ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/256278026

Métodos Estadísticos para Economía y Empresa

Book · October 2011

CITATIONS
0

READS
1,117

2 authors:

Ana Jesus Lopez-Menendez
University of Oviedo
133 PUBLICATIONS **655** CITATIONS
SEE PROFILE

Rigoberto Perez Suarez
University of Oviedo
96 PUBLICATIONS **243** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Monitoring Sustainable Development View project

Economic Forecasting View project

All content following this page was uploaded by Ana Jesus Lopez-Menendez on 25 February 2017.

The user has requested enhancement of the downloaded file.

Métodos estadísticos para Economía y Empresa

Rigoberto Pérez y Ana Jesús López
rigo@uniovi.es , *anaj@uniovi.es*

Octubre 2011

A nuestras familias y amigos

ISBN13 978-84-694-9009-9
Depósito Legal: AS-04398-2011
Edición 2011
Revisión V.1.0


This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 3.0 Unported License.

Breve reseña de autores

Los autores de este libro son profesores del Departamento de Economía Aplicada de la Universidad de Oviedo (Unidad de Estadística y Econometría).

`https://sites.google.com/a/uniovi.es/libros/meee`

Rigoberto Pérez Suárez es Catedrático de Universidad y su amplia experiencia docente incluye asignaturas de Estadística Econometría y Series temporales tanto en primer y segundo ciclo como en doctorados y másteres. Es autor de varios libros de texto (Nociones Básicas de Estadística, Análisis de datos económicos I: Métodos descriptivos, Análisis de datos económicos II: Métodos inferenciales) y del software docente ADE+, así como de numerosas publicaciones relativas a la innovación educativa y el e-learning.
También ha sido Director de Area de Innovación de la Universidad de Oviedo (2000-2006) y Director del Campus Virtual Compartido del grupo G9 (2004-2006).
En el ámbito investigador es autor de diversas publicaciones en revistas de impacto y ha dirigido numerosas tesis doctorales y proyectos de investigación, generalmente referidos a la predicción económica y al análisis de la desigualdad.

Ana Jesús López Menéndez es Profesora Titular de Universidad y su docencia abarca asignaturas de Estadística, Econometría y Series temporales, tanto en la Universidad de Oviedo como el el Campus Virtual Compartido G9. También ha sido profesora visitante en universidades de Reino Unido, Hungría y Cuba. Es autora de los manuales Análisis de datos económicos I: Métodos descriptivos y Análisis de datos económicos II: Métodos inferenciales, así como de numerosas publicaciones relativas a la innovación educativa y el e-learning.
En el ámbito investigador es autora de diversos artículos publicados en revistas de impacto, ha dirigido seis tesis doctorales y ha participado en numerosos proyectos de investigación.

Índice general

PRESENTACIÓN

La información económica forma parte de nuestra realidad cotidiana y afecta a nuestras vidas. Estadísticas como el Indice de Precios de Consumo (IPC), la tasa de paro o los índices bursátiles son referencias habituales en los medios de comunicación, por lo que resulta imprescindible conocer su significado y dominar las técnicas estadísticas necesarias para su correcta utilización.

Hace aproximadamente un año, con motivo de la celebración del primer Día Mundial de la Estadística (20-10-2010), la declaración institucional de Naciones Unidas destacaba la importancia de las estadísticas como herramienta para el desarrollo económico y social, y su trascendente papel en la adopción de decisiones gubernamentales, empresariales y personales en una sociedad moderna.

Convencidos de la importancia de la Estadística, presentamos ahora este texto, cuyo antecedente es el manual *Análisis de datos económicos II- Métodos inferenciales* publicado en 1997 por Ediciones Pirámide y actualmente descatalogado. Nuestro objetivo al elaborar *Métodos estadísticos para Economía y Empresa* es contribuir a la difusión de las técnicas estadísticas, animados por nuestras experiencias previas, los comentarios de nuestros colegas universitarios y las posibilidades que ofrecen las nuevas tecnologías para la elaboración de un manual digital y su difusión a través de la Red.

Este libro se estructura en un total 10 capítulos agrupados en tres partes, dedicadas respectivamente a Probabilidad (capítulos 1 a 4), Inferencia Estadística (capítulos 5 a 9) e Introducción a la Econometría (capítulos 9 y 10) y consideramos que sus contenidos pueden ser de utilidad tanto para estudiantes universitarios de diversos grados del ámbito de las Ciencias Sociales (Administración y Dirección de Empresas, Economía, Contabilidad y Finanzas, Relaciones Laborales y Recursos Humanos, Comercio, ...) como para profesionales interesados en las técnicas inferenciales de aplicación habitual en el contexto socioeconómico.

Con este ánimo, el manual *Métodos estadísticos para Economía y Empresa* estará a partir de ahora disponible en la Red en formato pdf, de forma libre y gratuita, accesible bajo licencia Creative Commons en el sitio web:

`https://sites.google.com/a/uniovi.es/libros/MEEE`

Gracias a todos los que, de un modo u otro, nos han acompañado en el camino que ha conducido a este libro. Confiamos en que sus contenidos resulten de utilidad y agradecemos de antemano cualquier comentario o sugerencia.

Parte I.

Probabilidad

1. Incertidumbre y probabilidad

La probabilidad forma parte de nuestros esquemas habituales de razonamiento, proporcionando un instrumento en el que a veces incluso inconscientemente nos apoyamos para emitir opiniones o tomar decisiones.

En efecto, vivimos en un mundo incierto en el que debemos conformarnos con cuantificar esa incertidumbre, habitualmente en términos de probabilidad, conociendo así el grado de creencia en nuestros resultados y conclusiones.

La probabilidad es el pilar básico en el que descansa todo el proceso inductivo. De ahí la importancia de abordar su estudio desde varias ópticas distintas: el concepto y significado de la probabilidad, su cuantificación numérica y la axiomática de la probabilidad, marco formal que posibilita una modelización matemática de los fenómenos aleatorios.

Cualquiera de los aspectos señalados puede resultar de gran trascendencia, y de hecho existe una bibliografía muy extensa sobre cada uno de ellos. Sin embargo, en nuestros estudios la probabilidad tiene un carácter instrumental y no constituye un fin en sí misma. Por ello, aun reconociendo la conveniencia de reflexionar sobre el significado de la probabilidad, prestaremos aquí una atención preferente a las reglas de funcionamiento, las posibilidades y los riesgos de esta poderosa herramienta, que acompañará como medida de credibilidad a nuestras conclusiones.

El origen de la probabilidad no es claro aunque los juegos de azar se practicaban desde muy antiguo y las leyes de la combinatoria elemental, imprescindibles para la cuantificación de probabilidades, eran conocidas por los árabes y los matemáticos del Renacimiento pero más como una rama del álgebra que en su contexto actual. En las obras de N.F. Tartaglia (1499-1557) y Galileo Galilei (1564-1642) se recogen problemas de probabilidad y combinatoria relacionados con juegos de azar y existe una abundante correspondencia entre B. Pascal (1623-1662) y P. Fermat (1601-1665) en la que, mediante el estudio de juegos de azar, ambos matemáticos sientan la base de los fundamentos de la probabilidad. El primer tratado sobre probabilidades publicado corresponde a Christian Huygens (1654-1705) con *On reasoning in Games of Chance*, obra que sirvió de estímulo a James Bernoulli, autor del texto *Ars Conjectandi*, publicado en 1705 y de clara influencia en todos los trabajos posteriores.

1.1. Definiciones de probabilidad

Los "usuarios" de la probabilidad no necesitan conocer con exactitud el concepto al que responde este término, del mismo modo que para ser un buen jugador de ajedrez o un excelente conductor no es necesario conocer la "filosofía" implícita en estas actividades, sino únicamente sus reglas de funcionamiento.

De hecho, a lo largo de su desarrollo histórico se ha generado una gran controversia no solucionada sobre el significado de la probabilidad.

1.1.1. Probabilidad clásica

La *teoría clásica* de la probabilidad, originada directamente en los juegos de azar, establece una definición conectada a su cuantificación. Este concepto, debido a Laplace establece:

Definición 1.1. La probabilidad de un suceso es el cociente del número de casos favorables al suceso entre el total de casos posibles, supuestos igualmente verosímiles.

Este concepto de probabilidad, que suele denominarse de Laplace, se remonta sin embargo al trabajo *The Doctrine of Chances* de De Moivre (1711) concebido como un manual para los interesados en juegos de azar. Por el contrario, Pierre Simon, marqués de Laplace (1749 1827) elaboró un total de 10 principios del cálculo de probabilidades, entre los que figura por primera vez la definición anterior, que no se han visto alterados desde su obra *Théorie Analitique des Probabilités* (1812).

El *concepto clásico*, que ha dominado hasta principios del presente siglo, ha sido objeto de diversas críticas debidas a su falta de rigor lógico (lo definido entra en la definición) y al supuesto de resultados igualmente verosímiles en el que se basa la teoría.

La justificación de esta hipótesis viene dada por el principio de indiferencia, que defiende la simetría u homogeneidad de resultados en la situación considerada, o bien por el principio de la razón insuficiente según el cual, si no existe razón que favorezca alguno de los resultados con respecto a los demás, admitiremos que todos tienen igual probabilidad. Sin embargo ninguno de estos principios soluciona las dificultades planteadas por la definición clásica, cuya aplicación práctica se limita a un ámbito muy reducido (experimentos con número finito de resultados equiprobables).

Pese a sus limitaciones el esquema clásico de probabilidad está muy arraigado, debido en gran medida a su conexión con los juegos de azar. Así, nadie duda de asignar probabilidad de $\frac{1}{6}$ a cada una de las caras de un dado, $\frac{1}{40}$ a cada carta de la baraja española o $\frac{1}{2}$ a los resultados "cara" y "cruz" al lanzar una moneda.

No obstante, es necesario ser prudente en la utilización de esta probabilidad ya que en caso contrario pueden cometerse abusos y llegar a resultados incorrectos. Así, a modo de ejemplo, la cuantificación de la probabilidad de seguir contratado se llevaría a cabo empleando este concepto clásico como cociente entre casos favorables (sólo uno) y posibles (sólo dos: seguir contratado o no).

El uso indiscriminado del concepto clásico para cuantificar probabilidades puede llevarnos, en el caso de que los resultados posibles no sean equiprobables, a conclusiones sorprendentes e incluso absurdas. De hecho, según este método, asignaríamos probabilidades del 50 % a sucesos del tipo "llegar a ser premio Nobel", "presenciar un terremoto" y en general a todos aquellos sucesos asociados a un experimento con dos resultados.

No siempre resulta sencillo definir resultados simétricos o equiprobables garantizando así la aplicabilidad de la definición clásica. Un ejemplo famoso es la discusión protagonizada por D'Alembert, Fermat y Pascal en torno a un juego sencillo: el lanzamiento de dos monedas, sobre el que se formula la apuesta "sacar al menos una cara". Según el razonamiento seguido por D'Alembert, la probabilidad de victoria sería $\frac{2}{3}$, ya que de los tres resultados posibles (ninguna cara, una cara, dos caras) dos son favorables. Sin embargo, es necesario tener presente el principio de simetría inherente a la

probabilidad clásica. Este principio exigiría describir los resultados del experimento mediante sucesos equiprobables: cara cruz, cara cara, cruz cara, cruz cruz y, dado que de estas cuatro posibilidades tres son favorables a la apuesta planteada, la probabilidad de éxito sería $\frac{3}{4}$.

En otras ocasiones las inexactitudes son más difíciles de detectar. Supongamos una situación más compleja que las anteriores, en la que una empresa concede a sus trabajadores ciertos permisos situados en días que la empresa denomina "comodín". Con el objeto de garantizar a todos sus trabajadores sea cual sea su horario y jornada laboral la posibilidad de disfrutar de este día, se acuerda que los "comodines" serán situados en meses seleccionados al azar pero siempre el día 13.

Si un trabajador se preguntan cuál es la probabilidad de que el "comodín" coincida en un viernes, permitiéndoles así disfrutar de un largo fin de semana, parece legítimo en un principio el supuesto de equiprobabilidad y simetría que justifica un resultado $P(Viernes) = \frac{1}{7}$, coincidente con el de cualquier otro día de la semana.

La aplicación de la probabilidad clásica no plantea en principio inconvenientes. Sin embargo, tras un razonamiento más sofisticado se aprecia que, debido a los ajustes horarios y la configuración de los calendarios, los distintos días de la semana como justificaremos más adelante no son equiprobables.

1.1.2. Probabilidad frecuencial

El *enfoque frecuencial* -también denominado *frecuentista*- de la probabilidad se sitúa en una perspectiva experimental.

Definición 1.2. Llamamos *probabilidad frecuencial* de un suceso al valor en torno al cual tiende a estabilizarse su frecuencia relativa.

Esta idea de probabilidad sólo es válida bajo el supuesto de fenómenos aleatorios experimentales (reproducibles bajo idénticas condiciones un número suficientemente elevado de veces) y que verifiquen el *principio de regularidad estadística*, según el cual las frecuencias relativas tienden a estabilizarse en torno a un cierto valor.

Esta noción de probabilidad, introducida por Venn en 1866 y desarrollada matemáticamente en los años 1920 por Von Mises y Reichenbach, es de uso generalizado, debido en gran parte al sencillo método de cálculo de probabilidades que lleva asociado. De hecho, la axiomática propuesta por Kolmogorov -que constituye una herramienta fundamental en el desarrollo del *Cálculo de probabilidades*- está inspirada en el comportamiento asintótico de las frecuencias.

Una de las críticas a la concepción frecuencial va referida al supuesto de que es posible repetir indefinidamente el experimento bajo condiciones idénticas, que excluye de su ámbito de aplicación gran parte de los fenómenos sociales y económicos. En estos casos, la única posibilidad de aplicar el concepto frecuencial de probabilidad sería admitir la cláusula "*ceteris paribus*".

Pese a las limitaciones teóricas que plantea, a menudo la probabilidad se aproxima directamente a través de la frecuencia relativa. De este modo, el resumen de la información pasada es utilizado como método de estimación de potencialidades futuras. Así, un graduado universitario puede usar los resúmenes de información referidos a las últimas promociones para calcular su probabilidad de encontrar trabajo, de obtener una beca, etc.

El concepto frecuentista permite resolver adecuadamente el problema relativo a los días "comodín", justificando que la probabilidad de que el comodín sea viernes es superior a la de cualquier otro día de la semana. En efecto, la determinación de las probabilidades de cada día de la semana exigiría conocer el número de repeticiones de cada resultado sobre el tiempo total de vigencia de nuestro

calendario (años 1600-2000). Si consideramos que de estos 400 años 97 son años bisiestos, el total de semanas resulta ser 20.871 y de ellas 4.800 fechas son día 13 de un mes. Aunque la enumeración es larga, puede observarse el día de la semana en que cada uno de ellos está situado, que resultan ser una cifra superior en el caso del viernes (688 días respecto a 684 para jueves y sábado, 685 para lunes y martes y 687 para domingo y miércoles). Una vez determinado este nuevo modelo, la probabilidad de viernes se situaría en $\frac{668}{4800} = 0,143$.

1.1.3. Probabilidad subjetiva

Las limitaciones de los enfoques anteriores han sugerido métodos alternativos de determinación de la probabilidad, en los que ésta aparezca desvinculada de la experimentación. Así, frente a las definiciones objetivas de probabilidad, que incluyen las dos anteriores, las *teorías subjetivas* consideran la probabilidad como "*grado de creencia*", resultando así aplicables a un conjunto más amplio de situaciones.

La utilización de esta acepción de probabilidad es muy frecuente, ya que a menudo se nos plantea la necesidad de cuantificar numéricamente el nivel de "verosimilitud" asignado a un hecho. Esta variante subjetiva introduce como rasgo diferencial respecto a las objetivas la participación directa del individuo que -en función de su situación particular- actúa como "asignador" de probabilidades.

El punto de vista subjetivista fue compartido por algunos de los precursores de la teoría del cálculo de probabilidades como J. Bernoulli, Bayes o Laplace. No obstante, es desde principios de este siglo cuando se le ha prestado mayor atención, siendo pionera la obra de Borel (1924) a la que siguieron, entre otras, las de Ramsey (1926), De Finetti (1931, 1937) y Savage (1954). Todos ellos consideran fundamental el comportamiento del individuo frente a la incertidumbre, que le conduce a asignar implícita o explícitamente un orden o una medida a la posibilidad de que los sucesos tengan lugar.

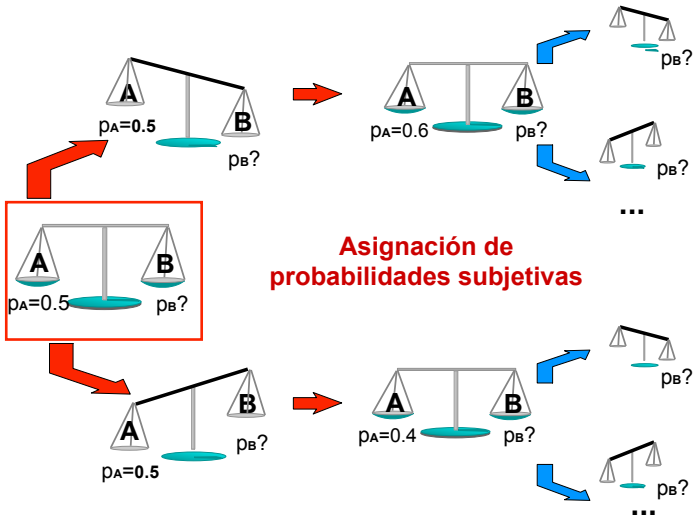
La corriente subjetivista no conduce a un concepto unívoco de la probabilidad, siendo posible distinguir varias acepciones. Algunos autores consideran que puede probarse la existencia de una función de probabilidad personal para cada individuo. Dicha probabilidad, denominada cualitativa o comparativa, se basa para cada individuo dado en comparaciones del tipo "un suceso no es más probable que otro".

Esta concepción logicista es debida, entre otros autores, a Keynes, Jeffreys, Koopman y Carnap, siendo su idea básica la extensión de los principios de la lógica matemática para establecer la probabilidad como medida en que una proposición (hipótesis) confirma a otra (experiencia).

Frente a las ventajas de este concepto de probabilidad, derivadas de su carácter formal, aparecen inconvenientes debidos a la dificultad de determinar probabilidades numéricas. Así, en el trabajo original de Keynes (1921) las probabilidades están tan sólo parcialmente ordenadas y no siempre son medibles numéricamente, no garantizando por tanto una medida numérica de la credibilidad racional.

Existen otros conceptos de la probabilidad entre los que, por su originalidad, queremos destacar el de sorpresa potencial de Shackle. En su obra *Decisión, orden y tiempo* (1966), este autor plantea de forma muy ingeniosa el problema de la decisión frente al tiempo, teniendo sólo en cuenta el momento presente. En este contexto, el concepto de probabilidad es sustituido por las ideas de "sorpresa potencial" y "grado de creencia" que resultan cuantificables si el individuo puede evaluar la distancia entre distintas sorpresas potenciales. Para Shackle todo decisor racional puede establecer su espacio de posibles resultados y asignar sus probabilidades, sin embargo el agente nunca es capaz de contemplar

Figura 1.1.: Probabilidad subjetiva



todas las posibilidades y puede producirse una sorpresa; de esta forma no existe el límite unitario a la probabilidad (podemos distribuir inicialmente una masa de probabilidad unitaria entre las distintas posibilidades y más tarde admitir una sorpresa no contemplada, con lo que al sumar esta probabilidad supera la unidad).

El resultado numérico de la probabilidad subjetiva aparece directamente ligado al individuo que lleva a cabo su cuantificación, resultando imprescindible que exista coherencia en el sistema de asignación (por ejemplo, transitividad).

La traducción numérica de los esquemas mentales individuales ("es muy probable que..." o "cierto hecho me parece más probable que algún otro") conlleva dificultades derivadas de la propia subjetividad. En realidad no existe un método de cuantificación de probabilidades subjetivas sino algunos mecanismos válidos para su aproximación.

Uno de estos mecanismos, defendido entre otros por De Finetti, es el estudio de las condiciones en las que un individuo se encontraría dispuesto a apostar por determinado suceso.

De este modo, si queremos calcular la probabilidad subjetiva de cierto suceso (que nos seleccionen para un puesto que hemos solicitado, que nuestro equipo favorito gane la liga, ...) podemos utilizar un sistema de apuestas, que resulta muy adecuado para aproximar probabilidades ya que refleja nuestra "creencia" en el suceso, asumiendo que existen racionalidad y sinceridad (es decir, obviamos el fanatismo, las apuestas irreflexivas o las de "farol").

Otro sistema de asignación de probabilidades subjetivas se tiene en los juegos: podemos idear un esquema de aproximación de la probabilidad del suceso que nos interesa mediante la comparación con ciertas "loterías".

Consideremos el ejemplo cuya ilustración aparece en la figura 1.1: a un individuo se le plantea un juego (B) mediante el cual se le otorgaría un premio en caso de que se produzca el suceso considerado (que sea seleccionado para un puesto, que su equipo gane la liga, etc). ¿Preferiría este juego u otro (A) en el que ganase el premio por el método clásico cara-cruz? Es evidente que esta sencilla decisión -que ilustramos mediante una balanza- no proporciona por sí misma una cifra de probabilidad, pero aporta nueva información: si el individuo responde que prefiere el juego B (condicionar el premio al suceso), ello refleja que considera este hecho "más probable" del 50 %. Como consecuencia la balanza se inclinaría en esta situación hacia la derecha, y viceversa en el caso contrario.

De este modo, tras la primera respuesta, según el lugar hacia el que se incline la balanza, deberemos alterar el contrapeso para la probabilidad desconocida. Así, si la primera respuesta es a favor del juego B, podríamos considerar una nueva alternativa A de probabilidad conocida y superior al 50 % (por ejemplo, una urna en la que el 60 % de bolas son éxitos). En el caso de que el individuo todavía prefiera jugar con la opción B frente a la urna, podemos concluir que la probabilidad subjetiva que le asigna al suceso también supera el 60 % y así sucesivamente.

Las compensaciones podrían ser efectuadas también en el otro sentido -tal y como recoge la figura 1.1- si las respuestas fuesen a favor del juego aleatorio. Siguiendo este esquema de razonamiento, el objetivo es aproximarse cada vez más a la cuantificación de la probabilidad subjetiva.

1.2. La probabilidad y su cuantificación

El análisis histórico de los estudios probabilísticos revela que los primeros esfuerzos fueron dirigidos a la solución de problemas concretos, esto es, a la cuantificación de la probabilidad, ignorándose su concepto. En efecto, el noble francés Antoine Gambaud, caballero de Méré, planteó al famoso matemático Blaise Pascal (1623-1662) uno de los problemas probabilísticos más antiguos y conocidos, relacionado con las apuestas a las que de Méré era aficionado. La discusión de estos problemas dio lugar a una extensa correspondencia entre los matemáticos Pascal y Pierre de Fermat, quienes nunca publicaron sus trabajos sobre probabilidad.

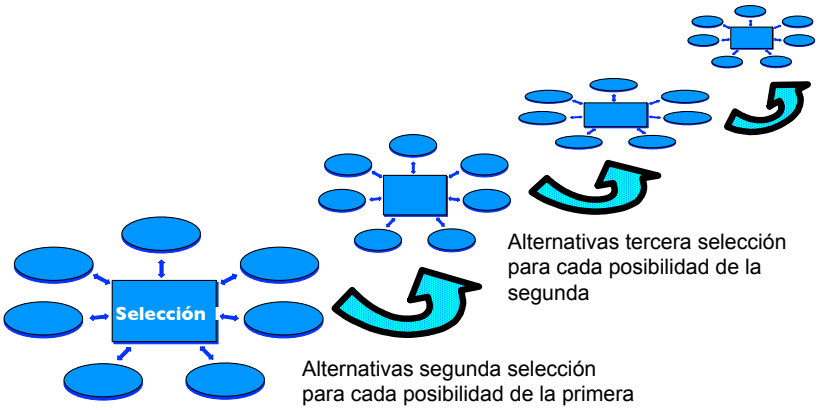
El problema planteado por de Méré, que habitualmente se designa "falacia de la Probabilidad" surgía al comparar dos apuestas, enunciadas como "Sacar al menos un 6 en 4 tiradas con un dado" y "Sacar al menos una suma 12 en 24 tiradas con dos dados". Según los cálculos llevados a cabo por de Méré, la probabilidad de éxito era idéntica en ambas apuestas, siendo las expresiones empleadas en su cálculo $P(E) = 4\frac{1}{6} = 0,66667$ para la primera y $P(E) = 24\frac{1}{36} = 0,66667$ para la segunda.

Fueron las diferencias de rentabilidad obtenidas con ambas apuestas las que llevaron a de Méré a consultar a Pascal.

[¿Por qué son incorrectos los cálculos propuestos? ¿qué concepto de probabilidad se aplica en estos casos?]

La cuantificación de probabilidades, que no es un tema completamente resuelto, aparece asociado a la teoría combinatoria. En efecto, para cuantificar el número de casos favorables y posibles asociados a determinada experiencia debemos identificar

Figura 1.2.: Variaciones con repetición. Selecciones sucesivas



las condiciones en las que ésta se realiza, evitando así confusiones como la que se esconde tras la falacia protagonizada por de Méré.

Una vez especificadas estas condiciones, la cuantificación de los casos (tanto favorables como posibles) se llevará a cabo mediante los conceptos de variaciones, permutaciones o combinaciones. Aunque no es nuestro objetivo efectuar un estudio detallado de teoría combinatoria, sí resulta conveniente señalar -mediante ilustraciones- las diferencias entre los conceptos y las fórmulas de cálculo asociadas a los mismos.

Imaginemos a modo de ejemplo que el dominical de un periódico decide incluir en cada ejemplar un cupón de sorteo con un número de 4 dígitos. ¿Cuántos cupones distintos existirán? Mediante un sencillo razonamiento -ilustrado en el gráfico 1.2- puede verse que las posibilidades son 10.000. En efecto, hay 10 opciones para el primer dígito (en el gráfico se han representado solamente 7 para no cargar excesivamente la figura) y, para cada uno de éstos, pueden a su vez seleccionarse 10 para el segundo. A su vez, para cada una de esas 100 posibilidades tendríamos otras 10 para el tercero y lo mismo para el cuarto dígito. El cálculo efectuado se corresponde con el caso de Variaciones con repetición de 4 elementos seleccionados entre 10 (también denominadas "de 10 elementos de orden 4"), y sus rasgos son la posibilidad de repetición (la selección de un dígito no le excluye para una nueva utilización) y la importancia del orden (es relevante en qué lugar está situado cada número).

Definición. Las *Variaciones con repetición* de m elementos de orden n se obtienen como: $VR_{m,n} = m.m \ldots m = m^n$

Una variante se obtiene cuando se excluye la posibilidad de repetición, apareciendo así las *Variaciones*.

Como es lógico, en esta situación disminuye el número de casos, ya que se eliminan posibilidades respecto al supuesto con repetición. De hecho, el número de cupones en los que no se repiten cifras son 10.9.8.7, variaciones de 4 elementos distintos seleccionados (sin repetición) entre 10.

Definición. Las *Variaciones de m elementos de orden n* se obtienen mediante la expresión:$V_{m,n} = m(m-1)(m-2)\cdots(m-n+1)$, que se denomina "factorial generalizado de m de orden n".

Las *Variaciones* aparecen con gran frecuencia en la práctica, donde es bastante habitual excluir un elemento ya seleccionado. Supongamos, por ejemplo, que en el dominical del periódico se desea 5 reportajes para otras tantas páginas, y deben decidir entre un total de 12 trabajos ya elaborados. Como parece evidente que cada reportaje sólo puede ser utilizado una vez, nos encontraríamos con Variaciones de 5 elementos seleccionados sin repetición entre 12, esto es, $V_{12,5} = 12{,}11{,}10{,}9{,}8$.

¿Qué sucedería si el número de reportajes disponibles fuese tan sólo 5? En esta situación las variaciones anteriores presentan un rasgo particular: estamos seguros de que todos los artículos aparecerán recogidos en el suplemento y el único rasgo diferenciador será por tanto el orden de los mismos. Nos encontramos así con un caso particular de variaciones: las *Permutaciones*, en este caso de 5 elementos, cuyo resultado sería 5.4.3.2.1=5!

Definición. Las *Permutaciones* de m elementos son las ordenaciones posibles de los mismos, y coinciden con las variaciones sin repetición de m elementos de orden m: $P_m = V_{m,m} = m(m-1)(m-2)\cdots 1 = m!$

En las permutaciones sólo aparece como elemento diferenciador el orden, supuesto que todos los elementos ordenados son distintos entre sí.

Imaginemos ahora que el suplemento dominical dedica su contraportada a publicidad, para lo cual dispone de un total de 6 casillas. En principio podríamos plantear que, una vez seleccionados los seis anunciantes, hay 6! formas de configurar la contraportada, según el modo en que se ordene la publicidad.

Sin embargo ¿qué sucedería si un mismo anuncio aparece dos veces, para enfatizar así el efecto publicitario? En este caso el lector verá dos veces la misma imagen sin distinguir posibles intercambios entre las casillas en las que se halla y en consecuencia las ordenaciones percibidas no serán ahora 6! sino sólo la mitad (dividimos entre 2! formas en las que pueden ordenarse los elementos repetidos).

De modo similar, podría haber más "anuncios repetidos", con la consiguiente reducción en las permutaciones [¿Qué sucedería por ejemplo si de los 6 anuncios hay 2 de un macroconcierto y otros 3 son la portada de un libro?]

Definición. Las *Permutaciones con repetición* recogen las ordenaciones de m elementos, donde a, b, c ... son repetidos entre sí. Su método de cálculo es:

$$P_m^{a,b,c} = \frac{P_m}{P_a P_b P_c} = \frac{m!}{a!b!c!}$$

Supongamos por último que el periódico decide aumentar el equipo de colaboradores del dominical, al cual se incorporarán 3 trabajadores más, seleccionados de entre los 7 nuevos contratados. ¿De cuántos modos pueden ser seleccionados estos nuevos miembros del equipo?

En situaciones como la descrita, denominadas *Combinaciones*, se trata de extraer subgrupos a partir de un total. Como consecuencia, el orden es un factor irrelevante ya que simplemente nos interesa qué nuevos trabajadores integran el equipo pero no qué lugar ocupan.

¿Cómo se cuantificarían las posibilidades de seleccionar esos 3 individuos entre los 7 candidatos? Un posible razonamiento sería distinguir dos grupos: uno de ellos integrado por los 3 que pasan al equipo y otro por los 4 trabajadores que se ocuparán de otras tareas.

En la situación planteada, lo único que nos interesa distinguir es si un individuo está o no en ese grupo. Por tanto, una aproximación válida consiste en partir del total de posibles ordenaciones de los 7 individuos (7!) y eliminar de ellas las ordenaciones dentro de cada grupo (no nos interesa cómo se ordenan los 3 que pasan al equipo ni tampoco las ordenaciones de los 4 que no entran en el mismo). Así se llegaría a la fórmula de cálculo de las combinaciones 7 sobre 3, esto es,

$$\begin{pmatrix} 7 \\ 3 \end{pmatrix} = \frac{7!}{3!4!}$$

Definición. Las *Combinaciones de m de orden n* son subconjuntos de n elementos seleccionados de un total de m: $C_{m,n} = \dfrac{m!}{n!(m-n)!}$

Esta expresión se corresponde con el número combinatorio, cuya representación en forma tabular es el triángulo de Pascal, en el que cada término es suma de los dos términos situados inmediatamente por encima de él.

La expresión de las combinaciones puede ser obtenida como caso particular de permutaciones con repetición. Para ello, basta tener presente que nos interesa únicamente la agrupación efectuada, por lo cual del total de ordenaciones de los m elementos (m!) ignoramos las ordenaciones de los elementos seleccionados para integrar los subconjuntos (n!) y también las de los no seleccionados$(m-n)$!. Se obtiene así: $C_{m,n} = P_m^{n,(m-n)}$.

Desde luego la teoría combinatoria abarca otras expresiones de cálculo que no hemos recogido aquí. Así, si por ejemplo distribuimos tres ejemplares del dominical entre 2 kioscos sin ninguna restricción (podrían ir todos al mismo, por ejemplo), la expresión de cálculo de las posibilidades vendría dada por Combinaciones con repetición, de aparición menos frecuente que las anteriores y cuya fórmula guarda relación con las combinaciones.

Las *Combinaciones con repetición* permiten cuantificar las posibilidades de repartir en m grupos un total de n elementos idénticos, a través de la expresión: $CR_{m,n} = \frac{(m+n-1)!}{n!(m-1)!)}$. En concreto, para el ejemplo propuesto se tendrían combinaciones en dos grupos con tres elementos repetidos, cuya fórmula viene dada por

$$CR_{2,3} = \frac{(2+3-1)!}{3!1!}$$

Las expresiones anteriores resultan útiles para solucionar el problema planteado por De Méré: El cálculo correcto para su primera apuesta viene dado por:

$$P(G) = \frac{c.f.}{c.p.} = \frac{\text{Resultados con al menos un 6}}{\text{Resultados en 4 tiradas}} = \frac{C_{4,1}VR_{5,3} + C_{4,2}VR_{5,2} + C_{4,3}VR_{5,1} + 1}{VR_{6,4}} = 0,52$$

donde la presencia de Combinaciones corresponde a los "huecos" o tiradas en las que aparece el resultado 6, mientras las Variaciones con Repetición del numerador recogerían los posibles números para completar las restantes tiradas.

Por su parte, la segunda apuesta sería resuelta en los siguientes términos:

$$P(G) = \frac{c.f.}{c.p.} = \frac{\text{Resultados con suma 12 en las 24 tiradas}}{\text{Resultados en 24 tiradas}}$$

cuya cuantificación resulta de mayor complejidad.

Como veremos en un apartado posterior, las probabilidades asociadas a estas dos apuestas pueden ser cuantificadas de forma más sencilla aplicando ciertas propiedades de la probabilidad.

1.3. Definición axiomática de la probabilidad

La caracterización axiomática de la probabilidad es una idealización matemática en la cual encajan las diferentes interpretaciones. De este modo, con independencia de cuál sea el concepto de probabilidad que utilicemos, la probabilidad será concebida como una cantidad numérica asociada con un suceso que posee ciertas propiedades básicas expresadas por medio de axiomas.

En las primeras décadas de este siglo se cuestionaba el significado de la probabilidad y habían surgido diferentes concepciones de la misma; parecía imperiosa la necesidad de formular un modelo teórico sobre el que fundamentar el desarrollo sistemático del Cálculo de Probabilidades y donde encajasen las diversas interpretaciones de la probabilidad fuera cual fuera su concepción.

Este modelo lo proporciona la teoría axiomática de la probabilidad, que sobre la base de un reducido número de axiomas permite desarrollar todo el Cálculo de Probabilidades, independientemente de cuál sea el significado de la probabilidad; si una concepción satisface la axiomática exigida, puede ser considerada una probabilidad y por lo tanto puede aplicársele todo el desarrollo posterior alcanzado por esa teoría. Además, otra característica de esta formalización es la de ser autónoma, es decir, que se adhiere al principio de que el cálculo de probabilidades es un método para transformar unas probabilidades en otras.

Las axiomáticas establecidas en el primer cuarto de este siglo no tenían un carácter de formalización, sino que trataban de caracterizar concepciones concretas de la probabilidad. Tras varios intentos, fue Kolmogorov quien en 1933 dio una axiomática para la probabilidad, hoy reconocida casi universalmente, sobre la que se fundamentó el Cálculo de Probabilidades. Esta axiomática se basa en dos conceptos fundamentales: álgebra o σ-álgebra de sucesos y la medida de probabilidad.

Para exponer la axiomática de Kolmogorov, consideremos un *fenómeno aleatorio*, y sea E el *conjunto de resultados posibles*, que también se denominan *casos o sucesos elementales*; a E se le llama *espacio muestral o suceso seguro*.

Consideremos a modo de ejemplo el lanzamiento de un dado. Teniendo en cuenta que su resultado no es predecible de una forma determinista, se tratará de un fenómeno aleatorio cuyos casos o sucesos elementales serán: {1}, {2}, ..., {6}, que describen los posibles resultados en el lanzamiento del dado.

El espacio muestral E o suceso seguro estará formado por todos los posibles resultados: E={1, 2, ..., 6}.

No siempre nos interesará conocer la probabilidad de sucesos elementales, sino que a veces estaremos interesados en la probabilidad de que ocurran determinadas combinaciones de estos sucesos.

En el ejemplo anterior puede interesarnos no sólo conocer la probabilidad de los resultados elementales, sino también cuantificar la probabilidad de que el resultado sea par, mayor que 4 o menor que 3, por ejemplo.

Por lo tanto tendremos que establecer una estructura que recoja estas combinaciones. Así, después de definir ciertos sucesos (suceso imposible (vacío), unión, intersección, complementario, diferencia y diferencia simétrica), acompañamos al espacio muestral E de una familia de sucesos (o subconjuntos de él), $\boldsymbol{\mathcal{A}}$, que tiene cierta estructura algebraica (σ-álgebra).

Definición 1.3. Toda *σ-álgebra* se caracteriza por verificar las tres condiciones siguientes:

1. El suceso imposible está en $\boldsymbol{\mathcal{A}}$
2. Si un suceso está en $\boldsymbol{\mathcal{A}}$, su complementario $A^c = \bar{A}$ también lo está
3. La unión numerable de conjuntos de $\boldsymbol{\mathcal{A}}$, pertenece a $\boldsymbol{\mathcal{A}}$, $\forall A_1, A_2, \ldots \in \boldsymbol{\mathcal{A}}, \cup_{i=1}^{\infty} A_i \in \boldsymbol{\mathcal{A}}$

Cuando se sustituye la condición 3) por la unión finita, el resultado es un *álgebra*. En espacios muestrales finitos ambos conceptos coinciden.

En el ejemplo del lanzamiento del dado, el álgebra de sucesos estaría formado por los sucesos elementales: {1}, {2}, ..., {6}, sus complementarios: {2,3,...,6}, {1,3,...,6}, ..., {1,2,...,5}, la unión de cada dos sucesos elementales: {1,1},{1,2}, ...,{1,6},{2,1}, ...,{2,6}, ...,{6,1},{6,2},....,{6,6}, los complementarios de estos sucesos, la unión de cada 3 sucesos elementales, sus complementarios, ..., las intersecciones, etc. [¿Cuántos elementos integrarán este álgebra?]

Definición 1.4. *A c*ada elemento de $\boldsymbol{\mathcal{A}}$ se le llama suceso, y al par $(E,\boldsymbol{\mathcal{A}})$ se le denomina *espacio probabilizable* (o espacio medible).

Sobre cada espacio probabilizable pueden definirse distintas medidas de probabilidad. Siguiendo la formulación dada por Kolmogorov, podemos establecer la siguiente definición:

Definición 1.5. Dado un espacio probabilizable $(E, \boldsymbol{\mathcal{A}})$, una *medida de probabilidad* es una aplicación de $\boldsymbol{\mathcal{A}}$ en $\Re$:

$$P : A \in \boldsymbol{\mathcal{A}} \rightarrow P(A) \in \Re$$

que verifica las siguientes propiedades:

1. $P(A) \geq 0\ ,\ \forall A \in \boldsymbol{\mathcal{A}}$
2. $P(E) = 1$
3. Si $A_1, A_2, \ldots, A_n, \ldots$ pertenecen a $\boldsymbol{\mathcal{A}}$ y son incompatibles dos a dos, entonces la probabilidad de la unión es la suma de probabilidades:

$$A_i \cap A_j = \emptyset,\ \forall i \neq j \Rightarrow P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

Cuando el espacio muestral es finito, la familia $\boldsymbol{\mathcal{A}}$ queda reducida a un álgebra (que puede ser partes de E), y el tercer axioma de la probabilidad puede ser sustituido por la *aditividad finita*. Sin embargo, cuando el espacio muestral es infinito, tenemos que definir la probabilidad sobre una σ-álgebra, ya que no podemos definir una medida sobre la familia de todos los subconjuntos de E, de forma que siga verificando la propiedad de aditividad numerable. La terna $(E, \boldsymbol{\mathcal{A}}, P)$ se denomina *espacio de probabilidad*.

Sobre el espacio muestral anterior, el álgebra de sucesos estará formada, entre otros, por los siguientes: $\boldsymbol{\mathcal{A}}$={{1},...,{6},....{2,4,6},...,{1,2},...,{5,6},...}. Sobre el espacio probabilizable $(E, \boldsymbol{\mathcal{A}})$ podemos definir distintas medidas de probabilidad, por ejemplo:

$$P : A \in \boldsymbol{\mathcal{A}} \rightarrow P(A) \in \Re$$

$P(\{1\}) = \frac{1}{6}, P(\{2\}) = \frac{1}{6}, \ldots, P(\{6\}) = \frac{1}{6}$

$P(\{2,4,6\}) = P(\{2\}) + P(\{4\}) + P(\{6\}) = \frac{3}{6}$ por tratarse de la probabilidad de una unión de sucesos incompatibles.

$P(\{5,6\}) = P(\{5\} \cup \{6\}) = P(\{5\}) + P(\{6\}) = \frac{2}{6}$

$P(\{1,2\}) = \frac{2}{6}$

Con lo cual, bajo la estructura de álgebra podemos responder a las preguntas que nos habíamos planteado como probabilidad de obtener un número par, mayor que 4 o inferior a 3. De igual forma podemos calcular la probabilidad de cualquier otro suceso que nos planteemos sobre los resultados del lanzamiento de un dado.

Esta función de probabilidad cumple los axiomas anteriores [Compruébese]

La cuantificación de probabilidades para sucesos compuestos se llevó a cabo a partir de la asignación hecha a los sucesos elementales. Si, por ejemplo, hacemos una nueva asignación de probabilidad a los sucesos elementales:

$$P(\{1\}) = \frac{2}{12}, P(\{2\}) = \frac{1}{12}, P(\{3\}) = \frac{2}{12}, P(\{4\}) = \frac{1}{12}, P(\{5\}) = \frac{2}{12}, P(\{6\}) = \frac{4}{12}$$

es fácil comprobar que nos conduciría a otra función de probabilidad diferente. [Obtener la probabilidad de los sucesos anteriores]. Observamos por tanto que sobre un espacio probabilizable pueden definirse diversos espacios de probabilidad.

Los axiomas anteriores están inspirados en las propiedades de las frecuencias y resultan aplicables en una amplia variedad de situaciones. Así, si un trabajador, ante la incertidumbre laboral, desea obtener la probabilidad de que su actual contrato sea prorrogado, el experimento tendría dos resultados posibles, que podemos denotar por T: "continuar contratado" y S: "ser despedido". El espacio

muestral E estará entonces formado por esos dos sucesos elementales: $E = \{T, S\}$ y para cuantificar la probabilidad de cada suceso podríamos basarnos en información frecuencial sobre renovación de contratos. A modo de ejemplo, si sabemos que el 75 % de los contratos han sido renovados se tendría: $f(T) = \frac{3}{4}$; $f(S) = \frac{1}{4}$.

Resulta evidente que las frecuencias obtenidas en ningún caso serán negativas. Por su parte la frecuencia del suceso seguro viene dada por:

$$f(E) = f(T \cup S) = \frac{\text{nº de veces que ocurre (T o S)}}{\text{nº de realizaciones del experimento}} = \frac{4}{4} = 1$$

Sobre esta expresión podemos comprobar que, dado que T y S no tienen intersección común, se verifica:

$$f(T \cup S) = \frac{\text{nº de ocurrencias de T+nº de ocurrencias de S}}{4} = \frac{3+1}{4} = f(T) + f(S)$$

por tanto, observamos que las frecuencias relativas verifican los tres axiomas exigidos a la probabilidad. [Comprobar que la probabilidad clásica también verifica los axiomas anteriores]

La axiomática de Kolmogorov fue posible gracias al gran desarrollo alcanzado por la teoría de la medida y la integral de Lebesgue; por otra parte, su desarrollo se debió en gran medida a la identificación entre sucesos y conjuntos, puesta de manifiesto por Stone en 1936, mediante el teorema que lleva su nombre. Esta circunstancia le permitió también aprovechar los conocimientos relativos a la teoría de conjuntos y gracias a este isomorfismo podemos utilizar indistintamente la terminología de sucesos (imposible, incompatibles, ...) o la relativa a conjuntos (vacío, disjuntos, ...).

En la axiomática original dada por Kolmogorov el axioma 3) se encontraba desdoblado en dos axiomas: aditividad finita y continuidad monótona en el vacío; sin embargo, algún tiempo después se demostró que estos supuestos eran equivalentes a la aditividad numerable.

1.3.1. Propiedades elementales de la probabilidad

Como consecuencia de los axiomas de Kolmogorov, dados los sucesos A y B se cumplen una serie de propiedades básicas:

1. $P(A^c) = 1 - P(A)$
2. $P(\emptyset) = 0$
3. Si $A \subset B$ entonces $P(A) \leq P(B)$
4. $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

La comprobación de estas propiedades es bastante inmediata:

1. $E = A \bigcup A^c$entonces $P(E) = P(A \cup A^c)$; teniendo en cuenta el segundo axioma la probabilidad del suceso seguro es la unidad y si además consideramos que un suceso y su complementario son disjuntos, el axioma 3 nos dice que esa probabilidad es la suma de las probabilidades de los dos sucesos, luego: $P(E) = 1 = P(A) + P(A^c)$; de donde se tiene: $P(A^c) = 1 - P(A)$
2. Este resultado se deduce del anterior teniendo en cuenta que $E^c = \emptyset$.

3. Como $A \subset B$, podemos expresar B como: $B = A \cup (B \cap A^c)$, siendo los dos sucesos que forman la unión (A y $B \cap A^c$) disjuntos [¿por qué?], de donde el axioma 3 asegura: $P(B) = P(A) + P(B \cap A^c)$, y como la probabilidad es una función no negativa ($P(B \cap A^c) \geq 0$), por tanto se tiene la proposición enunciada [¿por qué?].

4. Los sucesos A y B pueden ser expresados de la siguiente forma: $A = (A \cap B) \cup (A \cap B^c)$, $B = (A \cap B) \cup (A^c \cap B)$ siendo los sucesos intersección considerados en los dos casos disjuntos [¿por qué?], por lo cual se tiene: $P(A) = P(A \cap B) + P(A \cap B^c)$, $P(B) = P(A \cap B) + P(A^c \cap B)$. Por otra parte $A \cup B$ puede descomponerse como unión de sucesos disjuntos: $A \cup B = (A \cap B^c) \cup (A \cap B) \cup (A^c \cap B)$, con lo que su probabilidad puede obtenerse como: $P(A \cup B) = P(A \cap B^c) + P(A \cap B) + P(A^c \cap B)$ Teniendo en cuenta las expresiones anteriores y sustituyendo se llega al resultado enunciado. [Completar la justificación]

Esta última propiedad puede extenderse a un mayor número de sucesos; por ejemplo si C es otro suceso se tiene:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

[Justificar este enunciado]

Los valores extremos de probabilidad sugieren algunos comentarios. Como hemos comprobado el suceso imposible tiene probabilidad nula; sin embargo, algunas veces incluimos dentro del suceso imposible ciertos resultados que, aunque no tienen asignada probabilidad inicial, podrían llegar a ocurrir, hecho que contradice o bien la asignación de probabilidad nula o bien la especificación del experimento.

Un ejemplo muy intuitivo para ilustrar esta discusión es el experimento consistente en lanzar una moneda. Los resultados que consideramos posibles son cara (C) y cruz (F), por lo cual el suceso seguro será E={C,F} y su complementario será considerado como suceso imposible; sin embargo, al lanzar una moneda, ésta puede caer de canto, resultado que es complementario al suceso seguro y en cambio no es imposible.

Este mismo ejemplo nos sirve para reflexionar sobre la probabilidad unitaria: el hecho de que sea posible obtener resultados fuera de E nos exigiría asignar a éste una probabilidad inferior a la unidad.

La solución de estos problemas puede basarse en una revisión del espacio muestral, incluyendo sucesos de probabilidad nula (sería la probabilidad asignada a caer de canto una moneda); sin embargo, esta solución podría ser compleja cuando trabajamos con espacios muestrales infinitos. Otra solución posible iría en la línea de la "sorpresa potencial" de Shackle, donde la probabilidad puede alcanzar valores superiores a la unidad.

En el trabajo habitual no suele adoptarse ninguna de las soluciones propuestas, pero conviene ser conscientes de la posibilidad -aunque remota- de que aparezca este problema.

[Si al lanzar 1.000.000 veces una moneda la frecuencia relativa de que ésta quede de canto es prácticamente despreciable ¿convertiría este hecho en "imposible" al resultado?]

La axiomática dada por Kolmogorov es la habitualmente utilizada, pero no es la única. Otras axiomáticas importantes han sido desarrolladas por Renyi y Popper.

A. Renyi (1954) elaboró una axiomática de la probabilidad basada en el concepto de probabilidad condicionada que generaliza la introducida por Kolmogorov.

Por su parte, la teoría formal introducida por Popper en su gran obra La lógica de la investigación científica (1934) puede ser considerada "abstracta" y supera algunas limitaciones de las axiomáticas anteriores.

1.4. Probabilidad condicionada e independencia

Como ya hemos comentado, la asignación de probabilidades sólo en ciertas ocasiones puede ser efectuada de forma exacta, resultando habitual que existan distintas estimaciones de la verosimilitud de cierto suceso.

Una de las razones -pero no la única- que justifica estas diferencias en la asignación de la probabilidad es la subjetividad a la que ya hemos aludido: el hecho de que un individuo sea más o menos optimista, su propia opinión respecto a un tema, le pueden llevar a asignar probabilidades mayores a los hechos que considera deseables y viceversa para los que querría evitar.

En ciertos casos, las diferencias entre las probabilidades, pueden venir también justificadas por la información disponible, que conducirá a la asignación de probabilidad "condicionada" a dicha información.

A modo de ejemplo, al calcular la probabilidad de renovación de un contrato será interesante disponer de información lo más amplia y actualizada posible sobre las distintas posibilidades futuras, de modo que al incorporar esta información la probabilidad asignada será "condicionada".

A menudo tiene interés estudiar cómo un suceso puede condicionar a otro, de modo que la disponibilidad de información llega a alterar las probabilidades asignadas a los posibles resultados.

Consideremos un nuevo ejemplo, representado en la tabla que sigue: se trata de la distribución porcentual de los contratos, clasificados según un doble criterio: el tipo de contrato (clasificado en técnicos y de gestión) y el sector de actividad (industria o servicios).

↓ Sexo \ Sector→	Industria	Servicios
Técnicos	40	25
Gestión	10	25

Si a partir de la información anterior comparamos la proporción de contratos de los dos sectores de actividad -que sería equivalente a la probabilidad de cada uno de ellos- observamos que éstas son coincidentes ($\frac{c.f.}{c.p.} = 0,5$ en los dos casos). Sin embargo ¿qué ocurriría si nos interesasen específicamente los contratos técnicos? En este caso eliminaríamos de nuestro análisis los contratos de gestión (que representan un 35 % del total) y pasaríamos a considerar únicamente el 65 % restante (primera fila de la tabla).

En consecuencia, ahora la industria tendría una mayor proporción de contratos ($\frac{c.f.}{c.p.} = \frac{40}{65}$ frente a los servicios, donde se obtiene $\frac{c.f.}{c.p.} = \frac{25}{65}$).

1.4.1. Probabilidad condicionada

La información desempeña un papel relevante en la asignación de probabilidades.

Definición 1.6. La *probabilidad condicionada* por un suceso B (P(B)>0), se define para cada suceso $A \in \mathcal{A}$ como:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

Esta expresión puede ser aplicada en el ejemplo anterior, en el que denominaremos T, G, I y S a los sucesos "Técnicos", "Gestión", "Industria" y "Servicios", respectivamente. Las probabilidades iniciales serían entonces $P(I) = P(S) = \frac{50}{100}$, mientras las probabilidades condicionadas al perfil técnico se obtendrían, aplicando la definición anterior:

$$P(I/T) = \frac{P(I \cap T)}{P(T)} = \frac{\frac{40}{100}}{\frac{65}{100}} = \frac{40}{65} \; ; \; P(S/T) = \frac{P(S \cap T)}{P(T)} = \frac{\frac{25}{100}}{\frac{65}{100}} = \frac{25}{65}$$

De una manera más formal podemos establecer la siguiente definición:

Definición 1.7. Dado un espacio de probabilidad $(E, \boldsymbol{\mathcal{A}}, P)$ y un suceso B ($B \in \boldsymbol{\mathcal{A}}$) con probabilidad no nula ($P(B) > 0$), denominamos *probabilidad condicionada* por el suceso B a una aplicación P_B definida como:

$$P_B(\cdot) : A \in \boldsymbol{\mathcal{A}} \rightarrow P_B(A) = P(A/B) = \frac{P(A \cap B)}{P(B)} \in [0, 1]$$

[La función toma valores no negativos por ser el cociente de dos probabilidades, donde cada una de ellas es no negativa].

La función P_B cumple la axiomática de Kolmogorov.
En efecto, tendríamos que comprobar que:

1. $P_B(A) \geq 0$
2. $P_B(E) = 1$
3. Dada una colección de sucesos $A_1, A_2, \ldots, A_n$ ($A_i \in \boldsymbol{\mathcal{A}}$) disjuntos dos a dos, se tiene:

$$P_B\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\left(\bigcup_{i=1}^{n} A_i\right)/B\right) = \sum_{i=1}^{n} P(A_i/B)$$

El primer axioma se comprueba de modo inmediato ya que se tiene $P(A/B) = \frac{P(A \cap B)}{P(B)} \geq 0$, por serlo el numerador.

Por lo que se refiere al segundo axioma, se tiene: $P(E/B) = \frac{P(E \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1$

Consideremos ahora la colección de sucesos especificada en el tercer axioma; se verifica:

$$P_B\left(\bigcup_{i=1}^{n} A_i\right) = P\left(\left(\bigcup_{i=1}^{n} A_i\right)/B\right) = \frac{P\left(\left(\bigcup_{i=1}^{n} A_i\right) \cap B\right)}{P(B)} = \frac{P\left(\bigcup_{i=1}^{n} (A_i \cap B)\right)}{P(B)}$$

teniendo en cuenta que los sucesos $A_i \cap B$ son disjuntos dos a dos [¿por qué?], el axioma iii) de la caracterización de Kolmogorov, garantiza:

$$P_B\left(\bigcup_{i=1}^{n} A_i\right) = \frac{\sum_{i=1}^{n} P(A_i \cap B)}{P(B)} = \sum_{i=1}^{n} \frac{P(A_i \cap B)}{P(B)} = \sum_{i=1}^{n} P(A_i/B) = \sum_{i=1}^{n} P_B(A_i)$$

Puede extenderse sin ningún problema la justificación anterior al caso de una sucesión infinita de sucesos.

