleja una elasticidad-renta de la demanda de alimentos inferior a la unidad, y el canal de la demanda exige utilidades no homotéticas, en las que los consumos relativos dependen de la renta.

La transformación estructural es causa y consecuencia del crecimiento. La reasignación del trabajo desde la agricultura hacia sectores más productivos es una fuente de crecimiento. El cambio estructural explica buena parte del crecimiento asiático y ha frenado el de América Latina y África, donde el trabajo se desplazó a servicios informales de baja productividad (McMillan y Rodrik, 2011). A la vez, la enfermedad de Baumol reduce el crecimiento agregado a medida que los servicios estancados absorben el empleo, lo que enlaza con la desaceleración.

Aspectos positivos: son regularidades sólidas, comunes a países y periodos, que el marco unisectorial no captura. Explican, además, parte de las diferencias de crecimiento entre regiones.

Limitaciones: son correlaciones que no separan el canal de la demanda del de la oferta, y la productividad de los servicios es difícil de medir, lo que condiciona la lectura de la enfermedad de los costes.