Queda pues comprobado que la probabilidad condicionada P_B es una verdadera función de probabilidad.

En la definición anterior el suceso que condiciona (B) debe tener una probabilidad positiva para que el cociente de probabilidades esté definido. Pero podemos preguntarnos ¿qué ocurriría si B tuviese asignada probabilidad nula? En este caso caben dos tipos de argumentación: por un lado podríamos decir que no tiene sentido condicionar a un suceso imposible puesto que esta condición en la práctica no se va a dar; sin embargo, utilizando el concepto frecuencial de la probabilidad existen sucesos que nunca se han verificado, que por tanto tienen probabilidad nula, y que podrían servirnos para hacer simulaciones del tipo "¿qué habría ocurrido si ....?"; esto es, existen hechos sobre los que no disponemos de experiencia y que sin embargo de verificarse habrían alterado los resultados posteriores.

En tales supuestos resulta claro que la definición anterior de probabilidad condicionada no es adecuada. Una alternativa puede ser replantearse el espacio de probabilidad que sirvió para establecer la definición e introducir un nuevo álgebra $\boldsymbol{\mathcal{A}}_B$ de la forma siguiente:

$$\boldsymbol{\mathcal{A}}_B = \{A \cap B / \forall A \in \boldsymbol{\mathcal{A}}\}$$

[$\boldsymbol{\mathcal{A}}_B$ cumple las condiciones de álgebra o σ-álgebra si lo es $\boldsymbol{\mathcal{A}}$].

Ahora sobre el nuevo espacio probabilizable $(E, \boldsymbol{\mathcal{A}}_B)$ podemos definir la probabilidad condicionada como una función de probabilidad general.

Una idea de gran interés y trascendencia en los análisis estadísticos es el concepto de *independencia en probabilidad.*

1.4.2. Independencia en probabilidad

Este concepto admite una interpretación muy intuitiva:

Definición 1.8. Dados dos sucesos A, $B \in \boldsymbol{\mathcal{A}}$, se dice que A es independiente de B cuando la probabilidad de ocurrencia de A no se ve afectada por la de B, es decir: $P(A/B) = P(A)$.

[Establecer la definición "B independiente en probabilidad de A"]

Proposición 1.1. *La definición de independencia en probabilidad es equivalente a la relación:*

$$P(A \cap B) = P(A)P(B)$$

conocida como condición de independencia.

Demostración. En efecto, si A es independiente en probabilidad de B, se tiene: $P(A/B) = P(A)$

Por otra parte, teniendo en cuenta la definición de probabilidad condicionada se verifica: $P(A/B) = P(A \cap B)/P(B)$

Igualando las dos expresiones y despejando se obtiene: $P(A \cap B) = P(A)P(B)$.

Recíprocamente, si se verifica la condición de independencia, entonces A es independiente de B. En efecto, partimos ahora de la relación $P(A \cap B) = P(A)P(B)$, y por otra parte, según la probabilidad condicionada se tiene:

$P(A/B) = P(A \cap B)/P(B)$, de donde: $P(A \cap B) = P(A/B)P(B)$

Los primeros miembros son iguales con lo cual igualando los segundos se obtiene: $P(A) = P(A/B)$ lo que concluye la demostración □

Gracias a la condición de independencia y a las propiedades anteriormente vistas, estaríamos en condiciones de resolver de modo más rápido el problema planteado por de Méré.

En efecto, aplicando la propiedad relativa a la probabilidad del complementario, el cálculo correcto de la apuesta 1 vendría dado por:

$$\begin{aligned} P(G) = & \quad 1 - P(\text{Ningún 6 en las 4 tiradas}) = \\ = & \quad 1 - P(\text{No 6 en la } 1^{\underline{a}})P(\text{No 6 en la } 2^{\underline{a}})P(\text{No 6 en la } 3^{\underline{a}})P(\text{No 6 en la } 4^{\underline{a}}) = \\ = & \quad 1 - \left(\tfrac{5}{6}\right)^4 = 0,52 \end{aligned}$$

Por su parte, para la apuesta II se tiene el cálculo:

$$P(G) = 1 - P(\text{Ninguna suma 12 en las 24 tiradas}) = 1 - \left(\frac{36}{36}\right)^{24} = 0,491$$

La condición de independencia es una relación simétrica de los sucesos A y B; por tanto si A es independiente de B, B también será independiente de A. Por este motivo, en el futuro sólo hablaremos de sucesos independientes sin especificar ningún sentido para esa independencia.

Aunque no se haya hecho una mención explícita a ello, la definición de independencia se apoya en la probabilidad condicionada que exige que el suceso que condiciona tenga una probabilidad no nula; esto es, $P(B) > 0$; y como la independencia entre A y B implica la misma entre B y A, también debe cumplirse $P(A) > 0$.

La equivalencia probada entre la definición y la condición de independencia es válida para todo par de sucesos de probabilidades no nulas.

Si por ejemplo $P(B) = 0$ y A es otro suceso cualquiera, la definición de "A independiente de B" no puede aplicarse con lo cual no podremos afirmar nada al respecto. Sin embargo, aplicando la condición de independencia se tiene: $P(A \cap B) \leq P(B) = 0$, por tanto: $P(A \cap B) = 0$.

Por otra parte: $P(A)P(B) = P(A),0 = 0$, de donde se obtiene: $P(A \cap B) = P(A)P(B)$.

De la comprobación anterior se extraen dos consecuencias: la primera es que la condición de independencia es más general que la definición establecida y es aplicable a cualquier par de sucesos A y B; y la segunda es que un suceso de probabilidad nula siempre es independiente de cualquier otro.

Hemos expresado el concepto de independencia ligado al de probabilidad; sin embargo pueden establecerse diversos conceptos de independencia según cuál sea la referencia respecto a la que medimos esa independencia. De hecho, esta idea guarda relación con la independencia estadística, que viene expresada en términos de frecuencias relativas.

Otro concepto que guarda gran similitud con los anteriores es el de independencia en información; un suceso A es informativamente independiente de otro B, si la información que proporciona el primero no disminuye al conocerse la que puede suministrar el segundo. Si introducimos una medida I indicativa de la información que contiene un suceso, representamos por $I(A)$ la información suministrada por A y por $I(A/B)$ la información que permanece en A cuando se conoce B. Pues bien, A será independiente en información de B si $I(A) = I(A/B)$, es decir, la información que proporciona A cuando se conoce B es la máxima que puede suministrar y por tanto B no contiene información sobre A.

1.5. Probabilidad total y teorema de Bayes

En algunos casos, un mismo resultado puede tener lugar bajo distintas situaciones alternativas, por lo que su probabilidad debe ser cuantificada mediante una "fórmula compuesta".

Consideremos por ejemplo las perspectivas de renovación de contrato de un trabajador, que lógicamente estarán relacionadas con el contexto económico. Como consecuencia, la cuantificación de la probabilidad total de renovación del contrato exigirá tomar en cuenta los distintos escenarios o posibilidades que podrían presentarse.

Admitamos para simplificar que las perspectivas futuras se limitan a tres posibilidades recogidas en la tabla: la existencia de un crecimiento económico significativo que parece la situación más verosímil (digamos con probabilidad del 60 %) permitiría a la empresa renovar todos los contratos; una segunda posibilidad sería el estancamiento, al que se asigna una probabilidad del 10 % y en cuyo caso se renovarían el 80 % de los contratos y por último se contempla la posibilidad de crisis económica (con una verosimilitud del 30 %), escenario en el que sólo un 50 % de los contratos serían renovados.

La información disponible se resume en la tabla

Alternativa	Probabilidad	Renov. Contratos
Expansión (X)	0,6	100 %
Estancamiento (E)	0,1	80 %
Crisis (C)	0,3	50 %

Como es lógico, las tres alternativas de futuro X, E y C son incompatibles y cada una de ellas puede tener intersección con el resultado que nos interesa (renovación de contrato, R).

Se tiene entonces $P(R) = P((R \cap X) \cup (R \cap E) \cup (R \cap C)) = P(R \cap X) + P(R \cap E) + P(R \cap C)$ que, aplicando la fórmula de la probabilidad condicionada, pueden ser expresadas a su vez como:

$$P(R) = P(R/X)P(X) + P(R/E)P(E) + P(R/C)P(C) = 0,6 + 0,08 + 0,15 = 0,83$$

En ocasiones nos interesa conocer cuál es la probabilidad de que el suceso se haya producido bajo una situación concreta. Por ejemplo, si a un trabajador se le informa

de que su contrato será renovado ¿qué probabilidad asignaríamos a la existencia de expansión económica? ¿y a la crisis? Este tipo de razonamiento plantea un "ajuste" en el sistema de probabilidades iniciales a medida que incorporamos nueva información.

En concreto, con el supuesto planteado tendríamos para la situación concreta R la expresión condicionada: $P(X/R) = P(X \cap R)/P(R)$ cuyo denominador ha sido calculado mediante probabilidad total.

Así pues, el resultado de esta probabilidad sería $P(X/R) = \frac{0,6}{0,83} = 0,72$, esto es, una vez confirmada la renovación del contrato, estimaríamos en un 72 % la probabilidad de expansión económica.

De modo análogo se revisarían las probabilidades de estancamiento y crisis. Este ejemplo ilustra los teoremas de la probabilidad total y de Bayes, de gran importancia en la evolución de la estadística.

1.5.1. Sistema completo de sucesos

Antes de formalizar esos teoremas debemos introducir el concepto de *partición.*

Definición 1.9. Dado un sistema de sucesos $A_1, \ldots, A_n \in \boldsymbol{\mathcal{A}}$, se dice que forman una *partición* o un *sistema completo de sucesos*, si todos ellos son "factibles", esto es, tienen una probabilidad positiva de verificarse, son incompatibles dos a dos y su unión cubre todo el espacio muestral.

1. $P(A_i) > 0, \forall i = 1, \ldots, n$
2. $A_i \cap A_j = \emptyset,\ \forall i \neq j$
3. $\bigcup_{i=1}^{n} A_i = E$, que aplicando la función de probabilidad a ambos miembros y teniendo en cuenta 2) equivale a: $P(E) = 1 = \sum_{i=1}^{n} P(A_i)$

[Las alternativas de futuro del ejemplo anterior forman una partición. ¿Por qué?]

1.5.2. Teorema de la probabilidad total

Teorema 1.1. *Dado un espacio de probabilidad $(E, \boldsymbol{\mathcal{A}}, P)$ sobre el cual se puede establecer una partición $(A_1, \ldots, A_n)$, la probabilidad de un suceso cualquiera B, $(B \in \boldsymbol{\mathcal{A}})$, puede calcularse mediante la siguiente expresión:*

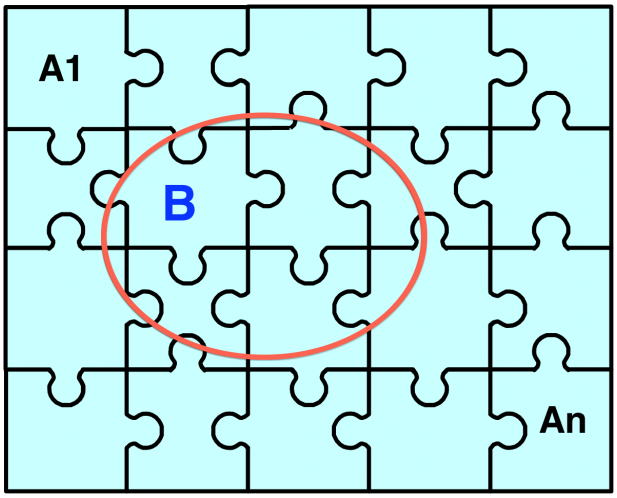
$$P(B) = \sum_{i=1}^{n} P(B/A_i)P(A_i)$$

Esta relación se conoce como *fórmula de la probabilidad total*, porque permite calcular la probabilidad total de un suceso a partir de las probabilidades de sus partes (intersecciones con los elementos de la partición).

Demostración. En efecto, podemos expresar: $B = B \cap E$

Como $(A_1, \ldots, A_n)$ forman un sistema completo de sucesos, la tercera característica de los mismos implica que $B = B \cap E = B \cap (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} (B \cap A_i)$. Los elementos

Figura 1.3.: Sistema completo de sucesos



de la partición son incompatibles dos a dos, $(B \cap A_i) \subset A_i$, por tanto esas intersecciones son también incompatibles dos a dos y en consecuencia si aplicamos la función de probabilidad, se tiene:

$P(B) = P\left(\bigcup_{i=1}^{n}(B \cap A_i)\right) = \sum_{i=1}^{n} P(B \cap A_i)$ [¿por qué?]

Por otra parte, de la fórmula de la probabilidad condicionada se sigue:

$P(B/A_i) = \frac{P(B \cap A_i)}{P(A_i)}$, y despejando: $P(B \cap A_i) = P(B/A_i)P(A_i)$

Por tanto sustituyendo en la expresión anterior, resulta:

$$P(B) = \sum_{i=1}^{n} P(B/A_i)P(A_i)$$

□

La interpretación del teorema de la probabilidad total aparece ilustrada en el esquema adjunto 1.3, donde los sucesos de la partición son representados como piezas de un puzzle (incompatibles y cuya unión es el espacio muestral). Según el enunciado de este teorema, la probabilidad de que se presente un determinado efecto final (suceso B) puede ser evaluada considerando las verosimilitudes del suceso B bajo las distintas alternativas (sucesos de la partición, A_i), debidamente ponderadas por la probabilidad asociada a cada alternativa.

Sin duda la propiedad de la probabilidad que ha alcanzado una mayor trascendencia es el siguiente resultado.

1.5.3. Teorema de Bayes

Teorema 1.2. *Dada una partición $A_1, \ldots, A_n$ de E y otro suceso B tal que $P(B) > 0$, el teorema de Bayes nos dice que entonces:*

$$P(A_i/B) = \frac{P(B/A_i)P(A_i)}{\sum_{i=1}^{n} P(B/A_i)P(A_i)}$$

Este teorema fue enunciado por primera vez por Bayes en 1763, aunque éste sólo lo demostró para el caso de equiprobabilidad de las causas (A_i). La demostración completa del teorema corresponde a Laplace (1812).

Demostración. Vamos a demostrar este resultado. Aplicando la definición de probabilidad condicionada se tiene:
$P(B/A_i) = \frac{P(B \cap A_i)}{P(A_i)}$, de donde $P(B \cap A_i) = P(B/A_i)P(A_i)$.

Por otra parte, $P(A_i/B) = \frac{P(A_i \cap B)}{P(B)}$ y sustituyendo el numerador por la expresión anterior y el denominador por la fórmula de la probabilidad total, obtenemos el resultado enunciado. □

Las probabilidades $P(A_i)$ se denominan *probabilidades iniciales o a priori* o *probabilidades de las causas*; las $P(A_i/B)$ se denominan *probabilidades finales o a posteriori*, probabilidades que se asignan a las causas después del conocimiento del suceso B; y las $P(B/A_i)$ se conocen como *verosimilitudes*. La fórmula de Bayes nos indica cómo la información proporcionada por el suceso B modifica las probabilidades iniciales que, mediante el empleo de la verosimilitud, transforma en probabilidades finales. También puede interpretarse este teorema en los siguientes términos: si consideramos el suceso B como un efecto (resultado de alguna observación o experimento), entonces la fórmula anterior nos indica la probabilidad de que A_i sea la causa de B.

Para el ejemplo propuesto, la introducción de información adicional (renovación o no del contrato) permitiría pasar de las probabilidades iniciales a probabilidades a posteriori según indica la tabla siguiente [Justificar cómo se ha obtenido cada uno de estos resultados]:

	Probabilidades		
Alternativa	A priori $P(A_i)$	Condicionadas a Renovación $P(A_i/R)$	Condicionadas a no Renovación $P(A_i/\bar{R})$
Expansión (X)	0,6	0,72	0
Estancamiento (E)	0,1	0,10	0,12
Crisis (C)	0,3	0,18	0,88

Es fácil comprobar cómo las "correcciones" incorporadas a las probabilidades a priori van en la dirección esperada según la información recibida: si sabemos que producirá la renovación de un contrato, tenderíamos a pensar que es más verosímil la expansión económica (cuya probabilidad pasa a ser del 72 %). Si por el contrario la información es que no se producirá la renovación, la probabilidad más afectada al alza es la de crisis económica que pasa a situarse en un 88 %.

La principal dificultad del teorema de Bayes estriba en calcular o definir las probabilidades iniciales. Este fue un tema de gran controversia que dio lugar a dos concepciones distintas de la estadística: la clásica y la bayesiana.

2. Magnitudes aleatorias

A menudo resulta interesante el estudio de magnitudes cuyo valor es imposible predecir de forma exacta. En estas situaciones, las técnicas estadísticas descriptivas -aunque útiles- son insuficientes, revelándose como imprescindible la utilización de probabilidades para cuantificar la potencialidad.

A modo de ejemplo, si consideramos la actualidad económica de una jornada, es posible que aparezcan noticias referidas a los beneficios de las entidades bancarias, el nivel de precios, los nuevos empleos que se generarán asociados a una inversión, el crecimiento del PIB, ...

Aunque todas estas características tienen elementos comunes, también se aprecian entre ellas algunos rasgos diferenciales. Así, para tratar la información relativa a los beneficios de entidades bancarias podríamos aplicar herramientas de estadística descriptiva, ya que estamos describiendo información pasada. Por el contrario, el planteamiento cambiaría si la información se refiere al número de empleos que de forma directa o indirecta serán generados por cierta inversión, ya que en este caso hablamos de un hecho futuro y por tanto existirá un componente de incertidumbre. De modo análogo, esta presencia de incertidumbre se manifiesta en hechos que, aun sin ser futuros, no pueden ser analizados desde un punto de vista determinista, al resultar imposible un análisis exhaustivo de los mismos. De ahí que las informaciones relativas al crecimiento del PIB o la inflación sean estimaciones, basadas en información parcial y que aparecerán acompañadas de alguna medida de su credibilidad, en términos de probabilidad.

2.1. Variable aleatoria. Variables discretas y continuas

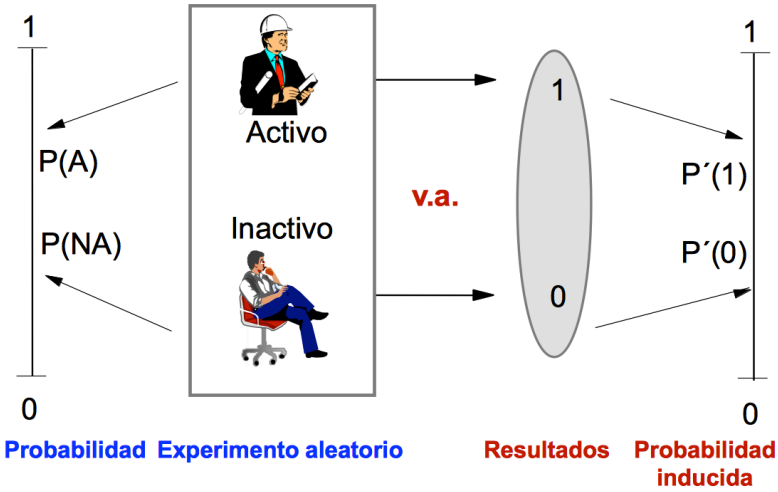
Cuando la realidad se estudia de forma descriptiva, las variables estadísticas resultan adecuadas, al describir o representar esa realidad mediante sus correspondientes valores y frecuencias.

Este mismo esquema puede ser trasladable a la descripción de magnitudes aleatorias, categoría en la que se incluyen gran parte de los fenómenos económicos, que raramente son predecibles de forma determinista ya que suelen contener una componente estocástica.

Aparece así el concepto de *variable aleatoria* (v.a.), de gran trascendencia en los análisis inferenciales, entendido como una función numérica de los resultados asociados a fenómenos aleatorios.

En ciertas ocasiones, esta realidad económica aparece directamente como una variable, esto es, descrita mediante valores. Por el contrario, otros fenómenos se presentan

Figura 2.1.: Variables aleatorias



en principio como categorías ("sector económico en el que se incluye una actividad", "situación laboral de un individuo", "contexto económico", ...) por lo cual resulta conveniente una transformación de las mismas, que lleve aparejada su descripción numérica.

Consideremos un experimento aleatorio que describe la situación laboral de un individuo w. El conjunto de posibles resultados será: $E = \{A = activo, NA = noactivo\}$.[1] Consideremos definida sobre este espacio muestral una variable X que asigna el valor 1 a A y 0 a NA. Está claro que esta función es una variable aleatoria, puesto que si elegimos al azar un individuo w no podemos conocer de forma exacta si el valor que le asignará X es 0 o 1, aunque sí podremos llegar a calcular la probabilidad con la que puede tomar esos valores.

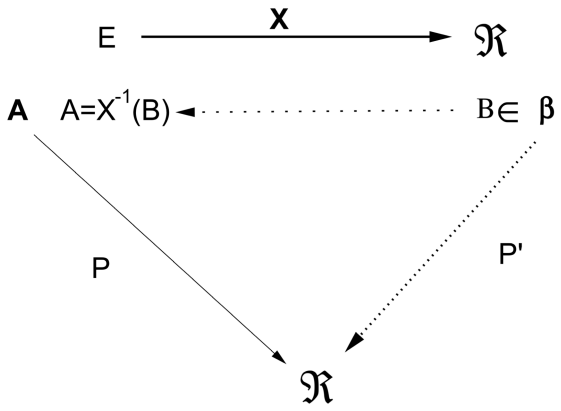
Una variable aleatoria debe transformar los resultados de un fenómeno aleatorio, elementos del espacio muestral E, en números reales, luego será una aplicación X de E en $\Re$. Como el resultado del fenómeno es incierto, el valor de la aplicación también lo será, pero la identificación de resultados con valores (reales), lleva implícita la identificación de sus probabilidades.

Sin embargo, el conocimiento de la probabilidad representa más un hecho cualitativo que cuantitativo, pues en el ejemplo anterior sólo nos informa sobre las posibilidades de encontrarse en cada una de esas categorías, sin detenerse en las valoraciones que éstas llevan implícitas.

[1] Según la Encuesta de Población Activa del INE, se definen como activas aquellas personas que realizan una actividad económica (caso de los empleados u ocupados) o que, no realizando tal actividad, están en condiciones y desean hacerlo (caso de los desempleados o parados).

En el suceso complementario NA consideramos por tanto la población inactiva, en la que se incluyen los estudiantes, los jubilados, las personas dedicadas exclusivamente al cuidado del propio hogar y las personas incapacitadas para trabajar.

Figura 2.2.: Probabilidad inducida



Supongamos una función aleatoria que asigna valor unitario a la situación de actividad. Tal y como muestra la figura 2.1todos los activos conducen al valor 1; por lo tanto tendremos tantas posibilidades de obtener 1 como de seleccionar un activo. Así pues, el sistema inicial de probabilidades sobre E induce un nuevo sistema P' sobre $\Re$, de manera que $P'(1) = P(A)$ y $P'(0) = P(NA)$. Esta probabilidad inducida viene inferida por la definición de la v.a. que especifica a qué valores se les asigna una probabilidad no nula y por el sistema inicial de probabilidades que permite la asignación de su cuantía.

De una forma más general, necesitamos establecer un espacio probabilizable sobre el cuerpo de los números reales, y a partir de él definir la probabilidad (inducida) que permita esa identificabilidad

El espacio nos lo proporciona $\Re$ y la σ-*álgebra de Borel*, $\boldsymbol{\beta}$, definida sobre $\Re$; esta σ-álgebra estará formada por todos los intervalos abiertos, semiabiertos, cerrados y sus intersecciones, uniones, ... Denotaremos por $(\Re, \boldsymbol{\beta})$ este espacio probabilizable.

Definición 2.1. Definimos una *variable aleatoria*, X, como una función de E en $\Re$ que sea medible; esto es, que la imagen inversa de todo boreliano (elemento de $\boldsymbol{\beta}$) sea un suceso (elemento de $\boldsymbol{\mathcal{A}}$).

En un sentido estricto tendríamos que distinguir entre el concepto de magnitud aleatoria (introducido en párrafos anteriores como v.a.) y la definición que acabamos de establecer de variable aleatoria. Podemos observar que esta definición es más restrictiva que la enunciada anteriormente, pues se pueden buscar contraejemplos de magnitudes aleatorias (asociadas a los resultados de fenómenos aleatorios) que no satisfacen la definición de variable aleatoria.

La definición dada responde satisfactoriamente a nuestros objetivos, pues nos permite definir de forma natural una probabilidad inducida sobre $\Re$ por P y X que garantice la identificación anterior como se recoge en la figura 2.2.

En efecto, dado un boreliano B en $\boldsymbol{\beta}$, la contraimagen de B por X será un suceso A en $\boldsymbol{\mathcal{A}}$, y además es el único suceso que la variable aleatoria transforma en B. Por tanto, como la probabilidad inducida,

P', de B tiene que coincidir con la probabilidad de A, la definición natural de la probabilidad inducida será: $P'(B) = P(X^{-1}(B)) = P(A)$.

Por otra parte, los intervalos semiabiertos de la forma $(-\infty, x]$, $x \in \Re$, generan la σ-álgebra de Borel, y por lo tanto, podemos reducir las definiciones anteriores a intervalos de este tipo, puesto que cualquier boreliano podrá ser expresado mediante una combinación de los intervalos anteriores. Así tenemos:

Definición 2.2. Una variable aleatoria X es una aplicación de E en $\Re$, que verifica: $X^{-1}(-\infty, x]$ es un elemento de $\boldsymbol{\mathcal{A}}$ para todo $x \in \Re$.

Definición 2.3. Definimos la *probabilidad inducida*, P', sobre $\boldsymbol{\beta}$ como aquella función de conjunto, que para todo x de $\Re$ verifica: $P'((-\infty, x]) = P\left(X^{-1}(-\infty, x]\right)$.

Proposición 2.1. *La probabilidad inducida así definida es una función de probabilidad.*

Demostración. En efecto, tendremos que comprobar que cumple los tres axiomas de la probabilidad; esto es:

- $P'(B) \geq 0$, $\forall B \in \boldsymbol{\beta}$
- $P'(E) = 1$
- Si $B_1, \ldots, B_n, \ldots$ es una sucesión de borelianos disjuntos dos a dos $B_i \cap B_j = \emptyset$, $\forall i \neq j$, entonces: $P'\left(\bigcup_{i=1}^{\infty} B_i\right) = \sum_{i=1}^{\infty} P'(B_i)$.

Los dos primeros supuestos son elementales [¿Por qué?]

Por otra parte, al ser X una v.a., se cumple que la imagen inversa de cualquier boreliano es un elemento de $\boldsymbol{\mathcal{A}}$; por tanto, $X^{-1}(B_i) = A_i \in \boldsymbol{\mathcal{A}}$, $\forall i = 1, \ldots, n, \ldots$. Puesto que la colección de borelianos son incompatibles dos a dos, existe en $\boldsymbol{\mathcal{A}}$ una sucesión $A_1, \ldots, A_n, \ldots$, también disjuntos [¿Por qué?], tales que:

$P'\left(\bigcup_i B_i\right) = P\left(X^{-1}\left(\bigcup_i B_i\right)\right) = P\left(\bigcup_i \left(X^{-1}(B_i)\right)\right) = P\left(\bigcup_i A_i\right)$ y por ser P una función de probabilidad y los A_i disjuntos se tiene:

$P\left(\bigcup_i A_i\right) = \sum_i P(A_i) = \sum_i P\left(X^{-1}(B_i)\right) = \sum_i P'(B_i)$ lo cual concluye la justificación.

□

En lo que sigue no haremos distinción terminológica entre la probabilidad inicial y la inducida.

Volviendo al ejemplo anterior de individuos activos e inactivos, es fácil comprobar que la magnitud definida cumple la definición de variable aleatoria. En efecto, la σ-álgebra $\boldsymbol{\mathcal{A}}$ en este caso viene determinada por: $\boldsymbol{\mathcal{A}} = \{\{\emptyset\}, \{A\}, \{NA\}, \{E\}\}$ (incluye todas las uniones, intersecciones y complementarios de sus elementos); luego el espacio $(E, \boldsymbol{\mathcal{A}})$ es un espacio probabilizable.

Sobre este espacio pueden establecerse infinitas medidas de probabilidad, por ejemplo: $P(A) = 0,75$ y $P(NA) = 0,25$. (P es una función definida sobre todos los elementos de $\boldsymbol{\mathcal{A}}$, evidentemente $P(\emptyset) = 0$ y $P(E) = 1$). De esta forma la terna $(E, \boldsymbol{\mathcal{A}}, P)$ constituye un espacio de probabilidad.

Para comprobar que X es una v.a. tendremos que ver que la imagen inversa de todo intervalo de la forma $(-\infty, x]$ pertenece a $\boldsymbol{\mathcal{A}}$; en efecto,

- $\forall x < 0,\ X^{-1}(-\infty, x] = \emptyset \in \boldsymbol{\mathcal{A}}$
- $\forall x \in [0, 1),\ X^{-1}(-\infty, x] = \{NA\} \in \boldsymbol{\mathcal{A}}$
- $\forall x \geq 1,\ X^{-1}(-\infty, x] = E \in \boldsymbol{\mathcal{A}}$ [¿por qué?]

Luego queda justificado que X es una variable aleatoria.

Especifiquemos ahora su función de probabilidad inducida. Los intervalos de la forma $(-\infty, x]$ que debemos estudiar se reducen a los tres casos anteriores, y la probabilidad inducida se cuantificará como sigue:

- $\forall x < 0,\ P'(-\infty, x] = P\left(X^{-1}(-\infty, x]\right) = P(\emptyset) = 0$
- $\forall x \in [0, 1),\ P'(-\infty, x] = P\left(X^{-1}(-\infty, x]\right) = P(NA) = 0,25$
- $\forall x \geq 1,\ P'(-\infty, x] = P\left(X^{-1}(-\infty, x]\right) = P\left(X^{-1}(-\infty, 0]\right) + P\left(X^{-1}(0, 1]\right) + P\left(X^{-1}(1, x]\right) = 0 + P(NA) + P(A) = 0 + 0,25 + 0,75 = 1$

La probabilidad inducida está inferida por la probabilidad inicial (podría ser otra, por ejemplo $P(A) = 0,5$ y $P(NA) = 0,5$) y por la variable aleatoria (así la v.a. que asignase $X(A) = 10$ y $X(NA) = -10$, induciría una nueva probabilidad).

Es interesante tener presente que la función de probabilidad es una función de conjunto; esto es, está definida sobre conjuntos, hecho que dificulta su manipulación matemática (sin ir más lejos, las funciones de conjunto no son representables). La introducción de la variable aleatoria y la probabilidad inducida permiten transformar esta función en otra que reduce esos conjuntos abstractos a intervalos de la recta real. Sin embargo, y a pesar de la considerable simplificación que representa, la manejabilidad de la función de probabilidad inducida sigue siendo escasa; como veremos en epígrafes posteriores resulta deseable pasar de este tipo de funciones a otra que sea una función de punto definida sobre números reales.

Las magnitudes aleatorias admiten una clasificación según los valores que pueden adoptar y así siguiendo criterios análogos a las variables estadísticas distinguimos las *variables discretas* de las *continuas*.

El caso de *variables discretas* se corresponde con aquellas magnitudes cuyo recorrido de valores posibles es finito o infinito numerable: trabajadores afectados por cierto convenio laboral, número de clientes que acuden a una entidad bancaria, población ocupada de cierto sector económico, ...

Cuando en cambio las magnitudes que analizamos pueden tomar un conjunto de valores infinito no numerable, estamos en el caso de *variables continuas*. En esta categoría se incluyen la Renta Nacional de cierto país, el nivel de inflación acumulada en un mes, el consumo de combustible en cierto proceso productivo, ...

Las dos categorías anteriores, aunque son las más frecuentes, no agotan todas las posibilidades. Así, consideremos por ejemplo la variable T="tiempo de espera de un conductor ante un semáforo", cuyo recorrido de valores viene dado por $\{\{0\}, \{[T_1, T_2]\}\}$

Esta magnitud aleatoria pertenece a la categoría de variables que podemos denominar mixtas, ya que adoptaría un primer tramo de recorrido discreto (valor $T = 0$ para aquellos casos en los que el semáforo está en verde y por tanto el conductor no debe esperar) y en los restantes casos se situaría en el tramo continuo representado por el intervalo $[T_1, T_2]$, cuyos extremos recogerían respectivamente los tiempos de espera mínimo (T_1, que indicaría el tiempo de reacción del conductor cuando el semáforo cambia inmediatamente a verde) y máximo (T_2, tiempo total de espera si el semáforo acaba de cambiar a rojo).

Aunque este último tipo de variable mixta es poco frecuente, merece ser tenido en consideración y podrían encontrarse algunas ilustraciones del mismo en la vida diaria (el tiempo que un individuo se ve obligado a esperar en la consulta de un médico, las tarifas de algunos servicios telefónicos o eléctricos, las ganancias obtenidas con ciertos juegos de azar...).

Los sucesos aleatorios se caracterizan por "poder ser" y no por "ser"; esta "potencialidad" es la diferencia básica entre una variable estadística y una aleatoria y entre sus correspondientes valores.

Podríamos plantearnos entonces si, una vez observada una variable aleatoria, ésta se transforma en estadística por el simple hecho de pasar de futuro a pasado. Evidentemente, la respuesta es negativa ya que la diferencia entre ambas categorías entraña algo más, referido al hecho de que la variable estadística se supone exenta de incertidumbre a diferencia de la variable aleatoria cuyos valores pudieron haber sido otros (sustituimos la certeza por la posibilidad o grados de posibilidad).

Ahora bien, una vez observada cierta variable aleatoria, si nos abstraemos de la incertidumbre que rodea a sus valores y los tomamos como ciertos, entonces podríamos efectuar sobre los mismos un estudio de tipo descriptivo.

Cuando desarrollamos un análisis sobre variables estadísticas, el ámbito se denomina estadística descriptiva. Teniendo en cuenta la identificación anterior, la estadística descriptiva también se puede desarrollar sobre los valores de variables aleatorias, reservando en este caso la probabilidad para el proceso de inducción posterior de resultados.

2.2. Distribución de probabilidad de una variable aleatoria

La descripción de una variable estadística se lleva a cabo mediante sus valores y las frecuencias con las que toma los mismos.

Para las v.a. tenemos que sustituir los valores y las frecuencias por el rango hipotético de valores que puede tomar y las probabilidades asociadas.

El rango de la variable se determina en función de lo que conceptualmente mida esa magnitud. Por lo que se refiere a la función de probabilidad, ésta podría definirse como:

$$p : x \in \Re \rightarrow p(x) = P(X = x) = P\left(\{w \in E / X(w) = x\}\right) \in [0, 1]$$

La cuantificación de esta probabilidad puntual no siempre tiene sentido, dependiendo la descripción probabilística de una variable de su carácter discreto o continuo.

Consideremos dos características asociadas a cierta entidad bancaria durante el próximo año: número de empleados por sucursal y volumen de beneficios de la misma. Dado que ambas magnitudes van referidas al futuro serán variables aleatorias. La primera de ellas, discreta, quedaría descrita si consiguiéramos aproximar mediante algún método la probabilidad de que el número de empleados de una sucursal sea 1, 2, ..., procedimiento que sin embargo no es generalizable al caso de los beneficios, como consecuencia de su carácter continuo.

En efecto, existen infinitas posibilidades de beneficio que son no numerables. Cada una de ellas tiene cierta posibilidad de ocurrencia -en principio todas ellas positivas-

y, por pequeñas que sean estas cantidades, su suma sería infinito, no verificando el segundo axioma de la probabilidad.

Supongamos que los beneficios pueden oscilar entre a y b, $a \leq X \leq b$, y cualquier valor de ese recorrido es posible, $P(x) > 0$, $\forall x \in [a, b]$. Si denotamos por p el ínfimo de estos valores:

$$p = inf\,\{P(x) : x \in [a, b]\}$$

entonces: $\sum_x P(x) \geq \sum_x p = \infty$

La justificación intuitiva de este comportamiento radica en que en el primer caso, cuando las variables son discretas, la suma está indexada en un subconjunto de los números naturales, mientras que en el segundo caso la suma se realiza sobre un campo continuo de valores.

2.2.1. Función de distribución

La probabilidad inducida asigna a cada intervalo de la forma $(-\infty, x]$ un valor de probabilidad que depende de x, y representa la probabilidad acumulada hasta ese valor. Así pues, podemos definir una función real de variable real, que a cada x le asigne su probabilidad acumulada. Esta función, que se denota por F, se denomina *función de distribución.*

Definición 2.4. La *función de distribución* de una v.a. X es una función definida como:

$$F : x \in \Re \to F(x) = P\left(X^{-1}(-\infty, x]\right) = P\left((-\infty, x]\right) = P(X \leq x) \in [0, 1]$$

Proposición 2.2. *Una función de distribución cumple las siguientes propiedades:*

1. *F es monótona no decreciente: $x_1 < x_2 \Rightarrow F(x_1) \leq F(x_2)$*
2. *$\lim_{x \to -\infty} F(x) = 0$ y $\lim_{x \to +\infty} F(x) = 1$*
3. *F es continua a la derecha. (Algunos autores, consideran como sistema de generadores de la σ álgebra $\boldsymbol{\beta}$ intervalos de la forma $[x, +\infty)$, en cuyo caso la continuidad será por la izquierda).*

Estas tres propiedades caracterizan a las funciones de distribución, de forma que toda función que satisfaga dichas propiedades será la función de distribución de una cierta variable aleatoria.

Demostración. Vamos a justificar estas propiedades:

Comencemos por comprobar que F es monótona no decreciente: Sean x_1 y x_2 dos números reales cualesquiera con $x_1 < x_2$, y denotemos por $[X \leq x]$ el suceso $\{w \in E / X(w) \leq x\}$. Podemos descomponer el suceso $[X \leq x_2]$ de la siguiente forma:

$$[X \leq x_2] = [X \leq x_1] \cup [x_1 < X \leq x_2]$$

es decir, mediante la unión de dos sucesos incompatibles; aplicando la función de probabilidad a los dos miembros, tendremos:

$$P([X \leq x_2]) = P([X \leq x_1]) + P([x_1 < X \leq x_2])$$

y en el segundo miembro los dos sumandos son no negativos, por lo cual se tiene: $P([X \leq x_2]) \geq P([X \leq x_1])$ o equivalentemente: $F(x_1) \leq F(x_2)$.

Comprobemos ahora la propiedad segunda: $\lim_{x \to +\infty} F(x) = 1$ y $\lim_{x \to -\infty} F(x) = 0$

$$\lim_{x \to +\infty} F(x) = F(+\infty) = P([X \leq \infty]) = P(\{w \in E / X(w) \leq +\infty\}) = P(E) = 1$$

$$\lim_{x \to -\infty} F(x) = F(-\infty) = P([X \leq -\infty]) = P(\{w \in E / X(w) \leq -\infty\}) = P(\emptyset) = 0$$

estas justificaciones son intuitivas aunque no totalmente rigurosas (obsérvese que hemos tratado el campo infinito como un punto, cuando tendríamos que tomar límites y calcular $\lim_{x \to +\infty} P([X \leq x])$.

Para la demostración completa de esta propiedad tendremos que considerar una sucesión arbitraria de valores x_n, con $x_n < x_{n+1}$ y $\lim_{n \to \infty} x_n = \infty$; sin pérdida de generalidad podemos suponer que esta sucesión es la de los números naturales $\{0, 1, \ldots, n, \ldots\}$ que cumplen las condiciones anteriores.

El suceso $[X \leq n]$ puede ser expresado como una unión de sucesos incompatibles de la forma siguiente:

$$[X \leq n] = [X \leq 0] \cup [0 < X \leq 1] \cup \cdots \cup [n-1 < X \leq n]$$

y como la probabilidad de la unión es la suma de probabilidades, tenemos:

$$F(n) = P([X \leq n]) = P([X \leq 0]) + P([0 < X \leq 1]) + \cdots + P([n-1 < X \leq n])$$

Por otra parte, el suceso seguro puede ser expresado como unión infinita de los sucesos disjuntos anteriores: $E = [X \leq 0] \bigcup_i [i-1 < X \leq i]$, por lo tanto:

$$P(E) = P([X \leq 0]) + \sum_i P([i-1 < X \leq i])$$

de donde se determina que la serie que ahí aparece es convergente y por tanto, $\forall \delta > 0$, podemos encontrar un n suficientemente grande tal que:

$$P([X \leq n]) = P([X \leq 0]) + P([0 < X \leq 1]) + \cdots + P([n-1 < X \leq n]) > 1 - \delta$$

Por tanto queda demostrado que $\lim_{n \to \infty} F(n) = 1$ y de forma general $\lim_{x \to +\infty} F(x) = 1$.

De forma análoga se demuestra $\lim_{x \to -\infty} F(x) = 0$.

Finalmente, pasemos a demostrar la propiedad tercera: F es continua a la derecha, esto es: $\lim_{h \to 0^+} F(x+h) = F(x)$

En efecto, podemos considerar la siguiente descomposición: $[X \leq x+h] = [X \leq x] \cup [x < X \leq x+h]$

Calculando la probabilidad en ambos miembros y sustituyendo la función de distribución, se tiene: $F(x+h) = F(x) + P([x < X \leq x+h])$ con lo cual: $\lim_{h \to 0^+} F(x+h) = F(x) + \lim_{h \to 0^+} P([x < X \leq x+h])$. Y el último sumando del segundo término es nulo porque ese suceso se reduce al suceso imposible [¿por qué?], lo cual concluye la demostración.

□

Si consideramos la variable aleatoria discreta que recoge la situación laboral de un individuo

$$X = \begin{cases} 1 & \text{si w es activo} \\ 0 & \text{si w es inactivo} \end{cases}$$

las probabilidades acumuladas serán nulas para todo valor de X inferior a 0. Para todo x positivo e inferior a 1 el único valor factible menor que 1 es el 0 ($P(X = 0) = 0,25$), por lo cual la probabilidad acumulada será 0,25, y finalmente para todo $x \geq 1$, los valores factibles menores o iguales a x son el 0 y el 1, por lo que la probabilidad acumulada será $0,25 + 0,75 = 1$.

Las características de la función de distribución (f.d.) son análogas para toda variable discreta: es una función escalonada que está definida en toda la recta real, antes del primer valor posible de la variable $F(x)$ es nula, la función experimenta un salto en cada uno de los valores factibles de la variable, la altura de dicho salto es igual a la probabilidad puntual de ese valor (que, por ser no negativa, da lugar a una función monótona no decreciente) y, por último, para x no inferiores al último valor de la variable, $F(x)$ permanece constante e igual a la unidad.

Como podemos observar en el comentario anterior quedan reflejadas las propiedades señaladas de la función de distribución. En este ejemplo se trata de una función de distribución no continua, con tantos puntos de discontinuidad como valores puede tomar la variable (para v.a. discretas serán un número finito o infinito numerable). Además cuando avanzamos hacia un valor de la variable por la izquierda la f.d. toma un valor constante que no coincide con el del punto:

$$\lim_{\delta \to 0^+} F(x_i - \delta) \neq F(x_i)$$

y la diferencia entre las dos cantidades será precisamente la probabilidad asociada a ese valor. Por tanto la función no es continua por la izquierda. [¿Lo es por la derecha?, razónese la respuesta] [¿en qué condiciones la función sería continua?]

La introducción en el modelo matemático asociado a un experimento aleatorio del concepto de variable aleatoria tiene como objetivo primordial facilitar el manejo de la probabilidad asociada al experimento. Este objetivo se logra al poder pasar de la probabilidad definida inicialmente sobre la σ-álgebra de los sucesos (que es una función real de conjuntos de naturaleza variada) a la probabilidad inducida (que es una función real de conjuntos reales, debido a la consideración de la σ-álgebra de Borel sobre $\Re$). Además, esta probabilidad inducida puede ser caracterizada mediante la función de distribución, que es una función de punto (función real de variable real) y por tanto más sencilla de interpretar y manejar desde un punto de vista matemático.

En general, las representaciones gráficas de las funciones de distribución correspondientes a v.a. discretas y continuas son las que se recogen en los gráficos siguientes: [Figura 2.3]

Estas representaciones gráficas nos sugieren establecer nuevas definiciones de variables aleatorias discretas y continuas. Podemos definir una v.a. discreta como aquélla cuya función de distribución es escalonada y una v.a. continua como aquélla cuya f.d. también lo es.

Estas últimas definiciones nos permiten comprobar gráficamente la existencia de variables mixtas, que se corresponderán con aquellas funciones de distribución que sin ser continuas tampoco sean escalonadas. [Figura 2.4]

Figura 2.3.: Función de Distribución

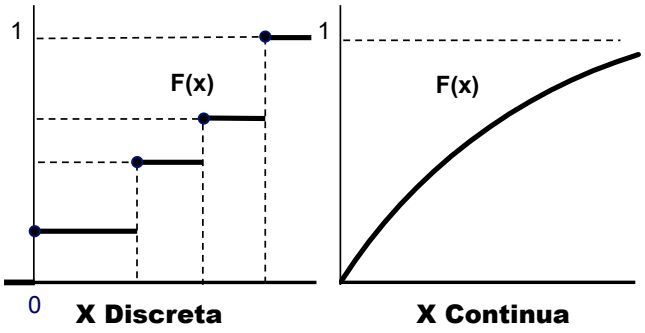
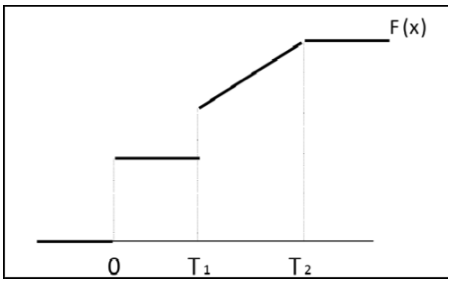


Figura 2.4.: Función de distribución de una v.a. mixta



El concepto de probabilidad acumulada resulta de gran interés ya que a menudo es muy útil conocer la probabilidad de que una magnitud alcance valores hasta uno dado o, de forma complementaria, valores superiores a él $(1 - F(x))$. Parece claro por tanto que la función de distribución proporciona una respuesta a preguntas del tipo ¿cuál es la probabilidad de que una sucursal elegida al azar tenga a lo sumo 8 empleados? o, de modo complementario, ¿con qué probabilidad encontraremos un volumen de beneficios superior a cierta cifra?

2.2.2. Probabilidades de intervalos

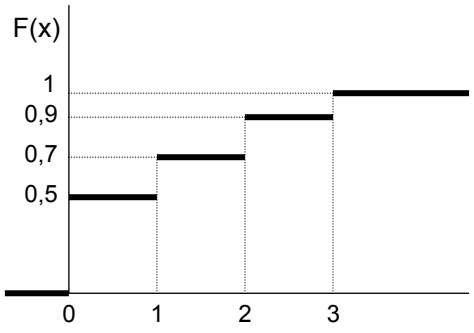
En algunas ocasiones resulta interesante conocer la probabilidad asociada a cierto recorrido (valores entre 4 y 6 empleados) o incluso una probabilidad puntual (¿cuál es la probabilidad de que una sucursal tenga exactamente 10 empleados?).

El primero de estos interrogantes aparece directamente conectado a la función de distribución, ya que para cualquier intervalo real $(a, b]$ se tiene:

$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$$

De este modo, a partir de los gráficos anteriores podemos identificar las probabilidades de intervalos, que vendrían representadas en el eje de ordenadas por la diferencia entre los correspondientes valores de la f.d.

Figura 2.5.: Función de distribución discreta



Cabe preguntarse si este esquema de razonamiento es extrapolable hasta llegar a identificar probabilidades de puntos concretos:

$$P(X=x)=\lim_{\delta\to 0}P(x-\delta<X\leq x+\delta)=\lim_{\delta\to 0}\left[F(x+\delta)-F(x-\delta)\right]=F(x^{+})-F(x^{-})=F(x)-F(x^{-})$$

debiéndose la última igualdad al hecho de que F es continua a la derecha.

2.2.3. Función de probabilidad

Cuando la v.a. es discreta, la función de distribución es escalonada y para dos valores consecutivos, x_{i-1}, x_i, se verifica: $F(x)=F(x_i-1)$, $\forall x\in[x_{i-1},x_i)$.

Entonces, a partir de la relación anterior, se tiene:

$$\begin{array}{ll} P(X=x)=F(x)-F(x^{-})=F(x_{i-1})-F(x_{i-1})=0 & si\, x_{i-1}<x<x_i \\ P(X=x_i)=F(x_i)-F(x_i^{-})=F(x_i)-F(x_{i-1})=P(x_i) & si\, x=x_i \end{array}$$

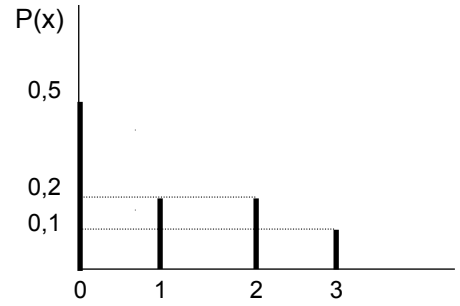
Por tanto hemos comprobado que a partir de la función de distribución podemos obtener la probabilidad de los valores de una v.a. discreta.

Consideremos a modo de ilustración la función de distribución asociada a la variable aleatoria discreta X="Número de medallas que obtendrá un deportista de élite en ciertos campeonatos" cuya representación gráfica aparece recogida en la figura 2.5.

Esta función permite calcular probabilidades acumuladas hasta cierto número de medallas. Así, a través de la gráfica de $F(x)$ podemos responder fácilmente a preguntas del tipo ¿cuál es la probabilidad de que un deportista no obtenga más de 2 medallas?, cuya respuesta $P(X\leq 2)=0,9$ se obtiene con sólo consultar la ordenada asociada al punto $x=2$ en el diagrama.

De modo análogo podríamos obtener probabilidades de intervalos. Por ejemplo, la probabilidad de que un deportista logre en la competición entre 1 y 3 medallas viene dada por $P(1\leq X\leq 3)=P(0<X\leq 3)=F(3)-F(0)=0,5$.

Figura 2.6.: Función de probabilidad



Dado que X es claramente discreta ¿cómo podríamos obtener la probabilidad de un valor concreto, digamos 3 medallas? Acudiendo a la expresión anteriormente obtenida para $P(X = x_i)$ se tiene:

$$P(X = 3) = F(3) - F(3^-) = F(3) - F(2) = 0,1$$

Es posible considerar otras magnitudes aleatorias discretas con recorrido mucho más amplio. Este sería el caso de la variable "Número de votos obtenidos por un partido político en las próximas elecciones municipales", cuya función de distribución sería -con la única salvedad del mayor recorrido- análoga a la anteriormente estudiada y daría así respuesta a preguntas del tipo ¿cuál es la probabilidad de que un partido no exceda los 2.500 votos?, ¿con qué probabilidad se situará entre 3.000 y 5.000 votos? o bien ¿cuál es la probabilidad de lograr exactamente 4.850 votos?

El procedimiento recogido para v.a. discretas, cuyo carácter numerable garantiza la identificación de los valores con probabilidad no nula, permite definir una *función de probabilidad* que asocia a cada valor de X una probabilidad no negativa, resultando evidente que la suma de todas ellas es -con independencia del concepto de probabilidad utilizado- la unidad.

Definición 2.5. La *función de probabilidad* de una variable aleatoria discreta, denominada a menudo *función de cuantía* o *función de masa de probabilidad*, viene dada por una aplicación $p : x \in \Re \to [0, 1]$ que cumple las condiciones:

$$p(x) \geq 0 \; y \; \sum_i p(x_i) = 1$$

La representación gráfica de la función de probabilidad asociada a la variable X="Número de medallas que obtendrá un deportista de élite en ciertos campeonatos" aparece recogida en la figura 2.6, y es análoga a los diagramas de barras utilizados para variables estadísticas discretas.

Proposición 2.3. *La f.d. $F(x)$ de una v.a. X puede expresarse como la suma de valores de probabilidad hasta ese valor.*

Demostración. En efecto, partiendo de la relación: $P(X = x_i) = F(x_i) - F(x_i^-) = F(x_i) - F(x_{i-1}) = p(x_i)$, podemos expresar:

$$F(x_i) = F(x_{i-1}) + p(x_i)$$

y procediendo de forma recursiva:

$$F(x_i) = F(x_{i-2}) + p(x_i) + p(x_{i-1}) = \cdots = F(x_1) + p(x_i) + p(x_{i-1}) + \cdots + p(x_2)$$

teniendo en cuenta por otra parte que para x_1 se cumple $F(x_1) = p(x_1)$, se obtiene:

$$F(x_i) = \sum_{j=1}^{i} p(x_j)$$

□

2.2.4. Función de densidad

Cuando la variable X considerada es continua y por tanto lo es también su función de distribución, se tiene:

$$P(X = x) = F(x) - F(x^-) = 0$$

por lo cual obtenemos que la probabilidad de cualquier punto es nula.

Así pues, en este caso no resulta posible aislar valores ni sus correspondientes probabilidades (la imposibilidad de enumerar los puntos hace que no tenga sentido hablar de la probabilidad de un punto aislado). En cambio, sí podríamos trabajar sobre un intervalo del que cuantificamos su probabilidad (de hecho, este es el modo de actuar en variables estadísticas con datos agrupados).

De esta forma, aunque carezca de sentido aislar la probabilidad de que los beneficios de una sucursal bancaria sean exactamente de 25 millones, sí podríamos considerar la probabilidad del intervalo $(20, 30]$, u otros cualesquiera $(a, b]$; conocida la forma explícita de la función de distribución obtendríamos $P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a)$.

Si dividimos el recorrido de la v.a. en k intervalos y asignamos a cada uno su probabilidad según el método anterior, siguiendo una analogía con la estadística descriptiva, podemos representar gráficamente [figura 2.7] esta distribución de probabilidad mediante un histograma, donde las áreas de los rectángulos son proporcionales a las probabilidades de los correspondientes intervalos que tienen como base.

Si consideramos subdivisiones cada vez más finas de ese recinto la silueta del histograma se va transformando y en el límite cuando la amplitud de los intervalos tiende a 0 podemos suponer que la representación [figura 2.8] corresponde a una cierta función,

Figura 2.7.: Histogramas de probabilidad

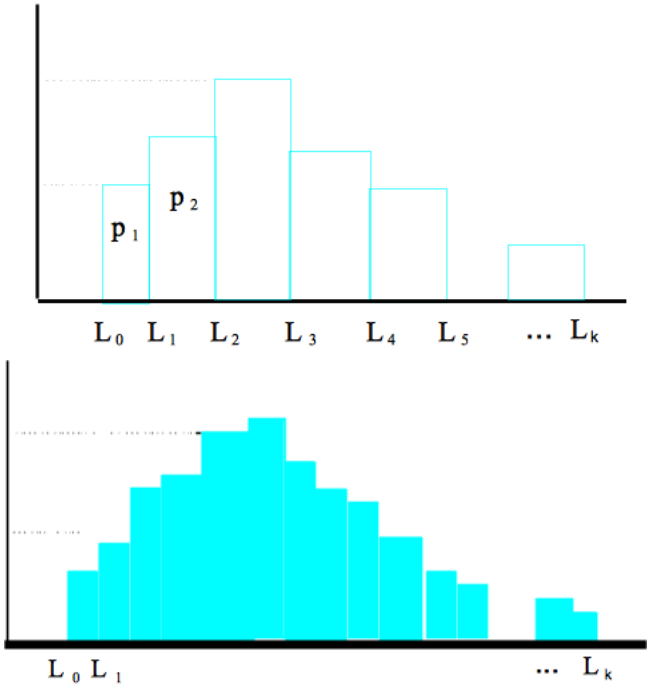
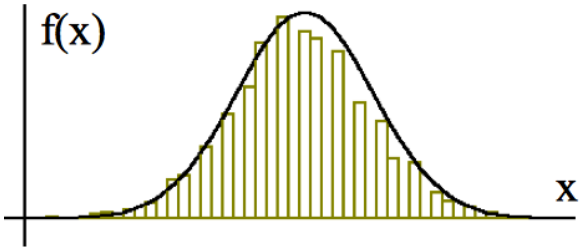


Figura 2.8.: Función de densidad



f, denominada función de densidad, y que a cada valor x real le asigna su densidad de probabilidad.

Para un intervalo cualquiera de amplitud $2h$, la probabilidad del intervalo viene dada por:

$$P(x-h < X \leq x+h) = F(x+h) - F(x-h)$$

Por otra parte, si la función $f(x)$ es integrable, entonces el teorema del valor medio del cálculo integral, nos dice que existe un punto intermedio y, de manera que el área de ese intervalo se puede expresar como: $P(x-h < X \leq x+h) = f(y) \cdot 2h$, de donde:

$$f(y) = \frac{P(x-h < X \leq x+h)}{2h}$$

y tomando límites cuando h tiende a cero, podemos escribir:

$$f(x) = \lim_{h \to 0} \frac{P(x-h < X \leq x+h)}{2h}$$

expresión que justifica el nombre que se le asigna a esta función: cuantifica la masa de probabilidad de un intervalo en relación a su amplitud, cociente que responde a la idea de "densidad", ya que para cada x puede ser interpretada como la densidad de probabilidad en un entorno infinitesimal de ese punto.

Proposición. *Dada la f.d. $F(x)$ de una cierta v.a. X, la función de densidad se obtiene como la derivada de $F(x)$.*

Demostración. Partiendo de la expresión de la función de densidad, si ahora expresamos la probabilidad del intervalo considerado en términos de la f.d., se tiene:

$$f(x) = \lim_{h \to 0} \frac{P(x-h < X \leq x+h)}{2h} = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h} = F'(x) = \frac{dF(x)}{dx}$$

con lo cual queda justificado que la función de densidad se obtiene por derivación de la función de distribución (obsérvese que esta relación $f(x) = F\,'(x)$ sería una extensión al caso continuo de la anteriormente vista para variables discretas $P(X = x_i) = F(x_i) - F(x_{i-1})$). □

La función $f(x)$ que recoge la *densidad de probabilidad* será no negativa pues, según comprobamos en las propiedades de la f.d., $F(x)$ es una función monótona no decreciente y por tanto su derivada (si existe) no puede ser negativa. Desde un punto de vista aún más intuitivo, si fuese negativa podríamos encontrar algún intervalo -aunque tuviese amplitud infinitesimal- con probabilidad negativa.

Consideremos por ejemplo una magnitud aleatoria continua X que cuantifica la distancia (expresada en miles de km.) recorrida semanalmente por un viajante, cuya función de distribución viene dada por la expresión:

$$F_X(x) = \begin{cases} 0 & si\, x < 2 \\ \frac{1{,}25x - 2{,}5}{x} & si\, 2 \leq x < 10 \\ 1 & si\, 10 \leq x \end{cases}$$

a partir de la misma, derivando, se obtiene fácilmente la función de densidad:

$$f(x) = \begin{cases} \frac{2{,}5}{x^2} & si\, 2 \leq x < 10 \\ 0 & \text{en otro caso} \end{cases}$$

[Efectuar los cálculos necesarios]

De igual forma que para v.a. discretas es posible obtener $F(x)$ mediante agregación de la función de probabilidad, en el caso de las variables continuas podemos obtener la f.d. $F(x)$ a partir de la función de densidad. En este caso la suma será sustituida por una integral, obteniéndose:

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Demostración. Bastaría tener en cuenta que podemos expresar $F(x)$ como una integral de Stieltjes-Lebesgue $F(x) = \int_{-\infty}^{x} dF(t)$ y que cuando la variable es continua se cumple: $dF(x) = f(x)dx$; por tanto se verificará:

$$F(x) = \int_{-\infty}^{x} dF(t) = \int_{-\infty}^{x} f(t)dt$$

Obsérvese que para variables discretas, las diferencias $dF(x)$ son nulas en casi todos los puntos salvo en los posibles valores de la variable, donde $dF(x_i) = F(x_i) - F(x_{i-1}) = p(x_i)$. □

Las propiedades de la f.d., $F(x) \geq 0$ y $\lim_{x\to\infty} F(x) = 1$ trasladan a la función de densidad las características de no negatividad y área unitaria $\int_{-\infty}^{\infty} f(x)dx$=1.

Por otra parte, hemos visto cómo se expresa la probabilidad de un intervalo en términos de la f.d.; sustituyendo ahora ésta por la función de densidad, se tendrá:

$$P(a < X \leq b) = F(b) - F(a) = \int_{-\infty}^{b} f(x)dx - \int_{-\infty}^{a} f(x)dx = \int_{a}^{b} f(x)dx$$

Definición 2.6. Llamamos *función de densidad*, si existe, a una aplicación $f : x \in \Re \to \Re^+$, que cumple las condiciones:

- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)dx = 1$
- $\forall a, b \in \Re$, $-\infty < a < b < +\infty$ se tiene: $P(a < X \leq b) = \int_{a}^{b} f(x)dx$

En la definición anterior hemos introducido la puntualización "si existe". Es necesario señalar que, buscando una mayor claridad intuitiva, en los párrafos anteriores hemos cometido algunos abusos de lenguaje.

En efecto, el hecho de que la f.d. de una v.a. continua también sea continua no implica que sea diferenciable ni que se pueda expresar como la integral de la función de densidad. La existencia de esta integral queda garantizada si le imponemos a $F(x)$ alguna otra restricción, como la de *continuidad absoluta*, en cuyo caso tendríamos que distinguir entre v.a. continuas y *absolutamente continuas*, siendo estas últimas las que permiten enlazar la f.d. con la de densidad.

Por otra parte, el teorema fundamental del cálculo integral establece que si $F(x) = \int_{-\infty}^{x} f(t)dt$, entonces $F(x)$ es una función continua y $F'(x) = f(x)$ en todo punto x de continuidad de f.

Así pues, para obtener la f.d. $F(x)$ a partir de la función de densidad será necesario que la primera sea absolutamente continua, y la función de densidad puede obtenerse como derivada de $F(x)$ sólo en los puntos de continuidad de la primera.

Una vez efectuadas estas puntualizaciones debemos, sin embargo, señalar que todos los modelos continuos considerados a lo largo de este texto son también absolutamente continuos. Por tanto no haremos tal distinción, utilizando en un abuso de lenguaje la denominación de v.a. continuas aunque en realidad nos estaremos refiriendo a las absolutamente continuas, que por tanto tienen siempre garantizada la existencia de la función de densidad.

Volviendo al ejemplo anterior, dada la función de densidad:

$$f(x) = \begin{cases} \frac{2,5}{x^2} & si\, 2 \leq x < 10 \\ 0 & \text{en otro caso} \end{cases}$$

es posible obtener mediante integración de esta función de densidad la probabilidad acumulada o función de distribución que, para cada recorrido kilométrico x ($2 \leq x < 10$), viene dada por expresión $F_X(x) = \frac{1,25x-2,5}{x}$ [comprobar cómo se ha obtenido esta función] [¿cuál es su valor para $x = 10$?]

La relación entre $f(x)$ y $F(x)$ permite calcular la probabilidad de un intervalo cualquiera $(a, b]$ mediante la expresión:

$$P(a < X \leq b) = F(b) - F(a) = \int_a^b f(x)dx$$

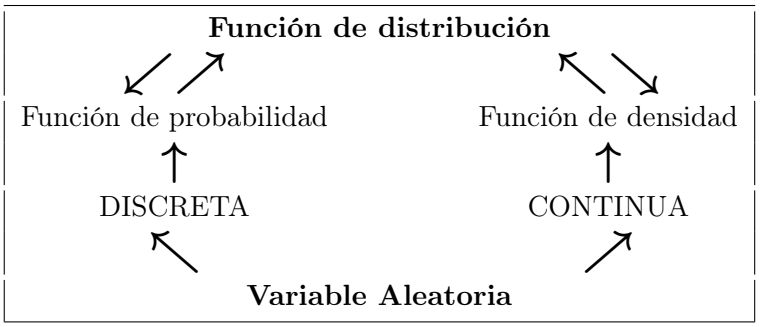
Así, por ejemplo, la probabilidad de que el recorrido del viajante oscile entre 2.000 y 4.000 km. semanales vendría dada por el valor $P(2 < X \leq 4) = 0,625$. Puede observarse que dicha probabilidad supera a la del recorrido entre 6.000 y 8.000 km., intervalo de igual amplitud que el anterior, pero que sin embargo resulta menos probable como consecuencia de la propia densidad de probabilidad $f(x)$ [comprobar que $P(6 < X \leq 8) \approx 0,10417$].

Una vez estudiadas las distintas vías a través de las cuales puede venir descrita una variable aleatoria, presentamos una síntesis de las mismas:

2.2.5. Variables aleatorias relacionadas: Cambio de variable

A menudo las magnitudes económicas se hallan relacionadas entre sí. Por ejemplo, supongamos que el número de empleados de una oficina bancaria es una v.a.; si cada

Figura 2.9.:



empleado trabaja h horas semanales, la variable que recoge el número total de horas es también aleatoria y vendría dada por $X^* = hX$.

De modo análogo, si la plantilla de cada sucursal aumentase en dos empleados, la v.a. que indica el nuevo número de empleados sería ahora $X' = X + 2$.

En cualquiera de estas dos situaciones nos enfrentamos a un cambio de variable. La magnitud aleatoria definida aparece conectada con la inicial, por lo cual resulta posible conocer su distribución de probabilidad a partir de la información sobre X. En concreto, para los ejemplos anteriores, por tratarse de variables discretas, bastaría con identificar el recorrido de valores de las nuevas variables X^* y X' y sus correspondientes probabilidades, que se obtienen a partir de las asociadas a X.

En los ejemplos propuestos, las transformaciones de X, X^* y X' vendrían descritas en los términos siguientes,

$F_{X^*}(x) = P(X^* \leq x) = P(hX \leq x) = P(X \leq \frac{x}{h}) = F_X(\frac{x}{h})$

$F_{X'}(x) = P(X' \leq x) = P(X + 2 \leq x) = P(X \leq x - 2) = F_X(x - 2)$

Es evidente que a menudo aparecen cambios de variable más sofisticados que los anteriormente descritos. Así, podríamos encontrarnos con cambios por tramos (por ejemplo, aumentos de 1, 2 o 3 empleados por sucursal según sus niveles iniciales) en cuyo caso la deducción de la distribución de la nueva variable sería más complicada.

Si consideramos ahora una magnitud continua (por ejemplo, los beneficios empresariales Y) y asumimos que la entidad debe pagar en concepto de impuestos un 15 % de sus beneficios tendríamos la variable $Y^* = 0,15Y$, cuya distribución de probabilidad podría ser obtenida a partir de la idea de probabilidad acumulada (recogida para los impuestos mediante la función de distribución de Y^*), conectando esta expresión con la probabilidad acumulada de los beneficios:

$$F^*(y^*) = P(Y^* \leq y^*) = P(0,15Y \leq y^*) = P\left(Y \leq \frac{y^*}{0,15}\right) = F\left(\frac{y^*}{0,15}\right)$$

En definitiva, se trataría en este caso de buscar qué recorrido de la variable beneficios se corresponde con uno dado de los impuestos, tarea sencilla siempre que la expresión de cambio de variable sea una función monótona y continua.

En el procedimiento genérico del cambio de variable distinguiremos según se trate de v.a. discretas o continuas.

1) Supongamos que X es una v.a. discreta que puede tomar un conjunto de valores $x_1, \ldots, x_n, \ldots$ con probabilidades respectivas $p, \ldots, p_n, \ldots$. Sea g una función definida en el conjunto imagen de X tal que $g(X)$ es una nueva v.a.:

$$E \to X(E) \subset \Re \to g(X(E)) \subset \Re$$

Entonces la variable $Y = g(X)$ es también discreta, y quedará especificada cuando conozcamos los valores que puede tomar y sus probabilidades respectivas. Por lo que se refiere a los valores, éstos serán $y_1 = g(x_1), \ldots, y_n = g(x_n), \ldots$ y sus probabilidades se cuantifican de la forma siguiente:

$$P(Y = y_i) = P\left(\{x_i \in \Re / g(x_i) = y_i\}\right) = P\left(\{x_i / x_i \in \{g^{-1}(y_i)\}\}\right) = P\left(\{x_i / x_i \in C_i\}\right) = \sum_{x_i \in C_i} p(x_i)$$

para obtener la última relación téngase en cuenta que C_i es un conjunto formado por un número finito o numerable de puntos (estos puntos son disjuntos y por tanto la probabilidad de C_i es la suma de las probabilidades de los puntos que lo integran).

A modo de ejemplo, reconsideremos la variable anterior X="Plantilla de una sucursal bancaria" y su transformación X^*="Número de horas trabajadas". Según el razonamiento expuesto se obtendría la probabilidad puntual de un número concreto de horas trabajadas como:

$$P(X^* = x_i^*) = P\left(\{x_i \in \Re / hx_i = x_i^*\}\right) = P\left(\left\{x_i / x_i = \frac{x_i^*}{h}\right\}\right)$$

Dado que la relación entre X y X^* es biyectiva, la distribución de probabilidad de ambas variables será coincidente (esto es, la correspondencia biyectiva entre las magnitudes aleatorias se traslada a sus distribuciones de probabilidad).

Consideremos ahora nuevamente la v.a. X="Número de votos obtenidos por los partidos políticos en las próximas elecciones municipales", a partir de la cual podemos definir Y="Número de concejales obtenidos por un partido político" según criterios del tipo:

$$Y = \begin{cases} 0 & si\, 0 \leq x < 1{,}000 \\ 1 & si\, 1{,}000 \leq x < 2{,}000 \\ 2 & si\, 2{,}000 \leq x < 3{,}000 \\ 3 & si\, 3{,}000 \leq x < 4{,}000 \\ \vdots & \vdots \end{cases}$$

En este caso, la correspondencia es sobreyectiva por lo cual la expresión genérica anterior nos proporcionaría la probabilidad de la variable Y como suma de probabilidades puntuales de varios valores de X.

2) Si la variable X es continua, la transformación $Y = g(X)$ puede ser discreta o continua. En el primer caso, resulta sencillo obtener la función de probabilidad de la nueva variable Y a partir de la función de densidad de X:

$$P(Y = y_j) = \int_{C_j} f(x)dx, \; siendo\, C_j = \{x / g(x) = y_j\}$$

Si por el contrario g es una función continua, será posible -siempre que g tenga inversa- obtener su f.d. a partir de la de X como:

- $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \leq g^{-1}(y)) = F_X\left(g^{-1}(y)\right)$ si $g(x)$ es monótona creciente
- $F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = P(X \geq g^{-1}(y)) = 1 - F_X\left(g^{-1}(y)\right)$ si $g(x)$ es monótona decreciente

Además en determinadas condiciones podemos encontrar una relación entre las funciones de densidad de X y de Y, como pone de manifiesto la siguiente propiedad:

Proposición 2.4. *Sea X una v.a. continua con función de densidad $f(x)$, la cual es estrictamente positiva en un intervalo $[a,b]$. Sea $Y = g(X)$ una transformación monótona y continua en el intervalo $[a,b]$, entonces Y es una v.a. continua cuya función de densidad viene dada por la expresión:*

$$f_Y(y) = \begin{cases} f_X\left(g^{-1}(y)\right)\left|\frac{dg^{-1}(y)}{dy}\right| & si\ y \in g\left([a,b]\right) \\ 0 & en\ otro\ caso \end{cases}$$

La justificación consiste en aplicar la regla de la cadena y distinguir los casos de monotonía creciente y decreciente.

A modo de ilustración de este último caso, retomando la variable X="Distancia kilométrica recorrida semanalmente por un viajante" podemos definir ahora Y="Dietas cobradas por desplazamientos" que viene dada por $Y = 24X$. Dado que se trata de una función monótona creciente de x es posible aplicar la expresión anterior para el cambio de variable, con lo cual se obtiene:

$$f_Y(y) = \begin{cases} f_X\left(\frac{y}{24}\right) \cdot \left(\frac{1}{24}\right) & si\ y \in [24, 240] \\ 0 & \text{en el resto} \end{cases}$$

2.3. Características asociadas a variables aleatorias. Valor esperado y varianza

El estudio de magnitudes económicas se simplifica considerablemente cuando se utilizan algunas características que ayudan a describir su distribución.

Así, en nuestro análisis de las entidades bancarias puede ser interesante resumir la información en datos sintéticos: "el valor esperado de beneficios es de 42 millones" o bien "el 75 % de las oficinas tienen menos de 12 empleados", ...

En definitiva, al igual que en estadística descriptiva estudiábamos un conjunto de medidas útiles para describir, desde distintas ópticas, una variable y que en consecuencia nos informaban sobre su valor central, dispersión, simetría, ... el razonamiento es válido para las magnitudes aleatorias, resultando conveniente definir una serie de características asociadas a su distribución.

En la práctica, estas características llegan a ser los rasgos identificadores de una variable aleatoria, siendo frecuente describir la variable mediante un "perfil" de la misma, que incluye el modelo probabilístico al que se adapta y ciertos parámetros relacionados con sus características de posición y dispersión.

Siguiendo un esquema similar al de las variables estadísticas definiremos dos tipos de características identificadoras de una variable aleatoria: una medida de posición (esperanza) y otras de dispersión (varianza y desviación típica).

El *valor esperado* o *esperanza* de una variable aleatoria se establece como un valor resumen de la misma, obtenido mediante una expresión sintética en la que intervienen tanto los valores de la variable como sus probabilidades.

Definición 2.7. Se define el *valor esperado* o *esperanza matemática* de una v.a. X, que denotamos por $E(X)$ o μ, como el valor, si existe, de la siguiente expresión:

$$E(X) = \mu = \int_{\Re} x dF(x)$$

Se trata de una integral de Stieltjes-Lebesgue, que no siempre será convergente.

Cuando la variable es continua, $dF(x) = f(x)dx$ y en consecuencia su valor esperado se expresará como:

$$E(X) = \mu = \int_{\Re} x f(x) dx$$

[Aplicando esta expresión, compruébese que, para el ejemplo del viajante, se obtiene un recorrido semanal esperado de 4.024 km. $(\mu = 4,024)$].

Para variables discretas, los únicos valores no nulos de las diferencias $dF(x)$ se corresponden con los valores de la variable, para los cuales se obtiene:

$$dF(x_i) = F(x_i) - F(x_{i-1}) = p(x_i) = p_i$$

y en consecuencia el valor esperado para este tipo de variables se transforma en una suma numerable:

$$E(X) = \mu = \sum_i x_i p_i$$

expresión similar a la media aritmética de variables estadísticas donde la probabilidad sustituye a la frecuencia relativa.

[Obténgase, a partir de la expresión anterior, la esperanza para la variable "Número de medallas obtenidas por un deportista", $\mu = 0,9$].

En la concepción frecuentista o frecuencialista de la probabilidad, ésta se define como límite de las frecuencias relativas, luego en el límite media aritmética y esperanza coincidirán. En efecto, cuando se llevan a cabo un número elevado de observaciones de la variable, la media aritmética de estas observaciones se aproxima considerablemente al valor esperado de la misma; en concreto demostraremos en el capítulo 4 que, $\forall \epsilon > 0$ se tiene:

$$\lim_{n \to \infty} P\left(\left|\bar{X} - \mu\right| > \epsilon\right) = 0$$

lo que asegura una convergencia en probabilidad entre ambas características.

A pesar de su denominación de "valor esperado" la esperanza no debe ser interpretada como el valor que esperamos que adopte la variable aleatoria. De hecho, en algunas ocasiones el cálculo de la esperanza da como resultado valores que la variable X no puede adoptar (esto sucede en el ejemplo de las medallas ganadas por el deportista, con esperanza de 0,9) e incluso puede ocurrir que la esperanza no exista, esto es, que se obtenga una suma infinita o una integral no convergente.

La esperanza matemática no siempre permite resolver de forma adecuada algunos problemas relativos a la ganancia esperada de un juego. La "paradoja de San Petersburgo" -denominada así por haber aparecido publicada en la *Revista de la Academia de San Petersburgo-* pone de manifiesto la diferencia entre esperanza matemática y esperanza moral.

Esta paradoja, que dio origen a numerosas discusiones en el siglo XVIII, puede ser planteada como sigue: "Dos jugadores A y B participan en un juego consistente en lanzar una moneda y apostar sobre el resultado considerado favorable (cara, por ejemplo). Las condiciones del juego son: el número de lanzamientos es ilimitado, el juego concluye cuando aparece una cara por primera vez y la apuesta se va duplicando con el número de lanzamientos. Así, por ejemplo, A pagará a B una cantidad x si sale cara en la primera tirada, $2x$ si no sale en la primera pero sí en la segunda, $4x$ si no sale hasta la tercera, etc.

Como consecuencia, la probabilidad que tiene B de ganar una cantidad x es $\frac{1}{2}$, de ganar $2x$ es $\frac{1}{4}$, la de $4x$ es $\frac{1}{8}$ y en general la probabilidad de ganar $2^n x$ es $\left(\frac{1}{2}\right)^{n+1}$. Su ganancia esperada en n pruebas vendrá dada por la expresión:

$$E(X) = \frac{1}{2}x + \frac{1}{2^2}2x + \frac{1}{2^3}2^2x + \cdots + \frac{1}{2^{n+1}}2^n x = nx\frac{1}{2}$$

que puede llegar a ser infinito si no limitamos el número de lanzamientos n y sin embargo ningún jugador estaría dispuesto a exponer en un juego como el descrito una suma importante de dinero, poniéndose así de relieve la limitación de la esperanza matemática.

Daniel Bernoulli introdujo en 1738 el concepto de "esperanza moral", germen de la moderna teoría de la utilidad marginal y donde además ya expresaba el principio de la utilidad marginal decreciente. Este concepto fue también analizado por Laplace (1814) quien comenta algunas aplicaciones del criterio de la expectativa moral o utilidad esperada.

La inexistencia del valor esperado en una v.a. continua se pone de manifiesto con la distribución de Cauchy, cuya función de densidad viene dada por la expresión:

$$f(x) = \frac{1}{(1+x^2)\,\pi}\ ,\ -\infty < x < \infty$$

Cuando nos interesa resumir una variable obtenida mediante una transformación de la variable original, podemos establecer la siguiente definición:

Definición. Dadas una v.a. X y una función g tal que $g(X)$ es de nuevo una v.a., se define la esperanza de esta nueva variable como:

$$E[g(X)] = \int_{\Re} g(x)dF(x)$$

Según que la nueva variable sea discreta o continua, este valor esperado podrá ser expresado con las formulaciones vistas anteriormente.

Como consecuencia de su carácter de operador lineal, la esperanza matemática cumple una serie de propiedades deseables.

Proposición 2.5. *La esperanza de cualquier v.a. X, presenta las siguientes propiedades para cualesquiera $a, c \in \Re$:*

1. $E(c) = c$
2. $E(aX) = aE(X)$
3. $E(X + c) = E(X) + c$
4. $E(aX + c) = aE(X) + c$

Demostración. En efecto,
$E(c) = \int_{\Re} cdF(x) = c\int_{\Re} dF(x) = c$
$E(aX) = \int_{\Re} axdF(x) = a\int_{\Re} xdF(x) = aE(X)$
$E(X + c) = \int_{\Re}(X + c)dF(x) = \int xdF(x) + c\int_{\Re} dF(x) = E(X) + c$
y a partir de los dos últimos resultados se obtiene: $E(aX + c) = aE(X) + c$
[Hemos aplicado que $\int_{\Re} dF(x) = \sum_i p_i = 1$ para variables discretas y para variables continuas $\int_{\Re} dF(x) = \int_{\Re} f(x)dx = 1$]

Proposición 2.6. *Para cualesquiera variables aleatorias X e Y, la esperanza de la suma puede ser obtenida como suma de sus esperanzas: $E(X + Y) = E(X) + E(Y)$*

□

Esta propiedad resulta de gran interés ya que, como veremos en capítulos posteriores, a menudo nos interesa trabajar con magnitudes aleatorias que se obtienen como agregados de otras.

La esperanza puede ser interpretada como "centro de gravedad" de una distribución de probabilidad ya que, si asumiésemos dicho valor como único representante de la población el error esperado sería nulo.

Es decir, si evaluásemos el error cometido al adoptar la esperanza como representante obtendríamos una nueva variable aleatoria $(X - E(X))$ resultando sencillo comprobar que su valor esperado es cero.

También podríamos considerar el error o desviación en forma complementaria; la esperanza puede ser un valor μ desconocido y podemos utilizar la variable o un conjunto de valores de ésta para aproximar ese valor desconocido. En este caso nos encontraríamos con que cada valor de X conduciría a un error o desviación, pero en síntesis estos errores se compensarían, obteniéndose una esperanza nula.

[Demuéstrese que $E(X - E(X)) = 0$ aplicando las propiedades anteriores y teniendo en cuenta que $E(X) = \mu = cte.$]

Las desviaciones o errores podrían ser definidos en términos más generales respecto a una característica M de la variable, con lo cual ya no queda garantizado un error esperado nulo. Este concepto de error o desviación $(X - M)$ va a jugar un papel muy importante en capítulos posteriores de este libro.

Además del "centro" de una distribución interesa también conocer la separación entre sus valores, ya que es necesario diferenciar las distribuciones en las que ciertos errores puntuales elevados resulten muy probables de aquellas otras donde estos mismos errores extremos tengan una pequeña probabilidad. Para diferenciar estas situaciones introduciremos la varianza como medida de dispersión que resume las desviaciones cuadráticas de una variable aleatoria respecto a su valor esperado.

Así, a partir de la distancia o desviación de un valor x al "centro" de la distribución $(X - \mu)$, definimos la varianza como la "desviación cuadrática esperada".

Definición 2.8. Dada una v.a. X definimos la *varianza*, que denotamos por σ^2 o $Var(X)$, como el valor, si existe, de la expresión:

$$\sigma^2 = Var(X) = E\left[X - E(X)\right]^2 = \int_{\Re} (x - \mu)^2 dF(x)$$

que en el caso continuo puede calcularse como:

$$\sigma^2 = \int_{\Re} (x - \mu)^2 f(x) dx$$

y en el caso discreto como:

$$\sigma^2 = \sum_i (x_i - \mu)^2 p_i$$

Como consecuencia del carácter aleatorio de la variable X, la varianza puede ser interpretada como una medida de riesgo en el sentido de que sintetiza los errores o desviaciones cuadráticas respecto a la esperanza, ponderando cada error o desviación por su correspondiente potencialidad (probabilidad). Es evidente que el mínimo riesgo posible se presentaría en el caso de varianza nula, esto es, para una variable cuyos valores coinciden con el esperado y que por tanto se denomina habitualmente "variable degenerada".

Si en vez de considerar los errores respecto a la media los considerásemos respecto a otra magnitud cualquiera M, $e = (X - M)$, entonces la desviación cuadrática esperada $E(e^2) = E(X - M)^2$, se denomina *error cuadrático medio respecto a* M.

Proposición 2.7. *A partir de la definición de varianza es posible obtener una expresión alternativa de esta medida:* $\sigma^2 = E(X^2) - \mu^2$, *que suele resultar más operativa para su cálculo.*

Demostración. Para comprobar la equivalencia entre las expresiones $E(X - \mu)^2$ y $E(X^2) - \mu^2$ basta desarrollar el cuadrado en la definición de varianza:

$$\sigma^2 = E(X - \mu)^2 = E\left(X^2 - 2\mu X + \mu^2\right) = E(X^2) + E(-2\mu X) + E(\mu^2)$$

Teniendo en cuenta las propiedades de la esperanza y que μ y -2 son constantes, se tiene:

$$\sigma^2 = E(X^2) - 2\mu E(X) + \mu^2 = E(X^2) - 2\mu\mu + \mu^2 = E(X^2) - \mu^2$$

□

Proposición 2.8. *La varianza de cualquier v.a.* X, *presenta las siguientes propiedades para cualesquiera* $b, c, M \in \Re$*:*

1. $\sigma^2 \geq 0$

2. $Var(X + c) = Var(X)$

3. $Var(bX) = b^2 Var(X)$

4. $Var(X) \leq E(X - M)^2$

Esta última propiedad permite calificar a la varianza de "medida cuadrática óptima" en el sentido de que esta expresión miminiza el valor de las desviaciones cuadráticas.

Demostración. La primera propiedad garantiza que la varianza es no negativa y se demuestra de forma inmediata a partir de la definición, en la que intervienen desviaciones al cuadrado que por tanto serán no negativas.

3.

$$\begin{aligned} Var(bX) = & \quad E\left[bX - E(bX)\right]^2 = E[bX - bE(X)]^2 = E[b(X - E(X))]^2 = \\ = & \quad b^2 E[X - E(X)]^2 = b^2 Var(X) \end{aligned}$$

En la demostración de esta expresión se ha hecho uso de la propiedad de la esperanza relativa al producto de una variable aleatoria por una constante. [Compruébese la propiedad 2, según la cual la varianza permanece inalterada ante cambios de origen en la variable aleatoria]

4. Sea M un valor real cualquiera; entonces se cumple que la desviación cuadrática respecto a M se hace mínima cuando dicho valor coincide con el esperado.

Para comprobar esta propiedad basta desarrollar la expresión genérica:

$$\begin{aligned} E(X - M)^2 = & \quad E(X - \mu + \mu - M)^2 = E\left[(X - \mu) + (\mu - M)\right]^2 = \\ = & \quad E\left[(X - \mu)^2 + 2(X - \mu)(\mu - M) + (\mu - M)^2\right] \end{aligned}$$

Teniendo en cuenta que el operador esperanza es lineal y que μ y M son constantes, se tiene:

$$E(X - M)^2 = E(X - \mu)^2 + 2(\mu - M)E(X - \mu) + (\mu - M)^2$$

y, como $E(X - \mu) = 0$, resulta:

$$E(X - M)^2 = E(X - \mu)^2 + \underbrace{(\mu - M)^2}_{\geq 0}$$

Al ser $(\mu - M)^2$ un cuadrado y por tanto no negativo, se tiene que a σ^2 hay que sumarle una cantidad mayor o igual a cero para alcanzar a $E(X - M)^2$ de donde: $E(X - M)^2 \geq \sigma^2$ □

La descomposición anterior separa el error cuadrático medio respecto a M en dos sumandos: la varianza de X y el cuadrado de la diferencia entre M y μ. La aleatoriedad de la variable afecta al primer sumando y no al segundo que es el cuadrado de una constante. Así pues, el primer sumando es un error aleatorio intrínseco a la

variable y el segundo un error determinista debido a la distancia entre μ y M.

Como consecuencia de su definición, la varianza es una medida de dispersión que analiza la proximidad de los valores de la variable a su valor esperado. En muchas ocasiones es útil construir un intervalo en torno a la esperanza donde probablemente se sitúe un alto porcentaje de valores de la variable; los límites de este intervalo se construyen sumándole y restándole a μ el nivel de dispersión. Sin embargo, nos encontramos con el inconveniente de que las unidades de μ son las mismas que las de la variable y en cambio, las de la varianza son unidades cuadráticas. Por este motivo es conveniente introducir nuevas medidas de dispersión.

Definición 2.9. Definimos la *desviación típica* o *estándar*, que denotamos por σ o $STD(X)$, como la raíz cuadrada (con signo positivo) de la varianza.

De este modo, se dispone de una medida de dispersión cuya información aparece como complementaria a la proporcionada por la esperanza. Esta medida permite acotar probabilidades de intervalos con independencia del modelo de probabilidad que siga la variable X ya que, como demostraremos en un epígrafe posterior, se cumple para cualquier $k > 0$:

$$P\left(\mu - k\sigma < X < \mu + k\sigma\right) \geq 1 - \frac{1}{k^2}$$

expresión conocida como *desigualdad de Chebyshev* y cuya interpretación es muy clara: la probabilidad de que un valor de la variable se sitúe en cierto entorno de su esperanza, determinado ese entorno por su desviación típica, es mayor que una cierta cantidad.

Así pues, gracias a la acotación anterior, podemos entender la desviación típica -y en su caso la varianza- como una medida de riesgo asociada a la v.a.

Además, en el caso de que la variable X siga una distribución conocida, las acotaciones anteriores pueden ser perfeccionadas. Más concretamente, si X se adapta a un modelo normal, se puede comprobar que el 66 % de los valores se sitúan dentro del intervalo $(\mu - \sigma, \mu + \sigma)$; esto es: $P(\mu - \sigma < X < \mu + \sigma) \geq 0,66$. Ampliando el intervalo observamos que la proporción de valores que se sitúan en él aumenta y así se tiene:

$$P(\mu - 2\sigma < X < \mu + 2\sigma) \geq 0,95 \text{ y } P(\mu - 3\sigma < X < \mu + 3\sigma) \geq 0,99$$

Las características anteriormente vistas -esperanza, varianza y desviación típica- son equivalentes a las correspondientes medidas descriptivas media, varianza y desviación típica, con la única salvedad de que ahora su rasgo característico es la potencialidad como consecuencia del carácter aleatorio de X.

Las características anteriores, aun siendo las más habituales, no agotan las posibilidades de describir una magnitud aleatoria. De hecho, las distintas medidas de tendencia definidas para variables estadísticas son generalizables al caso de variables aleatorias.

Una vez visto el paralelismo entre media y esperanza, la *Moda* podría ser identificada como valor que maximiza la probabilidad o densidad y, a través de la función de distribución podemos contemplar de modo inmediato cualquier cuantil (a modo de ejemplo, la *Mediana* sería aquel valor de la variable para el que se acumula una probabilidad del 50 %, esto es, se cumple $p\left(X \leq Me\right) = 0,5$).

Cuando se quieren comparar las dispersiones de varias v.a., la varianza y la desviación típica no son útiles porque muestran la variación respecto a su valor esperado y dependen de las unidades y de la magnitud de éste. En este tipo de problemas es conveniente introducir las medidas de dispersión relativas, de las cuales la de uso más generalizado es el *coeficiente de variación de Pearson*, definido como el valor de la expresión: $CV = \frac{\sigma}{\mu}$.

Las características de esperanza y varianza vistas anteriormente son casos particulares de otras de carácter más general que denominamos *momentos*. Un momento es una medida de desviación que viene caracterizada por dos parámetros M y r; el parámetro M indica el centro de referencia respecto al cual se calculan las desviaciones y r expresa la forma de medir esa desviación y sus unidades.

Definición 2.10. Se denomina *momento de orden r centrado respecto a M* al valor esperado, si existe, de la variable $(X - M)^r$:

$$\mu_{r,M} = E(X - M)^r$$

Para valores específicos de M y r esta expresión proporciona momentos concretos y así, en el caso $r = 2$ se obtiene el *error cuadrático medio respecto a M*. Por otra parte, las particularizaciones habituales de M son 0 y $E(X)$; así podemos establecer las siguientes definiciones:

Definición 2.11. Se denomina momento de orden r centrado respecto a $E(X)$ o simplemente *momento centrado de orden r*, que se denota por μ_r, al valor, si existe, de la expresión:

$$\mu_r = E\left[X - E(X)\right]^2$$

Se llama momento de orden r centrado respecto al origen (0) o simplemente *momento no centrado de orden r*, α_r, al valor, si existe, de la expresión:

$$\alpha_r = E(X - 0)^r = E(X)^r$$

[Compruébese que se cumple: $\alpha_0 = 1$, $\alpha_1 = E(X) = \mu$, $\alpha_2 = E(X)^2$, y también: $\mu_0 = 1$, $\mu_1 = 0$, $\mu_2 = \sigma^2$].

A veces la notación de los valores esperados puede llevarnos a confusión; conviene tener presente que $E(X^2) = E(X)^2$, $(E(X))^2 = E^2(X)$, $E[(X - \mu)^2] = E(X - \mu)^2$.

Dada una v.a. X, si existe el momento (centrado o no) de orden s, existen todos los de orden inferior a s y, de modo complementario, si no existe el de orden s tampoco existe el de cualquier otro orden superior a s.

La relación entre momentos centrados y no centrados viene dada por la siguiente expresión:

$$\mu_r = \alpha_r - C_{r,1}\mu^1\alpha_{r-1} + C_{r,2}\mu^2\alpha_{r-2} + \cdots + (-1)^r\mu^r$$

[La justificación es sencilla sin más que considerar la expresión de un binomio elevado a r]

Un caso particular de esta relación es la fórmula de cálculo deducida para la varianza: $\sigma^2 = \mu_2 = \alpha_2 - \alpha_1^2$.

Otras características importantes asociadas a las v.a. están referidas a la forma que presenta su función de probabilidad o de densidad; nos referimos a las características de asimetría y apuntamiento de la curva. Existen varios indicadores para medir estos parámetros, siendo los más usuales los denominados coeficientes γ_1 y γ_2 de Fisher.

El estudio de la forma de una distribución se efectúa habitualmente adoptando como referencia el modelo normal -que analizaremos con detalle en el capítulo siguiente- cuya representación es una curva simétrica campaniforme conocida como "campana de Gauss".

Definición 2.12. El *coeficiente de asimetría* γ_1 se define como el cociente: $\gamma_1 = \frac{\mu_3}{\sigma^3}$ y su valor se compara respecto al 0, resultando una distribución asimétrica positiva o a la derecha si su coeficiente γ_1 es positivo y asimétrica negativa o a la izquierda si éste es negativo. Cuando el resultado es nulo, decimos que la curva es simétrica.

El *coeficiente de apuntamiento* γ_2 se define como: $\gamma_2 = \frac{\mu_4}{\sigma^4} - 3$, y su resultado se compara también con el 0, valor de referencia que corresponde a una distribución normal estándar y que de presentarse permite calificar a una distribución de mesocúrtica. Los valores positivos de este índice se corresponden con un apuntamiento superior al normal (distribuciones calificadas de leptocúrticas) mientras que para valores negativos el apuntamiento es inferior al normal y las distribuciones se denominan platicúrticas.

En ocasiones nos interesará conocer la distribución del valor agregado total de v.a. X entre los elementos que componen la población. En este caso se utilizan las *medidas de concentración* y *desigualdad*.

Las medidas más utilizadas son la curva de Lorenz y el índice de *Gini -Lorenz* que lleva asociado, que en este caso formalizaremos en términos probabilísticos

Definición 2.13. Dada una variable aleatoria X con función de distribución $F(x)$ el índice de concentración de Gini-Lorenz viene dado por el resultado de la expresión:

$$L(X) = 1 - 2\int_0^1 F_1(x)dF(x)$$

donde la f.d. $F(x)$ representa la proporción de rentistas por debajo de una cantidad x y $F_1(x)$ se define como:

$$F_1(x) = \int_0^x t\frac{dF(t)}{\mu}$$

y representa la proporción de renta que reciben los rentistas anteriores.

La interpretación, propiedades e inconvenientes de este índice son idénticas a las recogidas en el caso de variables estadísticas y remitimos al lector al libro *Introducción a la Estadística Económica* (Rigoberto Pérez, Covadonga Caso, María Jesús Río y Ana J. López) donde se trata el tema con detalle.

Aunque la curva de Lorenz y el índice de Gini-Lorenz son las medidas más tradicionales, con carácter más reciente han sido introducidas medidas que solucionan sus limitaciones. En concreto, en trabajos anteriores hemos propuesto medidas de la desigualdad desde las ópticas individual (indicador asociado a la persona que sufre o repercute desigualdad) y colectiva (medida obtenida como síntesis de los indicadores individuales)[2].

[2]Un estudio detallado de estas medidas de desigualdad, que incluye también sus conexiones con los indicadores de pobreza y su análisis normativo, aparece recogido en el trabajo de López, A.J. y R. Pérez (1991): *Indicadores de desigualdad y pobreza. Nuevas alternativas* publicado como Documento de trabajo 037/1991 de la Facultad de CC. Económicas y Empresariales de la Universidad de Oviedo

Definición 2.14. Denominamos *índice de desigualdad individual* asociado a una renta x al valor de la expresión:

$$d(x) = \frac{\mu}{x}\text{-}1$$

Para x distinto de 0, este coeficiente es una nueva v.a. que recoge la desigualdad generada por cada renta individual. Como consecuencia, su valor esperado, si existe, será indicativo del nivel de *desigualdad colectiva*:

$$D = E(d) = E\left(\frac{\mu}{x} - 1\right) = \int_0^\infty \left(\frac{\mu}{x} - 1\right) dF(x)$$

Además de las funciones de probabilidad, de densidad y de distribución estudiadas existen otras funciones que podemos asociar a toda v.a. y que son importantes instrumentos de trabajo; a partir de ellas pueden obtenerse, por ejemplo, la función de densidad o los momentos de una distribución. Nos estamos refiriendo a la función generatriz de momentos y a la función característica.

Definición 2.15. Se denominada *función generatriz de momentos* (f.g.m.), de una v.a. X, si existe, al valor de la expresión:

$$M_X(t) = E(e^{tX})$$

donde t es una variable real.

Se trata de una función real de variable real, que identifica por completo una variable aleatoria. Se cumple así una condición de identidad, de modo que si la función generatriz de momentos existe puede demostrarse que es única y determina por completo a la distribución de probabilidad de X (a toda función de distribución corresponde una función generatriz de momentos y recíprocamente).

La función generatriz puede plantear problemas de existencia, pues puede ocurrir que para determinados valores de t el valor esperado $E(e^{tX})$ no exista. No obstante, para las distribuciones usuales la f.g.m. toma un valor finito por lo que la dificultad anterior puede ser obviada.

La justificación de la denominación función generatriz de momentos viene dada por la siguiente propiedad:

Proposición 2.9. *Si existe el momento de orden r, se tiene:*

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \alpha_r \ ; \ \forall r = 1, 2, \ldots$$

es decir, dada una variable aleatoria X es posible obtener sus momentos sucesivos a partir de $M_X(t)$ siempre que esta función y sus derivadas existan.

Demostración. La demostración de esta propiedad se lleva a cabo teniendo en cuenta que pueden intercambiarse los operadores diferencial y esperanza:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = \left.\frac{d}{dt}E\left(e^{tX}\right)\right|_{t=0} = \left.E\left[\frac{d}{dt}\left(e^{tX}\right)\right]\right|_{t=0} = \left.E\left[X\left(e^{tX}\right)\right]\right|_{t=0} = E(X) = \mu = \alpha_1$$

De modo análogo, para la segunda derivada se obtiene en el punto $t = 0$:

$$\left.\frac{d^2 M_X(t)}{dt^2}\right|_{t=0} = \left.\frac{d^2}{dt^2} E\left(e^{tX}\right)\right|_{t=0} = E\left[\left.\frac{d^2}{dt^2}\left(e^{tX}\right)\right]\right|_{t=0} = E\left[\left.\frac{d}{dt}\left(Xe^{tX}\right)\right]\right|_{t=0} = E\left[\left.X^2\left(e^{tX}\right)\right]\right|_{t=0} = E(X^2) = \alpha_2$$

y en general, para la derivada de orden r:

$$\left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0} = \left.\frac{d^r}{dt^r} E\left(e^{tX}\right)\right|_{t=0} = E\left[\left.\frac{d^r}{dt^r}\left(e^{tX}\right)\right]\right|_{t=0} = E\left[\left.X^r\left(e^{tX}\right)\right]\right|_{t=0} = E(X^r) = \alpha_r$$

□

Consideremos de nuevo la v.a. "Número de medallas obtenidas por un deportista" y su distribución de probabilidad:

$X = 0$	$P(X = 0) = 0,5$
$X = 1$	$P(X = 1) = 0,2$
$X = 2$	$P(X = 2) = 0,2$
$X = 3$	$P(X = 3) = 0,15$

Según la expresión vista para la función generatriz de momentos se tiene en este caso

$$M_X(t) = E\left(e^{tX}\right) = \left[0,5 + 0,2\left(e^t + e^{2t}\right) + 0,1e^{3t}\right]$$

cuya derivada respecto a t es:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = 0,2\left(e^t + 2e^{2t}\right) + 0,3e^{3t}$$

expresión que para $t = 0$ da lugar a:

$$\left.\frac{dM_X(t)}{dt}\right|_{t=0} = 0,2(1 + 2) + 0,3 = 0,9$$

valor coincidente con la $E(X)$ anteriormente calculada.

Proposición 2.10. *Otras propiedades importantes de la f.g.m. son las relativas a los cambios de origen y escala:*

1. $M_{c+X}(t) = e^{tc} M_X(t)$
2. $M_{bX}(t) = M_X(tb)$

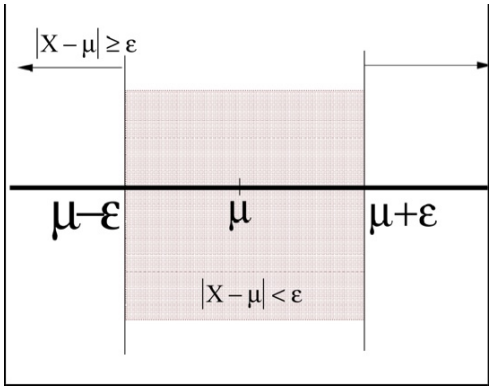
Demostración. En efecto:

$$M_{c+X}(t) = E\left[e^{t(c+X)}\right] = E\left(e^{tc}e^{tX}\right) = e^{tc}E\left(e^{tX}\right) = e^{tc}M_X(t)$$

$$M_{bX}(t) = E\left[e^{t(bX)}\right] = E\left[e^{(tb)X}\right] = M_X(tb)$$

□

Figura 2.10.: Acotación de la probabilidad



La *función característica* presenta como ventaja con respecto a la f.g.m. que siempre tiene garantizada su existencia.

Definición. Se define la función característica asociada a una v.a. X como una aplicación $\phi_X : t \in \Re \to \phi_x(t) \in \mathbb{C}$, dada por la expresión:

$$\phi_X(t) = E\left(e^{itx}\right) = E\left[\cos tx + i \cdot \sin tx\right]$$

Este valor esperado existe siempre para todo t real verificándose una identidad entre las funciones de densidad y característica: a toda función de densidad corresponde una única función característica y recíprocamente.

Como observamos, la función característica toma valores en el campo de los números complejos y los conocimientos de integración compleja exceden el nivel de formalización que seguimos en esta obra.

2.4. Desigualdad de Chebyshev

En epígrafes anteriores hemos estudiado el cálculo de probabilidades basado en la distribución probabilística de las variables aleatorias. Dado que esta distribución no siempre resulta conocida en la práctica, presentamos en este apartado una ampliación del cálculo de probabilidades a las situaciones con información más escasa.

A modo de introducción, consideremos una magnitud aleatoria X cuya distribución probabilística desconocemos. Ante esta ausencia de información debemos conformarnos con un cálculo aproximado de probabilidades, bien sea la correspondiente al interior de un intervalo o bien la complementaria (probabilidad de las colas).

Tal y como indica la figura 2.10, podríamos plantearnos evaluar la probabilidad de que la variable aleatoria X se encuentre fuera de la zona sombreada, esto es, discrepe de su valor esperado en cuantía superior a cierto margen ϵ. Dada la ausencia de información, deberemos limitarnos a garantizar que dicha probabilidad guarde relación directa con la dispersión de la variable ($Var(X)$) e inversa con el margen fijado (ϵ).

P.L. Chebyshev (1821-1894) y J. Bienaymé (1796-1878) desarrollaron de modo independiente la desigualdad generalmente conocida con el nombre del primero, en la que se establece una acotación superior para la probabilidad de las colas de un intervalo centrado en el valor esperado de una variable.

Proposición. *Consideremos una variable aleatoria X con esperanza y varianza finitas; entonces la* desigualdad de Chebyshev *permite afirmar que para cualquier número real ϵ positivo se verifica:*

$$P\left(|X - E(X)| \geq \epsilon\right) \leq \frac{Var(X)}{\epsilon^2}$$

La desigualdad de Chebyshev se obtiene como caso particular de la desigualdad básica:

Proposición. *Sea b una constante positiva y $h(X)$ una función no negativa, donde X es una v.a. Entonces siempre que $E(X)$ exista se cumple:*

$$P\left[h(X) \geq b\right] \leq \frac{1}{b} E[h(X)]$$

Más concretamente, la desigualdad de Chebyshev se correspondería con el caso en que $h(X) = [X - E(X)]^2$, $\epsilon = \sqrt{b}$. [Compruébese].

Demostración. Para demostrar la desigualdad básica definimos un conjunto $A = \{x/h(x) \geq b\}$ con $0 < P(A) < 1$. Podemos entonces expresar:

$$E\left[h(X)\right] = E\left[h(X)/A\right] P(A) + E\left[h(X)/A^c\right] P(A^c) \geq E\left[h(X)/A\right] P(A) \geq bP(A)$$

donde hemos aplicado para la primera desigualdad $h(x) \geq 0$ y para la segunda $h(x) \geq b$ para todo x de A.

□

En el caso de que nos interese acotar la probabilidad interior al intervalo (área sombreada en la figura anterior) se obtiene una garantía o cota inferior dada por la siguiente expresión, aplicable a cualquier variable aleatoria X con esperanza y varianza finitas, $\forall \epsilon > 0$:

$$P\left(|X - E(X)| < \epsilon\right) \geq 1 - \frac{Var(X)}{\epsilon^2}$$

Esta desigualdad se obtiene de forma inmediata ya que con sólo aplicar la propiedad de la probabilidad del complementario se tiene: $P\left(|X - E(X)| < \epsilon\right) = 1 - P\left(|X - E(X)| \geq \epsilon\right)$ expresión a la que aplicamos el primer enunciado de Chebyshev:

$$P\left(|X - E(X)| \geq \epsilon\right) \leq \frac{Var(X)}{\epsilon^2}$$

La desigualdad de Chebyshev en cualquiera de sus dos expresiones proporciona cotas para probabilidades de la variable aleatoria X, en función de su dispersión y del margen de error considerado ϵ.

En síntesis, estas desigualdades garantizan ciertas probabilidades mínimas para cualquier entorno de la esperanza, que aumentan con el margen considerado ϵ y disminuyen con la dispersión poblacional. Si por el contrario queremos aproximar la probabilidad fuera de ese entorno, la desigualdad de Chebyshev proporciona una cota superior, que guarda relación directa con la dispersión e inversa con el margen fijado.

Además, cuando el margen de error considerado se expresa como proporción de la desviación estándar ($\epsilon = k\sigma$) es posible llegar a formulaciones alternativas para las dos acotaciones de Chebyshev. Así, dada una variable aleatoria X con esperanza y varianza finitas se cumple para cualquier $k > 0$:

$$P\left(|X - E(X)| \geq k\sigma\right) \leq \frac{1}{k^2}$$

$$P\left(|X - E(X)| < k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Estos nuevos enunciados de la desigualdad de Chebyshev confirman la interpretación de la desviación estándar como medida de la dispersión. En efecto, resulta sencillo comprobar cómo cambia la acotación inferior del intervalo y superior de las colas a medida que aumenta el número de desviaciones típicas consideradas en nuestro margen de error:

Valor k	Cota inferior para $P\left(\lvert X - E(X)\rvert < k\sigma\right)$	Cota superior para $P\left(\lvert X - E(X)\rvert \geq k\sigma\right)$
1	0	1
2	0,75	0,25
3	0,89	0,11
4	0,9375	0,0625
5	0,96	0,04
10	0,99	0,01

Como puede apreciarse en la tabla anterior, la consideración de un margen no superior a una sola desviación estándar no resultaría en absoluto informativa, ya que para $k = 1$ la cota inferior es 0 y la superior 1, valores entre los cuales -como recoge la axiomática de Kolmogorov- se halla comprendida toda probabilidad. [¿Qué acotaciones se obtendrían si el margen considerado es de $\frac{\sigma}{2}$?].

3. Modelos de probabilidad

En nuestra realidad cotidiana nos encontramos diversas variables de carácter aleatorio que, tal y como hemos expuesto en el capítulo anterior, sólo pueden ser analizadas convenientemente si disponemos de información sobre su distribución de probabilidad. Supongamos a modo de ejemplo que una publicación está elaborando un reportaje sobre experiencias empresariales, en el que existen varios aspectos inciertos.

I *Se ha contactado con 20 empresarios de distintos perfiles a los que se desea entrevistar para el reportaje pero se desconoce cuántos de ellos accederán a ser entrevistados. Desde el equipo de redacción se asume que las respuestas de los distintos empresarios convocados son independientes y se confía en que, dado el prestigio de la publicación, un 80 % de ellos accederán finalmente a colaborar.*

II *El reportaje incluirá, además de las entrevistas, imágenes alusivas a la actividad de los empresarios. De las 15 fotografías seleccionadas se elegirán aleatoriamente 3 para la portada y se confía en que esté representado el empresariado tanto de sexo masculino como femenino.*

III *La entrevista, que se ajustará a un modelo ya diseñado, tendrá una duración aleatoria en función de las respuestas y talante del empresario. No obstante, se prevé una duración cercana a dos horas, resultando poco probable que la entrevista se desvíe considerablemente de este tiempo por exceso o por defecto.*

IV *El equipo responsable del reportaje confía en la calidad de su trabajo, por lo que se espera que apenas aparezcan errores tipográficos a lo largo de sus páginas.*

En las cuatro etapas descritas aparecen magnitudes de carácter aleatorio, con distintas distribuciones de probabilidad. No obstante, en la práctica muchas de estas variables aleatorias presentan comportamientos comunes que pueden ser descritos mediante pautas. Así, el esquema del número de empresarios que acceden a la entrevista del periódico es similar al de los potenciales clientes que finalmente compran un producto o al de los estudiantes que aprueban un examen.

De igual modo, los tiempos de duración de las entrevistas, aunque aleatorios, seguirán previsiblemente un modelo en forma de campana (mayores probabilidades para los valores centrales y menores para observaciones extremas). Este tipo de distribución -como justifica su denominación, "normal"- servirá para describir otras muchas características (la altura de los empresarios, su peso, ...).

En efecto, existen modelos probabilísticos cuyo interés reside en la capacidad de describir comportamientos genéricos de distintas magnitudes aleatorias que resultan

semejantes según ciertas pautas. Nos encontramos así con grandes "familias probabilísticas" designadas con nombres propios que incluyen como casos particulares numerosos fenómenos, incorporando sus rasgos diferenciales mediante parámetros.

Los modelos de probabilidad son idealizaciones probabilísticas de fenómenos aleatorios. Representamos los fenómenos o experimentos aleatorios mediante variables aleatorias, y estas variables proporcionan una partición de la clase de todos los fenómenos posibles, de modo que a cada fenómeno le corresponde una variable aleatoria, pero cada una de éstas representa a diversos fenómenos.

Por otra parte, cada magnitud aleatoria lleva asociada, como ya hemos visto, una función de distribución (que a veces también se denomina *ley de probabilidad*), y muchas de éstas mantienen estructuras comunes, salvo algún parámetro que las especifique. Entonces dividimos el conjunto de todas las distribuciones posibles en grupos (también llamados modelos), de forma que las distribuciones que integran cada grupo tengan la misma estructura.

En este nivel de abstracción, hemos pasado del conjunto de fenómenos a un conjunto de modelos matemáticos o probabilísticos más fáciles de analizar, de modo que un reducido número de estos modelos recogen una mayoría de los experimentos posibles, bien porque las características del experimento encajen plenamente en las condiciones del modelo o bien porque las observaciones se ajusten más o menos a los resultados teóricos que predice el modelo.

Naturalmente existen muchos modelos de probabilidad. Nos ocuparemos en este capítulo de los más importantes, que han sido estudiados con detalle y llevan nombre propio.

Cada modelo probabilístico representa una familia de funciones de distribución, que dependen de uno o más parámetros y cuyas especificaciones determinan las distribuciones particulares que integran la familia. El estudio de estos modelos se centra en analizar la ecuación general que representa a la familia, sus características y sus propiedades.

Por otra parte, nos interesa conocer la probabilidad con la que un fenómeno se presentará a través de determinados sucesos; conocida la variable aleatoria y el modelo que representa a ese fenómeno, tal probabilidad se puede calcular a partir de la ecuación general y los parámetros (conocidos o supuestos) de forma exacta o bien de forma aproximada, mediante el empleo de métodos numéricos. No obstante algunos de estos modelos se encuentran tabulados, lo que simplifica el cálculo de esas probabilidades.

En los apartados que siguen analizaremos aquellos modelos probabilísticos de uso más habitual -distribuciones binomial, geométrica, hipergeométrica, uniforme y normal- y también algunos otros que, aunque no son de utilización tan generalizada, resultan de interés por describir adecuadamente ciertos fenómenos relevantes en el ámbito económico: ocurrencia de sucesos "raros", tiempos de espera, distribuciones de renta, ...

En algunos casos se considera también como modelo probabilístico la denominada *distribución singular* que, por su gran sencillez, puede servir de introducción a las restantes distribuciones. Se trata de un modelo que concentra toda la masa de probabilidad -la unidad- en un único valor, por lo cual su aplicabilidad práctica es muy escasa.

De hecho, el modelo singular se correspondería con el caso de una variable degenerada, carente de interés para nuestros estudios al no manifestar variación alguna.

Una variable X presenta distribución singular cuando su probabilidad se concentra en un punto, x_0. Como consecuencia su función de probabilidad viene dada por:

$$P(X = x) = \begin{cases} 1 & para\, x = x_0 \\ 0 & para\, x \neq x_0 \end{cases}$$

Dado que la variable X presenta un único valor, resulta evidente que éste coincidirá con su esperanza ($\mu = x_0$). [¿Cuál será su varianza?]

Una variable X con distribución singular es "probabilísticamente" equivalente a una constante, pues toma con probabilidad 1 determinado valor x_0. Conviene sin embargo tener presentes las consideraciones sobre los sucesos de probabilidad nula y de probabilidad unitaria realizadas en el primer tema.

3.1. Modelo Binomial

Un modelo probabilístico de aplicación generalizada es el binomial, que aparece cuando al efectuar observaciones reiteradas analizamos en cuántos casos se han presentado determinados resultados, habitualmente denominados "éxitos".

Consideremos el ejemplo anterior, asumiendo que 20 empresarios han sido convocados para entrevistarles. Dado que no podemos anticipar la respuesta de cada uno de ellos, ésta puede ser identificada con una variable aleatoria con dos únicos resultados asociados a los sucesos "aceptar" (éxito) y su complementario "no aceptar" (fracaso).

Aunque la variable aleatoria "respuesta de un empresario" podría venir definida de múltiples formas, resulta habitual asignar el valor 1 al éxito y 0 al fracaso. De este modo, tendríamos una v.a. discreta cuya distribución de probabilidad quedaría perfectamente determinada una vez conocida la probabilidad de aceptación p.

Definición 3.1. Dada una *prueba dicotómica* (también llamada de Bernoulli), caracterizada por dos resultados mutuamente excluyentes (éxito y fracaso), indicando por p la probabilidad de éxito, la variable aleatoria definida como:

$$X = \begin{cases} 1 & \text{si ocurre éxito} \\ 0 & \text{si ocurre fracaso} \end{cases}$$

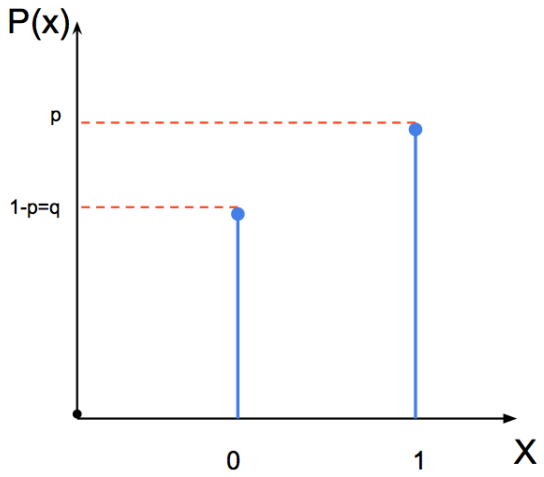
se denomina *modelo o distribución de Bernoulli* (o dicotómica) $\mathcal{B}(p)$.

Conocida la probabilidad de éxito p se tiene también la probabilidad de fracaso (complementario) $1 - p = q$ y la función de probabilidad de X vendría dada por:

$$P(X = 0) = q \; ; \; P(X = 1) = p$$

Este sencillo modelo discreto puede ser representado mediante un diagrama de barras como como recoge la figura 3.1, a partir del cual se obtiene de modo inmediato la probabilidad acumulada asociada a la función de distribución:

Figura 3.1.: Función de probabilidad del modelo de Bernoulli



$$F(x) = P(X \leq x) = \begin{cases} 0 & si\, x < 0 \\ q & si\, 0 \leq x < 1 \\ 1 & si\, 1 \leq x \end{cases}$$

Las características de este modelo probabilístico pueden ser también obtenidas con facilidad, ya que a partir de su definición se tiene:

$$\mu = E(X) = 1 \cdot p + 0 \cdot q = p$$

$$\sigma^2 = Var(X) = (1-p)^2 p + (0-q)^2 q = pq(q+p) = pq$$

expresión a la que puede llegarse también a partir de la fórmula alternativa $E(X^2) - \mu^2$. [Compruébese que en este caso se obtiene $E(X^2) = E(X)^2 = p$ y $E^2(X) = p^2$]

La interpretación de ambos parámetros es ilustrativa: la esperanza o "centro de gravedad" de la distribución (cuyos únicos valores son 0 y 1) correspondería a la probabilidad de éxito (p).

Por su parte, la varianza, al aproximar el riesgo, debe también tomar en consideración la probabilidad de fracaso q. De hecho, puede verse fácilmente, derivando la varianza respecto a p (condición necesaria de extremo), que el riesgo máximo se obtendría para el caso en que éxito y fracaso tuvieran probabilidades idénticas ($p = q = 0,5$). [¿Cuál sería el caso de riesgo mínimo? ¿por qué?].

Al exigirle a la variable que tome los valores 0 y 1, de alguna forma estamos normalizando dicha variable. Esta normalización resulta muy cómoda, ya que podemos observar que los valores no sólo intervienen en las funciones de probabilidad y de distribución, sino que también condicionan las características esperanza y varianza, por lo que variables dicotómicas con valores distintos de 0 y 1 no tendrían características comparables. [¿Cuánto valdrían la esperanza y la varianza de una variable que asignase los valores -10 y 10 a los sucesos fracaso y éxito respectivamente?].

La distribución de Bernoulli toma su nombre de Jakob Bernoulli (1654-1705) quien introdujo el modelo en su obra *Ars Conjectandi*. A este autor, perteneciente a una de las familias más relevantes

en la historia de la probabilidad, se deben numerosas aportaciones. Fue el primero que se preocupó por la extensión de la probabilidad a otros campos distintos de los juegos de azar; también introdujo el teorema de Bernoulli en el cual demostraba la convergencia de la frecuencia relativa a la probabilidad.

Ya nos hemos referido a otro de los miembros de la familia, Daniel Bernoulli, quien propuso una solución para la famosa "paradoja de San Petersburgo".

Si bien el modelo anterior permite describir gran cantidad de fenómenos, en general las investigaciones no se limitarán a una prueba única, siendo frecuentes en el ámbito económico los estudios donde se observa cierto resultado en una serie de pruebas repetidas.

Este es el caso del ejemplo propuesto, donde un total de 20 empresarios han sido convocados para la entrevista, pudiendo decidir cada uno de ellos si acudirá o no. Como consecuencia de la repetición de la experiencia, definiremos ahora una variable aleatoria que designe el número de éxitos obtenidos a lo largo de la investigación (en nuestro ejemplo "número de empresarios que acceden a la entrevista").

El rasgo distintivo de esta variable con respecto al modelo de Bernoulli es la importancia de un nuevo parámetro n, que representa el número de observaciones llevadas a cabo. Suponemos que estas n observaciones son independientes entre sí y asumimos como constante a lo largo de n pruebas la probabilidad de éxito recogida por el parámetro p. Ambos serán los rasgos característicos del modelo binomial, que suele designarse por $\mathcal{B}(n,p)$. La siguiente tabla ilustra varias situaciones en las que un modelo binomial describe adecuadamente los resultados que nos interesan:

Experiencia	Éxito	Probabilidad de éxito	Nº de pruebas
Lanzar un dado	"Sacar un 2"	$p = \frac{1}{6}$	3
Situación laboral	"Activo"	Tasa actividad	12
Sondeo	"Votar SI"	$p = 0,4$	24
Entrevista	"Acudir"	$p = 0,8$	20
↓		↘	↙
v.a. "Nº de éxitos en n pruebas"		$\mathcal{B}(n,p)$	

Si examinamos los rasgos que tienen en común los ejemplos propuestos podemos llegar a la conclusión de que existen ciertos requisitos para que una experiencia sea incluida dentro de la "familia binomial". Así, será relevante conocer si la probabilidad de éxito es constante en todas las pruebas (en los lanzamientos de un dado este supuesto parece evidente, pero sin embargo podrían existir diferentes tasas de actividad según los sectores económicos, distintas probabilidades de votar "sí" según la ideología política, o diferentes probabilidades de acudir a la entrevista según la fama o el nivel de ocupación del empresario).

Del mismo modo, nos interesará saber si las observaciones son independientes (en los ejemplos del sondeo y de las entrevistas a empresarios podría existir relación entre las respuestas, ya que sabemos que unas personas ejercen influencia sobre otras).

Recopilando las consideraciones anteriores:

Definición 3.2. Un modelo binomial se basa en los siguientes supuestos:

- Se llevan a cabo n pruebas u observaciones.
- Las n observaciones son independientes entre sí.
- La probabilidad de éxito p permanece constante en todas las pruebas.

En estas condiciones, la variable aleatoria que recoge el "número de éxitos en las n pruebas" se dice que sigue un *modelo binomial* $\mathcal{B}(n, p)$.

Toda sucesión de pruebas que verifican las condiciones anteriores se denominan pruebas de Bernoulli y dan lugar al proceso de Bernoulli. El proceso de Bernoulli surge de una manera natural cuando realizamos observaciones en una población infinita o cuando, tratándose de una población finita, las observaciones se seleccionan al azar con reposición. En estas situaciones los supuestos enunciados son fácilmente admisibles, ya que las observaciones no alteran la estructura poblacional (gracias al reemplazamiento o bien al tamaño poblacional infinito) quedando así garantizada la independencia entre observaciones y la constancia de la probabilidad de éxito p.

Una vez descritos sus rasgos básicos, la distribución de probabilidad de un modelo binomial se obtiene con facilidad.

Consideremos una variable $X \approx \mathcal{B}(n, p)$. En primer lugar, debemos definir su recorrido de valores que, por describir el número de éxitos, en ningún caso podrán superar el número de pruebas realizadas ni ser inferiores a cero.

Tendremos por tanto que los valores que la variable X puede adoptar son: $0, 1, 2, \ldots, n$ y podemos deducir gracias a los supuestos subyacentes en el modelo binomial la probabilidad asociada a cada uno de esos valores.

Para cada valor posible k, $P(X = k)$ recoge la probabilidad de que se presenten k éxitos a lo largo de n pruebas (esto es, que k empresarios acudan a la entrevista, que obtengamos k veces el resultado 2 al lanzar un dado, que k de los individuos seleccionados sean activos, ...).

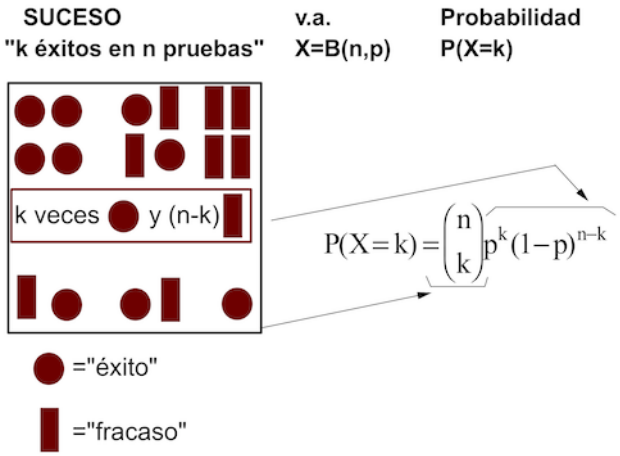
Para cuantificar la probabilidad anterior debemos responder a dos interrogantes: ¿de cuántas formas distintas podríamos obtener esos k éxitos? y ¿cuál es la probabilidad de cada una de ellas?

El primero de los interrogantes hace referencia a los casos favorables (cuya secuencia aparece en el esquema) mientras el segundo evalúa la verosimilitud de cada uno de ellos. Como cada una de las secuencias posibles excluye todas las demás, se trata de resultados incompatibles y por tanto la probabilidad de la unión viene dada por la suma de probabilidades. [Figura 3.2]

Para responder a la cuestión "formas de seleccionar k éxitos en n pruebas", podemos acudir -como ya hemos visto en un tema anterior- al análisis combinatorio. Concretamente se trata de combinaciones de n elementos de orden k, supuesto que también responde al caso de permutaciones de n elementos con repeticiones k (no distinguimos un éxito de los demás) y $(n - k)$ (no distinguimos los fracasos entre sí).

A partir de cualquiera de los razonamientos anteriores la expresión de cálculo sería $C_{n,k} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$, que cuantifica el número de secuencias de n observaciones con k éxitos.

Figura 3.2.: Esquema Binomial



Ahora bien, necesitaríamos además conocer la probabilidad de que se presente cada una de esas situaciones favorables, que se corresponde con el suceso "k éxitos y $n-k$ fracasos" . Se trata pues de la probabilidad de la intersección que -gracias al supuesto de independencia entre pruebas- se obtiene como producto de probabilidades, dando como resultado la expresión $p^k q^{n-k}$.

Una duda que podría plantearse es si todos los casos favorables, esto es, los que presentan k éxitos, son equiprobables. La respuesta es afirmativa ya que los supuestos del modelo nos permiten afirmar que las pruebas son independientes y la probabilidad de éxito p permanece constante. Como consecuencia la probabilidad de cualquier secuencia de resultados que incluya k éxitos y $n-k$ fracasos será la probabilidad de n sucesos independientes, dada por un producto en el que k términos son p y los restantes $(n-k)$ términos son $1-p=q$.

Una vez examinados los factores que intervienen en la probabilidad, estamos en condiciones de construir la función de probabilidad correspondiente a un modelo binomial $\mathcal{B}(n,p)$ que viene dada por:

$$P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}\ ;\ con\, k=0,1,\cdots,n \tag{3.1}$$

Para comprobar que la expresión anterior es una verdadera función de probabilidad basta verificar las condiciones de no negatividad y suma unitaria.

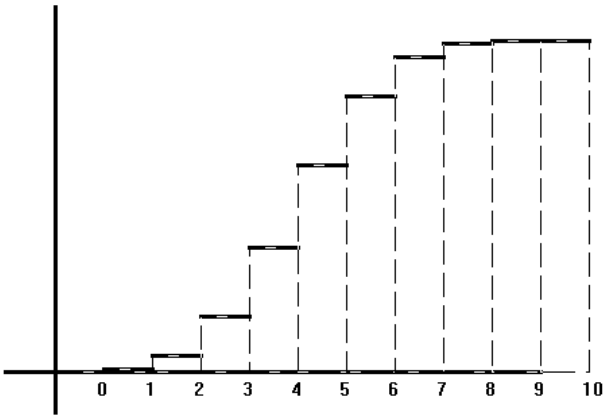
La primera de ellas es inmediata por ser no negativos todos los términos que aparecen en la expresión $P(X=k)$. Por lo que respecta a la segunda condición, se tiene

$$\sum_{k=0}^{n}P(X=k)=\sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k}$$

expresión que se corresponde con el desarrollo del Binomio de Newton $(p+q)^n$ cuyo valor es la unidad por ser $q=1-p$.

La denominación del modelo binomial se debe a su conexión con el binomio de Newton. De hecho,

Figura 3.3.: Función de Distribución Binomial



la expresión de la probabilidad binomial $\binom{n}{k}p^k(1-p)^{n-k}$ representa el k-ésimo término del desarrollo del binomio $(p+q)^n$.

Con independencia de cuál haya sido su proceso de generación, podemos afirmar que toda variable aleatoria discreta, cuya función de probabilidad venga dada por la expresión 3.1 sigue un modelo Binomial ($X \approx \mathcal{B}(n,p)$).

La expresión de la función de distribución para un modelo binomial $\mathcal{B}(n,p)$ es

$$F(x) = P(X \leq x) = \begin{cases} 0 & si\, x < 0 \\ \sum_{k=0}^{[x]} \binom{n}{k} p^k (1-p)^{n-k} & si\, 0 \leq x < n \\ 1 & si\, n \leq x \end{cases}$$

fácilmente deducible a partir de la correspondiente función de probabilidad y cuya representación gráfica se recoge en la figura 3.3.

Tanto la función de probabilidad como la de distribución -aunque sencillas- pueden resultar poco operativas para valores elevados de la variable. Para subsanar este inconveniente, las probabilidades del modelo binomial aparecen tabuladas para ciertos valores de los parámetros n y p.

El manejo de las tablas del modelo binomial, que aparecen recogidas a continuación, consiste en seleccionar la fila correspondiente al tamaño (n), a partir de la cual se elige el número de éxitos (k), y finalmente determinaremos la columna en la que se sitúa la probabilidad de éxito (p). En la intersección se obtienen las probabilidades puntuales de los valores correspondientes al modelo binomial $\mathcal{B}(n,p)$, esto es, $k = 0, 1, \ldots, n$.

A modo de ejemplo, calculemos la probabilidad de que, al lanzar 3 veces un dado, se obtengan dos resultados pares. Dado que los resultados de los lanzamientos son independientes y que las

Tabla 3.1.: Distribución Binomial $\mathcal{B}(n,p)$. Función de probabilidad

n	k/p	0,05	0,10	0,15	0,20	0,25	0,30	0,35	0,40	0,45	0,50
1	0	0,9500	0,9000	0,8500	0,8000	0,7500	0,7000	0,6500	0,6000	0,5500	0,5000
	1	0,0500	0,1000	0,1500	0,2000	0,2500	0,3000	0,3500	0,4000	0,4500	0,5000
2	0	0,9025	0,8100	0,7225	0,6400	0,5625	0,4900	0,4225	0,3600	0,3025	0,2500
	1	0,0950	0,1800	0,2550	0,3200	0,3750	0,4200	0,4550	0,4800	0,4950	0,5000
	2	0,0025	0,0100	0,0225	0,0400	0,0625	0,0900	0,1225	0,1600	0,2025	0,2500
3	0	0,8574	0,7290	0,6141	0,5120	0,4219	0,3430	0,2746	0,2160	0,1664	0,1250
	1	0,1354	0,2430	0,3251	0,3840	0,4219	0,4410	0,4436	0,4320	0,4084	0,3750
	2	0,0071	0,0270	0,0574	0,0960	0,1406	0,1890	0,2389	0,2880	0,3341	0,3750
	3	0,0001	0,0010	0,0034	0,0080	0,0156	0,0270	0,0429	0,0640	0,0911	0,1250
4	0	0,8145	0,6561	0,5220	0,4096	0,3164	0,2401	0,1785	0,1296	0,0915	0,0625
	1	0,1715	0,2916	0,3685	0,4096	0,4219	0,4116	0,3845	0,3456	0,2995	0,2500
	2	0,0135	0,0486	0,0975	0,1536	0,2109	0,2646	0,3105	0,3456	0,3675	0,3750
	3	0,0005	0,0036	0,0115	0,0256	0,0469	0,0756	0,1115	0,1536	0,2005	0,2500
	4	0,0000	0,0001	0,0005	0,0016	0,0039	0,0081	0,0150	0,0256	0,0410	0,0625
5	0	0,7738	0,5905	0,4437	0,3277	0,2373	0,1681	0,1160	0,0778	0,0503	0,0313
	1	0,2036	0,3281	0,3915	0,4096	0,3955	0,3602	0,3124	0,2592	0,2059	0,1563
	2	0,0214	0,0729	0,1382	0,2048	0,2637	0,3087	0,3364	0,3456	0,3369	0,3125
	3	0,0011	0,0081	0,0244	0,0512	0,0879	0,1323	0,1811	0,2304	0,2757	0,3125
	4	0,0000	0,0004	0,0022	0,0064	0,0146	0,0284	0,0488	0,0768	0,1128	0,1563
	5	0,0000	0,0000	0,0001	0,0003	0,0010	0,0024	0,0053	0,0102	0,0185	0,0313
6	0	0,7351	0,5314	0,3771	0,2621	0,1780	0,1176	0,0754	0,0467	0,0277	0,0156
	1	0,2321	0,3543	0,3993	0,3932	0,3560	0,3025	0,2437	0,1866	0,1359	0,0938
	2	0,0305	0,0984	0,1762	0,2458	0,2966	0,3241	0,3280	0,3110	0,2780	0,2344
	3	0,0021	0,0146	0,0415	0,0819	0,1318	0,1852	0,2355	0,2765	0,3032	0,3125
	4	0,0001	0,0012	0,0055	0,0154	0,0330	0,0595	0,0951	0,1382	0,1861	0,2344
	5	0,0000	0,0001	0,0004	0,0015	0,0044	0,0102	0,0205	0,0369	0,0609	0,0938
	6	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0018	0,0041	0,0083	0,0156
7	0	0,6983	0,4783	0,3206	0,2097	0,1335	0,0824	0,0490	0,0280	0,0152	0,0078
	1	0,2573	0,3720	0,3960	0,3670	0,3115	0,2471	0,1848	0,1306	0,0872	0,0547
	2	0,0406	0,1240	0,2097	0,2753	0,3115	0,3177	0,2985	0,2613	0,2140	0,1641
	3	0,0036	0,0230	0,0617	0,1147	0,1730	0,2269	0,2679	0,2903	0,2918	0,2734
	4	0,0002	0,0026	0,0109	0,0287	0,0577	0,0972	0,1442	0,1935	0,2388	0,2734
	5	0,0000	0,0002	0,0012	0,0043	0,0115	0,0250	0,0466	0,0774	0,1172	0,1641
	6	0,0000	0,0000	0,0001	0,0004	0,0013	0,0036	0,0084	0,0172	0,0320	0,0547
	7	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0006	0,0016	0,0037	0,0078
8	0	0,6634	0,4305	0,2725	0,1678	0,1001	0,0576	0,0319	0,0168	0,0084	0,0039
	1	0,2793	0,3826	0,3847	0,3355	0,2670	0,1977	0,1373	0,0896	0,0548	0,0313
	2	0,0515	0,1488	0,2376	0,2936	0,3115	0,2965	0,2587	0,2090	0,1569	0,1094
	3	0,0054	0,0331	0,0839	0,1468	0,2076	0,2541	0,2786	0,2787	0,2568	0,2188
	4	0,0004	0,0046	0,0185	0,0459	0,0865	0,1361	0,1875	0,2322	0,2627	0,2734
	5	0,0000	0,0004	0,0026	0,0092	0,0231	0,0467	0,0808	0,1239	0,1719	0,2188
	6	0,0000	0,0000	0,0002	0,0011	0,0038	0,0100	0,0217	0,0413	0,0703	0,1094
	7	0,0000	0,0000	0,0000	0,0001	0,0004	0,0012	0,0033	0,0079	0,0164	0,0313
	8	0,0000	0,0000	0,0000	0,0000	0,0000	0,0001	0,0002	0,0007	0,0017	0,0039

probabilidades de las distintas caras del dado son constantes, la variable aleatoria X:"número de resultados pares en tres lanzamientos de dado" sigue un modelo binomial con $n = 3$ y $p = P(\text{"}par\text{"}) = \frac{1}{2}$.

Seleccionando en la tabla ambos parámetros se obtienen las probabilidades correspondientes a los valores de la variable: 0, 1, 2 y 3. Para nuestro ejemplo concreto, se tiene por tanto $P(X = 2) = 0,375$.

El objetivo de las tablas es recoger en un espacio limitado información amplia sobre las probabilidades binomiales. Este ahorro de espacio se consigue limitando los recorridos de los valores n y p, y también aprovechando algunas propiedades del modelo binomial. Así, cuando el valor de p sea mayor que 0,5, la simetría entre éxitos y fracasos permite también obtener probabilidades de un modelo $\mathcal{B}(n,p)$ a partir de su relación con $\mathcal{B}(n,q)$. [Comprobar que si $X \approx \mathcal{B}(n,p)$ e $Y \approx \mathcal{B}(n,q)$, entonces: $P(X = k) = P(Y = n - k)$]

Por su parte, los valores de n contemplados no suelen exceder 10 o 12, hecho que se debe a que -como veremos más adelante- a medida que el tamaño n crece, el modelo binomial puede ser aproximado por la distribución normal.

La siguiente tabla recoge la función de distribución binomial para algunos valores de n y p.

El uso de esta tabla es análogo al descrito en el caso anterior, sin más que tener en cuenta que ahora para cada n y p se recogen en columna las probabilidades acumuladas hasta los valores $k = 0, 1, \ldots, n$. Esta tabla resulta muy adecuada para responder a preguntas del tipo: probabilidad de que el número de éxitos sea al menos x_1, probabilidad de obtener a lo sumo x_2 éxitos, ...

Las características esperanza y varianza del modelo binomial vienen dadas en función de sus parámetros a partir de las expresiones $\mu = np$ y $\sigma^2 = npq$.

La obtención de las características esperanza y varianza puede ser efectuada a partir del binomio de Newton. En efecto, el valor esperado se obtiene como:

$$\begin{aligned} E(X) = & \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} = \sum_{k=1}^{n}\frac{n!}{(k-1)!(n-k)!}p^k(1-p)^{n-k} = \\ = & \; np\sum_{k=1}^{n}\frac{(n-1)!}{(k-1)!(n-1-(k-1))!}p^{k-1}(1-p)^{n-1-(k-1)} \end{aligned}$$

y haciendo r=k-1 se tiene:

$$E(X) = np\sum_{r=0}^{n-1}\frac{(n-1)!}{r!(n-1-r)!}p^r q^{n-1-r} = np(p+q)^{n-1} = np$$

Por su parte, la varianza de la variable viene dada por $Var(X) = E(X)^2 - E^2(X)$, cuyo cálculo resulta más sencillo mediante la expresión $Var(X) = E[X(X-1)+X] - E^2(X)$

En efecto, se tiene mediante un procedimiento análogo al cálculo de la esperanza:

$$\begin{aligned} E[X(X-1)] = & \sum_{k=0}^{n} k(k-1)\frac{n!}{k!(n-k)!}p^k q^{n-k} = \sum_{k=2}^{m}\frac{n!}{(k-2)!(n-k)!}p^k q^{n-k} = \\ = & \; n(n-1)p^2\sum_{k=2}^{n}\frac{(n-2)!}{(k-2)!(n-k)!}p^{k-2}q^{n-k} \end{aligned}$$

que haciendo $r = k - 2$ conduce a:

Tabla 3.2.: Modelo Binomial $\mathcal{B}(n,p)$. Función de Distribución

n	$k\backslash p$	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
1	0	0,9000	0,8000	0,7000	0,6000	0,5000	0,4000	0,3000	0,2000	0,1000
	1	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
2	0	0,8100	0,6400	0,4900	0,3600	0,2500	0,1600	0,0900	0,0400	0,0100
	1	0,9900	0,9600	0,9100	0,8400	0,7500	0,6400	0,5100	0,3600	0,1900
	2	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
3	0	0,7290	0,5120	0,3430	0,2160	0,1250	0,0640	0,0270	0,0080	0,0010
	1	0,9720	0,8960	0,7840	0,6480	0,5000	0,3520	0,2160	0,1040	0,0280
	2	0,9990	0,9920	0,9730	0,9360	0,8750	0,7840	0,6570	0,4880	0,2710
	3	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
4	0	0,6561	0,4096	0,2401	0,1296	0,0625	0,0256	0,0081	0,0016	0,0001
	1	0,9477	0,8192	0,6517	0,4752	0,3125	0,1792	0,0837	0,0272	0,0037
	2	0,9963	0,9728	0,9163	0,8208	0,6875	0,5248	0,3483	0,1808	0,0523
	3	0,9999	0,9984	0,9919	0,9744	0,9375	0,8704	0,7599	0,5904	0,3439
	4	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
5	0	0,5905	0,3277	0,1681	0,0778	0,0312	0,0102	0,0024	0,0003	0,0000
	1	0,9185	0,7373	0,5282	0,3370	0,1875	0,0870	0,0308	0,0067	0,0005
	2	0,9914	0,9421	0,8369	0,6826	0,5000	0,3174	0,1631	0,0579	0,0086
	3	0,9995	0,9933	0,9692	0,9130	0,8125	0,6630	0,4718	0,2627	0,0815
	4	1,0000	0,9997	0,9976	0,9898	0,9688	0,9222	0,8319	0,6723	0,4095
	5	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
6	0	0,5314	0,2621	0,1176	0,0467	0,0156	0,0041	0,0007	0,0001	0,0000
	1	0,8857	0,6554	0,4202	0,2333	0,1094	0,0410	0,0109	0,0016	0,0001
	2	0,9841	0,9011	0,7443	0,5443	0,3438	0,1792	0,0705	0,0170	0,0013
	3	0,9987	0,9830	0,9295	0,8208	0,6562	0,4557	0,2557	0,0989	0,0158
	4	0,9999	0,9984	0,9891	0,9590	0,8906	0,7667	0,5798	0,3446	0,1143
	5	1,0000	0,9999	0,9993	0,9959	0,9844	0,9533	0,8824	0,7379	0,4686
	6	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
7	0	0,4783	0,2097	0,0824	0,0280	0,0078	0,0016	0,0002	0,0000	0,0000
	1	0,8503	0,5767	0,3294	0,1586	0,0625	0,0188	0,0038	0,0004	0,0000
	2	0,9743	0,8520	0,6471	0,4199	0,2266	0,0963	0,0288	0,0047	0,0002
	3	0,9973	0,9667	0,8740	0,7102	0,5000	0,2898	0,1260	0,0333	0,0027
	4	0,9998	0,9953	0,9712	0,9037	0,7734	0,5801	0,3529	0,1480	0,0257
	5	1,0000	0,9996	0,9962	0,9812	0,9375	0,8414	0,6706	0,4233	0,1497
	6	1,0000	1,0000	0,9998	0,9984	0,9922	0,9720	0,9176	0,7903	0,5217
	7	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
8	0	0,4305	0,1678	0,0576	0,0168	0,0039	0,0007	0,0001	0,0000	0,0000
	1	0,8131	0,5033	0,2553	0,1064	0,0352	0,0085	0,0013	0,0001	0,0000
	2	0,9619	0,7969	0,5518	0,3154	0,1445	0,0498	0,0113	0,0012	0,0000
	3	0,9950	0,9437	0,8059	0,5941	0,3633	0,1737	0,0580	0,0104	0,0004
	4	0,9996	0,9896	0,9420	0,8263	0,6367	0,4059	0,1941	0,0563	0,0050
	5	1,0000	0,9988	0,9887	0,9502	0,8555	0,6846	0,4482	0,2031	0,0381
	6	1,0000	0,9999	0,9987	0,9915	0,9648	0,8936	0,7447	0,4967	0,1869
	7	1,0000	1,0000	0,9999	0,9993	0,9961	0,9832	0,9424	0,8322	0,5695
	8	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

$$E[X(X-1)] = n(n-1)p^2 \sum_{r=0}^{n-2} \frac{(n-2)!}{r!(n-2-r)!} p^r q^{n-2-k} = (n(n-1)p^2(p+q)^{n-2} = n(n-1)p^2$$

Así pues, sustituyendo esta expresión se obtiene para la varianza:

$$Var(X) = E[X(X-1)] + E(X) - E^2(X) = n(n-1)p^2 + np - (np)^2 = np(1-p) = npq$$

Los cálculos anteriores resultan más sencillos a partir de la función generatriz de momentos, que para un modelo $\mathcal{B}(n,p)$ viene dada por la expresión $M_X(t) = (e^t p + q)^n$ para $-\infty < t < \infty$.

Partiendo de la definición de función generatriz de momentos se obtiene:

$$M_X(t) = E\left(e^{tX}\right) = \sum_{k=0}^{n} e^{tk} p^k q^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (e^t p)^k q^{n-k} = \left(e^t p + q\right)^n$$

Como ya hemos comprobado en el capítulo anterior, la función generatriz de momentos M proporciona los momentos de cualquier orden r como valor particular en el punto $t = 0$ de la correspondiente derivada de orden r. Es decir:

$$E\left(X^r\right) = \left.\frac{d^r M_X(t)}{dt^r}\right|_{t=0}$$

[Comprobar que $E(X) = \left.\frac{dM_X(t)}{dt}\right|_{t=0} = np$]

El valor esperado puede ser interpretado como "valor al que tiende el promedio de éxitos al aumentar indefinidamente en idénticas condiciones el número de pruebas". Por su parte, la varianza de X cuantifica el riesgo, aumentando con el número de pruebas y también a medida que se aproximan los valores p y q.

Así, si en el ejemplo de los empresarios convocados para la entrevista asumimos que la probabilidad de aceptación es de un 80 %, el número esperado de empresarios entrevistados será 16 y la varianza de 3,2.

Si por el contrario el valor de p fuese 0,56 se tendría $\mu = 11,2$ y $\sigma^2 = 4,928$.

[¿Qué cambios significativos se han producido en los parámetros?] [¿cómo se interpretaría el valor esperado 11,2 entrevistados?]

El interés del modelo binomial es considerable. Además de su generalizada aplicación en muchos casos prácticos, es posible establecer conexiones entre esta distribución y otros modelos probabilísticos que estudiaremos en apartados posteriores.

Así, podemos definir nuevos modelos probabilísticos con sólo alterar la definición de las variables (distribución geométrica y binomial negativa) o bien algunos de los supuestos en los que se basa el modelo binomial (distribución hipergeométrica).

Por otra parte, las probabilidades binomiales pueden ser aproximadas -para valores elevados de n y pequeños de p- por la distribución denominada de *Poisson* y es posible -bajo ciertos supuestos que analizaremos en temas posteriores- aproximar cualquiera de estas distribuciones por el *modelo normal*.

3.2. Distribuciones Geométrica y Binomial negativa

Los supuestos en que se basan los procesos de Bernoulli (independencia entre pruebas y valor constante de la probabilidad de éxito p) permiten definir nuevos modelos probabilísticos.

Así, podríamos suponer ahora que, en vez de prefijar el número de empresarios convocados a las entrevistas, se decide llevar a cabo consultas sucesivas hasta tener confirmado determinado número de entrevistas.

Las hipótesis de independencia entre las respuestas y probabilidad constante de aceptación (p) siguen siendo válidas, pero cambia sin embargo la variable aleatoria de interés, que vendría ahora definida como "número de consultas hasta obtener las entrevistas necesarias".

La situación más sencilla sería efectuar consultas hasta confirmar una entrevista, variable recogida por la *distribución geométrica*.

Definición 3.3. Dado un proceso de Bernoulli de pruebas independientes con sólo dos alternativas y probabilidad de éxito (p) constante, la magnitud aleatoria X definida como "número de pruebas necesarias hasta la obtención del primer éxito" sigue un modelo denominado *geométrico* que abreviadamente se representa por $\mathcal{G}(p)$.

Aunque presenta rasgos comunes con el modelo binomial, la distribución geométrica resulta mucho más sencilla, dado que sólo debemos cuantificar la probabilidad de que en cierta observación aparezca el primer éxito, sin preocuparnos de las observaciones anteriores (en las que todos los resultados fueron fracasos y por tanto idénticos entre sí).

Como consecuencia de estas características, el recorrido de una variable $X \approx \mathcal{G}(p)$ será infinito numerable: $1, 2, \ldots$ y su función de probabilidad viene dada por la expresión $P(X = k) = (1-p)^{k-1}p$, en la que se aprecia que p -probabilidad de éxito- es el parámetro característico del modelo.

Toda variable aleatoria discreta cuya función de probabilidad venga dada por la expresión:

$$P(X = k) = (1-p)^{k-1}p \; ; \; con\, k = 1, 2, \ldots$$

se dice que sigue un *modelo Geométrico o de Pascal* ($X \approx \mathcal{G}(p)$).

El matemático y filósofo Blaise Pascal (1623-1662), cuyo apellido se utiliza para designar la distribución geométrica, es -gracias en gran medida a su correspondencia con Pierre Fermat- autor de algunos de los fundamentos de la ciencia de la probabilidad, hasta el punto de que Laplace considera a ambos autores como precursores de la Teoría de la Probabilidad.

Vamos a analizar las características de la distribución geométrica. Para ello comencemos por justificar que la expresión vista para $P(X = k)$ es una verdadera función de probabilidad; esto es, se trata de una función no negativa (puesto que los factores que intervienen son no negativos) y además su suma es la unidad:

$$\sum_{k=0}^{\infty} P(X = k) = \sum_{k=1}^{\infty} pq^{k-1} = p\sum_{k=1}^{\infty} q^{k-1}$$

La última suma corresponde a una progresión geométrica de razón q, lo cual justifica el nombre que recibe esta distribución. Cuando la razón es menor que la unidad (q<1) la serie geométrica es convergente y su suma es el primer término de la serie partido por uno menos la razón. En este caso:

$$\sum_{k=0}^{\infty} P(X=k) = p\frac{q^0}{1-q} = \frac{p}{p} = 1$$

La función de distribución de esta variable será:

$$F(x) = P(X \leq x) = \sum_{k=0}^{[x]} P(X=k) = 1 - \sum_{k=[x]+1}^{\infty} pq^{k-1} = 1 - p\sum_{k=[x]+1}^{\infty} q^{k-1} = 1 - p\frac{q^{[x]}}{1-q} = 1 - q^{[x]}$$

En ciertas ocasiones se plantea una versión alternativa del modelo geométrico, definiendo la variable X' como "Número de fracasos antes del primer éxito". Resulta sencillo deducir la función de probabilidad, que en este caso viene dada por la expresión:

$$P(X'=k) = (1-p)^k p \; ; \; con\, k = 0, 1, 2, \ldots$$

a la que es también posible llegar mediante un cambio de variable (si designamos por X y X' a las variables "número de pruebas hasta el primer éxito" y "número de fracasos antes del primer éxito" respectivamente, se tendría $X = X' + 1$).

Una v.a. geométrica puede tomar infinitos valores y la probabilidad de la cola aumenta conforme va disminuyendo la probabilidad de éxito.

El valor esperado de esta distribución es $\mu = \frac{1}{p}$, expresión a la que se llega a partir del desarrollo $E(X) = \sum_{k=0}^{\infty} kpq^{k-1}$. Como es lógico, esta característica variará de modo inverso con la probabilidad de éxito. Por su parte, la varianza viene dada por la expresión $Var(X) = \frac{q}{p^2}$.

Para obtener el valor esperado, debemos tener en cuenta que la derivada de la suma es la suma de las derivadas y que la serie es convergente, con lo cual esta característica viene dada por:

$$\begin{aligned} E(X) = & \; p\sum_{k=1}^{\infty} kq^{k-1} = p\sum_{k=1}^{\infty} \frac{d}{dq}q^k = p\frac{d}{dq}\left(\sum_{k=1}^{\infty} q^k\right) = p\frac{d}{dq}\left(\frac{q}{1-q}\right) \\ = & \; p\frac{(1-q)+q}{(1-q)^2} = \frac{p}{p^2} = \frac{1}{p} \end{aligned}$$

Por lo que se refiere a la varianza, ésta puede ser expresada como:

$$Var(X) = E(X^2) - E^2(X) = E[X(X-1)] + E(X) - E^2(X)$$

donde todos los sumandos son conocidos a excepción de $E[X(X-1)]$ que puede ser calculado de forma análoga al caso de la esperanza, donde ahora aparecerá una deri-

vada segunda.

Tanto el valor esperado como la varianza pueden ser obtenidos fácilmente a partir de la función generatriz de momentos del modelo, que viene dada por

$$M_X(t) = E(e^{tX}) = \sum_{k=1}^{\infty} e^{tk} q^{k-1} p = \frac{p}{q} \sum_{k=1}^{\infty} \left(e^t q\right)^k = \frac{e^t p}{1 - e^t q}$$

a partir de la cual se obtiene $E(X) = \frac{d}{dt} M_X(t)|_{t=0} = \frac{1}{p}$.

[Compruébese de modo análogo que $E(X^2) = \frac{d^2}{dt^2} M_X(t)|_{t=0} = \frac{1+q}{p^2}$, y en consecuencia se obtiene de nuevo: $Var(X) = \frac{q}{p^2}$]

Las tablas 3.3 y 3.4 recogen la función de probabilidad y de distribución del modelo geométrico.

El manejo de tablas de la distribución geométrica es similar al descrito para la distribución binomial: en la primera columna se recoge el número de pruebas necesarias para obtener el primer éxito y en las restantes columnas se han seleccionado ciertos valores de p.

Como ya se comentó a la vista de los gráficos, las tablas confirman cómo al aumentar la probabilidad de éxito p la probabilidad de la cola se hace menor y así, aunque el número de pruebas para obtener un éxito pueden ser infinitas, se observa que con $p = 0,5$ obtenemos una probabilidad casi nula a partir de la prueba número 15.

Son numerosas las aplicaciones prácticas del modelo geométrico, que podría resultar útil para describir -bajo los supuestos de independencia y probabilidad constante- el número de apuestas efectuadas por un jugador hasta obtener premio, las visitas de un viajante hasta vender un artículo, las convocatorias de examen a las que acude un alumno hasta obtener un aprobado, los días que un individuo mira el buzón hasta recibir cierta carta, ...

[¿Sería adecuado en estos ejemplos el supuesto de p constante? ¿y el de independencia entre las observaciones?]

La hipótesis de independencia garantiza que la probabilidad de que sean necesarios más de k nuevos intentos para obtener el primer éxito no se ve afectada por el número de pruebas que ya llevemos realizadas. Esta propiedad se conoce como "pérdida de memoria" de la distribución geométrica.

El resultado anterior puede formalizarse como sigue: Si X es una v.a. $\mathcal{G}(p)$, entonces se cumple: $P(X > k + m/X > m) = P(X > k)$. En efecto:

$$\begin{aligned} P(X > k + m/X > m) = & \frac{P(X > k + m, X > m)}{p(X > m)} = \frac{P(X > k + m)}{P(X > m)} = \frac{1 - F_X(k + m)}{1 - F_X(m)} = \\ = & \frac{1 - \left(1 - q^{k+m}\right)}{1 - \left(1 - q^m\right)} = q^k = 1 - P(X \leq k) = P(X > k) \end{aligned}$$

La interpretación de esta expresión es la siguiente: la información de que hemos realizado ya m

Tabla 3.3.: Modelo Geométrico. Función de probabilidad

$k \backslash p$	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
1	0,0900	0,1600	0,2100	0,2400	0,2500	0,2400	0,2100	0,1600	0,0900
2	0,0810	0,1280	0,1470	0,1440	0,1250	0,0960	0,0630	0,0320	0,0090
3	0,0729	0,1024	0,1029	0,0864	0,0625	0,0384	0,0189	0,0064	0,0009
4	0,0656	0,0819	0,0720	0,0518	0,0312	0,0154	0,0057	0,0013	0,0001
5	0,0590	0,0655	0,0504	0,0311	0,0156	0,0061	0,0017	0,0003	
6	0,0531	0,0524	0,0353	0,0187	0,0078	0,0025	0,0005	0,0001	
7	0,0478	0,0419	0,0247	0,0112	0,0039	0,0010	0,0002		
8	0,0430	0,0336	0,0173	0,0067	0,0020	0,0004			
9	0,0387	0,0268	0,0121	0,0040	0,0010	0,0002			
10	0,0349	0,0215	0,0085	0,0024	0,0005	0,0001			
11	0,0314	0,0172	0,0059	0,0015	0,0002				
12	0,0282	0,0137	0,0042	0,0009	0,0001				
13	0,0254	0,0110	0,0029	0,0005	0,0001				
14	0,0229	0,0088	0,0020	0,0003					
15	0,0206	0,0070	0,0014	0,0002					
16	0,0185	0,0056	0,0010	0,0001					
17	0,0167	0,0045	0,0007	0,0001					
18	0,0150	0,0036	0,0005						
19	0,0135	0,0029	0,0003						
20	0,0122	0,0023	0,0002						
21	0,0109	0,0018	0,0002						
22	0,0098	0,0015	0,0001						
23	0,0089	0,0012	0,0001						
24	0,0080	0,0009	0,0001						
25	0,0072	0,0008							
26	0,0065	0,0006							
27	0,0058	0,0005							
28	0,0052	0,0004							
29	0,0047	0,0003							
30	0,0042	0,0002							
40	0,0015								
50	0,0005								
100									

Tabla 3.4.: Modelo Geométrico. Función de distribución

$k\backslash p$	0,10	0,20	0,30	0,40	0,50	0,60	0,70	0,80	0,90
1	0,1900	0,3600	0,5100	0,6400	0,7500	0,8400	0,9100	0,9600	0,9900
2	0,2710	0,4880	0,6570	0,7840	0,8750	0,9360	0,9730	0,9920	0,9990
3	0,3439	0,5904	0,7599	0,8704	0,9375	0,9744	0,9919	0,9984	0,9999
4	0,4095	0,6723	0,8319	0,9222	0,9688	0,9898	0,9976	0,9997	1,0000
5	0,4686	0,7379	0,8824	0,9533	0,9844	0,9959	0,9993	0,9999	1,0000
6	0,5217	0,7903	0,9176	0,9720	0,9922	0,9984	0,9998	1,0000	1,0000
7	0,5695	0,8322	0,9424	0,9832	0,9961	0,9993	0,9999	1,0000	1,0000
8	0,6126	0,8658	0,9596	0,9899	0,9980	0,9997	1,0000	1,0000	1,0000
9	0,6513	0,8926	0,9718	0,9940	0,9990	0,9999	1,0000	1,0000	1,0000
10	0,6862	0,9141	0,9802	0,9964	0,9995	1,0000	1,0000	1,0000	1,0000
11	0,7176	0,9313	0,9862	0,9978	0,9998	1,0000	1,0000	1,0000	1,0000
12	0,7458	0,9450	0,9903	0,9987	0,9999	1,0000	1,0000	1,0000	1,0000
13	0,7712	0,9560	0,9932	0,9992	0,9999	1,0000	1,0000	1,0000	1,0000
14	0,7941	0,9648	0,9953	0,9995	1,0000	1,0000	1,0000	1,0000	1,0000
15	0,8147	0,9719	0,9967	0,9997	1,0000	1,0000	1,0000	1,0000	1,0000
16	0,8332	0,9775	0,9977	0,9998	1,0000	1,0000	1,0000	1,0000	1,0000
17	0,8499	0,9820	0,9984	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000
18	0,8649	0,9856	0,9989	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000
19	0,8784	0,9885	0,9992	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
20	0,8906	0,9908	0,9994	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
21	0,9015	0,9926	0,9996	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
22	0,9114	0,9941	0,9997	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
23	0,9202	0,9953	0,9998	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
24	0,9282	0,9962	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
25	0,9354	0,9970	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
26	0,9419	0,9976	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
27	0,9477	0,9981	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
28	0,9529	0,9985	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
29	0,9576	0,9988	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
30	0,9618	0,9990	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
40	0,9867	0,9999	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
50	0,9954	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000
100	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

pruebas sin éxito ($X > m$), no altera la probabilidad de que aún necesitemos k pruebas más hasta obtener un éxito. Así pues, el modelo geométrico no tiene memoria.

Una extensión natural de la distribución geométrica se obtiene cuando nos interesa observar las pruebas de Bernoulli necesarias hasta obtener un número determinado de éxitos (por ejemplo, si se realizasen las llamadas necesarias hasta conseguir 6 entrevistas con empresarios). La variable aleatoria definida en este caso se adapta a la distribución denominada *binomial negativa*.

Definición 3.4. Bajo los supuestos de un proceso de Bernoulli, una variable aleatoria X definida como "número de pruebas hasta el r-ésimo éxito" se distribuye según un *modelo binomial negativo*, denotado por $\mathcal{BN}(r, p)$.

La función de probabilidad de la variable viene dada por $P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$ donde $k = r, r+1, \ldots .$, expresión que puede ser justificada con un razonamiento similar al de los modelos anteriores.

En este caso, la probabilidad de que sean necesarias k pruebas hasta los r éxitos se obtiene tomando en cuenta dos factores: la probabilidad de intersección de r éxitos y $k - r$ fracasos y las posibles secuencias en que estas situaciones pueden presentarse. Obsérvese que para este segundo factor ignoramos la última prueba, que necesariamente corresponde al éxito r-ésimo, con lo cual se calculan las combinaciones $\binom{k-1}{r-1}$.

La denominación de esta distribución queda justificada en el esquema que sigue, donde se recogen paralelamente las condiciones del modelo binomial negativo y las correspondientes a la distribución binomial.

Como puede observarse, los rasgos distintivos de ambas distribuciones se refieren al papel aleatorio que el modelo binomial asigna al número de éxitos (fijadas las pruebas) y viceversa para la binomial negativa.

	Binomial $\mathcal{B}(n, p)$	Binomial Negativa $\mathcal{BN}(r, p)$
Probabilidad éxito	p	p
Núm. de pruebas	n(dado)	X(aleatorio)
Núm. de éxitos	Y(aleatorio)	r(dado)
Func. de Probabilidad	$P(Y = k) = \binom{n}{k} p^k q^{n-k}$ $k = 0, 1, 2, \ldots$	$P(X = k) = \binom{k-1}{r-1} p^r q^{k-r}$ $k = r, r+1, \ldots$

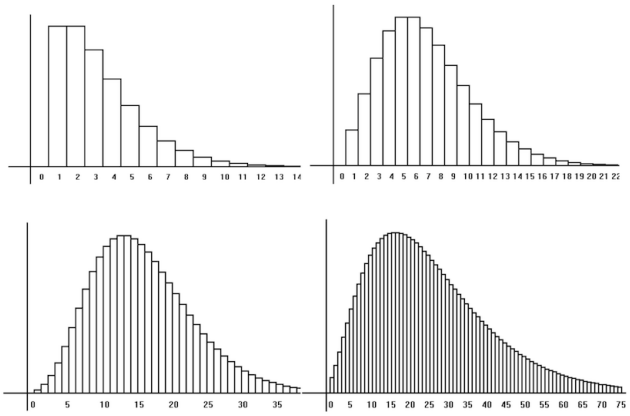
Como consecuencia de esta conexión entre ambas distribuciones de probabilidad puede comprobarse que dadas las variables $Y \approx \mathcal{B}(n, p)$ y $X \approx \mathcal{BN}(r, p)$ se cumple: $P(Y \geq r) = P(X \leq n)$ y también $P(Y < r) = P(X > n)$.

Por otra parte, podemos observar fácilmente que el caso particular $X \approx \mathcal{BN}(r = 1, p)$ coincide con la distribución geométrica $\mathcal{G}(p)$.

La gráfica de la función de probabilidad de la distribución binomial negativa viene condicionada por los dos parámetros r y p; así en la figura 3.4 representamos 4 situaciones que muestran estas diferencias:

El primero de los gráficos de esta figura recoge la distribución de una v.a. $\mathcal{BN}(3,\ 0,5)$, en cuya representación observamos que se mantiene la asimetría característica de la

Figura 3.4.: Modelo Binomial negativo. Función de probabilidad



distribución geométrica; en cambio en el siguiente gráfico (superior derecha) se recoge una variable $\mathcal{BN}(7,\, 0,5)$, que presenta un menor grado de asimetría.

El gráfico inferior izquierdo corresponde a una $\mathcal{BN}(7,\, 0,3)$, en la que se observa una mayor simetría, como consecuencia de la disminución en la probabilidad de éxito y finalmente, en el último gráfico se recoge una v.a. $\mathcal{BN}(3,\, 0,1)$. Podemos concluir a la vista de estos resultados que a medida que aumenta el número de éxitos (r) o disminuye la probabilidad de éxito (p), la representación se aproxima hacia una función campaniforme.

El número esperado de pruebas hasta obtener r éxitos viene dado por la expresión $E(X) = \dfrac{r}{p}$ y la varianza del modelo es $Var(X) = \dfrac{rq}{p^2}$.

En ocasiones la distribución binomial negativa se asocia a la variable $X\,'$="número de fracasos obtenidos antes del r-ésimo éxito", definición que conduce a expresiones distintas a las estudiadas tanto para la función de probabilidad como para las características del modelo. Más concretamente, se tendría en este caso una función de probabilidad dada por:

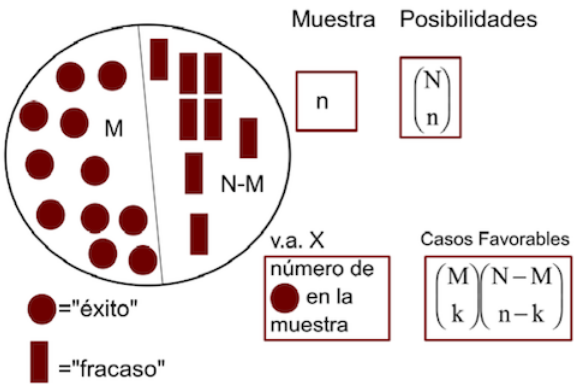
$$P(X' = k) = \binom{r+k-1}{k} p^r q^k \;;\; k = 0, 1, \ldots$$

observándose entonces con la variable X "número de pruebas hasta el r-ésimo éxito" la relación $X = X\,' + r$.

3.3. Modelo hipergeométrico

A menudo la realidad sobre la que efectuamos nuestras observaciones dista de los supuestos establecidos por los modelos probabilísticos. Así, recordando los ejemplos que recogíamos anteriormente podríamos plantearnos qué sucede si la respuesta de un

Figura 3.5.: Probabilidad Hipergeométrica



empresario afecta a las restantes, con lo cual los resultados de las pruebas dejan de ser independientes.

En estas situaciones se incumplen las hipótesis de independencia y de probabilidad constante asumidas en el proceso de Bernoulli, por lo cual, aun cuando nos siga interesando estudiar los elementos que presentan cierta característica, queda excluida la utilización del modelo binomial, resultando adecuada la *distribución hipergeométrica*.

Las condiciones en las que se define este modelo de probabilidad son las siguientes: consideramos una población total integrada por N elementos (empresarios, alumnos presentados a un examen, candidatos a un empleo, ...) sobre los que nos interesa estudiar determinada característica, que podríamos seguir denominando "éxito" (acceder a la entrevista, aprobar el examen, obtener el empleo, ...).

Definición 3.5. Supongamos clasificados los integrantes de la población según la característica de interés, tal y como indica el esquema 3.5: M elementos presentan el rasgo estudiado y $(N - M)$ no lo presentan.

Si de la población total seleccionamos aleatoriamente y sin reposición una muestra de n elementos, el número de ellos que presentan la característica analizada (éxitos) es una variable aleatoria que sigue una *distribución hipergeométrica* $\mathcal{H}(N, M, n)$.

Este modelo probabilístico aparece directamente asociado al análisis combinatorio ya que las condiciones del modelo equivalen a una selección aleatoria de n elementos extraídos simultáneamente (sin reposición) de una población de tamaño N.

Como consecuencia, la probabilidad de éxito no es constante y el número de posibilidades de selección coincide con los subconjuntos de n elementos extraídos sin reposición entre N, que pueden ser cuantificados mediante la fórmula de las combinaciones:

$$C_{N,n} = \binom{N}{n}$$

Dado que la selección es aleatoria, cada uno de estos grupos de tamaño n tiene idéntica probabilidad de ser seleccionado. Por tanto, se trata de sucesos equiprobables resultando aplicable la expresión de la probabilidad clásica. Así pues, se tiene:

$$P(X=k)=\frac{C_{M,k}C_{N-M,n-k}}{C_{N,n}}=\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}$$

[Justificar cómo ha sido obtenida la expresión del numerador][Si los n elementos se seleccionasen con reposición ¿cuál sería el modelo probabilístico para el número de éxitos?]

En la ilustración del reportaje sobre empresarios se puede encontrar un modelo hipergeométrico asociado a las fotografías de la portada. Supongamos que de un total de 15 entrevistados -de los cuales 6 eran mujeres- se seleccionarán aleatoriamente 3 fotografías para la portada.

La variable aleatoria X que describe el número de mujeres que aparecen en portada viene descrita por un modelo $\mathcal{H}(N=15, M=6, n=3)$. [¿Por qué es obvio que no se trata de un modelo binomial?]

Aunque en principio podría parecer razonable que el recorrido de esta variable oscilase entre 0 y n (y así sucede en nuestro ejemplo, ya que puede haber en portada de 0 a 3 mujeres) es necesario tener presente tanto el número de elementos seleccionados (n) como el tamaño de elementos de la población con el rasgo estudiado (M). Así, por ejemplo ¿qué sucedería si sólo hubiese dos mujeres entre los 15 entrevistados? Es evidente que X no podría exceder el valor 2. ¿Y si por el contrario 13 de los 15 entrevistados fueran mujeres?; podríamos asegurar en ese caso que $X \geq 1$.

En definitiva, resulta necesario contemplar situaciones en las que el colectivo de interés tenga menos de n elementos ($M < n$) en cuyo caso el recorrido de X no podría exceder el valor M. Además la cota inferior no necesariamente se sitúa en 0, pudiendo existir un número de elementos extraídos que obligatoriamente pertenecerán al colectivo (en concreto, el valor inferior de X es el máximo entre 0 y la diferencia $n-(N-M)$).

Como consecuencia de este razonamiento se tiene $\text{máx}\{0, n-(N-M)\} \leq k \leq \text{mín}\{n, M\}$, cumpliéndose

$$\sum_{k=\text{máx}\{0,n-(N-M)\}}^{\text{mín}\{n,M\}} \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = 1$$

La función de probabilidad definida por la variable hipergeométrica cumple las condiciones exigidas a éstas al ser no negativa [¿por qué?], y de suma unitaria. Para comprobar este segundo aspecto basta tener en cuenta una propiedad de los números combinatorios según la cual:

$$\binom{N}{n}=\sum_{x=0}^{n}\binom{M}{x}\binom{N-M}{n-x}$$

Las características del modelo hipergeométrico son las siguientes:

$$E(X) = n\frac{M}{n} = np$$

$$Var(X) = n\frac{M}{n}\left(1 - \frac{M}{N}\right)\left(\frac{N-n}{N-1}\right) = npq\left(\frac{N-n}{N-1}\right)$$

en las que pueden apreciarse similitudes con el modelo binomial: el valor esperado se obtiene de modo análogo (siendo $p = \frac{M}{N}$) y el riesgo disminuye como consecuencia del factor de corrección $\left(\frac{N-n}{N-1}\right)$ correspondiente al rasgo de selección sin reposición.

El factor de corrección (que será inferior a la unidad para $n > 1$) resulta de gran interés en el muestreo de poblaciones finitas, puesto que incorpora el ajuste en la dispersión de la variable que se produce como consecuencia del muestreo sin reposición, esto es, al eliminar el riesgo inherente a las observaciones repetidas.

Puede comprobarse que, a medida que el tamaño poblacional N aumenta, este factor de corrección se aproxima a la unidad, de tal modo que en poblaciones conceptualmente infinitas resulta irrelevante que el muestreo se efectúe con o sin reposición.

El cálculo del valor esperado se efectúa mediante el desarrollo siguiente:

$$E(X) = \sum_{k=\text{máx}\{0,n-(N-M)\}}^{\text{mín}\{n,M\}} k\frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}} = M\sum_{k=\text{máx}\{0,n-(N-M)\}}^{\text{mín}\{n,M\}} \frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N}{n}} =$$

$$= \frac{Mn}{N}\sum_{k=\text{máx}\{0,n-(N-M)\}}^{\text{mín}\{n,M\}} \frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N-1}{n-1}} = \frac{Mn}{N}$$

ya que se cumple

$$\sum_{k=\text{máx}\{0,n-(N-M)\}}^{\text{mín}\{n,M\}} \frac{\binom{M-1}{k-1}\binom{N-M}{n-k}}{\binom{N-1}{n-1}} = 1$$

Mediante un método similar se obtiene la varianza σ^2, característica que puede ser expresada como:

$$Var(X) = E\left[X(X-1)\right] + E(X) - E^2(X)$$

[Compruébese que se cumple $E\left[X(X-1)\right] = \frac{M(M-1)n(n-1)}{N(N-1)}$ y utilizando el resultado obtenido para la esperanza se tiene $Var(X) = npq\frac{N-n}{N-1}$, siendo $p = \frac{M}{N}$]

A medida que en un modelo discreto aumenta el número de parámetros, resulta más difícil resumir en unas tablas de probabilidad la función de cuantía o la de distribución. Por ejemplo, en el caso binomial se tienen dos parámetros además del valor x cuya probabilidad tratamos de calcular, motivo por el cual una representación tridimensional sería lo más indicado. En el modelo hipergeométrico surge un nuevo parámetro con lo que la representación es más difícil.

Una alternativa a las limitaciones anteriores sería el cálculo directo de las probabilidades a partir de la expresión de la función de probabilidad, pero esto puede dar lugar a errores de aproximación importantes. Por ejemplo, si el tamaño de la población es muy elevado, podemos tener problemas de desbordamiento de memoria por manejar cifras excesivamente altas o bajas; en este caso sería recomendable factorizar las expresiones de cálculo mediante cocientes parciales que permitirían mayor exactitud (esto quiere decir que deberíamos ir simultaneando operaciones de multiplicar y dividir para mantener los resultados dentro de un tamaño razonable).

Afortunadamente, este problema puede ser resuelto considerando otras alternativas.

La principal diferencia entre los modelos binomial e hipergeométrico estriba en el tamaño de la población, ya que si ésta fuese infinita las probabilidades de selección en cada observación permanecerían constantes y el modelo podría reducirse a uno binomial. Pues bien, aunque el tamaño poblacional no sea infinito, si es suficientemente grande la aproximación binomial puede resultar satisfactoria al proporcionarnos bajos márgenes de error respecto a las probabilidades hipergeométricas. Sin embargo, para que esta aproximación sea buena, debemos tener en cuenta una cosa más: el tamaño de la muestra.

En efecto, si el tamaño de la población es elevado, en las primeras observaciones las probabilidades prácticamente no se alteran, pero si la muestra llegase al 90 % de la población, en las últimas observaciones los casos posibles se reducen a poco más del 10 % de la población original por lo que el tamaño pudo haberse vuelto pequeño. Un criterio para la sustitución de las probabilidades de la hipergeométrica por la binomial sería $N > 50$, $n < 0,1N$.

Para comprobar empíricamente el efecto de esta aproximación, recordemos el supuesto que venimos considerando respecto al número de mujeres empresarias que aparecen en la portada de la publicación. Se trata de una distribución hipergeométrica: $\mathcal{H}(N = 15, M = 6, n = 3)$, sobre la que podemos estar interesados en conocer la probabilidad de que en la portada aparezca exactamente una mujer:

$$P(X = 1) = \frac{\binom{6}{1}\binom{9}{2}}{\binom{15}{3}} = 0,474725$$

Si aproximamos esta probabilidad por una binomial, se obtendría: $p = \frac{6}{15}$, $P(X = 1) = 0,432$, pudiendo apreciarse que las diferencias de probabilidad son del orden del 10 %.

Supongamos que multiplicamos por 10 el número de entrevistas y el de mujeres, con lo cual el modelo resultante sería: $\mathcal{H}(N = 150, M = 60, n = 3)$. En esta situación la probabilidad hipergeométrica sería: $P(X = 1) = 0,43587$ mientras que la de la binomial no cambiaría [¿por qué?] y se obtienen diferencias del orden del 0,7 %.

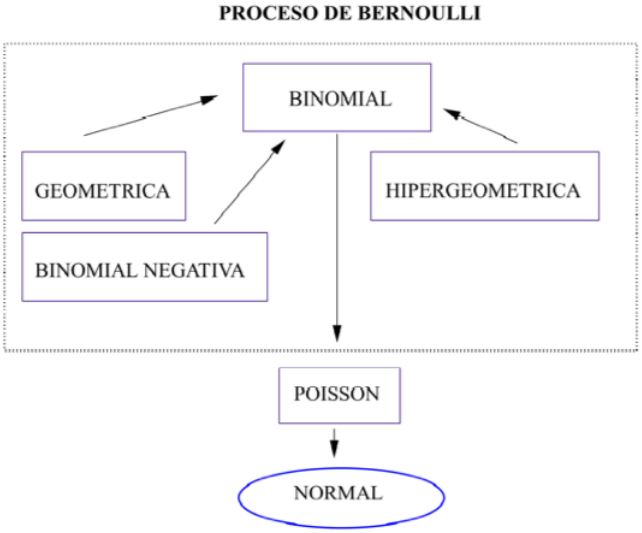
El proceso de Bernoulli puede ser considerado como punto de partida para la definición de los modelos probabilísticos analizados hasta ahora, que aparecen conectados entre sí según el esquema y presentan a su vez relaciones con otras distribuciones de probabilidad que estudiaremos en posteriores apartados.

3.4. Modelo Uniforme

En algunas magnitudes aleatorias no existe ninguna evidencia a favor de determinados resultados, por lo cual resulta aplicable el principio de indiferencia. Este sería el caso cuando lanzamos un dado, extraemos una bola de un bombo de lotería, o seleccionamos al azar una carta de la baraja.

También con frecuencia nos encontramos con que conocemos el recorrido de una

Figura 3.6.: Esquema de modelos discretos



magnitud aleatoria pero carecemos de cualquier información adicional. Esta situación, que puede darse tanto en variables discretas como continuas, conduce al modelo uniforme, cuya distribución se corresponde con un reparto equitativo de la probabilidad.

Supongamos, por ejemplo, que una persona se dispone a desplazarse utilizando el metro de su ciudad, y consultando el plano, observa que existen tres líneas alternativas que le conducen hasta su destino, con servicio cada 10 minutos.

En este ejemplo aparecen dos magnitudes aleatorias de distinta índole pero de características similares: una de ellas es la línea de metro elegida y la otra la hora a la que éste realiza su salida.

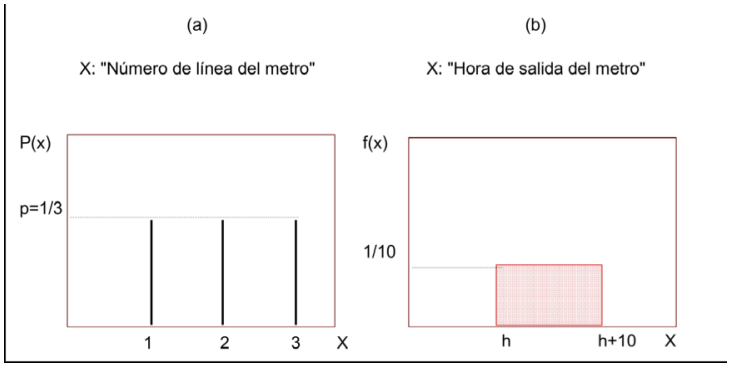
Tal y como representa la figura 3.7 la primera de estas características es una variable discreta a la que pueden asociarse valores 1, 2 y 3. Sin embargo, la segunda es continua dentro de un recorrido que, a partir de una hora genérica h, expresada en minutos, podemos denominar $(h, h + 10)$.

En ambos casos la información se limita a la ya recogida, ignorándose cómo se distribuye la probabilidad de las magnitudes. Ante esta falta de información, se adopta el principio de indiferencia, asumiendo equiprobabilidad de los resultados posibles. Como consecuencia de este supuesto aparece -en sus versiones discreta y continua- el modelo uniforme. Los gráficos adjuntos ilustran la distribución de probabilidad de ambas características: si la línea de metro es seleccionada al azar se tiene

$$P(X = 1) = P(X = 2) = P(X = 3) = \frac{1}{3}$$

Análogamente, la probabilidad se reparte de modo uniforme en cada intervalo de amplitud 10 minutos como el representado, pudiendo adoptar cualquier valor de dicho

Figura 3.7.: Modelo uniforme discreto y continuo



recorrido.

3.4.1. Caso discreto

Consideremos una variable aleatoria X con posibles valores $x_1, x_2, \ldots, x_n$ que se asumen indiferentes $p(x_1) = p(x_2) = \cdots = p(x_n)$. Podremos encontrarnos en esta situación si conocemos que dichos resultados son equiprobables o bien si la ausencia de información adicional nos lleva a admitir el supuesto de uniformidad.

Teniendo en cuenta que debe cumplirse $\sum_{i=1}^{n} p(x_i) = 1$, se tiene entonces $p(x_i) = \dfrac{1}{n}$ para cada uno de los valores $i = 1, 2, \ldots, n$.

Esta distribución de probabilidad correspondiente a un modelo uniforme conduce a la definición clásica de probabilidad como cociente entre casos favorables y casos posibles. En efecto, al carecer de información, asumimos que todos los posibles resultados de X son equiprobables y así se tiene para cualquier $k = 1, 2, \ldots$ que

$$P(X \leq x_k) = \sum_{i=1}^{k} p(x_i) = k\frac{1}{n} = \frac{k}{n}$$

Como consecuencia de los rasgos anteriores, para el modelo uniforme se obtiene un valor esperado coincidente con la media aritmética de los valores de la variable $\mu = \sum_{i=1}^{n} \frac{x_i}{n}$.

Esta expresión de la esperanza recoge el principio de indiferencia: asigna igual peso a cada uno de los valores de la variable aleatoria X como consecuencia de que sus correspondientes probabilidades son en este caso coincidentes.

3.4.2. Caso continuo

Las características del modelo uniforme anterior pueden extenderse de modo inmediato al caso continuo. En realidad, éste es el supuesto que subyace en las representaciones gráficas tipo histograma para datos agrupados, cuando sólo conocemos la frecuencia o la probabilidad de un intervalo y asumimos que este valor se reparte uniformemente en cierto recorrido genérico (a, b).

El modelo uniforme, que se representa abreviadamente $\mathcal{U}(a, b)$ se denomina también rectangular en alusión a su representación gráfica. Esta distribución -como consecuencia del principio de indiferencia o de la ausencia de información- asigna probabilidades idénticas a cualesquiera intervalos de igual amplitud. Así, en nuestro ejemplo del metro ilustrado en la figura 3.7, se observa que coinciden las probabilidades asociadas a cualquier intervalo de un minuto de amplitud (subintervalos $(h, h+1)$ y $(h+5, h+6)$ por ejemplo).

Definición 3.6. Dada una variable aleatoria continua X distribuida según un *modelo uniforme* $X \approx \mathcal{U}(a, b)$ su función de densidad viene dada por la expresión:

$$f(x) = \begin{cases} \dfrac{1}{b-a} & si\ a < x < b \\ 0 & \text{en otro caso} \end{cases}$$

La expresión de $f(x)$ puede ser deducida fácilmente de modo gráfico, teniendo en cuenta que dicha función asigna una densidad constante a cada punto del intervalo y que -según la definición de $f(x)$- el área del rectángulo de base $(b-a)$ debe ser unitaria. Se tiene así que:

$$P(a \leq X \leq b) = \int_a^b f(x)dx = \int_a^b k dx = k(b-a) = 1$$

con lo cual $f(x) = k = \dfrac{1}{b-a}$ para todo $a < x < b$.

[Obtener la función de distribución de X, que vendrá dada por la expresión: $F(x) = \dfrac{x-a}{b-a}$ para $a \leq x < b$] [¿cuál sería su representación gráfica?]

Las características del modelo uniforme vienen dadas en función de los extremos del correspondiente intervalo. Así, dada $X \approx \mathcal{U}(a, b)$ puede obtenerse fácilmente $\mu = \dfrac{a+b}{2}$, centro de gravedad del recorrido de la variable. [¿Cuál es la hora esperada para el metro del ejemplo anterior?]

Por su parte, la varianza viene dada por la expresión $\sigma^2 = \dfrac{(b-a)^2}{12}$, que depende únicamente del recorrido de la variable considerada y se puede obtener por diferencia de $E(X^2) = \dfrac{b^3 - a^3}{3(b-a)}$ y el cuadrado de la esperanza (μ^2).

Ambas características pueden también ser deducidas a partir de la función generatriz de momentos, que para este modelo adopta la expresión $M_X(t) = \dfrac{e^{tb} - e^{ta}}{t(b-a)}$.

3.5. Modelo Normal

El supuesto de continuidad resulta adecuado para numerosas magnitudes económicas, que frecuentemente pueden adoptar cualquiera de los infinitos valores de su campo de variación.

El único modelo continuo que hemos analizado hasta ahora, la distribución uniforme, puede resultar adecuado si no tenemos razones para pensar que ciertos valores de la variable sean más probables que otros. No obstante, a menudo aparecen distribuciones cuya representación viene dada por una curva campaniforme, esto es, cuyo recorrido central concentra gran parte de la probabilidad. La generalidad de este tipo de magnitudes justifica su denominación como *modelo normal*.

La distribución normal fue obtenida inicialmente por De Moivre en 1733. Sin embargo, habitualmente se conoce como modelo de Gauss, o de Gauss-Laplace por ser estos autores quienes, durante el siglo XVIII, analizaron sus propiedades e impulsaron su utilización.

Aunque la distribución normal se revela como un modelo probabilístico sumamente útil para la descripción de numerosos fenómenos económicos, los trabajos iniciales de Gauss (1777-1855), que dieron lugar a la curva normal, iban referidos a errores de medida en observaciones astronómicas, cuya distribución era de tipo campaniforme.

Por su parte, Pierre Simon, marqués de Laplace (1749-1827) obtuvo este modelo como aproximación de otras distribuciones. Este resultado, de gran trascendencia en las técnicas inferenciales, se conoce como Teorema Central del Límite y será analizado con detalle en un capítulo posterior.

A modo de ilustración de este modelo normal, consideremos de nuevo el ejemplo inicial y supongamos que los empresarios son convocados para realizar la entrevista a las 4 de la tarde, realizando el reportaje fotográfico a su finalización. Así pues, y teniendo en cuenta que la duración esperada de las entrevistas es de 2 horas se ha convocado al equipo fotográfico a las 6 de la tarde, pero es evidente que, por motivos diversos, no todas las entrevistas tendrán exactamente la misma duración, sino que puede haber ligeras desviaciones, anticipándose o retrasándose la hora de finalización.

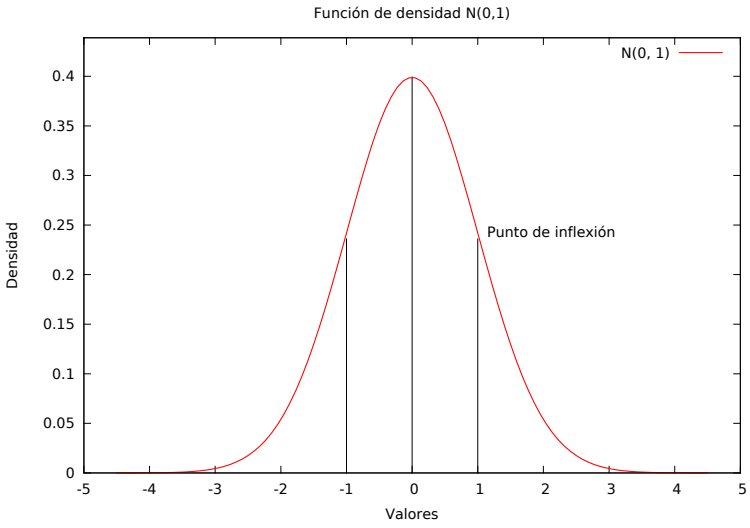
Como consecuencia, el "retraso respecto a la hora prevista" será una variable aleatoria continua, cuya representación podría ser como sigue: una curva aproximadamente simétrica, que acumula la mayor probabilidad en torno a la observación central (entrevistas que finalizan a la hora prevista, con retrasos aproximadamente nulos) siendo despreciable la probabilidad de entrevistas muy breves (valores negativos extremos) o muy largas (valores elevados con signo positivo).

3.5.1. Modelo Normal estándar

El tipo de representación comentado se corresponde con una *distribución normal tipificada o estándar* , denotada como $\mathcal{N}(0,1)$ y que sirve como modelo de referencia por ser su esperanza nula y su desviación típica unitaria.

Definición 3.7. Se dice que una variable aleatoria X sigue una distribución normal estándar, que denotamos $X \approx \mathcal{N}(0,1)$, si su función de densidad viene dada por la

Figura 3.8.: Modelo normal estándar



expresión:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \; ; \; -\infty < x < +\infty$$

La representación gráfica de esta función corresponde a una curva simétrica, que alcanza su valor máximo en el punto $x = 0$, presenta dos puntos de inflexión (en -1 y $+1$) y una asíntota horizontal en el eje de abscisas.

Proposición 3.1. *La anterior $f(x)$ es una verdadera función de densidad dado que su expresión es no negativa y además su integral es la unidad.*

Demostración. En efecto,

$$\int_{-\infty}^{+\infty} f(x)dx = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} 2 \int_{0}^{+\infty} e^{-\frac{x^2}{2}} dx$$

y haciendo el cambio $t = \frac{x^2}{2}$, $dx = \frac{1}{\sqrt{2t}} dt$, se obtiene:

$$\int_{-\infty}^{+\infty} f(x)dx = \frac{2}{\sqrt{2\pi}\sqrt{2}} \int_{0}^{\infty} t^{-\frac{1}{2}} e^{-t} dt = \frac{1}{\sqrt{\pi}} \int_{0}^{\infty} t^{\frac{1}{2}-1} e^{-t} dt = \frac{1}{\sqrt{\pi}} \Gamma\left(\frac{1}{2}\right) = 1$$

donde la última integral es la función matemática $\Gamma\left(\frac{1}{2}\right)$ cuyo valor es $\sqrt{\pi}$.

□

Proposición 3.2. *Las características de esta distribución son $\mu = 0$, $Var(X) = \sigma^2 = 1$, que coinciden con los parámetros del modelo normal.*

Demostración. En efecto:

$$\mu = E(X) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x e^{-\frac{x^2}{2}}\, dx = -\frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty} = 0$$

Por otra parte:

$$E(X^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}}\, dx$$

que integrando por partes:

$$\begin{bmatrix} u = x & du = dx \\ dv = x e^{-\frac{x^2}{2}} dx & v = \int_{-\infty}^{+\infty} dv = -e^{-\frac{x^2}{2}} \end{bmatrix}$$

se obtiene:

$$E(X^2) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} x^2 e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \left[\underbrace{-x e^{-\frac{x^2}{2}} \Big|_{-\infty}^{+\infty}}_{=0} - \int_{-\infty}^{+\infty} -e^{-\frac{x^2}{2}}\, dx \right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2}}\, dx = 1$$

□

Proposición. *La función generatriz de momentos de una distribución $\mathcal{N}(0,1)$, viene dada por la expresión: $M_X(t) = e^{\frac{t^2}{2}}$, $-\infty < t < +\infty$, a partir de la cual se podrían obtener las características anteriores.*

Demostración. Esta función se obtiene como:

$$M_X(t) = E\left(e^{tx}\right) = \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{tx} e^{-\frac{x^2}{2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{x^2}{2} + tx} dx$$

ahora construimos un cuadrado perfecto en el exponente para lo cual sumamos y restamos $-\frac{t^2}{2}$,

$$\begin{aligned} M_X(t) = & \ \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{\left(-\frac{x^2}{2} + tx - \frac{t^2}{2}\right) + \frac{t^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{(x-t)^2}{2} + \frac{t^2}{2}} dx = \\ = & \ \frac{1}{\sqrt{2\pi}} e^{\frac{t^2}{2}} \int_{-\infty}^{+\infty} e^{-\frac{(x-t)^2}{2}} dx = e^{\frac{t^2}{2}} \left(\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-\frac{z^2}{2}} dz \right) \end{aligned}$$

La expresión que figura entre paréntesis en el último término, en la que hemos hecho el cambio de variable $z = x - t$, se corresponde con la integral de la función de densidad de una $\mathcal{N}(0,1)$ cuyo valor es unitario; por tanto se obtiene: $M_X(t) = e^{\frac{t^2}{2}}$.

□

Al presentar esperanza nula y dispersión unitaria, la interpretación del modelo $\mathcal{N}(0,1)$ resulta muy intuitiva: cada valor de la variable X mide el número de desviaciones estándar que dicho valor se separa de su valor esperado.

El modelo normal sirve además de referencia en cuanto a las características de forma: simetría y curtosis. De hecho, el coeficiente de apuntamiento habitualmente utilizado es el propuesto por Fisher, que es resultado de comparar para cada distribución el ratio $\frac{\mu_4}{\sigma^4}$ con el valor 3, asociado al apuntamiento del modelo normal estándar.

Por lo que se refiere a las áreas acumuladas bajo la curva normal, que se corresponden con la función de distribución $F(x)$, éstas vendrán dadas por la expresión: $\mathrm{F(x)}=\int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$, cuyo cálculo debe ser efectuado por métodos de integración numérica.

En la notación anterior aparece un abuso de lenguaje al denotar por x tanto el punto donde nos situamos para calcular la probabilidad acumulada como la variable de integración en ese recorrido. Si queremos ser más precisos podemos diferenciar ambos papeles de x, expresando: $F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$.

Como consecuencia, las probabilidades asociadas a este modelo aparecen tabuladas, siendo posible así calcular para cada punto x el correspondiente valor de probabilidad acumulada $F(x)$.

La estructura habitual de las tablas, que viene recogida en la tabla 3.5, sólo proporciona los valores de la función de distribución correspondientes al recorrido positivo de la variable, seleccionando para ésta el valor entero y dos decimales. Sin embargo, gracias a la simetría del modelo resulta inmediata la obtención de probabilidades acumuladas para valores negativos.

Por otra parte, si nos interesase calcular probabilidades tipo "mayor que" bastaría aplicar la idea de complementario a las probabilidades recogidas en tablas ("menor o igual a"). Los mismos planteamientos anteriores son también válidos para obtener las probabilidades de intervalos.

Manejo de tablas N(0,1)

El manejo de estas tablas consiste simplemente en compatibilizar nuestras necesidades con la información que aparece recogida en ellas, esto es, el valor de la función de distribución $F(x)$. Así, en el esquema siguiente proponemos algunos ejemplos -que aparecen ilustrados gráficamente- de cómo se llevaría a cabo este proceso de adecuación:

Las probabilidades de los intervalos considerados, que aparecen ilustradas gráficamente en la cuarta columna, pueden ser obtenidas mediante las expresiones indicadas en la columna tercera. Puede apreciarse que la expresión final de cálculo incluye $F(a)$ y $F(b)$ cuando los valores a y b son positivos, mientras que en el caso de que sean negativos aparecen respectivamente $F(-a)$ y $F(-b)$.

Conviene tener presente que no todas las tablas tienen el mismo formato. En concreto, aunque las

Tabla 3.5.: Modelo normal. Función de distribución

x	0	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0	0,5000	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,1	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,2	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,3	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,4	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,5	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,6	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,7	0,7580	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,8	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,9	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,1	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,2	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,3	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,4	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,1	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,2	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,3	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,4	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,5	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,6	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,7	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,8	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,9	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986
3	0,9987	0,9987	0,9987	0,9988	0,9988	0,9989	0,9989	0,9989	0,9990	0,9990
3,1	0,9990	0,9991	0,9991	0,9991	0,9992	0,9992	0,9992	0,9992	0,9993	0,9993
3,2	0,9993	0,9993	0,9994	0,9994	0,9994	0,9994	0,9994	0,9995	0,9995	0,9995
3,3	0,9995	0,9995	0,9995	0,9996	0,9996	0,9996	0,9996	0,9996	0,9996	0,9997
3,4	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9997	0,9998
3,5	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998	0,9998
3,6	0,9998	0,9998	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999	0,9999

Figura 3.9.:

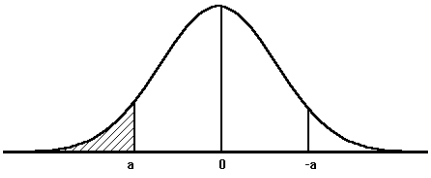
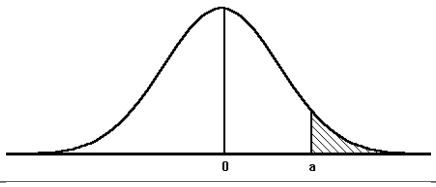
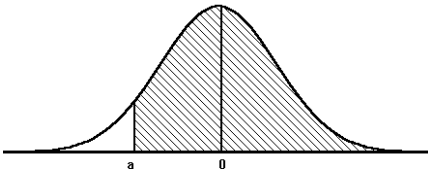
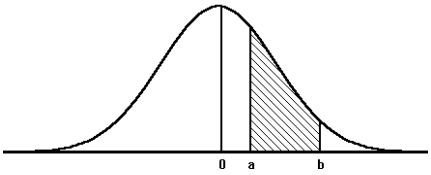
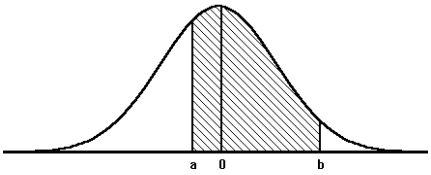
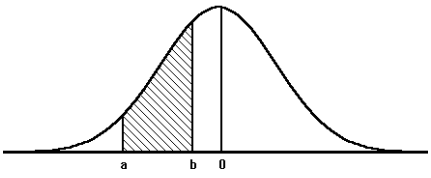
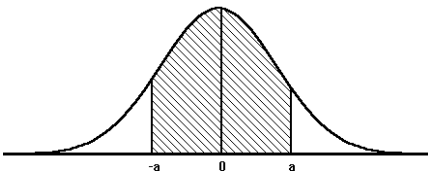
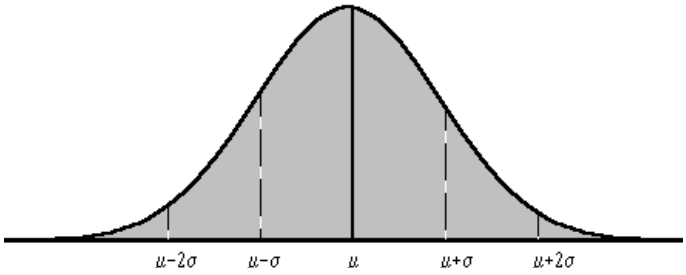
Información necesaria	Situación	Tratamiento de Información de Tablas	Representación Gráfica
$P(X \leq a)$	$a < 0$	$F(a) = 1 - F(-a)$	
$P(X > a)$	$a \geq 0$	$1 - F(a)$	
	$a < 0$	$F(-a)$	
$P(a < X < b)$	$0 < a < b$	$F(b) - F(a)$	
	$a < 0 < b$	$F(b) - (1 - F(a)) =$ $F(b) - 1 + F(a)$	
	$a < b < 0$	$F(-a) - F(-b)$	
$P(-a \leq X \leq a) =$ $P(\lvert X \rvert \leq a)$	$0 < a$	$F(a) - F(-a)$ $= F(a) - (1 - F(a))$ $= 2F(a) - 1$	

Figura 3.10.: Modelo $\mathcal{N}(\mu, \sigma)$. Función de densidad



más habituales son las que incluyen valores de la función de distribución, podrían resultar también útiles otros tipos de tablas, como las que recogen el área central entre cada valor considerado y el origen.

En este caso la tabla proporciona para cada $a > 0$ el valor de $P(0 < X < a) = P(-a < X < 0)$, probabilidades que aparecen relacionadas con las comentadas anteriormente. Se dispone así de una expresión alternativa de las probabilidades que en algunos casos resulta más directa que la función de distribución. A modo de ejemplo, para el caso $a < 0 < b$ escribiríamos ahora $P(a < X \leq b) = P(a < X \leq 0) + P(0 < X \leq b)$.

3.5.2. Modelo Normal general

Gracias a las tablas de probabilidad comentadas, el modelo normal estándar es la referencia obligada para gran número de investigaciones. Sin embargo parece claro que muchas magnitudes cuya descripción podría adaptarse al modelo "normal" no presentarán sus características $\mu = 0\, y\, \sigma = 1$.

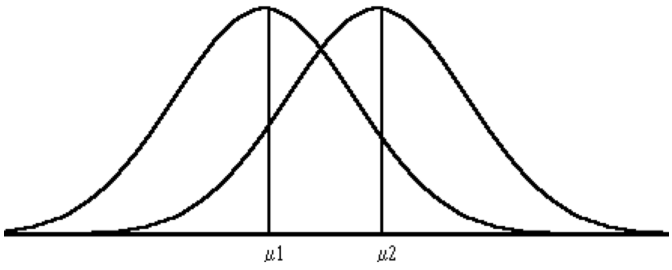
Así, en la ilustración inicial de las entrevistas realizadas a empresarios, la duración se concentrará en torno al tiempo esperado μ no nulo (2 horas, por ejemplo); su gráfica será campaniforme, centrada en μ y más o menos apuntada según la dispersión en las duraciones de las diferentes entrevistas.

Otros ejemplos de variables que pueden ser ajustadas al modelo normal son: los beneficios de una empresa, las ventas en establecimientos comerciales, el peso de los niños recién nacidos, la demanda de determinado producto, ... En todos estos casos el centro de la distribución se situará en una cantidad μ no nula que representa el valor esperado de la correspondiente variable.

Las características gráficas de esta distribución corresponden a la curva campaniforme, positiva, simétrica respecto a la recta $x = \mu$ (en la cual alcanza su valor máximo) y con colas asintóticas al eje de abscisas. Dicha curva es creciente para $x < \mu$ y decreciente para $x > \mu$, presentando puntos de inflexión en $\mu \pm \sigma$; es cóncava en el intervalo $(\mu - \sigma, \mu + \sigma)$ y convexa en el resto de su recorrido.

Definición 3.8. Se dice que una variable aleatoria X sigue una *distribución normal de parámetros* μ *y* σ, que denotamos como $X \approx \mathcal{N}(\mu, \sigma)$, si su función de densidad viene dada por la expresión:

Figura 3.11.: Modelo $\mathcal{N}(\mu, \sigma)$. Cambio de origen



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \; ; \; -\infty < x < \infty$$

La esperanza y la desviación típica de esta distribución coinciden precisamente con los parámetros que caracterizan esta población. Su función generatriz de momentos viene dada por la expresión:

$$M_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

Las demostraciones en este caso se realizan de forma análoga a las desarrolladas para la normal estándar considerando previamente el cambio $z = \dfrac{x-\mu}{\sigma}$.

Si sobre una v.a. Z con distribución normal estándar, $Z \approx \mathcal{N}(0,1)$ efectuamos una transformación del tipo $X = \mu + \sigma Z$, entonces la variable aleatoria X resultante se distribuye según un modelo normal general $\mathcal{N}(\mu, \sigma)$.

A modo de recíproco, si $X \approx \mathcal{N}(\mu, \sigma)$, entonces:

$$z = \frac{X-\mu}{\sigma} \approx \mathcal{N}(0,1)$$

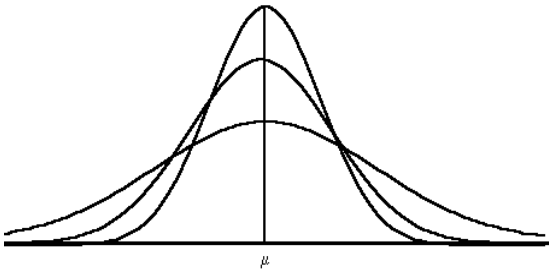
Los parámetros del modelo normal general μ y σ representan respectivamente características de posición y de escala, tal como indican las figuras 3.11 y 3.12. Cambios en μ suponen desplazamientos del eje de simetría de la curva a lo largo del eje de abscisas (Figura 3.11), mientras que las alteraciones en σ afectan a la dispersión, esto es, a la forma de la curva (Figura 3.12).

Como ya hemos comentado, la distribución normal estándar $\mathcal{N}(0,1)$ presenta la gran ventaja de hallarse tabulada, hecho que garantiza un cálculo sencillo de probabilidades. Sin embargo aparece ahora el problema de calcular probabilidades asociadas a una distribución normal general, que se resuelve mediante un proceso de tipificación de la variable para reducirla a su forma estándar.

El procedimiento de tipificación se revela como imprescindible si observamos que, para diferentes modelos de partida, las probabilidades de un mismo intervalo difieren.

Consideremos por ejemplo dos distribuciones A y B, ambas normales, sobre las que deseamos calcular la probabilidad de un mismo intervalo $(10, 20)$.

Figura 3.12.: Modelo $\mathcal{N}(\mu, \sigma)$. Cambio de escala



Para cuantificar la probabilidad del intervalo resulta necesario traducir las distribuciones A y B a una normal tipificada $\mathcal{N}(0, 1)$, cuyos valores aparecerán perfectamente tabulados.

Por lo que respecta al mecanismo para llevar a cabo la tipificación, éste será análogo al de variables estadísticas. A partir de cualquier variable $X \approx \mathcal{N}(\mu, \sigma)$ es posible obtener un modelo $Z \approx \mathcal{N}(0, 1)$ con sólo eliminar de la primera los efectos de sus parámetros, esto es, operar sobre ella el cambio .

La aplicación del procedimiento de tipificación a las distribuciones A y B conduce a los intervalos señalados sobre la distribución estándar, que se obtienen como resultado de eliminar de los intervalos iniciales los parámetros μ y σ. De este modo, para la variable $A \approx \mathcal{N}(9, 5)$ el intervalo inicial $(10, 20)$ quedaría transformado en $(0,2\ ,\ 2,2)$ una vez tipificada dicha variable $\left(Z_A = \dfrac{A-9}{5}\right)$, mientras el mismo proceso aplicado a $B \approx \mathcal{N}(15;\ 2,5)$ daría lugar al intervalo estandarizado $(-2, 2)$.

Conviene insistir en que la tipificación tiene como único objetivo referir las variables a un modelo estándar, permitiendo el cálculo de probabilidades. En cambio, este proceso elimina el propio significado de la magnitud inicial, impidiendo por tanto hacer interpretaciones sobre la misma.

En nuestro ejemplo, una vez tipificados los correspondientes recorridos es posible calcular las probabilidades correspondientes a los intervalos con ayuda de las tablas $\mathcal{N}(0, 1)$, obteniéndose los resultados 0,4068 y 0,9544 respectivamente. [Compruébese]

A pesar de que, como hemos visto, el modelo normal es adecuado para la descripción de numerosos fenómenos, la distribución de muchas magnitudes económicas (como la renta, la riqueza, los salarios, ...) no es simétrica, ya que la densidad se reparte de forma distinta en los estratos bajos que en niveles elevados.

Sin embargo, este hecho se resuelve a menudo con una transformación logarítmica de la variable, de modo que la distribución de $Y = \ln X$ sí se aproxima a un modelo normal.

En estas situaciones, la distribución de la variable X se denomina *logaritmo normal* y resulta muy adecuada para la descripción de magnitudes económicas como la renta, en especial para los niveles más bajos de ingreso.

3.6. Algunos modelos especiales de probabilidad

Entre el amplio abanico de posibilidades que se presentan para modelizar variables aleatorias estudiaremos a continuación algunas distribuciones que, aunque no son de utilización tan generalizada como las de apartados anteriores, resultan sin embargo muy adecuadas para describir ciertos fenómenos de interés.

3.6.1. Sucesos raros: modelo de Poisson

A menudo nos interesa estudiar sucesos que, aunque no resultan frecuentes, pueden presentarse en el transcurso del tiempo o del espacio. Estas situaciones del tipo "número de casas incendiadas en un año", "erratas en la página de un periódico", "llamadas de teléfono equivocadas", "errores de un equipo informático", "atracos en una sucursal bancaria", ... se adaptan bien al modelo probabilístico denominado de Poisson o "ley de los sucesos raros".

Esta distribución fue analizada por S.D. Poisson en un libro publicado en 1837 con el título *Investigación sobre la probabilidad de juicios en materia criminal y civil*, lo cual en cierto modo justifica sus dos denominaciones.

Por su parte, L.Bortkiewicz (1868-1931) fue el primero en observar que las ocurrencias de sucesos con pequeñas frecuencias en una población amplia pueden ajustarse mediante una distribución de Poisson, lo que denominó "ley de los pequeños números".

Bortkiewicz estudió el número de soldados fallecidos anualmente por coces de caballo en el ejército prusiano. Se examinaron 14 cuerpos durante 20 años, observando que estos 280 datos se ajustaban bien por un modelo de Poisson (de hecho, del total de cuerpos estudiados, se observaron 144 en los que no se registró ninguna muerte por la causa investigada).

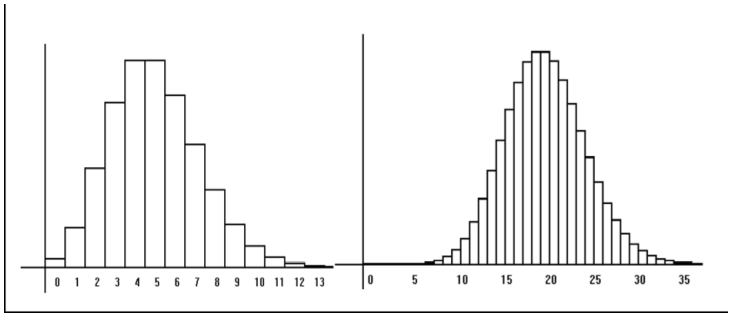
Otros conocidos ejemplos históricos de ajustes a un *modelo de Poisson* corresponden a las observaciones de estallidos de guerras mundiales entre los años 1500 y 1931, y los impactos de bombas alemanas sobre el área de Londres durante la segunda guerra mundial.

Volviendo a las ilustraciones iniciales, podríamos plantearnos analizar el número de errores tipográficos que se producirán a lo largo del reportaje sobre los empresarios. Examinando el recorrido de esta variable aleatoria, parece claro que podría llegar a tomar valores elevados, pero sin embargo, el número esperado de errores será bajo, al resultar poco probable que éstos tengan lugar en un intervalo determinado (página, sección, ...).

Las características anteriores describen un modelo de Poisson, asociado al proceso del mismo nombre que se basa en los siguientes supuestos:

- El proceso es estable, al producir un número medio de sucesos λ constante por unidad de tiempo o espacio.
- Los sucesos se presentan aleatoriamente y de modo independiente, es decir, el número de sucesos observados en un intervalo no condiciona los resultados de otro intervalo disjunto del anterior.
- La probabilidad de que el suceso estudiado se presente dos o más veces en un intervalo pequeño es aproximadamente nula.

Figura 3.13.: Modelo de Poisson. Función de probabilidad



(a) $P(\lambda{=}5)$ (b) $P(\lambda{=}20)$

Bajo las condiciones descritas, la variable aleatoria X que recoge "el número de sucesos en un intervalo de determinada amplitud" se distribuye según un *modelo de Poisson*, representado abreviadamente por $P(\lambda)$. Los valores que puede tomar esta variable son: $0, 1, 2, \ldots$ y su función de probabilidad viene dada por:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

La figura 3.13 recoge la representación de esta función de probabilidad para valores de $\lambda = 5$ y $\lambda = 20$.

Podemos observar cómo cuando aumenta λ (figura 3.13b) la gráfica tiende a ser campaniforme, lo que nos sugiere que para valores elevados del parámetro esta distribución podrá ser aproximada por el modelo normal.

Esta distribución viene caracterizada por un único parámetro λ que representa el número medio de sucesos por unidad de tiempo o espacio. Como consecuencia, el valor del parámetro cambia según cuál sea la "unidad" adoptada, esto es, en función de la amplitud del intervalo espacial o temporal en el que nos movemos.

Definición 3.9. De un modo general, toda v.a. discreta X que puede adoptar valores $0, 1, 2, \ldots$ con probabilidades dadas por la expresión $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$, se dice que sigue un *modelo de Poisson* $\mathcal{P}(\lambda)$.

Esta función de probabilidad puede ser obtenida como límite de un modelo binomial $\mathcal{B}(n, p)$, cuando se aumenta indefinidamente el número de pruebas n y la probabilidad p tiende a 0. Bajo estas condiciones el modelo binomial se aproxima a una distribución de Poisson con $\lambda = np$, resultando estas aproximaciones adecuadas cuando $np < 5$ y $p < 0,1$.

Haciendo $\lambda = np$ se tiene $p = \frac{\lambda}{n}$ y $q = 1 - \frac{\lambda}{n}$ con lo cual:

$$\lim_{n\to\infty} P(X=k) = \lim_{n\to\infty}\left[\binom{n}{k}p^k q^{n-k}\right] = \lim_{n\to\infty}\left[\binom{n}{k}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k}\right] =$$
$$= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\frac{n(n-1)\cdots(n-k+1)\left(1-\frac{\lambda}{n}\right)^n}{n^k\left(1-\frac{\lambda}{n}\right)^k}\right] =$$
$$= \frac{\lambda^k}{k!}\lim_{n\to\infty}\left[\left(1-\frac{\lambda}{n}\right)^n\frac{1\cdot\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{\lambda}{n}\right)^k}\right] = \frac{\lambda^k}{k!}e^{-\lambda}$$

por ser $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$ y $\lim_{n\to\infty}\frac{1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{k-1}{n}\right)}{\left(1-\frac{\lambda}{n}\right)^k} = 1$

Manejo de tablas $\mathcal{P}(\lambda)$

Con el objetivo de facilitar su cálculo, las probabilidades de la distribución de Poisson aparecen tabuladas para distintos valores de λ. En la tabla 3.6 se recoge la función de probabilidad de esta distribución.

Por lo que se refiere a las características de este modelo, su rasgo más destacado es la coincidencia de esperanza y varianza con el valor del parámetro λ.

Dado que λ es un parámetro determinante del modelo de Poisson, cabría preguntarse cómo se llega a conocer su valor en la práctica. La respuesta es la información pasada, ya que -asumidos los supuestos de estabilidad del proceso- el promedio de éxitos que se han producido en un intervalo de determinada amplitud permite conocer la esperanza de la variable.

La esperanza y la varianza pueden ser obtenidas a partir de la función generatriz de momentos, que en esta distribución viene dada por la expresión $M_X(t) = e^{\lambda(e^t-1)}$ que se obtiene como sigue:

$$M_X(t) = \sum_{k=0}^{\infty} e^{tk}\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\left(e^t\lambda\right)^k}{k!} = e^{-\lambda}e^{e^t\lambda} = e^{\lambda(e^t-1)}$$

3.6.2. Tiempos de espera: modelo exponencial

En el apartado anterior hemos visto algunos ejemplos de magnitudes discretas que podría describirse según un modelo de Poisson (errores tipográficos, llamadas equivocadas, fallos de un equipo informático...). Si en cambio estudiásemos el tiempo que transcurre hasta que se produce el siguiente error o fallo, la variable -aunque relacionada en cierto modo con la anterior- sería continua.

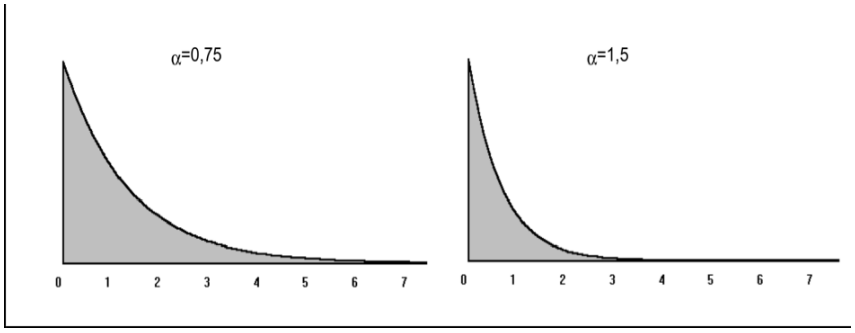
Este nuevo planteamiento resulta habitual en el ámbito económico-empresarial, en el que frecuentemente interesa conocer el período de tiempo necesario hasta que se presenta determinado acontecimiento: aprobación de presupuestos, salida de una empresa a bolsa, contratación de un trabajador, ...

Estas magnitudes aleatorias se adaptan a un modelo continuo denominado exponencial cuya función de densidad es decreciente como indica la figura 3.14

Tabla 3.6.: Modelo de Poisson. Función de probabilidad

$\lambda \backslash x$	0	1	2	3	4	5	6	7	8	9
0,1	0,9048	0,0905	0,0045	0,0002						
0,2	0,8187	0,1637	0,0164	0,0011	0,0001					
0,3	0,7408	0,2222	0,0333	0,0033	0,0003					
0,4	0,6703	0,2681	0,0536	0,0072	0,0007	0,0001				
0,5	0,6065	0,3033	0,0758	0,0126	0,0016	0,0002				
0,6	0,5488	0,3293	0,0988	0,0198	0,0030	0,0004				
0,7	0,4966	0,3476	0,1217	0,0284	0,0050	0,0007	0,0001			
0,8	0,4493	0,3595	0,1438	0,0383	0,0077	0,0012	0,0002			
0,9	0,4066	0,3659	0,1647	0,0494	0,0111	0,0020	0,0003			
1	0,3679	0,3679	0,1839	0,0613	0,0153	0,0031	0,0005	0,0001		
1,1	0,3329	0,3662	0,2014	0,0738	0,0203	0,0045	0,0008	0,0001		
1,2	0,3012	0,3614	0,2169	0,0867	0,0260	0,0062	0,0012	0,0002		
1,3	0,2725	0,3543	0,2303	0,0998	0,0324	0,0084	0,0018	0,0003	0,0001	
1,4	0,2466	0,3452	0,2417	0,1128	0,0395	0,0111	0,0026	0,0005	0,0001	
1,5	0,2231	0,3347	0,2510	0,1255	0,0471	0,0141	0,0035	0,0008	0,0001	
1,6	0,2019	0,3230	0,2584	0,1378	0,0551	0,0176	0,0047	0,0011	0,0002	
1,7	0,1827	0,3106	0,2640	0,1496	0,0636	0,0216	0,0061	0,0015	0,0003	0,0001
1,8	0,1653	0,2975	0,2678	0,1607	0,0723	0,0260	0,0078	0,0020	0,0005	0,0001
1,9	0,1496	0,2842	0,2700	0,1710	0,0812	0,0309	0,0098	0,0027	0,0006	0,0001
2	0,1353	0,2707	0,2707	0,1804	0,0902	0,0361	0,0120	0,0034	0,0009	0,0002
2,1	0,1225	0,2572	0,2700	0,1890	0,0992	0,0417	0,0146	0,0044	0,0011	0,0003
2,2	0,1108	0,2438	0,2681	0,1966	0,1082	0,0476	0,0174	0,0055	0,0015	0,0004
2,3	0,1003	0,2306	0,2652	0,2033	0,1169	0,0538	0,0206	0,0068	0,0019	0,0005
2,4	0,0907	0,2177	0,2613	0,2090	0,1254	0,0602	0,0241	0,0083	0,0025	0,0007
2,5	0,0821	0,2052	0,2565	0,2138	0,1336	0,0668	0,0278	0,0099	0,0031	0,0009
2,6	0,0743	0,1931	0,2510	0,2176	0,1414	0,0735	0,0319	0,0118	0,0038	0,0011
2,7	0,0672	0,1815	0,2450	0,2205	0,1488	0,0804	0,0362	0,0139	0,0047	0,0014
2,8	0,0608	0,1703	0,2384	0,2225	0,1557	0,0872	0,0407	0,0163	0,0057	0,0018
2,9	0,0550	0,1596	0,2314	0,2237	0,1622	0,0940	0,0455	0,0188	0,0068	0,0022
3	0,0498	0,1494	0,2240	0,2240	0,1680	0,1008	0,0504	0,0216	0,0081	0,0027
3,1	0,0450	0,1397	0,2165	0,2237	0,1733	0,1075	0,0555	0,0246	0,0095	0,0033
3,2	0,0408	0,1304	0,2087	0,2226	0,1781	0,1140	0,0608	0,0278	0,0111	0,0040
3,3	0,0369	0,1217	0,2008	0,2209	0,1823	0,1203	0,0662	0,0312	0,0129	0,0047
3,4	0,0334	0,1135	0,1929	0,2186	0,1858	0,1264	0,0716	0,0348	0,0148	0,0056
3,6	0,0273	0,0984	0,1771	0,2125	0,1912	0,1377	0,0826	0,0425	0,0191	0,0076
3,8	0,0224	0,0850	0,1615	0,2046	0,1944	0,1477	0,0936	0,0508	0,0241	0,0102
4	0,0183	0,0733	0,1465	0,1954	0,1954	0,1563	0,1042	0,0595	0,0298	0,0132
5	0,0067	0,0337	0,0842	0,1404	0,1755	0,1755	0,1462	0,1044	0,0653	0,0363
6	0,0025	0,0149	0,0446	0,0892	0,1339	0,1606	0,1606	0,1377	0,1033	0,0688
7	0,0009	0,0064	0,0223	0,0521	0,0912	0,1277	0,1490	0,1490	0,1304	0,1014
8	0,0003	0,0027	0,0107	0,0286	0,0573	0,0916	0,1221	0,1396	0,1396	0,1241
9	0,0001	0,0011	0,0050	0,0150	0,0337	0,0607	0,0911	0,1171	0,1318	0,1318
10		0,0005	0,0023	0,0076	0,0189	0,0378	0,0631	0,0901	0,1126	0,1251

Figura 3.14.: Modelo exponencial. Función de densidad



Definición 3.10. Dada una variable aleatoria X se dice que se distribuye según un *modelo exponencial* de parámetro α cuando su función de densidad viene dada por:

$$f(x) = \alpha e^{-\alpha x} \; ; \; x > 0, \alpha > 0$$

La probabilidad acumulada para este modelo viene dada por la función de distribución $F(x) = 1 - e^{-\alpha x}$.

Las principales características del modelo exponencial vienen expresadas en función del parámetro α. Así se tiene una esperanza $\mu = \frac{1}{\alpha}$, que permite interpretar α como la inversa del tiempo medio de espera hasta la aparición de un suceso.

Por lo que respecta a la dispersión se tiene $\sigma^2 = \frac{1}{\alpha^2}$ [Compruébese].

Los parámetros esperanza y varianza de este modelo exponencial guardan claras similitudes con los correspondientes a la distribución geométrica, que es la "traducción" al caso discreto del modelo exponencial. La función generatriz de momentos por su parte viene dada por la expresión:

$$M_X(t) = \left(1 - \frac{t}{\alpha}\right)^{-1} = \frac{\alpha}{\alpha - t}$$

El modelo exponencial presenta dos características destacables. La primera es que se trata de una distribución sin memoria, esto es, la probabilidad de que no se produzca un suceso durante un intervalo es independiente de que haya tenido lugar antes. Este rasgo se expresa como: $P(X > k + m/X > m) = P(X > k)$ donde k y m son dos números reales ambos positivos.

Esta propiedad de pérdida de memoria se demuestra de modo similar a la vista para el modelo geométrico, ya que se tiene:

$$P(X > k+m/X > m) = \frac{P(X > k+m, X > m)}{P(X > m)} = \frac{P(X > k+m)}{P(X > m)} = \frac{e^{-\alpha(k+m)}}{e^{-\alpha m}} = e^{-\alpha k} = P(X > k)$$

Además, la distribución exponencial aparece conectada con la de Poisson en los siguientes términos: dada una variable $Y \approx \mathcal{P}(\lambda)$ que recoge el número de veces que se presenta un suceso en cierto intervalo, entonces el intervalo X transcurrido entre dos sucesos se distribuye según un modelo exponencial.

Si consideramos la variable Y: "número de veces que se presenta cierto suceso por unidad de tiempo" $Y \approx \mathcal{P}(\lambda)$ y definimos ahora X: "tiempo transcurrido hasta la primera aparición del suceso", entonces X será una variable aleatoria continua para la cual podemos calcular probabilidades gracias a su conexión con la variable Y.

En efecto, la probabilidad de que el tiempo necesario hasta la aparición del suceso sea superior a x coincide con la probabilidad de que en un intervalo temporal de amplitud x no se haya producido el suceso. Así pues, la variable Y_X: "número de veces que se presenta cierto suceso en un intervalo de amplitud x" vendrá caracterizada por el parámetro λx (es decir, $Y_X \approx \mathcal{P}(\lambda x)$) y en consecuencia se tiene: $P(X > x) = P(Y_X = 0) = e^{-\lambda x}$.

La función de distribución de X puede también ser obtenida como:

$$F(x) = P(X \leq x) = 1 - P(X > x) = 1 - P(Y_X = 0) = 1 - e^{-\lambda x}, x > 0$$

El modelo exponencial puede también ser estudiado como caso particular (cuando $p = 1$) del modelo generalizado gamma $\gamma(p, a)$. Se trata de una distribución continua biparamétrica que, por resultar adecuada para la modelización de rentas, estudiaremos en el epígrafe siguiente.

3.6.3. Modelos de distribución de la renta

El comportamiento de fenómenos económicos como la renta o la riqueza resulta difícil de describir mediante modelos probabilísticos. De hecho, a menudo se emplean con carácter complementario varios de los modelos que analizaremos en este apartado, o bien generalizaciones de los mismos.

La modelización probabilística de las rentas resulta de gran interés para poder aproximar la proporción de rentistas incluidos en determinado estrato de rentas. Entre las posibilidades que ofrece esta modelización se encuentran la realización de interpolaciones y extrapolaciones del número de rentistas en determinados grupos, la estimación -a partir de los parámetros característicos del modelo- de ciertos indicadores de desigualdad y pobreza, o la realización de simulaciones de políticas redistributivas de renta con la consiguiente evaluación de resultados.

Desde finales del siglo pasado las teorías estocásticas proporcionan una alternativa a los estudios deterministas de distribución de la renta. Los autores pioneros de estos estudios -McAlister (1879), Pareto (1897)- abrieron una corriente de investigación que ha alcanzado un notable desarrollo.

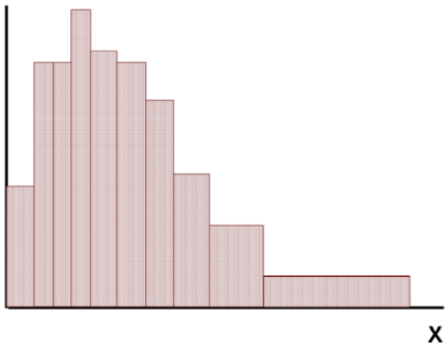
Supongamos que nos interesa conocer la distribución de la renta en una población. Desde una óptica descriptiva, esta variable vendría representada por una tabla estadística con valores $x_1, x_2, \ldots, x_k$ y sus correspondientes frecuencias $n_1, n_2, \ldots, n_k$.

Sin embargo, para conseguir una mayor operatividad en su manejo, estos datos aparecen frecuentemente agrupados en intervalos que representan en realidad decilas de ingresos. Este es el tipo de información recogida habitualmente por las Encuestas de Presupuestos Familiares, y su representación podría ser efectuada mediante un histograma (figura 3.15).

Este histograma representa las decilas de hogares según sus ingresos. Cada uno de los rectángulos que componen el histograma tendría una frecuencia relativa -o proporción de familias- del 10 %.

Sin embargo, para llegar a una descripción más completa de la población investigada, deberíamos analizar también la distribución de ingresos dentro de las decilas. En este sentido una primera opción sería -reconociendo nuestras limitaciones de información- asumir el modelo uniforme, esto es, considerar como válido el histograma, donde los rectángulos construidos sobre cada una de las

Figura 3.15.: Histograma. Decilas de ingresos



decilas recogen un 10 % de probabilidad, repartido igualitariamente entre los hogares que componen ese intervalo, con las consecuencias que ello conlleva.

[A modo de ejemplo ¿cuál sería la esperanza de ingresos en cada decila? ¿resulta adecuado este representante?]

Parece claro que el supuesto de uniformidad puede ser mejorado, buscando modelos que describan de forma más realista la distribución de la renta. En concreto, las distribuciones más habituales en la modelización de rentas, ingresos y gastos son la logaritmo normal, el modelo de Pareto y la distribución gamma.

3.6.3.1. Distribución logaritmo normal

Las rentas no se distribuyen habitualmente de modo simétrico, pero resulta posible llevar a cabo una transformación lineal de éstas mediante logaritmos, apareciendo así el modelo log-normal.

La *distribución logaritmo normal* (o simplemente log-normal) aparece cuando los logaritmos de los valores de la magnitud analizada se distribuyen normalmente, y se describe indicando los parámetros μ y σ de la variable en logaritmos.

Consideremos una v.a Y log-normal y sea $X = \ln Y$. Calculemos en primer lugar la f.d. de Y

$$F_Y(y) = P(Y \leq y) = P(\ln Y \leq \ln y) = P(X \leq \ln y) = F_X(\ln y)$$

donde X sigue un modelo normal $\mathcal{N}(\mu, \sigma)$.

Para calcular la función de densidad de Y derivemos su f.d.:

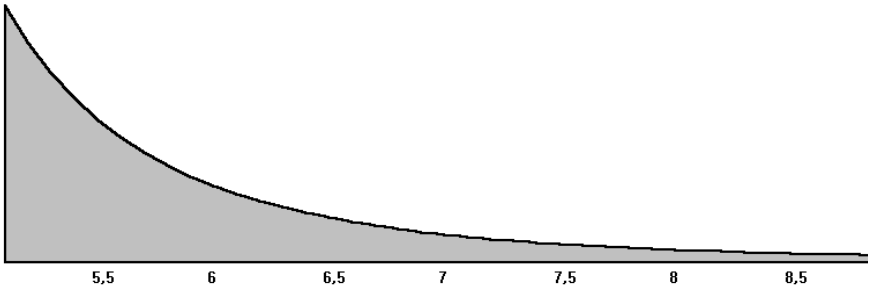
$$f_Y(y) = \frac{dF_Y(y)}{dy} = \frac{dF_X(\ln y)}{d(\ln y)} \frac{d(\ln y)}{dy} = f_X(\ln y) \frac{1}{y}$$

Así pues, teniendo en cuenta la expresión de f_X, se tiene:

Definición 3.11. Decimos que una v.a. Y sigue una *distribución log-normal*, si su función de densidad viene dada por la expresión:

$$f_Y(y) = \frac{1}{\sigma y \sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln y - \mu}{\sigma}\right)^2} \; ; \; \forall y > 0$$

Figura 3.16.: Modelo de Pareto. Función de densidad



Debemos observar que los parámetros μ y σ que aparecen en las expresiones anteriores corresponden al modelo normal y no al log-normal. Las características de la distribución logaritmo normal son las siguientes:

$$E(Y) = e^{\mu + \frac{\sigma^2}{2}} \; ; \; Var(Y) = e^{2\mu}\left(e^{2\sigma^2} - e^{\sigma^2}\right)$$

El modelo log-normal resulta aplicable cuando numerosos factores pequeños presentan un efecto multiplicativo. Esta ley, denominada "ley del efecto proporcional", fue introducida por McAlister (1879), si bien es conocida habitualmente como Ley de Gibrat debido a que fue este autor quien en su obra *Les Inegalités Economiques* (1931) la estudió de modo exhaustivo y la aplicó como modelo de renta.

Aunque el modelo logaritmo normal resulta adecuado para describir el comportamiento probabilístico de los tramos bajos de renta, suelen aparecer problemas en los tramos altos, para los que habitualmente esta distribución subestima las proporciones de rentistas.

3.6.3.2. Distribución de Pareto

En un análisis distributivo de la renta parece deseable tener en cuenta la existencia de un mínimo necesario para subsistir. Este valor (umbral que denominamos x_0) podría ser el gasto en alimentación, el salario mínimo interprofesional, la subvención a hogares pobres,...) y como consecuencia, la distribución de la variable podría venir representada por una curva como la recogida en la figura, correspondiente a un *modelo de Pareto* (en este caso $P(X_0 = 5, \alpha = 3)$).

El modelo de Pareto, introducido por este autor a finales del siglo pasado, se ha revelado históricamente útil en la descripción de la distribución de la renta y la riqueza. Dicho modelo se basa en que el número de personas que reciben una renta superior a cierta cantidad R es inversamente proporcional (aunque no de forma lineal) al citado valor.

Esta distribución viene caracterizada por dos parámetros: el ya comentado "nivel mínimo" x_0 y una constante α, ambos no negativos.

La función de densidad de este modelo y su representación gráfica, denominada curva de Pareto e ilustrada en la figura, 3.16 indican cómo a medida que aumentan los niveles de X disminuye su densidad de probabilidad. A partir de ella es posible obtener la proporción de personas con renta superior a un valor dado x como $\left(\frac{x_0}{x}\right)^{\alpha}$.

El modelo de Pareto es un caso particular de distribución truncada, que se presenta con cierta frecuencia en estadística económica.

Una distribución truncada es aquélla elaborada a partir de otra distribución, al darle un corte a la altura de cierto valor de la variable aleatoria e ignorando la parte derecha o izquierda de la misma (también podría considerarse un doble truncamiento e ignorar las dos bandas, quedándonos sólo con la parte central).

La ley de Pareto fue introducida por este autor a finales del siglo pasado, al estudiar la distribución de la renta y la riqueza. Según su propia formulación la distribución de renta viene dada por: $N = \frac{A}{x^{\alpha}}$, donde N es el número de personas por encima de un cierto valor R, y A y α son constantes.

Suponiendo que la renta x se sitúa por encima de un mínimo x_0, esta ley se reduce a un truncamiento de la distribución exponencial negativa en el punto $\ln x_0$.

Definición 3.12. Decimos que una variable aleatoria sigue la *ley de Pareto* de parámetros α y x_0, siendo $\alpha>0$, $x_0>0$, si su función de densidad viene dada por:

$$f(x) = \begin{cases} \frac{\alpha x_0^{\alpha}}{x^{\alpha+1}} & si\, x \geq x_0 \\ 0 & \text{en otro caso} \end{cases}$$

La esperanza matemática de este modelo existirá para $\alpha > 1$ y viene dada por la expresión: $E(X) = \frac{x_0 \alpha}{\alpha - 1}$. Por su parte, la varianza existirá para todo $\alpha > 2$: $Var(X) = \frac{\alpha x_0^2}{(\alpha-2)(\alpha-1)^2}$

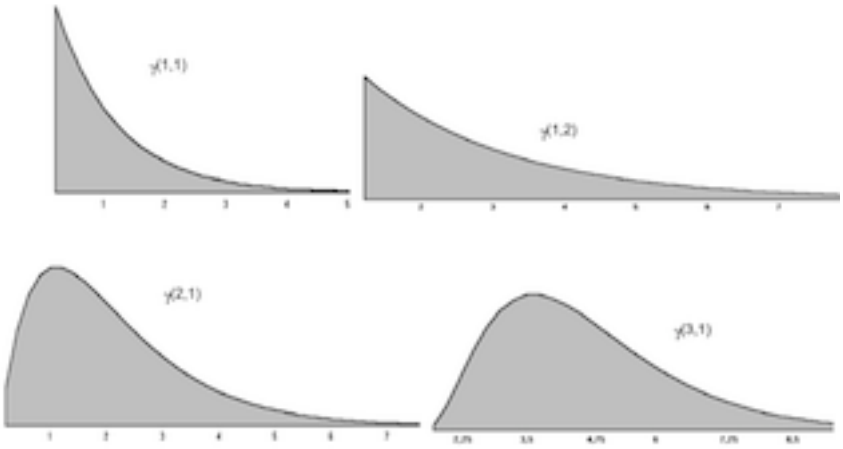
[Deducir las expresiones anteriores. Para la varianza puede comprobarse previamente que se cumple $E(X^2) = \frac{\alpha x_0^2}{\alpha-2}$]

Esta distribución admite una interpretación muy sencilla por cuanto que el ratio entre su valor esperado y la renta mínima, $\frac{E(X)}{x_0} = \frac{\alpha}{\alpha-1}$, puede ser considerado como una medida de desigualdad. De hecho, los estudios de Wilfredo Pareto aplicando su ley empíricamente a varios países y en distintos períodos de tiempo arrojaban resultados muy estables de α, lo que llevó a este autor a defender la incapacidad de las políticas más progresistas para reducir el nivel de desigualdad.

Debemos tener en cuenta sin embargo que la elección del valor mínimo condiciona el valor de α. Como consecuencia, el modelo de Pareto sólo describe la distribución de la renta -y por tanto mide la correspondiente desigualdad- para rentas superiores a la adoptada como umbral.

El interés del modelo de Pareto en economía se justifica por su validez para ajustar distribuciones empíricas, excepto en los estratos inferiores de renta. Como consecuencia, esta ley se complementa muy bien con la distribución logaritmo normal, en el sentido de que cuando una no se ajusta bien a la distribución de la renta, la otra suele dar resultados satisfactorios, y, viceversa. De forma global (las dos colas) las distribuciones de renta también suelen ajustarse, entre otros, a través de modelos Gamma.

Figura 3.17.: Modelo Gamma. Funciones de densidad



3.6.3.3. Distribución Gamma

El *modelo gamma* es otra distribución continua utilizada para describir la renta. Este modelo, que depende de dos parámetros (p y a), viene representado gráficamente por una curva que suele adaptarse bien a los distintos niveles de rentas.

Definición 3.13. Decimos que una variable aleatoria X sigue una *distribución Gamma* con parámetros p y a, que se denota por $\gamma(p, a)$, si su función de densidad viene dada por:

$$f(x) = \begin{cases} \dfrac{a^p}{\Gamma(p)} x^{p-1} e^{-ax} & si\, x > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Los parámetros característicos del modelo gamma p y a adoptan siempre valores positivos y recogen características de forma y escala respectivamente. En consecuencia, cambios en el parámetro p alteran el perfil o forma gráfica del modelo, mientras que el parámetro a viene relacionado con la unidad de medida de la variable tal y como muestra la figura 3.17.

Un caso particular de esta distribución es la expresión correspondiente a $p = 1$, modelo que recibe la denominación de exponencial de parámetro a y ha sido estudiado anteriormente.

Por lo que se refiere a las características de la distribución gamma, su valor esperado viene dado por $E(X) = \frac{p}{a}$ y su varianza es $Var(X) = \frac{p}{a^2}$.

La distribución gamma es una contribución de L. Euler (1707-1783) pero fue O. Ammon (1895) el primero en proponerla como modelo descriptivo de la distribución de la renta.

Otras aplicaciones de este modelo se deben a March (1898), Salem y Mount (1974) y Bartels (1977). Existen además generalizaciones de la distribución gamma como la propuesta por Amoroso (1924) y varios modelos probabilísticos conectados con éste.

Con el objetivo de aumentar la capacidad descriptiva de los modelos, algunos autores han introducido nuevas distribuciones probabilísticas de la renta. Entre ellas se encuentran la de Singh-Maddala (1976) que se obtiene como caso particular de una generalización de la familia beta, y la de Dagum (1977) que ha mostrado una buena adaptación a las distribuciones de renta tanto en países desarrollados como en otros en vías de desarrollo. No obstante, algunos de los modelos más recientes que se han revelado como muy adecuados para la descripción de la renta presentan expresiones muy complejas en las que intervienen varios parámetros que no resultan sencillos de estimar.

Todas estas distribuciones persiguen una descripción adecuada del comportamiento probabilístico de las rentas. Además, es interesante señalar que los parámetros característicos de estos modelos aparecerán conectados con los indicadores de la desigualdad de renta.

Sin entrar aquí en un análisis detallado de la desigualdad, presentamos a modo de resumen las expresiones que adoptan la medida clásica de Gini-Lorenz y el índice de desigualdad colectiva bajo los modelos probabilísticos más habituales para las rentas: Pareto, Log-normal y Gamma.

Índice	Pareto	Log-normal	Gamma
Indice Gini-Lorenz	$\frac{1}{2\alpha-1}$	$2F_{\mathcal{N}(0,1)}\left(\frac{\sigma}{\sqrt{2}}\right)-1$	$\frac{\Gamma\left(\alpha+\frac{1}{2}\right)}{\Gamma(\alpha+1)\sqrt{\pi}}$
Desigualdad colectiva	$\frac{1}{\alpha^2-1}$	$e^{Var(X)}-1$	$\frac{1}{\alpha-1}$

Si la renta es una variable aleatoria X cuya función de densidad de probabilidad es $f(x)$, los indicadores de desigualdad anteriores vienen dados por las expresiones siguientes:

- Indice de Gini-Lorenz: $L(X)=1-2\int_0^{+\infty}F_1(x)f(x)dx$, donde $F_1(x)=\frac{1}{\mu}\int_0^{+\infty}f(t)dt$
- Indice de desigualdad colectiva: $D(X)=\int_0^{+\infty}\left(\frac{\mu}{x}-1\right)f(x)dx$

4. Vectores aleatorios y distribuciones de agregados

La realidad socioeconómica es compleja y de ahí la necesidad de introducir los conceptos y expresiones que permitan un tratamiento simultáneo de k variables aleatorias.

Supongamos a modo de ejemplo que estamos interesados en conocer la probabilidad de que los beneficios de un centro comercial al finalizar este año superen a los del anterior. Lógicamente, este suceso aparecerá relacionado con el comportamiento de numerosas variables como el nivel de precios, la renta familiar disponible, las campañas publicitarias, la competencia de otros centros comerciales cercanos, ... Así pues, la variable X, "beneficios netos" aparecería relacionada con otro conjunto de variables Y, Z, W, ... con lo cual una probabilidad referida a X (por ejemplo, que los beneficios superen los 2.000 euros) podría ser planteada en los siguientes términos:

$$P(X > 2.000) = P(Y \geq 0,12,\ Z \geq 15.000, ...)$$

y para poder determinar la probabilidad pedida sería necesario conocer la distribución conjunta de ese vector de variables aleatorias.

Además, al realizar nuestro análisis podríamos también tener en cuenta que los beneficios totales del centro comercial son resultado de agregar los beneficios de cada uno de los establecimientos que lo integran, esto es, $X = \sum_i X_i$. Evidentemente, los distintos establecimientos podrían registrar evoluciones dispares, pero si su número es elevado, un comportamiento anómalo en uno de ellos no afectará demasiado al agregado (esta desviación podría verse compensada por otras de sentido contrario). Como consecuencia, resultará posible efectuar afirmaciones relativas al total ignorando cada uno de los sumandos, ya que este beneficio agregado tendrá un comportamiento "normal".

La consideración de vectores aleatorios integrados por k variables aleatorias permite abordar el estudio exhaustivo (marginal o condicionado) de una de ellas, resultando perfectamente aplicables los contenidos de capítulos anteriores. Sin embargo, nos interesarán preferentemente los análisis conjuntos, que permitirán conocer el modo en el que las variables se relacionan, si son o no independientes, ...

Al igual que sucedía para el estudio individualizado de variables aleatorias, existen infinitas distribuciones de probabilidad k-dimensionales. De entre ellas, podemos destacar ciertos modelos probabilísticos habituales para describir fenómenos económicos que resultan de generalizar al caso k-dimensional los modelos conocidos: así, la generalización del modelo binomial conducirá a la distribución de probabilidad multinomial y la extensión del hipergeométrico a la distribución multihipergeométrica. De forma

similar, introduciremos los modelos de Poisson multivariante y normal multivariante, resultado de generalizar al caso k-dimensional las distribuciones de Poisson y normal respectivamente.

Por otra parte, en el ámbito económico son numerosos los ejemplos de magnitudes aleatorias que se obtienen como resultado de la actuación conjunta de diversas causas. En estas situaciones -por ejemplo, los beneficios globales de los establecimientos del centro comercial, la producción final del sector industrial, la demanda agregada de un producto agrícola- el estudio se centra a menudo en el efecto final, para el cual es posible garantizar -bajo ciertos supuestos- una distribución aproximadamente normal.

4.1. Vectores aleatorios. Distribuciones k-dimensionales

La formalización del estudio de vectores aleatorios k-dimensionales permite ampliar el análisis de fenómenos aleatorios. Sin ánimo de llevar a cabo un estudio exhaustivo de las distribuciones k-dimensionales, dedicamos este epígrafe a recoger las principales definiciones y conceptos asociados al estudio de vectores aleatorios.

Centrándonos en el caso bidimensional, si investigamos conjuntamente dos características aleatorias X e Y, podemos disponer la información relativa a ambas variables mediante una tabla como la representada a continuación, en la que aparecen las probabilidades asociadas a los pares (x_i, y_j).

Y/X	x_1	x_2	$\cdots$	x_k	$p_{\cdot j}$
y_1	p_{11}	p_{21}	$\cdots$	p_{k1}	$p_{\cdot 1}$
y_2	p_{12}	p_{22}	$\cdots$	p_{k2}	$p_{\cdot 2}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
y_h	p_{1h}	p_{2h}	$\cdots$	p_{kh}	$p_{\cdot h}$
$p_{i\cdot}$	$p_{1\cdot}$	$p_{2\cdot}$	$\cdots$	$p_{k\cdot}$	$p_{\cdot\cdot} = 1$

4.1.1. Variable aleatoria bidimensional

Definición 4.1. Dadas dos v.a. unidimensionales X e Y definidas sobre el mismo espacio de probabilidad $(E, \mathcal{A}, P)$, se denomina *variable aleatoria bidimensional, que denotamos por el par* (X, Y), a la observación conjunta de dos variables:

$$(X, Y) : w \in E \to (X(w), Y(w)) \in \Re^2$$

De una manera más formalizada, si X e Y son dos v.a. definidas sobre el mismo espacio de probabilidad $(E, \mathcal{A}, P)$, y denotamos por $\boldsymbol{\beta}_2$ la σ-álgebra de Borel sobre $\Re^2$, construida con todas las uniones, intersecciones, complementarios, ... de rectángulos de $\Re^2$, definimos una variable aleatoria bidimensional (X, Y) como una aplicación:

$$(X, Y) : w \in E \to (X(w), Y(w)) \in \Re^2$$

tal que la imagen inversa de cualquier boreliano de $B \in \boldsymbol{\beta}_2$ sea un elemento de la σ-álgebra $\mathcal{A}$.

En este caso se define la probabilidad inducida como una aplicación: $P' : B \in \boldsymbol{\beta}_2 \rightarrow P'(B) = P\left[(X,Y)^{-1}(B)\right] \in \Re$

La σ-álgebra de Borel $\boldsymbol{\beta}_2$ está generada por rectángulos de la forma $(-\infty, x] \times (-\infty, y]$ (todo elemento $B \in \boldsymbol{\beta}$ puede ser expresado mediante operaciones de los rectángulos anteriores); si denotamos por $[X \leq x, Y \leq y] = \{w \in E / -\infty < X(w) \leq x, -\infty < Y(w) \leq y\}$, para comprobar que la variable bidimensional (X,Y) es aleatoria basta comprobar que para todo $(x,y) \in \Re^2$, se verifica: $[X \leq x, Y \leq y] = \{w \in E / X(w) \leq x, Y(w) \leq y\} \in \mathcal{A}$.

De la misma forma, la probabilidad inducida puede establecerse como: $P'\left[(-\infty, x] \times (-\infty, y]\right] = P(X \leq x, Y \leq y)$.

Decimos que una v.a. bidimensional (X,Y) es discreta si las variables X e Y que la integran son discretas. De igual manera diremos que es continua si sus componentes lo son.

En esta clasificación de las v.a. bidimensionales, la definición dada para variables continuas (que lo sean sus componentes) es en realidad una condición necesaria pero no suficiente; pueden encontrarse contraejemplos en los que tanto X como Y son continuas y en cambio la variable conjunta no lo es.

4.1.1.1. Función de distribución bidimensional

Definición 4.2. Dada una v.a. bidimensional (X,Y) definimos la *función de distribución conjunta* asociada a esta variable como:

$$F : (x,y) \in \Re^2 \rightarrow F(x,y) = P(X \leq x, Y \leq y) \in [0,1]$$

Esta función cumple las propiedades exigidas a las f.d.. Su representación gráfica tendría forma de un estereograma escalonado en el espacio para variables discretas (con puntos de salto en los pares (x_i, y_j)) y sería una superficie continua para variables continuas.

A partir de la función de distribución bidimensional podemos calcular la probabilidad de cualquier rectángulo $(a,b] \times (c,d]$, mediante la expresión:

$$P(a < X \leq b, c < Y \leq d) = F(b,d) - F(a,d) - F(b,c) + F(a,c)$$

4.1.1.2. Función de probabilidad bidimensional

Dada una v.a. bidimensional discreta, ésta podrá tomar un conjunto numerable de valores (x_i, y_j), $i,j = 1,2,\ldots$. Como caso particular de la expresión anterior podemos obtener la probabilidad conjunta de cada par de valores a partir de la función de distribución como:

$$P(X = x_i, Y = y_j) = F(x_i, y_j) - F(x_i, y_{j-1}) - F(x_{i-1}, y_j) + F(x_{i-1}, y_{j-1}) = p_{ij}$$

Definición. Podemos definir la *función de probabilidad conjunta* de una v.a. bidimensional (X,Y) como aquélla que asigna a cada posible resultado (x_i, y_j) una masa de probabilidad que verifica: $p_{ij} \geq 0$, y $\sum_{ij} p_{ij} = 1$.

A partir de la función de probabilidad es posible obtener la f.d. en un punto (x_i, y_j), como suma de las probabilidades de todos los pares con coordenadas no superiores al punto considerado; esto es:

$$F(x_i, y_j) = \sum_{i_1 \leq i} \sum_{j_1 \leq j} p_{i_1 j_1}$$

4.1.1.3. Función de densidad bidimensional

Si consideramos dos variables aleatorias continuas, podemos utilizar el mismo razonamiento anterior partiendo de una agrupación por intervalos de estas variables; pero este método perdería validez si planteásemos el problema a partir de valores individuales [¿por qué?].

Definición. Dada una v.a. bidimensional continua (X, Y), si existe una función $f(x, y)$, tal que:

$$f(x, y) = \frac{\partial^2 F(x, y)}{\partial x \partial y}$$

la denominaremos *función de densidad bidimensional (o conjunta) de X e Y.*

Obsérvese que para que exista la función de densidad de una v.a. bidimensional continua es necesario que exista la derivada parcial de segundo orden respecto a x e y.

A modo de operación inversa, podemos obtener la función de distribución conjunta a partir de la de densidad como:

$$F(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(x, y) dx dy$$

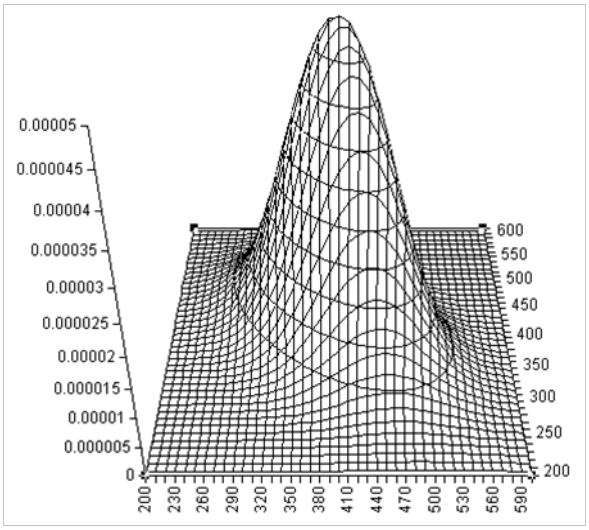
Por otra parte, a partir de la f.d. $F(x, y)$, podemos calcular la probabilidad de cualquier rectángulo $(a, b] \times (c, d]$, mediante la expresión:

$$P(a < X \leq b,\, c < Y \leq d) = \int_{a}^{b} \int_{c}^{d} f(x, y) dx dy$$

La probabilidad de que una v.a. se encuentre en un intervalo representa el área en ese intervalo por debajo de la función de densidad. En el caso bidimensional, la probabilidad de encontrarse en un rectángulo $(a, b] \times (c, d]$, será el volumen que sobre el mismo establece la función de densidad conjunta.

Así pues, podemos definir la función de densidad conjunta $f(x, y)$, si existe, como una aplicación $f : (x, y) \in \Re^2 \to \Re^+$, tal que:

Figura 4.1.: Función de densidad bidimensional



$\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$ y

$$P(a < X \leq b, c < Y \leq d) = \int_a^b \int_c^d f(x,y)dxdy$$

Las condiciones de existencia de la función de densidad bidimensional son equivalentes a las comentadas para el caso unidimensional con la extensión lógica de una a dos variables.

Las variables continuas que utilizaremos en este libro son absolutamente continuas por lo que identificaremos ambos términos, en un abuso de lenguaje, y por tanto consideraremos que dada una v.a. bidimensional continua, su función de densidad siempre existe.

4.1.1.4. Vectores aleatorios k-dimensionales

Habitualmente nos interesará recoger más de dos características de los elementos de la población. Podemos extender entonces los conceptos anteriores al caso k-dimensional, considerando el vector aleatorio $(X_1, X_2, \ldots, X_k)$.

Supongamos que cada componente del vector anterior es una v.a., en cuyo caso se dice que se trata de un vector aleatorio, y que son observadas conjuntamente para cada elemento de la población, de forma que cada elemento w proporciona un vector de información: $(X_1(w), X_2(w), \ldots, X_k(w))$; esto es, la variable aleatoria k-dimensional puede entenderse como:

$$(X_1, X_2, \ldots, X_k) : w \in E \rightarrow (X_1(w), X_2(w), \ldots, X_k(w)) \in \Re^k$$

Definida una σ-álgebra de Borel sobre $\Re^k$, generada por cubos k-dimensionales de la forma $(-\infty, x_1] \times (-\infty, x_2] \times \cdots \times (-\infty, x_k]$, la probabilidad inducida nos permite establecer una función de distribución k-dimensional $F(x_1, x_2, \ldots, x_k)$:

$$F : (x_1, x_2, \ldots, x_k) \in \Re^k \rightarrow F(x_1, x_2, \ldots, x_k) \in [0, 1]$$

Consideraremos una v.a. k-dimensional discreta o continua cuando lo sean sus componentes.

Dada una v.a. k-dimensional discreta $(X_1, X_2, \ldots, X_k)$, definimos la función de probabilidad k-dimensional como aquella que a cada posible valor de la variable, $(x_1, x_2, \ldots, x_k)$, le asigna una masa de probabilidad, que verifica:

$$p_{i_1,i_2,\ldots,i_k} = P(x_{i_1}, x_{i_2}, \ldots, x_{i_k}) = P(X_1 = x_{i_1}, X_2 = x_{i_2}, \ldots, X_k = x_{i_k})$$

$$p_{i_1,i_2,\ldots,i_k} \geq 0 \; y \; \sum_{i_1=0}^{+\infty} \sum_{i_2=0}^{+\infty} \cdots \sum_{i_k=0}^{+\infty} p_{i_1 i_2 \cdots i_k} = 1$$

La función de distribución se puede obtener mediante agregación de la función de probabilidad y ésta por diferencias de la función de distribución.

Para el caso de una v.a. k-dimensional continua $(X_1, X_2, \ldots, X_k)$, definimos la función de densidad como una aplicación f de $\Re^k$ en $\Re$, que cumple: $f(x_1, x_2, \ldots, x_k) \geq 0$, $\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k = 1$ y $P(a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2, \ldots, a_k < X_k \leq b_k) = \int_{a_1}^{b_1} \int_{a_2}^{b_2} \cdots \int_{a_k}^{b_k} f(x_1, x_2, \ldots, x_k) dx_1 dx_2 \cdots dx_k$

Podemos obtener la función de distribución a partir de la de densidad como una integral múltiple de ésta, y de forma inversa, la densidad conjunta se obtiene a partir de la función de distribución como la derivada parcial de orden k respecto a sus componentes $x_1, x_2, \ldots, x_k$.

4.2. Distribuciones marginales y condicionadas

Una variable aleatoria bidimensional (X, Y) es un vector formado por dos v.a. unidimensionales. Abordamos en este epígrafe la caracterización probabilística de cada una de estas componentes considerando un comportamiento libre de la otra variable o bien exigiéndole determinada condición.

Cuando imponemos alguna restricción al rango de la segunda variable nos encontraremos con distribuciones condicionadas, mientras que en caso contrario hablaremos de distribuciones marginales.

4.2.1. Distribuciones marginales

Para introducir el concepto de marginalidad consideremos una variable estadística bidimensional genérica (X, Y) que representamos por la siguiente tabla:

$\downarrow Y/X \rightarrow$	2	4	8	$f_{.j}$
1	0,1	0,2	0,1	0,4
2	0,05	0,05	0,1	0,2
3	0,1	0,1	0,2	0,4
$f_{i.}$	0,25	0,35	0,4	1

Si nos preguntamos por la frecuencia relativa con la que la variable X toma el valor 4 independientemente del comportamiento de la variable Y, la respuesta es 0,35, resultado obtenido como suma de las frecuencias relativas de $X = 4$ con todas las diferentes alternativas de la variable Y. Pues bien, podemos extender este concepto

de forma inmediata a las variables aleatorias.

Dada una v.a. bidimensional (X, Y), denominamos distribución marginal de X a la distribución de probabilidad de la v.a. X cuando consideramos un comportamiento libre de Y. Esto es, denotando por $F_X(x)$ a su función de distribución, se tiene:

$$F_X(x) = P(X \leq x) = P(X \leq x, Y < +\infty) = \lim_{y \to +\infty} F(x, y)$$

Distinguiendo para variables discretas y continuas, se obtiene:

- $F_X(x_i) = \lim_{y \to +\infty} F(x_i, y) = \sum_{h \leq i} \sum_{j=1}^{+\infty} p_{hj}$
- $F_X(x) = \lim_{y \to +\infty} F(x, y) = \int_{-\infty}^{x} \left(\int_{-\infty}^{+\infty} f(x, y) dy \right) dx$

A partir de la función de distribución obtenemos las funciones de probabilidad o de densidad mediante la diferencia o la derivada de la primera, obteniendo en cada caso la expresión que figura entre paréntesis ; es decir:

- $p_X(x_i) = F_X(x_i) - F_X(x_{i-1}) = \sum_{j=1}^{+\infty} p_{ij}$
- $f_X(x) = F_X^{'}(x) = \int_{-\infty}^{+\infty} f(x, y) dy$

Definición 4.3. Si (X, Y) es una v.a. bidimensional discreta definimos la *función de probabilidad marginal de* X como:

$$p_X(x) = \sum_{j=1}^{+\infty} p_{ij}$$

Si la variable es continua, definimos la *función de densidad marginal de* X, que denotamos por $f_X(x)$, como:

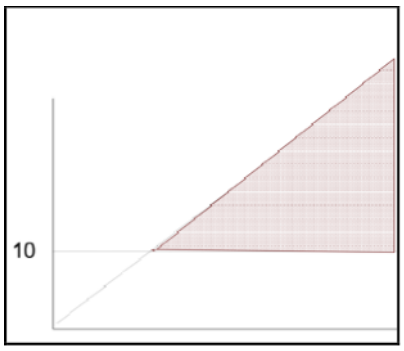
$$f_X(x) = \int_{-\infty}^{+\infty} f(x, y) dy$$

Si ahora consideramos la tabla inicial como una distribución probabilística bidimensional donde aparecen los valores que pueden adoptar las variables X e Y y las probabilidades p_{ij} con las que pueden asumir conjuntamente los valores (x_i, y_j), podemos obtener la probabilidad marginal $p_X(4) = 0,2 + 0,05 + 0,1 = 0,35$.

Razonando de modo similar para los restantes valores de X se llega a la distribución marginal recogida a continuación:

X	$p_X(x)$
2	0,25
4	0,35
8	0,4

Figura 4.2.: Campo de variación bidimensional



[Comprobar que se trata de una verdadera distribución de probabilidad. ¿Cuál sería la distribución marginal de la variable Y?]

El caso continuo incorpora algunos rasgos diferenciales, que podemos examinar a través de una ilustración. Consideremos un reciente estudio sobre los márgenes comerciales del sector de electrodomésticos, en el que se observaron conjuntamente los precios de venta (X) y su coste (Y), ambos en miles de unidades monetarias, obteniéndose la siguiente función de densidad:

$$f(x,y) = \begin{cases} \dfrac{200}{x^2y^2} & si\ 10 < y < x \\ 0 & \text{en otro caso} \end{cases}$$

cuyo campo de variación -representado en la figura 4.2- supone que el precio de venta tiene que ser mayor que su coste ($y < x$). [Compruébese que $f(x,y)$ es no negativa y se cumple $\int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$].

Si ahora queremos obtener la distribución marginal del precio de venta, podemos utilizar las expresiones anteriormente vistas para $f_X(x)$:

$$f_X(x) = \int_{-\infty}^{+\infty} f(x,y)dy = \int_{10}^{x} \frac{200}{x^2y^2}dy = \frac{200}{x^2}\left[-\frac{1}{y}\right]_{10}^{x} = \frac{20}{x^2} - \frac{200}{x^3}\ ,\ \forall x > 10$$

[Obtener de modo similar la distribución marginal del coste $f_Y(y)$]

Podemos comprobar que las funciones de probabilidad y de densidad marginales verifican los requisitos exigidos a estas expresiones; esto es, que son funciones no negativas cuya suma o integral es la unidad.

En el caso de la función de densidad marginal tendríamos que es un valor no negativo puesto que a cada punto le asigna el área de la sección determinada por la superficie $f(x,y)$ con un plano paralelo al eje de la variable que marginamos, y como tal área no puede adoptar valores negativos.

En segundo lugar tendríamos que comprobar que $\int_{-\infty}^{+\infty} f_X(x)dx = 1$ (en el caso que se tratase de la marginal de X). En efecto:

$$\int_{-\infty}^{+\infty} f_X(x)dx = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)dxdy = 1$$

obteniéndose la última igualdad por ser$f(x,y)$ una función de densidad.

[Comprobar que la función de probabilidad marginal $p_X(x)$ verifica las condiciones de no negatividad y suma unitaria]

La *función de distribución marginal* $F_X(x)$, puede obtenerse a partir de las correspondientes funciones de probabilidad o de densidad, según se trate de v.a. discretas o continuas, mediante suma o integración de las mismas:

$$F_X(x) = \sum_{x_i \leq [x]} p_X(x_i) = \sum_{x_i \leq [x]} \sum_{j=1}^{\infty} p(x_i, y_j)$$

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(x,y)dxdy$$

Las expresiones de cálculo de las probabilidades de intervalos, $[a < X \leq b]$, de valores $[X = x_i]$ o la obtención de la función de densidad marginal a partir de la función de distribución son iguales que en el caso unidimensional, teniendo en cuenta que ahora se trata de una distribución marginal.

Características marginales y medidas de correlación

Dado que las distribuciones marginales son variables unidimensionales es posible definir las características marginales asociadas a las mismas. En el caso continuo, la esperanza y la varianza marginal de X vendrían dadas por las siguientes expresiones:

$$\begin{aligned} E(X) = & \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} x f(x,y)dxdy = \int_{-\infty}^{+\infty} x \left[\underbrace{\int_{-\infty}^{+\infty} f(x,y)dy}_{=f_X(x)} \right] dx = \\ = & \int_{-\infty}^{+\infty} x f_X(x)dx = \mu_X \end{aligned}$$

$$\begin{aligned} \sigma_X^2 = & E(X-\mu_X)^2 = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} (x-\mu_X)^2 f(x,y)dxdy = \\ = & \int_{-\infty}^{+\infty} (x-\mu_X)^2 \left[\underbrace{\int_{-\infty}^{+\infty} f(x,y)dy}_{=f_X(x)} \right] dx = \int_{-\infty}^{+\infty} (x-\mu_X)^2 f_X(x)dx \end{aligned}$$

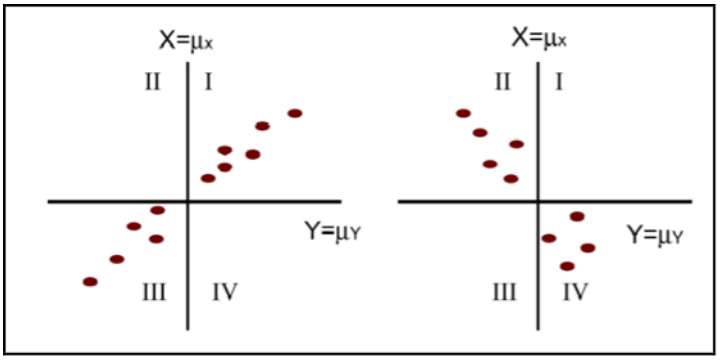
[¿Cómo se expresarían las características de la distribución marginal de Y?, ¿cuáles serían las expresiones anteriores para el caso de v.a. discretas?]

Entre las características marginales se encuentran los momentos centrados de orden 2, que se corresponden con las varianzas marginales. Nos interesan también otros momentos, mixtos, de orden 1 en la variable X y 1 en la variable Y.

Definición 4.4. Dadas dos variables aleatorias X e Y llamamos Covarianza, que denotamos por $Cov(X,Y)$ o σ_{XY} al valor de la expresión, si existe:

$$\sigma_{X,Y} = Cov(X,Y) = E\left[(X-\mu_X)(Y-\mu_Y)\right]$$

Figura 4.3.: Correlación X, Y



Sus fórmulas de cálculo para los casos discreto y continuo son las siguientes:

- $\sigma_{X,Y} = \sum_{i=1}^{\infty}\sum_{j=1}^{\infty}(x_i - \mu_X)(y_j - \mu_Y)p_{ij}$
- $\sigma_{X,Y} = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty}(x - \mu_X)(y - \mu_Y)f(x,y)dxdy$

La covarianza es una medida de la correlación lineal existente entre dos variables aleatorias. Así, si las variables X e Y presentan una relación lineal directa, se observará una nube de puntos como la representada en la parte izquierda de la figura 4.3, esto es, distribuida a lo largo de los cuadrantes I y III.

En tal situación, se observa que las desviaciones $(X - \mu_X) > 0$ aparecen asociadas a $(Y - \mu_Y) > 0$ y del mismo modo las desviaciones $(X - \mu_X) < 0$ se presentan con $(Y - \mu_Y) < 0$. Ambas posibilidades dan como resultado productos positivos por lo cual la covarianza presentará signo positivo.

Razonando de modo similar para el caso de relación lineal negativa, se observarían desviaciones negativas para una de las variables junto a desviaciones positivas para la otra (que se agruparían en los cuadrantes II y IV, tal y como recoge la gráfica representada a la derecha).

Puede observarse que la expresión de la covarianza es simétrica y por tanto la covarianza de X con Y es igual a la obtenida de Y con X.

Proposición. *La covarianza de una variable sobre sí misma coincide con la varianza marginal:*

$$\sigma_{X,X} = \sigma_X^2 = E\left[(X - \mu_X)(X - \mu_X)\right] = E\left(X - \mu_X\right)^2$$

Al igual que la varianza, la covarianza admite una expresión de cálculo más simple. En concreto, la covarianza puede obtenerse como diferencia entre la esperanza del producto y el producto de las esperanzas marginales:

$$\sigma_{X,Y} = E(XY) - E(X)E(Y)$$

[Compruébese]

La covarianza presenta la importante limitación de no encontrarse acotada; el valor resultante de la covarianza nos permitirá establecer el tipo de dependencia lineal (nula, directa o inversa), pero no así el nivel de la relación. Este inconveniente motiva la utilización del *coeficiente de correlación lineal* definido como cociente entre la covarianza y las desviaciones típicas de las variables:

$$\rho_{X,Y} = \frac{\sigma_{X,Y}}{\sigma_X \sigma_Y}$$

expresión que se encuentra acotada entre -1 y 1, y resulta adecuada como medida de correlación lineal entre dos variables aleatorias, puesto que su valor absoluto expresa el grado de la relación lineal.

Si consideramos el vector fila aleatorio $\mathbf{x} = (X, Y)$, podemos construir un vector de valores esperados y matrices de varianzas-covarianzas y de correlaciones como sigue:

$$E(\mathbf{x}) = E(X, Y) = (E(X), E(Y)) = (\mu_X, \mu_Y)$$

$$Cov(\mathbf{x}) = \begin{bmatrix} \sigma_X^2 & \sigma_{Y,X} \\ \sigma_{Y,X} & \sigma_Y^2 \end{bmatrix}$$

$$Corr(\mathbf{x}) = \begin{bmatrix} 1 & \rho_{Y,X} \\ \rho_{Y,X} & 1 \end{bmatrix}$$

Al igual que hemos visto para las variables unidimensionales, es posible generar los momentos bidimensionales a partir de una función generatriz.

La función generatriz de momentos de una v.a.bidimensional se define como el valor, si existe, de la expresión:

$$M_{(X,Y)}(t_1, t_2) = E\left(e^{t_1 X + t_2 Y}\right)$$

pudiendo comprobarse fácilmente a partir de esta expresión: $M_{X+Y}(t) = M_{(X,Y)}(t, t)$.

Distribuciones marginales en variables k-dimensionales

Podemos extender los conceptos anteriores de distribuciones marginales y sus matrices características al caso de variables k-dimensionales sin más complejidad que la derivada de su terminología.

En efecto, si denotamos por $\mathbf{x}$ un vector aleatorio con k componentes, $\mathbf{x} = (X_1, X_2, \ldots, X_k)$, la distribución marginal del componente X_j, será la distribución unidimensional de esta variable con independencia de lo que ocurra con los restantes componentes; esto es:

$$F_{X_j}(x_j) = \lim_{x_1 \to \infty} \cdots \lim_{x_{j-1} \to \infty} \lim_{x_{j+1} \to \infty} \cdots \lim_{x_k \to \infty} F(x_1, \ldots, x_{j-1}, x_j, x_{j+1}, \ldots x_k)$$

Para variables discretas y continuas obtenemos las funciones de probabilidad y densidad, respectivamente:

- $P_{X_j}(x_i) = \sum_{i_1=0}^{\infty} \cdots \sum_{i_{j-1}=0}^{\infty} \sum_{i_{j+1}=0}^{\infty} \cdots \sum_{i_k=0}^{\infty} P(x_{i_1}, \ldots, x_{i_{j-1}}, x_{i_j}, x_{i_{j+1}}, \ldots, x_{i_k})$

- $f_{X_j}(x_j) = \int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} f(x_1, \ldots, x_j, \ldots, x_k) dx_1 \cdots dx_{j-1} dx_{j+1} \cdots dx_k$

Las distribuciones marginales son de carácter unidimensional y por tanto su función de distribución se obtiene sumando o integrando las correspondientes funciones de probabilidad o de densidad marginales. [A partir de una distribución k-dimensional ¿cuántas distribuciones marginales pueden obtenerse?]

El vector de esperanzas y las matrices de covarianzas y correlaciones vienen en este caso dados por:

$$\mu = E(\mathbf{x}) = E(X_1, X_2, \ldots, X_k) = (E(X_1), E(X_2), \ldots, E(X_k)) = (\mu_1, \mu_2, \ldots, \mu_k)$$

$$Cov(\mathbf{x}) = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_k^2 \end{bmatrix}$$

$$Corr(\mathbf{x}) = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1k} \\ \rho_{21} & 1 & \ldots & \rho_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ \rho_{k1} & \rho_{k2} & \ldots & 1 \end{bmatrix}$$

La *función generatriz de momentos de una v.a. k-dimensional* se define como el valor, si existe, de la expresión:

$$M_{(X_1, X_2, \ldots, X_k)}(t_1, t_2, \ldots, t_k) = E\left(e^{t_1 X_1 + t_2 X_2 + \cdots + t_k X_k}\right)$$

A partir de la definición anterior es inmediato comprobar:

$$M_{\sum_{i=1}^k X_i}(\mathbf{t}) = M_{(X_1, X_2, \ldots, X_k)}(t, t, \ldots, t)$$

4.2.2. Distribuciones condicionadas

Otras distribuciones interesantes que se derivan de la observación conjunta de variables aleatorias son las *distribuciones condicionadas*.

Supongamos en primer lugar un vector bidimensional (X, Y). En el análisis marginal buscábamos la distribución de una variable con independencia del comportamiento de la otra; abordamos ahora un planteamiento complementario: la distribución de una variable cuando la otra adopta un comportamiento determinado; esto es, por ejemplo, la distribución de Y cuando X toma cierto valor $\mathbf{x}$, un conjunto B de posibles valores o cualquier valor no superior a $\mathbf{x}$.

Consideremos este último caso, analizando la distribución de Y cuando X adopta valores no superiores a $\mathbf{x}$. Si denotamos por $Y/X \leq x$ esta variable condicionada, su recorrido es idéntico al de Y, cambiando únicamente la distribución de probabilidad

asociada a la misma.

Dada una v.a. bidimensional (X,Y), con función de distribución conjunta $F(x,y)$, definimos la función de distribución de Ycondicionada a $X \leq x$, como la aplicación:

$$F_{Y/X\leq x}(./X \leq x) : y \in \Re \rightarrow F(y/X \leq x) = P(Y \leq y/X \leq x) = \frac{P(X \leq x, Y \leq y)}{P(X \leq x)} = \frac{F(x,y)}{F_X(x)} \in [0,1]$$

En lugar de condicionar al intervalo $(-\infty, x]$ podríamos hacerlo a cualquier otro conjunto B de posibles resultados o incluso a un único valor x. Cuando la restricción se limita a un único valor pueden surgir algunos problemas, puesto que para variables continuas es nula la probabilidad de un punto concreto (que aparecería en el denominador) por lo que no podríamos determinar la expresión de $F_{Y/X=x}$.

No obstante, estos problemas pueden resolverse sustituyendo el valor por un intervalo infinitesimal y tomando límites cuando la amplitud de éste tiende a cero.

Como en casos anteriores, podemos pasar de la función de distribución a la función de probabilidad o de densidad, según se trate de variables discretas o continuas, calculando la diferencia o la derivada de la función de distribución.

Definición 4.5. Dada una v.a. bidimensional discreta (X,Y) y un resultado $[X = x_i]$ de probabilidad no nula, definimos la *función de probabilidad condicionada* como:

$$P(Y = y_j/X = x_i) = \frac{P(x_i, y_j)}{P_X(x_i)} = \frac{P(x_i, y_j)}{\sum_{j=1}^{\infty} P(x_i, y_j)}$$

De forma similar, para variables aleatorias continuas, podemos definir la *función de densidad condicionada* a un valor x como:

$$f(y/x) = \frac{f(x,y)}{f_X(x)} = \frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy}$$

expresión que permite asignar a cada valor de Y su densidad de probabilidad condicionada, siempre que se cumpla $f_X(x) > 0$.

A modo de ilustración, consideremos que en el ejemplo anterior nos piden la distribución de Y condicionada a que X tome el valor 4. La probabilidad marginal de que se haya presentado el valor $X = 4$ es 0,35 como ya vimos en un apartado anterior, y los valores que puede presentar la variable condicionada $Y/X = 4$ son {1,2,3} (es decir, el mismo recorrido que la variable marginal Y). La probabilidad del valor $Y = 1$ condicionado a $X = 4$ vendrá dada por:

$$P(Y = 1/X = 4) = \frac{P(X = 4, Y = 1)}{P(X = 4)} = \frac{0,2}{0,35} = 0,5714$$

y de modo análogo se obtiene $P(Y = 2/X = 4) = 0,1429$ y $P(Y = 3/X = 4) = 0,2857$.

Las expresiones anteriores cumplen las condiciones exigidas a las funciones de probabilidad y de densidad. En efecto, dada la variable bidimensional continua (X, Y), la densidad de Y condicionada a $X = x$ es una función:

$$f(./x) : y \in \Re \rightarrow f(y/x) \in \Re$$

Según la definición anterior esta función $f(y/x)$ será no negativa, al obtenerse como cociente entre un valor de la función de densidad conjunta y un área, ambos no negativos.

Por otra parte, se tiene:

$$\int_{-\infty}^{+\infty} f(y/x)dy = \int_{-\infty}^{+\infty} \left[\frac{f(x,y)}{\int_{-\infty}^{+\infty} f(x,y)dy} \right] dy = \frac{\int_{-\infty}^{+\infty} f(x,y)dy}{\int_{-\infty}^{+\infty} f(x,y)dy} = 1$$

[Justificar que en el caso discreto, la probabilidad condicionada es una función de probabilidad]. [Definir las funciones de densidad y probabilidad de X condicionada a un valor $Y = y$]

* La función de distribución condicionada $F_{Y/X=x}(y) = P(Y \leq y/X = x)$, presenta las siguientes fórmulas de cálculo:

- $F(y/X = x_i) = \sum_{y_j < [y]} P(y_j/x_i) = \sum_{y_j < [y]} \frac{P(x_i, y_j)}{\sum_{j=1}^{\infty} P(x_i, y_j)} = \frac{\sum_{y_j < |y|} P(x_i, y_j)}{\sum_{j=1}^{\infty} P(x_i, y_j)}$
- $F(y/X = x) = \int_{-\infty}^{y} f(t/x)dt = \int_{-\infty}^{y} \frac{f(x,t)}{\int_{-\infty}^{+\infty} f(x,y)dy} dt$

A modo de ejemplo, a partir de la distribución bidimensional recogida en la tabla inicial, se obtendría la siguiente función de distribución para la variable Y condicionada al valor $X = 4$:

$$F(y/X = 4) = \begin{cases} 0 & para\, y < 1 \\ 0,2857 & para\, 1 \leq y < 2 \\ 0,4286 & para\, 2 \leq y < 3 \\ 1 & para\, 3 \leq y \end{cases}$$

[Obtener de modo similar la función de distribución $F(x/Y = 2)$]

Por lo que se refiere a las características de las distribuciones condicionadas, éstas pueden ser obtenidas de forma análoga al caso de las distribuciones marginales, teniendo en cuenta que ahora se utilizará la función de densidad o de probabilidad condicionada.

Distribuciones condicionadas en variables k-dimensionales

Si consideramos un vector k-dimensional $(X_1, X_2, \ldots, X_k)$ podemos establecer a partir de él condiciones muy diversas; por ejemplo: X_i/X_j, X_i/X_jX_k, X_iX_k/X_j. Puede entonces considerarse un mayor número tanto de variables condicionadas como condicionantes; y en el caso de estas últimas las condiciones pueden ir referidas a valores determinados o bien a rangos de la variable; respetando siempre la condición de que la probabilidad o densidad (según corresponda) del suceso que condiciona sea positiva.

No contemplamos aquí la casuística posible, que se desarrolla a partir de la probabilidad condicionada y sería generalización de las expresiones correspondientes a dos variables. [Proponemos al lector la obtención de algunos de estos resultados. Por ejemplo, ¿cómo se definiría la distribución de probabilidad de X_3 condicionada a $X_1 = x_1^*$ y $X_4 = x_4^*$?]

4.3. Modelos probabilísticos k-dimensionales

4.3.1. Distribución Multinomial

Entre los modelos unidimensionales discretos, la distribución binomial ocupa sin duda un papel relevante al describir el número de resultados favorables obtenidos en experiencias sucesivas.

La distribución binomial aparecía asociada a una experiencia dicotómica en la que sólo distinguimos dos posibles resultados, denominados éxito y fracaso. No obstante, a menudo nos interesan experiencias que presentan más de dos modalidades, cuya descripción por tanto no se adapta a dicho modelo. Este sería, por ejemplo, el caso de un estudio del sector económico en el que desarrolla su actividad una empresa, el estrato de edad al que pertenece un individuo o la intención de voto ante las próximas elecciones.

En todas estas situaciones, existen múltiples modalidades observables en la magnitud objeto de estudio, por lo cual es preciso definir una variable aleatoria capaz de cuantificar la intensidad con la que se ha presentado cada modalidad, surgiendo así el *modelo multinomial* o *polinomial* .

Considerando uno de los ejemplos citados, supongamos que estamos interesados en investigar la actividad económica a la que se dedican n empresas, seleccionadas aleatoriamente y con reposición. Si adoptamos la clasificación sectorial convencional: agricultura, industria, construcción y servicios, definiremos el vector aleatorio (X_1, X_2, X_3, X_4), cuyas componentes recogen el número de empresas dedicadas a cada una de las actividades económicas.

Supongamos un experimento aleatorio que repetimos n veces. En cada una de estas n repeticiones -que asumimos independientes entre sí- el resultado del experimento será uno y sólo uno de los resultados (modalidades) $A_1, A_2, \ldots, A_k$ y designaremos por p_j la probabilidad de que en una de las realizaciones independientes ocurra A_j, asumiendo que dichas probabilidades se mantienen constantes en las n pruebas.

Los sucesos $A_1, A_2 \ldots, A_k$ constituyen una partición de E: $\bigcup_{i=1}^{k} A_i = E$, $A_i \cap A_j = \emptyset$, $\forall i \neq j$ y se cumple por tanto $\sum_{i=1}^{k} p_i = 1$. Sean las variables aleatorias $X_1, X_2, \ldots, X_k$ que recogen el número de veces que han sido observados los sucesos $A_1, A_2, \ldots, A_k$ en las n realizaciones del experimento. Entonces el vector aleatorio k dimensional $((X_1, X_2, \ldots, X_k)$ recibe el nombre de variable aleatoria multinomial de parámetros $n, p_1, p_2, \ldots, p_k$ y se representa como $(X_1, X_2, \ldots, X_k) \approx \mathcal{M}(n, p_1, p_2, \ldots, p_k)$

La función de masa de probabilidad del vector aleatorio $(X_1, X_2, \ldots, X_k) \approx \mathcal{M}(n, p_1, p_2, \ldots, p_k)$ será la siguiente:

$$P(X_1 = x_1, \ldots, X_k = x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k} \; ; \; con \sum_{j=1}^{k} x_j = n$$

Estas probabilidades son no negativas y su suma es la unidad, por lo que se cumplen las condiciones de una función de probabilidad.

Justificábamos el nombre de la distribución binomial porque sus probabilidades se correspondían con los sumandos del binomio $(p+q)^n$. Del mismo modo, en la distribución actual las probabilidades

se corresponden con los sumandos correspondientes del multinomio $(p_1 + p_2 + \cdots + p_k)^n$, hecho que por una parte justifica su nombre y por otra garantiza que la suma de la función de probabilidad es precisamente $(p_1 + p_2 + \cdots + p_k)^n = 1^n = 1$.

Por lo que se refiere a las características de esta distribución, los momentos marginales nos proporcionan las medias y varianzas marginales dadas por las expresiones:

$$E(X_j) = np_j \ ; \ Var(X_j) = np_j(1 - p_j) \ ; \ \forall j = 1, \ldots, k$$

Además, por tratarse de un modelo k dimensional nos interesa conocer las expresiones de algunas características de correlación entre variables unidimensionales. Así la covarianza entre dos variables X_i y X_j, viene dada por:

$$\sigma_{X_i,X_j} = E\left[(X_i - np_i)(X_j - np_j)\right] = -np_ip_j \ ; \ \forall i \neq j = 1, \ldots, k$$

y el correspondiente coeficiente de correlación lineal:

$$\rho_{X_i,X_j} = \frac{\sigma_{X_i,X_j}}{\sigma_{X_i}\sigma_{X_j}} = \frac{-np_ip_j}{\sqrt{np_i(1-p_i)}\sqrt{np_j(1-p_j)}} = -\sqrt{\frac{p_ip_j}{(1-p_i)(1-p_j)}}$$

expresión que como vemos depende de las probabilidades de los sucesos considerados pero no del número de pruebas.

En ocasiones la definición de la distribución multinomial se lleva a cabo excluyendo una de las categorías que se adopta como referencia. Se tendría en este caso un vector k-1 dimensional $(X_1, \ldots, X_{k-1})$ siendo $\sum_{j=1}^{k-1} x_i \leq n$.

4.3.2. Distribución Multihipergeométrica

Al igual que el modelo hipergeométrico, la distribución multihipergeométrica (también llamada polihipergeométrica) aparece asociada al muestreo sin reposición, si bien en este caso se observan simultáneamente varias características. Se trata de una generalización de la distribución hipergeométrica similar a la que el modelo multinomial establece del binomial.

Si en la ilustración anterior de la actividad económica de las empresas asumimos ahora que el muestreo se realiza sin reposición, la situación podría describirse en los siguientes términos: partimos de una población de N empresas que pertenecen a 4 sectores distintos y al seleccionar sin reposición una muestra de n empresas investigamos cuántas de ellas se adscriben a cada sector, obteniendo así una variable aleatoria distribuida según un *modelo multihipergeométrico*.

Consideremos una población de tamaño N en la cual existen N_i elementos con las características de interés $A_1, A_2, \ldots, A_k$, que son excluyentes entre sí. Al extraer de esta población muestras de tamaño n sin reposición, definimos el vector aleatorio $(X_1, X_2, \ldots, X_k)$ donde X_i representa el número de elementos de la muestra con la característica A_i.

La función de masa de probabilidad de $(X_1, X_2, \ldots, X_k)$ viene dada por la expresión:

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k) = \frac{\binom{N_1}{x_1}\binom{N_2}{x_2}\cdots\binom{N_k}{x_k}}{\binom{N}{n}}$$

que corresponde a una *distribución de probabilidad multihipergeométrica* de parámetros $N, N_1, N_2, \ldots, N_k, n$ y se representa $(X_1, X_2, \ldots, X_k) \approx \mathcal{MH}(N, N_1, N_2, \ldots, N_k, n)$

Las características de la distribución multihipergeométrica guardan relación con las vistas para el modelo hipergeométrico obteniéndose:

$$E(X_i) = n\frac{N_i}{N} = np_i \; ; \; Var(X_i) = n\frac{N_i}{N}\left(1 - \frac{N_i}{N}\right)\frac{N-n}{N-1} = np_i\acute{(}1-p_i)\frac{N-n}{N-1}$$

donde por p_i representamos la proporción de casos favorables a la categoría A_i.

Por lo que se refiere a la covarianza, su expresión mantiene cierta similitud con la correspondiente a la distribución multinomial:

$$\sigma_{X_i,X_j} = \frac{nN_iN_j}{N^2}\frac{N-n}{N-1} = np_ip_j\frac{N-n}{N-1} \; ; \; \forall i \neq j = 1, 2, \ldots k$$

y conduce al coeficiente de correlación lineal

$$\rho_{X_i,X_j} = -\sqrt{\frac{N_i}{N-N_i}\frac{N_j}{N-N_j}} = -\sqrt{\frac{\frac{N_i}{N}}{\frac{N-N_i}{N}}\frac{\frac{N_j}{N}}{\frac{N-N_j}{N}}} = -\sqrt{\frac{p_i}{1-p_i}\frac{p_j}{1-p_j}}$$

4.3.3. Distribución Normal Multivariante

La distribución normal ocupa un papel fundamental en el análisis de datos, al tratarse de un modelo probabilístico adecuado para la descripción de numerosas magnitudes. En el caso de que estudiemos simultáneamente varias variables aleatorias, la idea de normalidad sigue siendo válida, extendida en este caso al ámbito multivariante.

Consideremos ahora un vector columna k-dimensional $\mathbf{x}$, para el cual suponemos que existe su vector de esperanzas $\boldsymbol{\mu}$ y su matriz de varianzas-covarianzas, que denotamos por $\boldsymbol{\Sigma}$:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{pmatrix} \; ; \; \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{21} & \sigma_2^2 & \ldots & \sigma_{2k} \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \ldots & \sigma_k^2 \end{bmatrix}$$

se dice que $\mathbf{x}$ sigue una *distribución normal multivariante* de parámetros $\boldsymbol{\mu}$ y $\boldsymbol{\Sigma}$, que representamos por $\boldsymbol{x} \approx \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, si su función de densidad viene dada por la expresión:

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{(2\pi)^{\frac{k}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left[(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]}$$

donde $|\boldsymbol{\Sigma}|$ denota el determinante de la matriz de covarianzas y $(\mathbf{x} - \boldsymbol{\mu})'$ el vector transpuesto de $(\mathbf{x} - \boldsymbol{\mu})$.

En el caso $k = 1$, se tiene: $\mathbf{x} = X_1, \boldsymbol{\mu} = \mu_1, \boldsymbol{\Sigma} = \sigma_1^2 = |\boldsymbol{\Sigma}|$, con lo que al sustituir obtenemos la función de densidad de la normal general univariante.

En el caso $k = 2$, se obtienen las siguientes expresiones para el vector de esperanzas, la matriz de covarianzas y su determinante:

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \; ; \; \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{pmatrix} \; ; \; |\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2 - \sigma_{12}^2$$

donde al calcular el determinante se tiene en cuenta que la covarianza es simétrica: $\sigma_{12} = \sigma_{21}$.

El determinante podemos expresarlo como: $|\boldsymbol{\Sigma}| = \sigma_1^2\sigma_2^2\left(1 - \frac{\sigma_{12}^2}{\sigma_1^2\sigma_2^2}\right) = \sigma_1^2\sigma_2^2(1-\rho^2)$, y la matriz inversa de $\boldsymbol{\Sigma}$ resulta:

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\Sigma}\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}$$

Sustituyendo ahora en la expresión general, se obtiene:

$$\begin{aligned} f(x_1,x_2) = & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}(x_1-\mu_1, x_2-\mu_2)\begin{pmatrix} \sigma_2^2 & -\sigma_{12} \\ -\sigma_{12} & \sigma_1^2 \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}} \\ = & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\left[(x_1-\mu_1)\sigma_2^2-(x_2-\mu_2)\sigma_{12}, -(x_1-\mu_1)\sigma_{12}+(x_2-\mu_2)\sigma_1^2\right]\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix}} \\ = & \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2}+\frac{(x_2-\mu_2)^2}{\sigma_2^2}-\frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]} \end{aligned}$$

que es la función de densidad de la normal bivariante.

La función generatriz de momentos de la normal multivariante viene dada por:

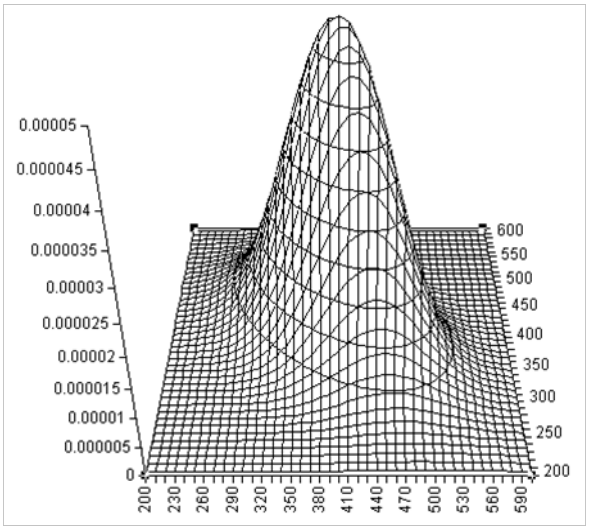
$$M_{(X_1,\ldots X_k)}(t_1,\ldots,t_k) = e^{(t_1,\ldots,t_k)\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_k \end{pmatrix} + \frac{(t_1,..,t_k)\begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{pmatrix}\begin{pmatrix} t_1 \\ \vdots \\ t_k \end{pmatrix}}{2}} = e^{\mathbf{t}'\boldsymbol{\mu} + \frac{\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}}{2}}$$

donde el último miembro está expresado en notación matricial.

Dada una variable normal multivariante, las funciones de densidad marginales coinciden con las funciones de densidad univariantes de distribuciones $\mathcal{N}(\mu_i, \sigma_i)$, con lo que se tiene: $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$.

Si las variables unidimensionales que componen la normal multivariante tienen esperanza 0 y varianza 1, las expresiones anteriores se simplifican considerablemente. A

Figura 4.4.: Modelo normal bivariante. Función de densidad



modo de ejemplo, la función de densidad de la normal bivariante resultaría:

$$f(x_1, x_2) = \frac{1}{2\pi\sqrt{1-\rho^2}} e^{-\frac{1}{2}\frac{1}{1-\rho^2}\left(x_1^2+x_2^2-2\rho x_1 x_2\right)}$$

Por lo que se refiere a la representación del modelo, para el caso bidimensional se obtendría un gráfico como el representado en la figura 4.4.

4.4. Variables aleatorias independientes

Cuando abordamos el estudio conjunto de varias variables aleatorias parece lógico interrogarnos sobre las ventajas que presenta el análisis del vector $(X_1, X_2, \ldots, X_k)$ con respecto al estudio marginal de las variables X_1, X_2,..., X_k.

La respuesta a este interrogante aparece conectada con el tipo de relaciones existentes entre las variables analizadas y resultará por tanto reveladora de las posibilidades de llevar a cabo posteriores estudios de regresión o modelizaciones causales.

Consideremos a modo de ilustración una serie de magnitudes económicas de carácter aleatorio: X_1:"Valor añadido bruto del sector industrial", X_2:"Tipo de interés a largo plazo", X_3:"Recaudación total del impuesto sobre sociedades", X_4:"Tasa de paro", planteándonos su análisis conjunto. Si el estudio del vector aleatorio (X_1, X_2, X_3, X_4) aporta información de la que no dispondríamos analizando separadamente las cuatro características (esto es, en sus distribuciones marginales) estamos detectando la existencia de interrelaciones entre las magnitudes (por ejemplo, que la evolución de los tipos de interés condiciona el VAB industrial, que éste último afecta a la tasa de paro, etc.). En esta situación, el hecho de imponer determinadas condiciones a una o varias

de las variables afectará a la distribución probabilística de las restantes, resultando por tanto de interés el planteamiento de estudios condicionados, que -como veremos en un capítulo posterior- aparece muy ligado a los análisis de regresión.

En definitiva, el objetivo último al que nos enfrentaríamos consiste en explicar o modelizar una característica a partir de las restantes, e incluso llegar a plantear modelos de ecuaciones simultáneas capaces de recoger las interrelaciones existentes entre el conjunto de magnitudes analizadas.

Si por el contrario llegásemos a la conclusión de que resulta indiferente estudiar las características de forma conjunta o marginal, estaríamos en situaciones donde el análisis conjunto no aporta nada a nuestro estudio, como consecuencia de la no existencia de interconexiones entre las variables estudiadas. Así pues, nos encontraríamos ante características aleatorias independientes, por lo cual carecería de interés cualquier análisis condicionado o intento de modelización causal.

La formalización del concepto de independencia para el caso bidimensional podría ser efectuada en los siguientes términos:

Definición 4.6. Sea (X, Y) una variable aleatoria bidimensional con función de distribución conjunta $F(x, y)$, y sean $F_X(x)$, $F_Y(y)$ las funciones de distribución marginales de X e Y respectivamente. En estas condiciones diremos que las variables X e Y son *independientes* si se verifica: $F(x, y) = F_X(x)F_Y(y)\ ,\ \forall (x, y) \in \Re^2$.

Dado que las variables aleatorias representan sucesos asociados a determinada experiencia aleatoria, el concepto de independencia de variables aparece conectado a la independencia de sucesos. Así, diremos que dos variables aleatorias X e Y son independientes si los sucesos $[a < X \leq b]$ y $[c < Y \leq d]$ son independientes para cualesquiera valores reales a, b, c, d.

En efecto, para comprobar que estos sucesos son independientes tendríamos que probar la siguiente igualdad:

$$P(a < X \leq b, c < Y \leq d) = P(a < X \leq b)P(c < Y \leq d)$$

y para ello partimos de la definición anterior de independencia. Utilizando la función de distribución conjunta podemos expresar:

$$\begin{aligned} P(a < X \leq b, c < Y \leq d) &= F(b, d) - F(b, c) - F(a, d) + F(a, c) = \\ &= F_X(b)F_Y(d) - F_X(b)F_Y(c) - F_X(a)F_Y(d) + F_X(a)F_Y(c) \end{aligned}$$

donde en la última igualdad hemos aplicado la independencia de las v.a., y si ahora sacamos factor común se tiene:

$$\begin{aligned} P(a < X \leq b, c < Y \leq d) = &\ F_X(b)[F_Y(d) - F_Y(c)] - F_X(a)[F_Y(d) - F_Y(c)] = \\ = &\ [F_X(b) - F_X(a)][F_Y(d) - F_Y(c)] = \\ = &\ P(a < X \leq b)P(c < Y \leq d) \end{aligned}$$

De modo recíproco, si se verifica la relación anterior para cualquier par de sucesos $[a < X \leq b]$, $[c < Y \leq d]$, entonces se tiene la condición de independencia enunciada. En efecto, para todo $(x, y) \in \Re^2$ bastará considerar los intervalos de tipo $(-\infty, x]$, $(-\infty, y]$ y la comprobación resulta inmediata.

$$F(x, y) = P(-\infty < X \leq x, -\infty < Y \leq y) = P(-\infty < X \leq x)P(-\infty < Y \leq y) = F_X(x)F_Y(y)$$

La condición de independencia puede también ser recogida mediante expresiones equivalentes en las que aparecen las funciones de probabilidad o de densidad, según que la variable considerada sea discreta o continua. Así, diremos que X es independiente de Y si y sólo si se verifica:

- $p(x_i, y_j) = p(x_i)p(y_j)\,,\ \forall (x_i, y_j) \in \Re^2$, para (X, Y) discreta
- $f(x, y) = f_X(x)f_Y(y)\,,\ \forall (x, y) \in \Re^2$, para (X, Y) continua

Para comprobar el resultado discreto, el proceso es similar al desarrollado anteriormente para intervalos, considerando ahora la probabilidad de un punto.

Por lo que se refiere al caso continuo, se tiene la siguiente expresión:

$$\begin{aligned} f(x,y) = & \frac{\partial^2 F(x,y)}{\partial x \partial y}\left[F_X(x)F_Y(y)\right] = \left[\frac{\partial}{\partial x}F_X(x)\right]\left[\frac{\partial}{\partial y}F_Y(y)\right] = \\ = & f_X(x)f_Y(y) \end{aligned}$$

A partir de cualquiera de las definiciones anteriores de independencia podemos demostrar que dos v.a. son independientes si y sólo si las distribuciones marginales y condicionadas coinciden. Aparece así un nuevo significado de la independencia; X es independiente de Y si su distribución no se ve afectada por los hipotéticos valores de Y que puedan haberse verificado.

En efecto, teniendo en cuenta la definición de probabilidad condicionada, para todo x_i y para todo y_j con $p(y_j) > 0$ se tiene:

$$p(x_i, y_j) = p(x_i/y_j)p(y_j)$$

y por ser independientes se cumple: $p(x_i, y_j) = p(x_i)p(y_j)$. Comparando miembro a miembro las dos ecuaciones resulta: $p(x_i) = p(x_i/y_j)$, y esto para todo x_i y para todo y_j.

[Compruébese que la implicación en sentido recíproco también es cierta]

También podemos expresar la independencia entre variables en términos de la f.g.m., y en este caso se tiene el siguiente resultado:

Proposición. *Dos variables X e Y son independientes si y sólo si $M_{(X,Y)}(t_1, t_2) = M_X(t_1)M_Y(t_2)$*

Sin duda, la independencia entre variables aleatorias es un concepto de gran trascendencia porque de esta característica se derivan propiedades de interés en el análisis conjunto de variables aleatorias.

Antes de pasar a estudiar esas propiedades, conviene comentar que, como consecuencia de su definición, el concepto de independencia entre variables aleatorias es simétrico. Este rasgo -ya comentado al estudiar la independencia entre sucesos- resulta sumamente intuitivo, ya que si la variable aleatoria X es independiente de Y se cumplirá también que Y es independiente de X.

Propiedades de la independencia de variables aleatorias

Proposición 4.1. *Dadas X e Y independientes se cumple $E(XY) = E(X)E(Y)$*

Demostración. La comprobación de esta propiedad resulta muy sencilla con sólo aplicar la condición de independencia anteriormente vista. Así, en el caso de una variable (X, Y) discreta, si el rango de valores de X es $x_1, \ldots, x_k$, y el de $Y, y_1, \ldots, y_h$, se tiene:

$$\begin{aligned} E(XY) = & \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j p(x_i, y_j) = \sum_{i=1}^{k}\sum_{j=1}^{h} x_i y_j p(x_i)p(y_j) = \\ = & \left[\sum_{i=1}^{k} x_i p(x_i)\right]\left[\sum_{j=1}^{h} y_j p(y_j)\right] = E(X)E(Y) \end{aligned}$$

[Comprobar que la propiedad se cumple en el caso continuo] □

Proposición 4.2. *Para todo par de variables X e Y independientes, se cumple $Cov(X, Y) = 0$ y $\rho_{XY} = 0$*

Demostración. La comprobación es inmediata como consecuencia de la propiedad anterior y de la expresión alternativa de la covarianza: $Cov(X, Y) = E(XY) - E(X)E(Y)$.

[Comprobar que en este caso $\rho_{XY} = 0$] □

Proposición 4.3. *Dadas X e Y independientes, se cumple $Var(X+Y) = Var(X) + Var(Y)$*

Demostración. La demostración se efectúa desarrollando la expresión de la varianza de la suma:

$$\begin{aligned} Var(X+Y) = & \; E\left[(X+Y) - E(X+Y]^2 = E\left[(X - E(X)) + (Y - E(Y))\right]^2 = \right. \\ = & \; E(X - E(X))^2 + E(Y - E(Y))^2 + 2E\left[(X - E(X))(Y - E(Y))\right] = \\ = & \; Var(X) + Var(Y) + 2Cov(X, Y) \end{aligned}$$

y aplicando ahora la propiedad anterior de covarianza nula, se tiene:

$$Var(X+Y) = \quad Var(X) + Var(Y)$$

[Comprobar que para la diferencia de v.a. independientes se obtendría de modo similar: $Var(X - Y) = Var(X) + Var(Y)$] □

Una generalización de la propiedad anterior puede establecerse tal y como sigue:

Proposición 4.4. *Dadas X e Y independientes, se cumple para cualesquiera a y b reales:*

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$$

Demostración. Para comprobar esta propiedad basta desarrollar la varianza de la combinación lineal hasta obtener:

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X, Y)$$

Puede comprobarse fácilmente que la varianza de la diferencia se corresponde con el caso particular de esta expresión cuando $a = 1$, $b = -1$ □

Proposición 4.5. *Para todo par de variables independientes X e Y y para cualesquiera f y g funciones medibles-Borel de esas variables, se cumple que las variables aleatorias $f(X)$ y $g(Y)$ son también independientes.*

Demostración. En efecto, se tiene:

$$\begin{aligned} P\{f(X) \leq x, g(Y) \leq y\} = & \quad P\left\{X \in f^{-1}(-\infty, x], Y \in g^{-1}(-\infty, y]\right\} = \\ = & \quad P\left\{X \in f^{-1}(-\infty, x]\right\} P\left\{Y \in g^{-1}(-\infty, y]\right\} = \\ = & \quad P\left\{f(X) \leq x\right\} P\left\{g(Y) \leq y\right\} \end{aligned}$$

desarrollo que puede efectuarse por ser f y g funciones medibles-Borel. □

Proposición 4.6. *Si X e Y son dos variables aleatorias independientes se verifica:*

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$

siendo $M_X(t)$, $M_Y(t)$ y $M_{X+Y}(t)$ las funciones generatrices de momentos de X, de Y y de su suma respectivamente.

Esta propiedad es un caso particular de la condición de independencia enunciada anteriormente en términos de la función generatriz de momentos k-dimensional, pues se tiene: $M_{X+Y}(t) = M_{(X,Y)}(t, t)$. Se trata de la particularización $t_1 = t$ y $t_2 = t$, que es una condición necesaria de la independencia pero no suficiente.

Demostración. En la comprobación de esta propiedad utilizamos la anterior, ya que si X e Y son v.a. independientes también lo serán sus funciones e^{tX} y e^{tY}:

$$M_{X+Y}(t) = E\left[e^{t(X+Y)}\right] = E\left(e^{tX} e^{tY}\right) = E\left(e^{tX}\right) E\left(e^{tY}\right) = M_X(t) M_Y(t)$$

□

Hemos visto que cuando dos variables son independientes entonces son incorreladas. La implicación simétrica no es cierta en general, pero sin embargo se verifica:

Proposición 4.7. *Si X e Y son variables aleatorias normales e incorreladas, entonces son independientes.*

Demostración. Supongamos $X \approx \mathcal{N}(\mu_1, \sigma_1)$ e $Y \approx \mathcal{N}(\mu_2, \sigma_2)$. La función generatriz de momentos de estas variables vienen dadas por:

$$M_X(t_1) = e^{t_1\mu_1 - \frac{1}{2}t_1^2\sigma_1^2} \; ; \; M_Y(t_2) = e^{t_2\mu_2 - \frac{1}{2}t_2^2\sigma_2^2}$$

Por otra parte, en el apartado anterior hemos visto la expresión de la f.g.m. para el modelo normal multivariante, que en el caso particular bivariante ($k = 2$) vendrá dada por:

$$M_{(X,Y)}(t_1, t_2) = e^{t_1\mu_1 + t_2\mu_2 + \frac{1}{2}(t_1^2\sigma_1^2 + t_2^2\sigma_2^2 + 2t_1t_2\sigma_{12})}$$

Si las variables son incorreladas $\sigma_{12} = 0$ y en consecuencia esta función puede expresarse como:

$$M_{(X,Y)}(t_1, t_2) = e^{t_1\mu_1 + \frac{1}{2}(t_1^2\sigma_1^2)} e^{t_2\mu_2 + \frac{1}{2}(t_2^2\sigma_2^2)} = M_X(t_1)M_Y(t_2)$$

□

4.4.1. Reproductividad

Además de las propiedades anteriores, algunos modelos probabilísticos cumplen la propiedad denominada *reproductividad* que resulta muy intuitiva y de gran interés práctico. A grandes rasgos esta propiedad garantiza que, dadas dos variables aleatorias independientes distribuidas según cierto modelo, la variable suma sigue también ese modelo probabilístico.

A modo de ilustración de esta propiedad, consideremos que un individuo apuesta al resultado "sacar 2" en 3 lanzamientos sucesivos de un dado. Como hemos estudiado en un capítulo anterior, la variable aleatoria que recoge el número de éxitos viene descrita por un modelo binomial con parámetros $n = 3$ y $p = \frac{1}{6}$.

Si consideramos ahora que un amigo del individuo anterior realiza la misma apuesta para 5 lanzamientos sucesivos de dado, ¿qué podríamos afirmar sobre el número de éxitos conseguidos por los dos amigos? La respuesta es que se tiene ahora la suma de dos variables independientes (los resultados del primer individuo no afectarán a los obtenidos por su amigo) y con $p = \frac{1}{6}$ constante, por lo cual la variable "número total de éxitos" también será binomial, en este caso con parámetros $n = 8$ y $p = \frac{1}{6}$.

Definición 4.7. Dada una familia de variables aleatorias $\mathbf{\Psi}$ se dice que ésta es *reproductiva* si y sólo si para todo par de variables aleatorias independientes $X_1, X_2 \in \mathbf{\Psi}$ se cumple $X_1, +X_2 \in \mathbf{\Psi}$.

Este enunciado genérico del requisito de reproductividad es aplicable a muchos de los modelos estudiados con anterioridad. Gracias a la reproductividad, podemos garantizar que la suma de variables aleatorias independientes distribuidas según cierto modelo sigue ese mismo modelo, presentando además parámetros relacionados con los de las variables iniciales. Dicho con otras palabras, un modelo reproductivo, permite

"reproducirse" dentro de él; es cerrado para la operación de sumar siempre que las variables sean independientes.

La reproductividad es un concepto relativo; si la familia $\mathbf{\Psi}$ depende de un vector de parámetros $\mathbf{v}$, puede ser reproductiva respecto a algunas componentes de este vector y no serlo respecto de otras.

Tal y como recoge la tabla 4.1, los modelos binomial, binomial negativo, normal, Poisson y gamma son reproductivos, pudiendo expresarse formalmente esta propiedad en los siguientes términos:

Proposición 4.8. *Si X e Y son variables aleatorias independientes que siguen distribuciones binomiales $\mathcal{B}(n_X, p)$ y $\mathcal{B}(n_Y, p)$ respectivamente, entonces la variable suma $X + Y$ también sigue una distribución binomial $\mathcal{B}(n_X + n_Y, p)$.*

Demostración. La comprobación de la reproductividad binomial se efectúa a partir de la función generatriz de momentos:

$$\begin{aligned} M_{X+Y}(t) = & \quad E\left[e^{t(X+Y)}\right] = E\left(e^{tX}e^{tY}\right) = E(e^{tX})E(e^{tY}) = \\ = & \quad \left(e^t p + q\right)^{n_X}\left(e^t p + q\right)^{n_Y} = \left(e^t p + q\right)^{n_X + n_Y} \end{aligned}$$

correspondiendo esta última expresión a la función generatriz de momentos de una variable binomial $\mathcal{B}(n_X + n_Y, p)$. □

Proposición 4.9. *Si X e Y son variables aleatorias independientes que siguen distribuciones binomiales negativas $\mathcal{BN}(r_X, p)$ y $\mathcal{BN}(r_Y, p)$ respectivamente, entonces la variable suma $X+Y$ también sigue una distribución binomial negativa $\mathcal{BN}(r_X+n_Y, p)$.*

Demostración. La comprobación se lleva a cabo de modo similar al modelo anterior, partiendo de la expresión de la función generatriz de momentos para una distribución binomial negativa $X \approx \mathcal{BN}(r, p)$:

$$M_X(t) = \left(\frac{e^t p}{1 - e^t p}\right)^r$$

[Compruébese que, aplicando esta propiedad en el caso particular $r = 1$, es posible obtener una distribución binomial negativa $X \approx \mathcal{BN}(r, p)$ como suma de r variables aleatorias independientes, cada una de ellas con distribución geométrica $\mathcal{G}(p)$] □

Proposición 4.10. *Si X e Y son variables aleatorias independientes que siguen distribuciones normales $\mathcal{N}(\mu_X, \sigma_X)$ y $\mathcal{N}(\mu_Y, \sigma_Y)$ respectivamente, entonces la variable suma $X + Y$ también sigue una distribución normal .*

[Comprobar, a partir de la función generatriz de momentos, que el modelo normal es reproductivo respecto a las características esperanza y varianza]

En el caso particular de que las distribuciones iniciales de X e Y se encuentren estandarizadas, puede verse fácilmente que la suma $X + Y$ seguirá un modelo $\mathcal{N}\left(0, \sqrt{2}\right)$.

Tabla 4.1.: Modelos y reproductividad

V.A. independientes	Variable Suma	Modelo	Reproductiviad
$X \approx \mathcal{B}(n_X, p)$ $Y \approx \mathcal{B}(n_Y, p)$	$X+Y \approx \mathcal{B}(n_X+n_Y, p)$	Binomial	respecto a n (p constante)
$X \approx \mathcal{BN}(r_X, p)$ $Y \approx \mathcal{BN}(r_Y, p)$	$X+Y \approx \mathcal{BN}(r_X+r_Y, p)$	Binomial Negativa	respecto a r (p constante)
$X \approx \mathcal{N}(\mu_X, \sigma_X)$ $Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$	$X+Y \approx \mathcal{N}\left(\mu_X+\mu_Y, \sqrt{\sigma_X^2+\sigma_Y^2}\right)$	Normal	respecto a μ y σ^2
$X \approx \mathcal{P}(\lambda_X)$ $Y \approx \mathcal{P}(\lambda_Y)$	$X+Y \approx \mathcal{P}(\lambda_X+\lambda_Y)$	Poisson	respecto a λ
$X \approx \gamma(p_X, a)$ $Y \approx \gamma(p_Y, a)$	$X+Y \approx \gamma(p_X+p_Y, a)$	Gamma	respecto a p

Proposición 4.11. *Si X e Y son variables aleatorias independientes que siguen distribuciones de Poisson $\mathcal{P}(\lambda_X)$ y $\mathcal{P}(\lambda_Y)$ respectivamente, entonces la variable suma $X+Y$ también sigue una distribución de Poisson $\mathcal{P}(\lambda_X+\lambda_Y)$.*

Si X e Y son variables aleatorias independientes que siguen distribuciones gamma $\gamma(p_X, a)$ y $\gamma(p_Y, a)$ respectivamente, entonces la variable suma $X+Y$ también sigue una distribución gamma $\gamma(p_X+p_Y, a)$.

Hasta ahora hemos abordado la independencia y sus propiedades en el caso bidimensional. Sin embargo, cabe extender este concepto al caso k-dimensional en los siguientes términos:

Definición. Se dice que las variables $X_1, X_2, \ldots, X_k$ son independientes si y sólo si se cumple:

$$F(x_1, x_2 \ldots, x_k) = \prod_{i=1}^{k} F_{X_i}(x_i),\ \forall (x_1, x_2, \ldots, x_k) \in \Re^k$$

donde $F(x_1, x_2 \ldots, x_k)$ es la función de distribución de la variable k-dimensional y $F_{X_i}(x_i)$ la función de distribución marginal de la variable unidimensional X_i, $(i = 1, 2, \ldots, k)$.

De la misma forma pueden establecerse las equivalencias de independencia k-dimensional en términos de la función de probabilidad, función de densidad o función generatriz de momentos.

Todas las propiedades anteriores pueden generalizarse al caso k-dimensional con la única dificultad derivada del cambio de notación.

A partir de este concepto de independencia pueden efectuarse las siguientes afirmaciones:

- Si $X_1, X_2, \ldots, X_k$ son variables unidimensionales e independientes, entonces toda subcolección de $X_1, X_2, \ldots, X_k$ es también una colección de variables aleatorias independientes.
- Si $\mathbf{x} = (X_1, X_2, \ldots, X_k)$ e $\mathbf{y} = (Y_1, Y_2, \ldots, Y_h)$ son dos vectores aleatorios independientes, entonces la variable aleatoria unidimensional X_i y la variable aleatoria unidimensional Y_j (componentes i-ésimo y j-ésimo de los vectores $\mathbf{x}$ e $\mathbf{y}$ respectivamente) son variables aleatorias independientes.

- Si $\mathbf{x} = (X_1, X_2, \ldots, X_k)$ e $\mathbf{y} = (Y_1, Y_2, \ldots, Y_h)$ son dos vectores aleatorios independientes, y $f(\mathbf{x})$ y $g(\mathbf{y})$ son funciones medibles-Borel de los vectores aleatorios, entonces los vectores $f(\mathbf{x}) = f(X_1, X_2, \ldots, X_k)$ y $f(\mathbf{y}) = f(Y_1, Y_2, \ldots, Y_h)$ son independientes.

4.5. Agregación de variables aleatorias

Son numerosos los ejemplos de magnitudes económicas que se obtienen a partir de variables individuales. En algunos casos, la nueva variable se obtiene como suma de las iniciales (la demanda agregada de un bien, por ejemplo) y en otras ocasiones como promedio, generalmente una media aritmética simple (el número anual medio de parados en cierto sector económico), si bien en ocasiones se calcula también una media ponderada (índices de precios o de producción).

Obsérvese que en estos casos el planteamiento es distinto al de epígrafes anteriores: hasta ahora nos ocupábamos de analizar conjuntamente k características mediante un vector k-dimensional mientras que en este apartado nuestro objetivo consiste en estudiar una característica aleatoria sobre n unidades (hogares, empresas, países,...). Obtendremos así una sucesión de variables aleatorias $X_1, X_2, \ldots, X_n$ que podemos identificar con un vector n-dimensional, a partir del cual podemos definir nuevas expresiones aleatorias.

Dadas las variables aleatorias unidimensionales $X_1, X_2, \ldots, X_n$ es posible definir a partir de las mismas las siguientes expresiones aleatorias:

- Suma o valor total: $S_n = \sum_{i=1}^{n} X_i$
- Media o valor medio: $\bar{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$
- Media ponderada $\bar{X}_W = \sum_{i=1}^{n} w_i X_i$ con pesos o ponderaciones w_i constantes, $0 < w_i < 1$ y $\sum_{i=1}^{n} w_i = 1$

Nos planteamos ahora el estudio de estas nuevas magnitudes aleatorias para lo cual analizaremos en primer lugar sus características esperanza y varianza.

Consideremos el vector aleatorio $(X_1, X_2, \ldots, X_n)$, con vector de esperanzas y matriz de varianzas-covarianzas finitos. Entonces las características de las magnitudes aleatorias suma, media y media ponderada pueden ser obtenidas a partir de las correspondientes características del vector n-dimensional.

Así, las esperanzas vendrían dadas por las siguientes expresiones:

$$E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu_i$$

$$E\left(\bar{X}_n\right) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n}\sum_{i=1}^{n}\mu_i$$

$$E\left(\bar{X}_W\right) = E\left(\sum_{i=1}^{n} w_i X_i\right) = \sum_{i=1}^{n} w_i \mu_i$$

en cuya obtención hemos aplicado únicamente las propiedades de la esperanza como operador lineal.

Por su parte, para las varianzas se obtiene:

$$Var(S_n) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) + \sum_{i\neq j}^{n} Cov(X_i, X_j) = \sum_{i=1}^{n}\sigma_i^2 + \sum_{i\neq j}^{n}\sigma_{ij}$$

expresión que, asumiendo el supuesto de independencia entre las variables podría escribirse como:

$$Var(S_n) = \sum_{i=1}^{n}\sigma_i^2$$

(obsérvese que bastaría con que las variables fuesen independientes dos a dos, ya que en ese caso se cumpliría $\sigma_{ij} = 0$, $\forall i \neq j = 1, 2, \ldots n$).

Aplicando el mismo razonamiento a las expresiones de la media simple y ponderada se obtendría, bajo el supuesto de independencia:

$$Var\left(\bar{X}_n\right) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 \; ; \; Var\left(\bar{X}_W\right) = \sum_{i=1}^{n} w_i^2\sigma_i^2$$

[Compruébense estos resultados]

Aunque en algunas ocasiones nos podría interesar tan sólo conocer las características de una magnitud agregada, en general nuestro objetivo será más amplio, ya que nos interesará obtener probabilidades referidas a esos agregados o incluso, si es posible, determinar por completo su distribución de probabilidad.

La cuantificación o aproximación de probabilidades relativas a agregados dependerá del nivel de información disponible en cada situación. Comenzando por el caso más favorable, podemos considerar como primera posibilidad aquélla en que las variables individuales X_i son independientes y distribuidas según cierto modelo probabilístico reproductivo. Bajo estos supuestos, es posible determinar con exactitud el modelo probabilístico de la suma o la media, así como los correspondientes parámetros.

A modo de ejemplo, si se consideran n variables aleatorias independientes $X_i \approx \mathcal{N}(\mu_i, \sigma_i)$ las distribuciones de los agregados suma y media serían respectivamente:

$$S_n \approx \mathcal{N}\left(\sum_{i=1}^{n}\mu_i, \sqrt{\sum_{i=1}^{n}\sigma_i^2}\right) \; ; \; \bar{X}_n \approx \mathcal{N}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_i, \frac{1}{n}\sqrt{\sum_{\mathbf{i=1}}^{\mathbf{n}}\sigma_{\mathbf{i}}^{\mathbf{2}}}\right)$$

lo que permitiría el cálculo exacto de probabilidades.

Es necesario tener presente que la propiedad de reproductividad se refiere a la suma de variables, por lo cual no permite efectuar afirmaciones sobre el modelo probabilístico de la media ponderada. Así, el hecho de que X_i se distribuya según un modelo binomial o de Poisson no permite afirmar que la expresión w_iX_i se adapte a dichos modelos.

Como caso particular, para el modelo normal sí podría garantizarse bajo el supuesto de independencia:

$$X_i \approx \mathcal{N}(\mu_i, \sigma_i) \Rightarrow w_iX_i \approx \mathcal{N}(w_i\mu_i, w_i\sigma_i) \Rightarrow \sum_{i=1}^{n} w_iX_i \approx \mathcal{N}\left(\sum_{i=1}^{n} w_i\mu_i, \sqrt{\sum_{i=1}^{n} w_i^2\sigma_i^2}\right)$$

En general no dispondremos de información exacta sobre la distribución de las variables individuales X_i, por lo cual deberemos conformarnos con aproximaciones a las probabilidades. Más concretamente, siempre que el vector n-dimensional tenga características finitas conocidas (esperanzas y matriz de varianzas-covarianzas) es posible obtener acotaciones de las probabilidades mediante la desigualdad de Chebyshev que aplicada respectivamente a las magnitudes suma, media y media ponderada da lugar a las expresiones que siguen:

$$P\left(|S_n - E(S_n)| \geq \epsilon\right) \leq \frac{Var(S_n)}{\epsilon^2}$$

$$P\left(\left|\bar{X}_n - E(\bar{X}_n)\right| \geq \epsilon\right) \leq \frac{Var(\bar{X}_n)}{\epsilon^2}$$

$$P\left(\left|\bar{X}_W - E(\bar{X}_w)\right| \geq \epsilon\right) \leq \frac{Var(\bar{X}_w)}{\epsilon^2}$$

Si sustituimos ahora las expresiones de estas magnitudes y sus correspondientes valores esperados y asumimos además la hipótesis de independencia, se tiene:

$$P\left(\left|\sum_{i=1}^{n} X_i - \sum_{i=1}^{n}\mu_i\right| \geq \epsilon\right) \leq \frac{\sum_{i=1}^{n}\sigma_i^2}{\epsilon^2}$$

$$P\left(\left|\frac{\sum_{i=1}^{n} X_i}{n} - \frac{\sum_{i=1}^{n}\mu_i}{n}\right| \geq \epsilon\right) \leq \frac{\sum_{i=1}^{n}\sigma_i^2}{n\epsilon^2}$$

$$P\left(\left|\sum_{i=1}^{n} w_i X_i - \sum_{i=1}^{n} w_i \mu_i\right| \geq \epsilon\right) \leq \frac{\sum_{i=1}^{n} w_i^2 \sigma_i^2}{\epsilon^2}$$

Tal y como hemos visto al enunciar la desigualdad de Chebyshev para una variable individual, a partir de las desigualdades anteriores es posible obtener formulaciones alternativas donde la cota ϵ sea proporcional a la desviación estándar y/o pasando a los complementarios.

Un caso particular de interés sería que las variables aleatorias $X_1, X_2, \ldots, X_n$ fueran independientes e identicamente distribuidas. Se obtendría entonces para cualquier i: $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$, características que conducen a las siguientes expresiones para la esperanza, la varianza y la cota de Chebyshev de las magnitudes suma, la media simple y ponderada:

Magnitud	Esperanza	Varianza	Acotación Chebyshev
Suma	$E(S_n) = n\mu$	$Var(S_n) = n\sigma^2$	$P\left(\lvert S_n - n\mu\rvert \geq \epsilon\right) \leq \frac{n\sigma^2}{\epsilon^2}$
Media	$\mathrm{E}\big(\bar{X}_n\big) = \mu$	$Var(\bar{X}_n) = \dfrac{\sigma^2}{n}$	$P\left(\big\lvert \bar{X}_n - \mu\big\rvert \geq \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}$
Media ponderada	$E(X_W) = \mu$	$Var\left(\bar{X}_W\right) = \sigma^2 \sum_{i=1}^{n} w_i^2$	$P\left(\big\lvert \bar{X}_W - \mu\big\rvert \geq \epsilon\right) \leq \frac{\sigma^2 \sum_{i=1}^{n} w_i^2}{\epsilon^2}$

Son numerosas las magnitudes económicas generadas mediante procesos aditivos como las anteriormente analizadas (sumas, promedios, medias ponderadas). Sin embargo, es también posible encontrar ciertas características originadas mediante procesos multiplicativos que, aunque menos frecuentes, resultan interesantes en el ámbito económico.

Así, en los modelos de mercados empresariales se asume en ocasiones la presencia de múltiples factores cuya interrelación origina un efecto final. Se trataría por tanto de expresiones multiplicativas del tipo , en las que el efecto de cada variable interactúa con los restantes factores.

Este tipo de esquema de composición fue investigado por autores como Gibrat (1931), quien en sus estudios sobre ingresos suponía que los valores de estas variables se hallan afectados por gran cantidad de factores aleatorios independientes, de varianza finita, que operan de modo multiplicativo y no aditivo. Esta propiedad se conoce como "ley del efecto proporcional" y bajo dicha ley la aplicación del teorema central del límite a los logaritmos de los elementos aleatorios conduce, en el límite, a una distribución logarítmo normal.

Sean n variables aleatorias unidimensionales $X_1, X_2, \ldots, X_n$, independientes con esperanza μ_i y varianza σ_i^2 para $i = 1, \ldots, n$. Entonces se tiene:

$$E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i) = \prod_{i=1}^{n} \mu_i$$

$$Var\left(\prod_{i=1}^{n} X_i\right) = E\left(\prod_{i=1}^{n} X_i^2\right) - \prod_{i=1}^{n} E^2(X_i) = \prod_{i=1}^{n} E\left(X_i^2\right) - \prod_{i=1}^{n} E^2(X_i)$$

si las variables tienen igual esperanza e igual varianza marginal, resulta:

$$E\left(\prod_{i=1}^{n} X_i\right) = \mu^n \; ; \; Var\left(\prod_{i=1}^{n} X_i\right) = \alpha_2^n - \mu^{2n}$$

4.6. Teoremas límites

Los teoremas límites son considerados como los resultados teóricos más trascendentales de la teoría de la probabilidad. Bajo este epígrafe se incluyen dos tipos distintos de resultados: las leyes de los grandes números y el teorema central del límite.

Las *leyes de los grandes números* hacen referencia a las condiciones bajo las cuales la media de una sucesión de variables aleatorias converge en algún sentido a la media poblacional. La importancia de estas leyes se debe a que justifican teóricamente el concepto frecuencialista de probabilidad y son de aplicación generalizada en inferencia estadística, cuando estudiamos muestras de tamaño elevado.

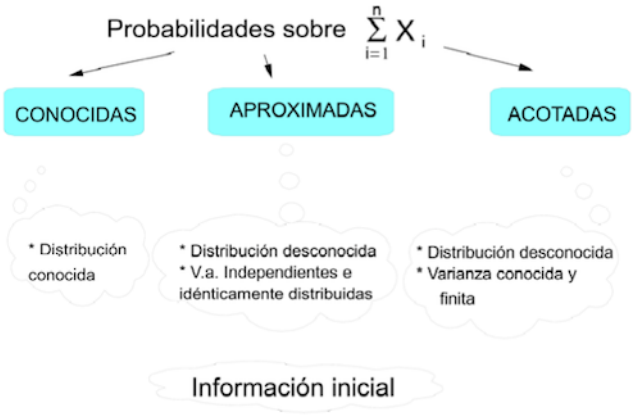
El *teorema central del límite*, por su parte, va referido a las condiciones bajo las cuales la suma de un número elevado de variables aleatorias tiene una distribución de probabilidad que es aproximadamente normal.

Los teoremas límites suponen una nueva etapa en nuestra aproximación a la distribución de los agregados económicos. Tal y como ilustra la figura 4.5, las situaciones estudiadas hasta el momento se corresponderían con dos casos extremos: el primero sería un conocimiento perfecto de la distribución (gracias, por ejemplo, a la aplicación de la reproductividad a cada una de las variables aleatorias que intervienen en el agregado) mientras que el extremo opuesto se correspondería con situaciones donde sólo es aplicable la acotación de probabilidades mediante la desigualdad de Chebyshev.

Como su propia denominación indica, los teoremas límites permiten establecer conclusiones referidas a los comportamientos asintóticos de sucesiones de variables aleatorias, por lo cual resulta interesante concretar el concepto de convergencia con el que estamos trabajando.

Cuando consideramos una sucesión numérica $\{x_n\}$ convergente a un valor x_0, el concepto de límite es unívoco. Pero cuando consideramos una sucesión de v.a. $\{X_n\}$ cada elemento de la sucesión presenta cierta aleatoriedad y converge a otra variable X_0 que también es aleatoria, por lo cual el concepto de convergencia admitirá diversos planteamientos según dónde pongamos el énfasis. Así podemos considerar la incertidumbre asociada a la convergencia numérica (Plim: probabilidad del límite) o bien el límite de las discrepancias aleatorias (LimP: límite de la probabilidad); también podemos considerar las convergencias de los modelos de probabilidad de X_n al de X_0 o considerar la convergencia en promedio de las desviaciones de cualquier orden.

Figura 4.5.: Probabilidades de Agregados



Las afirmaciones más fuertes que podemos llegar a efectuar responden al concepto de *convergencia fuerte o casi-segura* (c.s.) que se define en los siguientes términos:

Definición 4.8. Se dice que la sucesión $\{X_n\}$ converge a X *casi-seguro*, lo que representamos por $X_n \xrightarrow{c.s.} X_o$, si y sólo si $P(\lim_{n\to\infty} X_n = X) = 1$ o bien $\forall \epsilon > 0$, $\exists n_0$ tal que $n > n_0$, entonces:

$$P\left(|X_n - X| > \epsilon\right) = P\left(\{w \in E / |X_n(x) - X(w)| > \epsilon\}\right) = 0$$

Por su parte, la convergencia en probabilidad (P) -denominada habitualmente débil en contraposición a la anterior- se define como sigue:

Definición 4.9. Se dice que la sucesión $\{X_n\}$ converge a X en *probabilidad*, $X_n \xrightarrow{p} X_o$, si y sólo si:

$$\forall \epsilon > 0, \lim_{n\to\infty} P(|X_n - X| \geq \epsilon) = 0$$

o bien: $\forall \delta > 0$ y $\forall \epsilon > 0$, $\exists n_0$ tal que $n > n_0$, entonces:

$$P\left(|X_n - X| \geq \epsilon\right) = P\left(\{w \in E / |X_n(w) - X(w) \geq \epsilon) < \delta\right.$$

En el primer tipo de convergencia estamos garantizando que en el límite ambas variables coinciden salvo a lo sumo en un conjunto de probabilidad nula; intuitivamente, para un n suficientemente grande, la probabilidad de que X_n diste de de la variable límite más de cierto número ϵ es nula, esto es, X_n coincide casi-seguro con X. Con otras palabras, la convergencia casi-segura (que también se puede denominar convergencia con probabilidad uno) nos indica que para casi todos los resultados elementales $(w \in E)$ se verifica:

$$\lim_{n\to\infty} X_n(w) = X(w)$$

si denotamos por $E' \subset E$ el conjunto de resultados para los que se verifica el límite anterior, se tiene que $P(E') = 1$; el complementario podría no ser vacío pero su probabilidad es nula.

En el segundo tipo de convergencia garantizamos que el límite de la probabilidad de los conjuntos de discrepancia es nulo; de nuevo de forma intuitiva, fijada una constante δ arbitraria, podemos

encontrar un n suficientemente grande, tal que la probabilidad de que X_n diste de la variable límite más de cierto número ϵ es menor que la constante prefijada.

Otra forma de expresar esta convergencia sería:

$$\forall \epsilon > 0, \ \lim_{n\to\infty} P\left(|X_n - X| < \epsilon\right) = 1$$

lo cual quiere decir que si denotamos por $E'_n \subset E$ el conjunto de resultados donde $|X_n(w) - X(w)| < \epsilon$, se tiene:

$$\lim_{n\to\infty} P(E_n) = 1$$

La *convergencia en promedio* podemos enunciarla como sigue:

Definición 4.10. Dada una sucesión de v.a. $\{X_n\}$ se dice que *converge en media r-ésima* a la variable X si:

$$\lim_{n\to\infty} E\left[|X_n - X|^r\right] = 0$$

Los dos valores de r más utilizados son 1 y 2. Cuando $r = 1$, decimos que la convergencia es en media y cuando $r = 2$ se denomina convergencia en media cuadrática.

Proposición 4.12. *Si la sucesión* X_n converge en media cuadrática *a* X *entonces también converge en probabilidad.*

Demostración. En efecto, teniendo en cuenta la desigualdad de Chebyshev podemos expresar:

$$P\left(|X_n - X| \geq \epsilon\right) \leq \frac{E(X_n - X)^2}{\epsilon^2}$$

por tanto:

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) \leq \lim_{n\to\infty} \frac{E(X_n - X)^2}{\epsilon^2}$$

Si se verifica la convergencia cuadrática el segundo miembro es cero; entonces se tiene:

$$\lim_{n\to\infty} P\left(|X_n - X| \geq \epsilon\right) \leq 0$$

y como la probabilidad no puede ser negativa ese límite tiene que ser nulo y por tanto $\{X_n\}$ converge en probabilidad a X.

□

En principio no podemos establecer implicaciones entre la convergencia en media cuadrática y la convergencia casi-segura salvo que añadamos alguna hipótesis adicional.

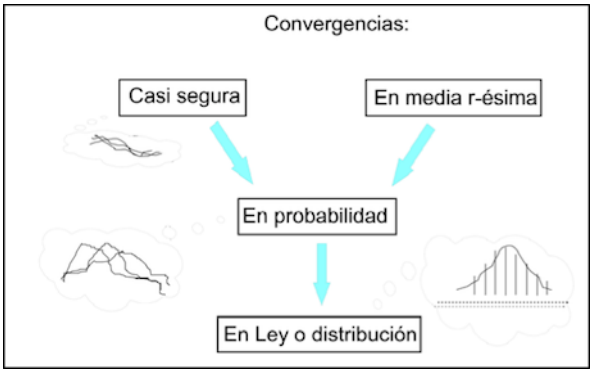
Por último la convergencia entre los modelos de probabilidad, que denominamos en *ley* (L) o *distribución* y representamos $X_n \xrightarrow{L} X$ podemos expresarla como sigue:

Definición 4.11. Se dice que una sucesión de v.a. $\{X_n\}$, cuyas funciones de distribución representamos por F_n, *converge en ley o distribución* a otra v.a. X, con f.d. F, si:

$$\lim_{n\to\infty} F_n(x) = F(x)\,, \forall x \in \Re$$

donde F es continua.

Figura 4.6.: Relación entre convergencias



La convergencia en distribución puede expresarse en términos de la función generatriz de momentos (el criterio habitualmente utilizado para analizar esta convergencia). La convergencia en probabilidad implica la convergencia en ley.

Ya hemos estudiado algunos modelos entre los cuales podemos establecer una convergencia en ley. Así, el modelo binomial converge al modelo de Poisson. Tendríamos que demostrar que la función de distribución binomial converge a la función de distribución de Poisson. Por ser ambas distribuciones discretas, para todo $x \in \Re$, sus f.d. constan de los mismos sumandos $\sum_{x_i \leq x} p(x_i)$ y ya hemos demostrado al estudiar estas distribuciones la convergencia de cada sumando de la binomial a la de Poisson, con lo cual se verifica la convergencia enunciada.

De modo similar, la distribución hipergeométrica converge en ley a la distribución binomial.

A modo de síntesis, en la figura 4.6 recogemos la relación entre los distintos tipos de convergencia:

4.6.1. Leyes de los grandes números

Las leyes de los grandes números resultan de gran interés, ya que justifican la concepción frecuentista de la probabilidad y avalan la utilización de la media muestral como aproximación del valor esperado de una población.

Teorema 4.1. *Sea $\{X_n\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas (i.i.d.), con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$. Si definimos la variable media*

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

entonces se cumple: $X_n \xrightarrow{P} \mu$.

Esto es, para cualquier $\epsilon > 0$ se cumple:

$$\lim_{n \to \infty} P\left(\left|\bar{X}_n - \mu\right| \geq \epsilon\right) = 0$$

Este enunciado se denomina habitualmente *ley débil de los grandes números*, dado que se trata de una convergencia débil o en probabilidad. (En el enunciado anterior puede sustituirse por su equivalente).

Demostración. La comprobación de este enunciado puede llevarse a cabo a partir de la desigualdad de Chebyshev, asumiendo que las variables tienen una varianza finita σ^2. En ese caso, la aplicación de la desigualdad de Chebyshev a la media proporciona la expresión:

$$P\left(\left|\frac{X_1+X_2+\cdots+X_n}{n}-\mu\right|\geq\epsilon\right)\leq\frac{\sigma^2}{n\epsilon^2}$$

a partir de la cual, con sólo tomar límites, se llega al resultado enunciado.

□

La primera demostración de la ley de los grandes números aparece recogida en la obra Ars Conjectandi de Jacob Bernoulli (1832). Este autor demostró la ley para el caso particular de variables dicotómicas:

Corolario. *Supongamos que se realizan n pruebas independientes de un experimento aleatorio en el que se observa el suceso A. Si denotamos por $f(A)$ la frecuencia relativa de este suceso y por $p=P(A)$ su probabilidad, que se asume constante a lo largo de las n pruebas, se cumple: $f(A)\xrightarrow{P}p$; esto es, para cualquier $\epsilon>0$:*

$$\lim_{n\to\infty}P\left(|f_n(A)-p|\geq\epsilon\right)=0$$

El enunciado de esta propiedad es equivalente a considerar una sucesión $\{X_n\}$ de pruebas independientes de Bernoulli, con probabilidad de éxito constante. La suma de estas variables indica el número de éxitos en las n pruebas y si calculamos la media reflejaremos la frecuencia relativa del éxito o suceso A:

$$\bar{X}_n=\frac{X_1+X_2+\cdots+X_n}{n}=f(A)$$

y por otra parte $E(X_i)=\mu=p$. Así pues, la expresión de la ley débil de los grandes números nos conduce al enunciado de esta propiedad.

Dado que dicha prueba es anterior a la desigualdad de Chebyshev, Bernoulli necesitó una metodología muy ingeniosa para llegar a su demostración de la ley.

El enunciado anterior puede ser generalizado al caso en que no se verifique la igualdad de esperanzas y varianzas.

Teorema 4.2. *Sea $\{X_n\}$ una sucesión de variables aleatorias independientes con $E(X_i)=\mu_i$ y $Var(X_i)=\sigma_i^2$. Si definimos la media de estas variables*

$$\bar{X}_n=\frac{\sum_{i=1}^{n}X_i}{n}\ y\ \mu=\frac{\sum_{i=1}^{n}\mu_i}{n}$$

entonces se cumple $\bar{X}_n\xrightarrow{P}\mu$.

Demostración. La comprobación de este enunciado puede ser efectuada aplicando la desigualdad de Chebyshev a la sucesión de medias, teniendo en cuenta que $E\left(\bar{X}_n\right)=\mu$ y

$$Var\left(\bar{X}_n\right)=\frac{1}{n^2}\left[Var(X_1)+Var(X_2)+\cdots+Var(X_n)\right]=\frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2\leq\frac{\sigma^2}{n}$$

donde $\sigma^2 = \text{máx}_i\{\sigma_i^2\}$.

□

En los enunciados anteriores hay una circunstancia que nos puede llamar la atención, y es el hecho de que para asegurar la convergencia de la media debamos establecer una hipótesis sobre un momento de orden superior (varianza). En este sentido, una demostración general de la ley débil de los grandes números fue establecida por el matemático ruso Khintchine (1929). Además de él, numerosos autores contribuyeron a la generalización de estas leyes, entre ellos Laplace, Chebyshev, Kolmogorov, Levy, Cramer, Gnedenko y Feller.

Además de la ley débil enunciada, que hace referencia a la convergencia en probabilidad, existen leyes fuertes de los grandes números, referidas a la convergencia casi-segura. Entre ellas, el enunciado más simple es el siguiente:

Teorema 4.3. *Sea $\{X_n\}$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, con la misma esperanza y varianza μ y σ^2 respectivamente, finitas. Entonces se verifica: $\bar{X}_n \xrightarrow{c.s.} \mu$.*

Este enunciado fue generalizado por Kolmogorov en dos sentidos: para el caso de sumas infinitas y para la convergencia a cualquier constante C.

4.6.2. Teorema central del límite

Como su propio nombre indica, el *Teorema central del límite* (TCL) ocupa un papel central en estadística. A grandes rasgos, este postulado garantiza que la suma de un número elevado de variables aleatorias independientes presenta una distribución aproximadamente normal; por tanto su aportación es doble: en primer lugar, permite el cálculo aproximado de probabilidades para tamaños elevados de muestra y además de ello proporciona una explicación a la generalidad con la que aparecen distribuciones campaniformes -aproximadamente normales- en los estudios empíricos.

Una de las formulaciones más sencillas del teorema central del límite es la versión de Levy-Lindeberg, que puede ser expresada en los siguientes términos:

Teorema 4.4. *Sea $\{X_n\}$ una sucesión de n variables aleatorias independientes e idénticamente distribuidas, con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ finitas y consideremos la suma de estas variables*

$$S_n = \sum_{i=1}^{n} X_i$$

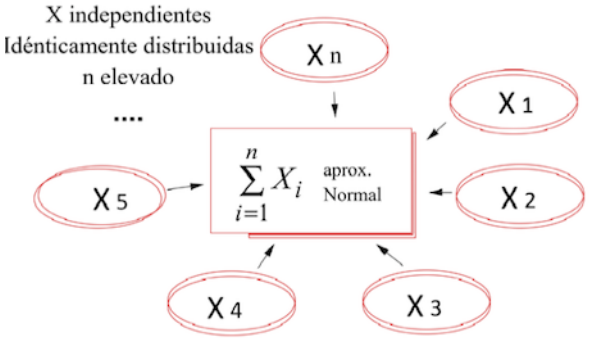
Entonces se cumple:

$$S_n \xrightarrow{L} \mathcal{N}\left(n\mu, \sigma\sqrt{n}\right)$$

o equivalentemente

$$\frac{S_n - n\mu}{\sigma\sqrt{n}} \xrightarrow{L} \mathcal{N}(0,1)$$

Figura 4.7.: Interpretación TCL



En términos generales, este teorema garantiza que la variable aleatoria definida como suma o como media de una sucesión de variables independientes e idénticamente distribuidas $X_1, X_2, \ldots, X_n$ presenta una forma que se aproxima al modelo normal a medida que el tamaño de muestra aumenta. En general, dicha aproximación se considera válida para tamaños muestrales superiores a $n = 30$.

La interpretación del teorema central del límite, que aparece ilustrada por la figura 4.7, puede ser efectuada en los siguientes términos: si se produce una actuación conjunta de numerosas causas individuales independientes entre sí, con distribución idéntica y escaso peso cada una de ellas, entonces el efecto total de estas causas es aproximadamente normal.

El teorema central del límite puede también ser aplicado a la media de las observaciones $\bar{X}_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, con el siguiente enunciado:

Teorema 4.5. *Dada una sucesión $\{X_n\}$ de v.a. independientes e idénticamente distribuidas, con $E(X_i) = \mu$ y $Var(X_i) = \sigma^2$ finitas, entonces se cumple:*

$$\bar{X}_n \xrightarrow{L} \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

o equivalentemente

$$\frac{\bar{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{L} \mathcal{N}(0,1)$$

considerándose estas aproximaciones válidas para $n > 30$.

La demostración de este teorema, que no vamos a realizar, puede ser efectuada en términos de la función generatriz de momentos.

La primera versión del TCL fue establecida por Abraham De Moivre (1667-1754) para variables de Bernoulli.

Teorema 4.6. *Si $\{X_n\}$ es una sucesión de v.a. de Bernoulli de parámetro p e independientes, se verifica:*

$$\frac{\sum_{i=1}^{n} X_i - np}{\sqrt{npq}} \xrightarrow{L} \mathcal{N}(0,1)$$

Demostración. La comprobación de este resultado es inmediata a partir del enunciado del TCL anterior, pues bastaría tener en cuenta que

$$E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = np \text{ y } Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = npq$$

□

Corolario 4.1. *Como conclusión de este resultado se tiene que la distribución binomial $\mathcal{B}(n,p)$ converge a la normal $\mathcal{N}\left(np, \sqrt{npq}\right)$.*

Demostración. Bastaría tener en cuenta que la suma de n v.a. independientes de Bernoulli de parámetro p es una binomial, $\mathcal{B}(n,p)$. En efecto, sean $X_1, X_2, \ldots, X_n$, v.a. independientes $\mathcal{B}(p)$, entonces la f.g.m de cada una de ellas será: $M_{X_i}(t) = \left(e^t p + q\right)$.

La f.g.m. de la suma será:

$$M_{\sum_{i=1}^{n} X_i}(t) = \prod_{i=1}^{n} M_{X_i}(t) = \left(M_{X_i}(t)\right)^n = \left(e^t p + q\right)^n$$

que es la f.g.m. de una binomial $\mathcal{B}(n,p)$.

Por tanto, en el enunciado de De Moivre sustituyendo la suma de Bernoulli por la correspondiente binomial, obtendremos la propiedad enunciada.

□

Laplace generalizó el enunciado dado por De Moivre, para el caso de variables discretas y simétricas. Paul Levy lo extendió a v.a. independientes idénticamente distribuidas con momentos de segundo orden finitos.

Corolario 4.2. *Para valores elevados de λ la distribución de Poisson $\mathcal{P}(\lambda)$ converge a la normal .*

Demostración. De modo intuitivo, la justificación podría efectuarse en los siguientes términos: consideremos n v.a. independientes, $X_1, X_2, \ldots, X_n$, distribuidas según un modelo de Poisson de parámetro $\frac{\lambda}{n}$. Se verifica entonces $E(X_i) = \frac{\lambda}{n}$ y $Var(X_i) = \frac{\lambda}{n}$.

Al ser la distribución de Poisson reproductiva y las variables $X_1, X_2, \ldots, X_n$ independientes, se tiene para la suma:

$$S_n \approx \mathcal{P}\left(\frac{\lambda}{n} + \frac{\lambda}{n} + \cdots + \frac{\lambda}{n}\right) = \mathcal{P}(\lambda)$$

Por otra parte, se trata de una sucesión de v.a. i.i.d. con esperanza y varianza finitas; por tanto en virtud del TCL su suma S_n converge a una normal con media $E(S_n)$ y varianza $Var(S_n)$, características que vienen dadas por las expresiones:

$$E(S_n) = E\left(\sum_{i=1}^{n} X_i\right) = n\frac{\lambda}{n} = \lambda \; ; \; Var(S_n) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = n\frac{\lambda}{n} = \lambda$$

Así pues, se cumple para tamaños suficientemente elevados de n:

$$S_n \xrightarrow{L} \mathcal{N}\left(\lambda, \sqrt{\lambda}\right)$$

con lo cual la distribución de Poisson puede ser aproximada por el modelo normal, siendo adecuada dicha aproximación para valores elevados del parámetro λ .

□

La demostración se llevaría a cabo de forma análoga para cualquier otra distribución que sea reproductiva.

Lvóvich Chebyshev (1821-1894) y Andrei A. Markov (1856-1922), ambos autores pertenecientes a la escuela rusa de San Petersburgo, establecieron la validez del teorema central del límite en condiciones más generales, como la alteración del supuesto de independencia. M.A. Liapunov (1857-1918) estableció una condición suficiente para que se produzca una convergencia a la normal aunque las variables X_i no tengan idéntica distribución.

Lindeberg dio en 1922 unas condiciones suficientes para la validez del teorema y William Feller en 1935 demostró en cierto sentido la necesidad de tales condiciones.

Como hemos visto, el teorema central del límite resulta aplicable a las magnitudes originadas mediante agregación de variables individuales. Este sería por ejemplo el caso de los errores de distinta índole, que se superponen hasta dar lugar al error total observado en un estudio, para el cual se obtienen habitualmente representaciones campaniformes aproximadamente normales.

Cabe por último señalar que, en el caso de que el teorema central del límite sea aplicado a variables aleatorias discretas, se plantean dudas sobre el valor puntual a partir del cual debe ser efectuada la aproximación. Para solucionar este inconveniente se introduce la *corrección de continuidad*, consistente en calcular la probabilidad sobre el valor medio de dos observaciones consecutivas de la variable.

En efecto, la aplicación del TCL podrá dar lugar a resultados distintos según el punto a partir del cual se cuantifique la probabilidad. Así, la probabilidad $P(X > x_i)$ podría ser también planteada como $P(X \geq x_{i+1})$, expresión que conduciría a un resultado inferior al anterior.

Para resolver este problema, se introduce la corrección de continuidad, consistente en enunciar las dos alternativas para la probabilidad buscada, adoptando finalmente como aproximación la correspondiente al valor intermedio,

$$x^* = \frac{x_i + x_{i-1}}{2}$$

que no pertenecerá al recorrido de la variable discreta.

Se tendría entonces como aproximación $P(X \geq x_i^*)$ tanto si la probabilidad inicialmente enunciada es $P(X > x_i)$ como si ésta fuese $P(X \geq x_{i+1})$. A modo de

ilustración, supongamos que deseamos conocer la probabilidad de que durante el próximo trimestre llueva más de 15 días. Si conocemos la probabilidad de que un día cualquiera llueva ($p = 0,2$, por ejemplo) tendríamos inicialmente una distribución binomial $X \approx \mathcal{B}(90,\ 0,2)$ que como hemos visto anteriormente puede ser aproximada por un modelo normal con esperanza $np = 18$ y varianza $npq = 14,4$.

Si calculamos la probabilidad pedida como $P(X > 15)$ se obtendría mediante la aproximación normal $P(X > 15) = P(Z > -0,7906) = 0,7854$ [compruébese]. Sin embargo, dado que el número de días con lluvia es una v.a. discreta, esta probabilidad debería coincidir con $P(X \geq 16)$, hecho que no se verifica si utilizamos la aproximación normal sobre este valor [compruébese que $P(X \geq 16) = 0,7009$]. Para solucionar este tipo de situaciones, la corrección de continuidad nos llevaría a aproximar ambas probabilidades como: $P(X \geq 15,5) = P(Z \geq -0,6588) = 0,745$.

Parte II.

Inferencia estadística

5. Muestras y estimadores

Las variables económicas no pueden ser conocidas de forma determinista ni siquiera de forma probabilística, ya que el tiempo, el espacio y otros factores contribuyen a que no tengan carácter estático y como consecuencia nuestros análisis se desarrollarán en un contexto dinámico e incierto. Así pues, las poblaciones que investigamos serán cambiantes, y generalmente nos encontraremos con un cierto desconocimiento sobre algunos parámetros y características de la población, o incluso sobre su propia estructura.

En la práctica nuestras decisiones deben ser adoptadas a partir de información parcial sobre la población investigada. Así, los estudios de mercado se basan en muestras de clientes, las pruebas de control de calidad examinan muestras del producto analizado, e incluso algunas de las principales estadísticas oficiales como el Indice de Precios de Consumo (IPC) o la tasa de paro se basan en la información procedente de encuestas muestrales: la Encuesta de Presupuestos Familiares (EPF) y la Encuesta de Población Activa (EPA), ambas realizadas por el Instituto Nacional de Estadística (INE).

Teniendo en cuenta la importancia de las investigaciones muestrales en el ámbito socioeconómico, en este capítulo recogemos una breve introducción a la selección de muestras y sus errores, para posteriormente centrarnos en el estudio de los estimadores, las propiedades básicas que deben cumplir y los principales métodos para su obtención.

5.1. Estudios muestrales. Conceptos básicos

5.1.1. Población

En estadística descriptiva el concepto de población es entendido como el conjunto de personas o cosas a las que va referida una investigación.

En este sentido estamos identificando los conceptos de *población* y *universo*. Sobre el colectivo de personas o cosas investigadas, habitualmente observamos una variable que puede tomar un conjunto de valores con una distribución de probabilidad determinada.

En el ámbito de la inferencia estadística la población suele identificarse no con el universo sino con la propia variable aleatoria, y así es habitual hablar de "una población X que se distribuye normalmente", "la esperanza poblacional", "la distribución de la población", ... afirmaciones que en realidad irían inicialmente referidas a la variable aleatoria pero que resultará más cómodo plantear en términos de la población.

En inferencia estadística esta segunda acepción suele ser la habitual y en algunas ocasiones se mezclan ambas terminologías. Así en poblaciones finitas hablamos indistintamente de una población

E (formada por unidades: personas o cosas) sobre la que se diseña el plan de muestreo y de una población X (v.a.) sobre la que estimamos el total o la media.

Cuando observamos varias variables partimos de una población identificada como universo. Sin embargo, cuando identificamos la población como v.a. podemos encontrarnos con una población k-dimensional o con k poblaciones unidimensionales.

Se denomina *tamaño poblacional* al número de elementos u observaciones que integran una población.

El tamaño de una población puede ser finito o infinito. En la mayor parte de las aplicaciones reales las poblaciones investigadas son finitas, como el número de lectores de un periódico o el número de automóviles que sale de una factoría. En otras ocasiones el tamaño poblacional puede ser infinito, como sucede por ejemplo para la población generada por los lanzamientos sucesivos de un dado o el conjunto de números reales pertenecientes al intervalo $[0, 1]$. A menudo, las poblaciones de interés, aunque finitas, son de tamaño tan elevado que en teoría se asumen como infinitas (por ejemplo, la población mundial, el parque móvil europeo, ...).

¿Sería preferible una población finita o una infinita? Hemos hecho esta pregunta durante muchos cursos a nuestros alumnos y la respuesta siempre es "finita": los números finitos nos parecen siempre más tangibles y conocidos (podemos alcanzarlos) y por tanto de más fácil manejo. El infinito es un campo cuyo tratamiento y comprensión requiere imaginación y que, en cualquier caso, se percibe como lejano. Sin embargo, ya hemos comentado que los modelos son idealizaciones matemáticas y en ellas lo infinito y lo continuo, en contra de nuestra intuición, tienen un importante papel simplificador.

La mayor parte del aparato matemático desarrollado hasta la actualidad es de carácter continuo, la matemática discreta o finita se encuentra menos desarrollada y por tanto su aplicación para resolver problemas reales (finitos) es limitada.

Por este motivo, nos interesará que las poblaciones sean infinitas y a ser posible continuas o aproximables por éstas.

La información necesaria para llevar a cabo el estudio estadístico se encuentra disponible en los elementos que componen la población (universo), a los que se denomina *unidades elementales* o *unidades primarias*. En estas unidades la información se encuentra en estado puro, completa y por tanto a partir de ellas la información se irá transmitiendo, agregando y sintetizando, de manera que cada proceso constituye un filtro donde la información va perdiendo fiabilidad.

La transmisión de la información desde los elementos de la población se realiza mediante *encuestas* que pueden ser *censales* o *muestrales*.

A pesar de que los avances informáticos nos permiten procesar volúmenes de información que hace unos años resultaban impensables, en general no será posible analizar exhaustivamente las poblaciones, como consecuencia de las limitaciones de recursos (tiempo, presupuesto, e incluso imposibilidad física cuando las poblaciones que investigamos son infinitas o perecederas). De ahí que debamos conformarnos con efectuar estudios parciales, llevando a cabo posteriormente una generalización de los resultados obtenidos.

5.1.2. Muestras

En el caso de que las poblaciones que estudiamos sean finitas -supuesto más habitual en la práctica- podría parecer en un principio que la investigación exhaustiva conlleva mayor fiabilidad que los estudios muestrales. Sin embargo, ello no es necesariamente cierto, puesto que la disminución del número de unidades investigadas permite aumentar el detalle con que éstas se analizan y en consecuencia la calidad de los resultados.

Esta ventaja, junto con el ahorro en tiempo y costes, justifica el interés que tienen en estadística las investigaciones muestrales.

Así, si deseamos llevar a cabo un análisis sobre la penetración de cierto producto en el mercado nos encontraremos con que un estudio exhaustivo de todos los puntos de venta muy probablemente desbordaría las posibilidades de cualquier empresa, debido al personal necesario, los desplazamientos del mismo, las consiguientes necesidades en cuanto a tiempo y costes ...

Estos mismos argumentos servirían para justificar la necesidad de tomar muestras en una amplia variedad de situaciones, en las que resultará recomendable limitar nuestro análisis a algunas unidades de la población investigada. De hecho, este tipo de estudio será inevitable cuando el análisis realizado afecte a las unidades investigadas, como en el caso de los procesos destructivos.

Los procesos destructivos de investigación justifican plenamente las técnicas muestrales, ya que en este tipo de estudios un análisis exhaustivo conllevaría el deterioro o destrucción de la población investigada. Ejemplos claros son las pruebas de control alimentario que incluyen degustación de productos, las experiencias científicas con elevado riesgo, las pruebas bélicas, etc.

Además de las características señaladas anteriormente, existen poblaciones en las que, por sus especiales características, se hacen más patentes las ventajas del muestreo: se trata de colectivos en los que existe homogeneidad respecto a las características investigadas. A modo de ejemplo, si se desea publicar una fe de erratas de determinada obra bastaría examinar un ejemplar, dado que la tirada sería una población perfectamente homogénea, generada como reproducción de un único original.

Algunos ejemplos extremos de poblaciones homogéneas se tienen en las pruebas clínicas (análisis de sangre, por ejemplo) o bien de cocina (temperatura o sabor de una sopa, ...), situaciones en las que una mínima parte de la población resulta suficientemente informativa del total.

Una vez analizadas sus ventajas, conviene señalar también algunos inconvenientes del muestreo. Entre ellos, el más inmediato es la posible introducción de errores asociados a la propia selección de la muestra, que no siempre es posible evitar.

Además, los estudios muestrales requieren una mayor cualificación personal, ya que aumenta considerablemente la complejidad del aparato estadístico necesario tanto

para el diseño de la muestra como para el tratamiento de la información.

Este balance de ventajas e inconvenientes de los estudios muestrales aparece sintetizado en el esquema siguiente. Su observación nos conduce a la conclusión recogida por el profesor Francisco Azorín -uno de los mayores impulsores del muestreo en nuestro país- quien afirma "las muestras en sentido amplio no sólo podrían ser importantes y en ocasiones necesarias, sino que generalmente son inevitables"[1] .

ESTUDIO CENSALES

- Imprescindibles en recuentos e investigaciones
- Exigen amplios recursos humanos y materiales
- Largo tiempo de ejecución
- Costes elevados

ESTUDIOS MUESTRALES

- Exigen trabajo especializado
- Permiten ahorro económico y de trabajo de campo
- Convenientes en poblaciones homogéneas
- Imprescindibles en procesos destructivos y poblaciones infinitas

Con el objetivo de aprovechar al máximo las ventajas del muestreo, en la actualidad este tipo de técnicas se incorporan incluso a algunas investigaciones censales. Así, el Censo de Población y Viviendas 2011 realizado por el INE, con el objetivo de ser más rápido, eficiente y económico, combina por primera vez el uso de registros administrativos con trabajo de campo, que incluye un censo de edificios y una gran encuesta muestral para conocer las características de personas y viviendas.

5.1.3. Subpoblaciones o estratos

Una vez justificada la conveniencia de limitarnos a análisis parciales de una población, podríamos preguntarnos si todo estudio parcial es muestral, es decir, si todo subconjunto de una población puede ser considerado como una muestra.

Obviamente, la respuesta es negativa, ya que una muestra estará constituida por elementos seleccionados de una población con el fin de representar a todo el colectivo. Se distingue así este concepto de otros subconjuntos poblacionales -habitualmente denominados *subpoblaciones o estratos*- integrados por elementos que presentan alguna característica en común.

La selección de subpoblaciones se lleva a cabo atendiendo a ciertos criterios, que garantizan la homogeneidad entre los elementos que integran cada subpoblación. Así, en algunas estadísticas oficiales los hogares son estratificados atendiendo a criterios socioeconómicos, los establecimientos se agrupan en subpoblaciones según el número de trabajadores, ...

[1] AZORIN, F. (1988): Curso breve de muestreo en poblaciones finitas. Curso de doctorado "Información y esquemas difusos", Universidad de Oviedo.

Si por el contrario seleccionásemos algunos hogares a partir del callejero o de la guía telefónica el resultado sería una muestra, ya que es previsible que en ese subconjunto de la población estuviesen incluidos muy distintos tipos de hogares. Esta heterogeneidad, equivalente a la que se observa en la población, es el rasgo característico de las muestras: así, un colegio podría ser considerado como una muestra de la población infantil, un hospital como una muestra de una población de enfermos o un establecimiento como una muestra de una población de trabajadores. No obstante, las muestras utilizadas en la investigación estadística suelen ser resultado de procesos de selección más complejos.

A modo de ejemplo, supongamos que deseamos realizar un estudio sobre la cuota de mercado de un producto y, una vez descartado por las razones anteriormente expuestas un estudio exhaustivo, debemos concretar el ámbito de la encuesta.

Una primera posibilidad sería realizar la encuesta sólo en una parte de la población (digamos una capital como Madrid). Sin embargo este método parece poco recomendable dado que cada ciudad presenta unos rasgos específicos (volumen de población, tipo de actividad a la que se dedican, dotaciones de servicios, ...) que la hacen distinta por ejemplo de las zonas rurales. Dichas características configuran a las ciudades como subpoblaciones, mientras que nuestro objetivo sería la extracción de muestras.

Resultaría interesante por tanto llegar a disponer de un núcleo representativo de la población, algo similar a una "micropoblación robot" cuyos rasgos serían los siguientes: un volumen moderado de habitantes, distribuidos según una pirámide poblacional similar a la de la población global, una estructura productiva equivalente a la global (en cuanto a proporción de población dedicada a cada sector productivo), la misma renta percápita e igualmente distribuida, una reproducción a escala de las ideologías, religiones, razas, etc.

Esta idea de "micropoblación robot" -que sería una fuente perfecta de información sobre la población total- resulta sin embargo demasiado ambiciosa, por lo cual en la práctica debemos contentarnos con muestras que presentan -consideradas globalmente- características similares a las del colectivo de interés. En definitiva, dado que nuestro estudio tendrá por objetivo una o varias características de la población, trataremos de que la aproximación que proporciona la muestra sea adecuada, es decir, que no se produzcan demasiadas discrepancias entre muestra y población.

5.1.4. Muestreo probabilístico

Un segundo interrogante referido al muestreo, sería si toda muestra debe necesariamente ser aleatoria o probabilística. Nuevamente la respuesta es negativa, puesto que el concepto genérico de muestra hace referencia a su finalidad (representar adecuadamente al conjunto de la población) pero no a su método de obtención, entendiendo como tal el criterio mediante el cual se procede a la elección de las unidades de la población. Este criterio permitirá distinguir entre muestras aleatorias (aquéllas seleccionadas al azar) y no aleatorias.

En un abuso del lenguaje utilizamos aquí el término "aleatorio" como sinónimo de "probabilístico". En realidad, "aleatorio" se aplica habitualmente a todo suceso que depende del azar y por tanto no puede ser conocido de antemano, mientras los términos "estocástico" o "probabilístico" indican que es posible asignar probabilidades de realización a los sucesos, esto es, cuantificar su incertidumbre.

Definición 5.1. Decimos que un proceso de selección es *probabilístico* o *aleatorio* cuando es posible asignar a cada muestra una probabilidad de ser elegida.

Cuando la muestra es aleatoria podemos asignarle una cierta función de distribución y sus correspondientes funciones de probabilidad o densidad según que la población sea discreta o continua; gracias a ello podemos establecer una distribución de probabilidad de los errores o un coeficiente de fiabilidad de los resultados asociados a estas muestras. En cambio, si la muestra no es aleatoria, las estimaciones pueden ser muy buenas, pero nunca tendremos garantías porque no es posible calcular ninguna medida de bondad asociada a la muestra. Por este motivo, desarrollaremos nuestros análisis inferenciales sobre las muestras aleatorias.

Función de distribución muestral

Nuestro objetivo central serán los procesos de muestreo aleatorio y las posibilidades inferenciales derivadas de los mismos.

Supongamos una población X y seleccionemos a partir de ella una muestra de tamaño unitario, que denotamos por X_1. Antes de realizar la selección, el valor que puede aparecer es uno cualquiera de la población y la probabilidad de que salga un valor determinado será la que dicho valor tenga en la población. Por tanto la distribución de X_1 será idéntica a la de X, de modo que, denotando por F la distribución de X y por F_{X_1} la de X_1 se tiene:

$$F_{X_1}(x) = F(x)$$

Hablamos de muestra genérica cuando ésta aún no se ha concretado en una realización, sino que se trata de una muestra potencial. En el caso anterior se trata de una muestra de tamaño uno que podemos identificar con la variable muestral X_1 y de la misma forma a F_{X_1} la denominaremos distribución de la muestra.

Supongamos ahora que se toma una muestra de tamaño dos. En la primera selección puede obtenerse un valor aleatorio, X_1 y en la segunda extracción de nuevo se puede obtener un valor aleatorio X_2; por tanto la muestra puede identificarse con una v.a. bidimensional (X_1, X_2). Utilizando la fórmula de la probabilidad condicionada, la función de distribución de la muestra en este caso será:

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2/X_1=x_1}(x_2)$$

Si la población es infinita el conocimiento de la primera unidad seleccionada no tiene influencia en la probabilidad de la segunda y lo mismo ocurriría si la población es finita y el proceso de selección conlleva la reposición de cada unidad observada. En estas condiciones las variables X_1 y X_2 son independientes y la distribución anterior puede ser simplificada:

$$F_{X_1,X_2}(x_1, x_2) = F_{X_1}(x_1)F_{X_2}(x_2)$$

Ya hemos visto que la primera componente tiene la misma distribución que X. Además, dado que consideramos que las dos extracciones son independientes, al devolver la unidad a la población para la segunda selección, la composición poblacional vuelve a ser la original y por tanto la distribución de X_2 también coincide con la de X, obteniéndose:

$$F_{X_1,X_2}(x_1, x_2) = F(x_1)F(x_2)$$

en este caso se dice que las variables muestrales son idénticamente distribuidas (i.d.).

Los supuestos de independencia e idéntica distribución (i.i.d.) son hipótesis simplificadoras del tratamiento estadístico; por tal motivo la inferencia estadística trabaja bajo estos supuestos y en los desarrollos que siguen, salvo que se especifique lo contrario, supondremos que las poblaciones son infinitas o bien que la selección se realiza con reemplazamiento.

Definición 5.2. Se denomina *muestra aleatoria simple* (m.a.s.), a aquélla que es seleccionada bajo los supuestos de independencia e idéntica distribución.

Si consideramos una muestra aleatoria simple de tamaño n, ésta puede ser identificada con una v.a. n-dimensional $(X_1, X_2, \ldots, X_n)$ cuyas componentes, bajo los supuestos asumidos, son independientes e idénticamente distribuidas. Así pues, la función de distribución de la muestra viene dada por:

$$F_{X_1,X_2,\ldots,X_n}(x_1, x_2, \ldots, x_n) = F(x_1)F(x_2)\cdots F(x_n) = \prod_{i=1}^{n} F(x_i)$$

La comprobación de esta expresión es una extensión de las desarrolladas en los párrafos precedentes. [¿Qué expresiones adoptaría la f.d. si se alterara alguna de las hipótesis anteriores?].

Puesto que la distribución de cada componente de la muestra genérica coincide con la de X, la variable $(X_1, X_2, \ldots, X_n)$ será discreta o continua según lo sea X. A partir de la f.d. anterior podemos obtener en su caso la función de probabilidad o densidad de la muestra genérica, denominada *función de verosimilitud*. Sin embargo, dado que el estudio de esta función -de suma importancia en todo el proceso inferencial- surge ligada a algún parámetro poblacional, posponemos al siguiente apartado su definición e interpretación.

Consideremos ahora una muestra particular $(x_1, x_2, \ldots, x_n)$, algunos de cuyos valores aparecerán repetidos y representemos en una tabla de frecuencias cada valor muestral x_i con su correspondiente frecuencia relativa $f(x_i)$. La aplicación que a cada valor observado le asigna su frecuencia relativa acumulada se denomina distribución de frecuencias de la muestra $F^*(x_i)$.

Es de suma importancia diferenciar entre los conceptos de muestra genérica y muestra concreta, y también entre la distribución probabilística de la muestra y su distribución de frecuencias. En los primeros casos existen las componentes de potencialidad o incertidumbre características de las variables aleatorias, mientras que en los segundos se trata de problemas descriptivos. Las diferencias son equivalentes a las que existen entre probabilidad y frecuencia o entre las características esperanza y media aritmética.

Para aclarar los conceptos anteriores, consideremos un ejemplo sencillo consistente en extraer bolas de la urna en la que hay un total de diez bolas, de las que seis son de color blanco y las cuatro restantes son negras.

El resultado de la extracción de una bola de la urna puede ser identificado con una v.a. X dicotómica (1 para el suceso "Blanco" y 0 para "Negro", por ejemplo) que vendrá caracterizada por la probabilidad $p = 0,6$.

Si de esta urna se realizan dos extracciones sucesivas con reposición, se obtiene una muestra aleatoria simple que describimos mediante la variable aleatoria (X_1, X_2), cuya distribución de probabilidad puede ser obtenida fácilmente, teniendo en cuenta las condiciones de independencia e idéntica distribución:

Observaciones muestrales	(x_1, x_2)	$p(x_1, x_2)$
(B,B)	(1,1)	0,36
(B,N)	(1,0)	0,24
(N,B)	(0,1)	0,24
(N,N)	(0,0)	0,16

Puede comprobarse que la función $p(x_1, x_2)$ es una verdadera función de probabilidad, por cumplir los requisitos de no negatividad y suma unitaria.

[Definir la variable aleatoria asociada a la extracción de tres bolas de la urna]

5.2. Errores y diseño de encuestas

Las encuestas tienen por objetivo la investigación de poblaciones para llegar a conocer ciertas características o parámetros poblacionales que denominaremos *valores verdaderos* (por ejemplo la tasa de paro, la renta media o el nivel de inflación de un país). Estos valores verdaderos serán desconocidos y nunca podrán ser cuantificados de una forma exacta, por lo cual deberemos ser conscientes de que el resultado de cualquier investigación -y especialmente las de tipo social- vendrá afectado por errores.

El estudio de la cuantificación de los errores está mucho más desarrollado en las ciencias naturales que en las ciencias sociales. Para constatar esta afirmación, basta recordar que en sus orígenes la distribución normal aparece asociada a los errores de medición en astronomía, investigaciones realizadas por Gauss a finales del siglo XVIII y principios del XIX.

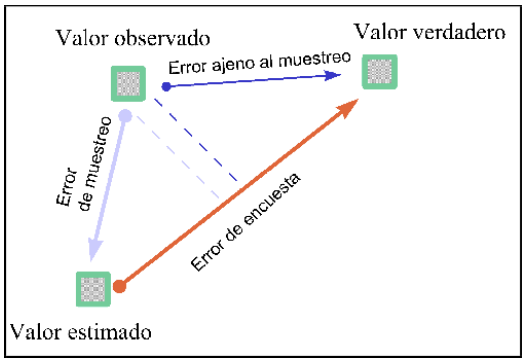
Sin embargo, en las ciencias sociales los problemas de la medición se encuentran menos desarrollados debido a la mayor dificultad que entraña en ellas la cuantificación. Ello no debe llevarnos a pensar que los errores en las ciencias sociales sean de menor cuantía, pues como recoge O. Morgenstern , en un amplio estudio sobre la exactitud de los datos económicos, en el ámbito de las ciencias sociales están presentes, al menos, todas las causas de error de las ciencias naturales.

En muchas ocasiones la necesidad de facilitar datos de carácter económico inspira cierto recelo en los agentes (sobre todo por las implicaciones de tipo jurídico o fiscal), lo que puede conducir a un falseamiento deliberado de la información. Así la falta de exactitud que pueden presentar los datos económicos facilitados por las empresas y los consumidores aconsejan una confrontación entre datos obtenidos por distintas vías de captación.

Otro tipo de dificultades son las relacionadas con la utilización de distintas fuentes, ya que con frecuencia existen discrepancias entre la información facilitada por varios organismos. La existencia de una pluralidad de observadores de una misma realidad introduce por sí misma elementos de error en la información cuantitativa porque los observadores pueden tener objetivos diferenciados (piénsese, por ejemplo, en las discrepancias sobre el paro estimado en la Encuesta de Población Activa (EPA) del INE y el paro registrado por los Servicios Públicos de Empleo).

Del mismo modo, pueden aparecer problemas de homogeneidad por parte de las unidades de observación, debido a la utilización de definiciones diferentes, a cambios en los criterios de clasificación, a desfases temporales en las magnitudes consideradas, etc.

Figura 5.1.: Errores de encuesta



5.2.1. Errores de encuesta

Aunque habitualmente nos centramos en un único error global, entendido como diferencia entre el valor verdadero y el resultado de nuestra investigación, conviene tener presente que dicho error habrá sido generado por varias fuentes, que difieren en cuanto a su importancia y sus posibilidades de control.

Aun analizando todos los elementos de la población, los errores de observación, como fallos en los instrumentos de medida, listados no actualizados, ..., nos llevarán a tomar como verdaderos ciertos valores que denominamos *valores observados*. Es evidente que el error o discrepancia entre los valores verdaderos y los observados no podrá ser cuantificado de una forma exacta aunque existan controles de calidad de las encuestas censales.

A partir de la información facilitada por una encuesta muestral, podemos obtener aproximaciones al valor verdadero que conocemos como *valor estimado*. Normalmente designamos por error de encuesta la discrepancia total existente entre las características poblacionales investigadas (valores verdaderos) y los resultados inferidos a partir de la muestra (valores estimados).

La figura 5.1 recoge las distintas fuentes de error que acompañan a una encuesta muestral y que conforman el *error de la encuesta*. Así, el hecho de seleccionar una parte de la población ya puede introducir un *error de muestreo* (¿por qué las unidades seleccionadas y no otras que nos conducirían a resultados distintos?). Si la selección muestral es aleatoria podemos conocer a priori la probabilidad que tiene la muestra de ser elegida y por tanto la probabilidad de cometer un determinado error de muestreo; es decir, nos interesan muestras probabilísticas para poder acotar el riesgo de estos errores en términos de probabilidad.

A continuación, una vez seleccionadas las unidades muestrales aparecen nuevas fuentes de error: aquéllas que se refieren a la observación de las unidades y que se denominan *errores ajenos al muestreo*, porque van asociados a la observación de las unidades con independencia de que el estudio sea muestral o poblacional. En esta categoría recogemos errores de tipo diverso como posibles deficiencias del marco o el

cuestionario, influencias del agente encuestador, ... que estudiaremos con detalle en un capítulo posterior y que habitualmente introducen sesgos en las conclusiones de nuestros estudios.

Decimos que un diseño (o la estimación derivada del mismo) es sesgado cuando las desviaciones o errores que origina tienen carácter sistemático. El sesgo puede ser debido a diversos factores, como el diseño de la encuesta, los instrumentos de medida o las respuestas y sus consecuencias son habitualmente subestimaciones o sobreestimaciones de las características investigadas.

5.2.2. Acuracidad y precisión

Como consecuencia de su carácter sistemático, los errores ajenos al muestreo resultan más fácilmente identificables que los muestrales aunque su control sólo será posible dentro de ciertos límites. Por el contrario, la aleatoriedad de los errores muestrales hace que sea necesaria para su cuantificación una sofisticada herramienta matemática.

La búsqueda de la mayor calidad posible en nuestros resultados aconseja minimizar las desviaciones entre valores verdaderos y estimados, esto es, el *error de encuesta*, con sus dos componentes. Según dónde se sitúe el énfasis aparecen los conceptos de *precisión y exactitud o acuracidad*.

El requisito de precisión exige limitar el error debido al muestreo, esto es, las oscilaciones de carácter aleatorio. Por su parte, la idea de exactitud o acuracidad va referida a todo el error de encuesta, por lo cual resulta más ambiciosa (además del requisito de precisión, exige un control de los errores ajenos al muestreo).

A modo de ilustración, pensemos en una balanza que en su posición normal se encuentra inclinada, de modo que pesa siempre algunos gramos de más.

Esto significaría que el instrumento de peso que estamos utilizando es sesgado. Sin embargo, puede ser muy preciso en el sentido de detectar cualquier diferencia de peso por reducida que ésta sea. El instrumento de medida en este caso será preciso y poco acurado, pues el peso de cualquier objeto se encuentra desviado respecto a su verdadero valor.

5.2.3. Diseño de encuestas y selección muestral

Como ya hemos comentado, es impensable llegar a realizar un estudio exento de errores. Sin embargo, los objetivos de exactitud y precisión serán más accesibles en la medida en que nuestro estudio tenga una base sólida. De ahí que el diseño de encuestas -cuyas etapas resumimos en el esquema siguiente- constituya una fase decisiva, que condiciona en gran medida la calidad de los resultados obtenidos.

DISEÑO DE ENCUESTAS

- Fase preliminar: objetivos del estudio
- Determinación del marco
  - Unidades elementales
  - Unidades complementarias
- Selección muestral
- Transmisión de la información
  - Contexto del estudio
  - Trabajo de campo
- Tratamiento de la información
  - Tabulación y síntesis
  - Técnicas inferenciales
- Evaluación de resultados

Dentro del diseño de encuestas incluimos desde las etapas previas al estudio (definición de objetivos y determinación de la población y sus unidades) hasta el trabajo de campo y los análisis posteriores (publicación y evaluación de resultados), siendo especialmente interesante desde la óptica estadística la etapa de selección muestral.

Todas estas etapas serán analizadas con detalle en un capítulo posterior, dedicado al muestreo en poblaciones finitas, por lo cual nos limitaremos aquí a describir cómo se seleccionan en la práctica muestras aleatorias o probabilísticas.

El término aleatorio, que consideramos sinónimo de probabilístico, suele ser utilizado de forma abusiva en el lenguaje coloquial, para indicar que una selección no está expresamente dirigida. Así por ejemplo, frases como "un encuestador de televisión ha salido a la calle preguntando la opinión de personas seleccionadas aleatoriamente" no serían estrictamente correctas. En efecto, el hecho de que el encuestador intente que su opinión subjetiva no afecte a la selección no basta para calificar a una muestra de aleatoria ya que, aunque los resultados muestrales son imprevisibles (interviene el azar), no es posible asignarles probabilidades.

Así pues, solamente denominaremos aleatorios (o estocásticos o probabilísticos) a aquellos procesos en los que podemos determinar la probabilidad de selección para cada muestra concreta.

Consideremos el total de hogares sobre los que deseamos analizar las pautas de lectura o bien el total de puntos de venta del periódico, para los cuales disponemos de un listado correctamente numerado. Una vez decididos a extraer una muestra aleatoria o probabilística ¿qué mecanismo podemos emplear para introducir azar en la selección?

Los ejemplos más conocidos son los sorteos: extracción al azar de bolas numeradas de una urna o un bombo de lotería, de tal forma que los elementos de la población cuyos números se correspondan con los extraídos pasan a integrar la muestra. Este

mecanismo, muy popular gracias a los sorteos de lotería nacional, resulta sin embargo impracticable para tamaños muestrales elevados debido al coste material y de tiempo que conlleva.

Como consecuencia de estas limitaciones, el método general consiste en acudir a *tablas de números aleatorios* generadas por distintos procedimientos físicos y matemáticos.

Estas tablas recogen los resultados de un proceso que genera dígitos decimales aleatorios, asociados a variables aleatorias independientes con valores 0, 1, ..., 9, que cumplen las siguientes propiedades:

- Cualquier dígito de la tabla tiene probabilidad $\frac{1}{10}$ de presentar valores 0, 1, ... , 9, es decir, corresponden a realizaciones de una v.a. discreta con distribución uniforme.
- Los distintos dígitos de la tabla son independientes.

Como consecuencia de su carácter aleatorio, los dígitos de las tablas no presentarán ningún orden concreto. Para ilustrar esta idea, E.U. Condon, director del Bureau of Standards afirmaba que la obtención de una secuencia lógica al seleccionar números aleatorios resulta tan probable como que los saltos de un mono sobre una máquina de escribir reproduzcan un párrafo de Shakespeare.[2]

Por lo que se refiere a la utilización de estas tablas, las únicas dudas podrían ir referidas a cuántos dígitos seleccionar y en qué orden.

El punto de arranque es arbitrario dado el propio carácter aleatorio de las tablas. Una vez situados en ciertas coordenadas, efectuaremos selecciones sucesivas de números avanzando por filas o por columnas.

Es importante además tener presente que cualquier elemento de la población debe ser candidato a formar parte de la muestra. Para garantizar esta potencialidad, el número de columnas seleccionadas en la tabla debe coincidir con el número de dígitos del tamaño poblacional N.

Una de las primeras tablas de números aleatorios fue elaborada en 1927 por L.H.C. Tippett, quien construyó una tabla de 41.600 dígitos a partir de datos del censo británico sobre las áreas parroquiales, eliminando en cada caso los dígitos primero y último.

En 1943 Fisher y Yates publicaron una tabla con 15.000 números, correspondientes a los dígitos que ocupaban el orden 15 y 19 en tablas logarítmicas de 20 dígitos.

El antecedente de las actuales rutinas generadoras de números aleatorios fue un método puesto en marcha en 1939 por Kendall y Babington-Smith, quienes generaron una tabla de 100.000 números con ayuda de una máquina electrónica que simulaba lanzamientos de un cuerpo geométrico de 10 caras, numeradas del 0 al 9. En la actualidad, la generación y contraste de números aleatorios sigue siendo un campo de investigación.

[2] Esta anécdota aparece recogida en Youden, W.J. (1957): "Random Numbers aren´t Nonsense" Industrial and Engineering Chemistry, 49, n. 10, 89 A

Como sucede en otros muchos campos, el avance de la informática ha simplificado considerablemente la selección de números aleatorios: hoy día gran parte de los programas estadísticos e incluso las hojas de cálculo contienen procedimientos para generar números aleatorios según distintos modelos de probabilidad (uniforme, normal, ...).

5.3. Estadísticos y estimadores

Imaginemos que nuestro objetivo fuese aproximar el valor verdadero de la renta esperada de una población de hogares, esto es, el parámetro poblacional $\mu = E(X)$. Una vez que dispongamos de información muestral (es decir, de rentas observadas para ciertos hogares, seleccionados aleatoriamente), debemos preocuparnos de llevar a cabo un tratamiento adecuado de la misma, diseñando procedimientos de síntesis muestral que se adapten a nuestros objetivos de estimación.

Parece claro que en ningún caso llegaremos a conocer con exactitud la verdadera renta esperada de la población. Sin embargo, podemos considerar algunas restricciones respecto a los errores que estamos dispuestos a asumir: en general, toleraremos los errores de tipo aleatorio (inevitables si tenemos en cuenta que es el azar quien determina qué hogares forman parte de la muestra) pero no será admisible sin embargo una sobrevaloración (o una subvaloración) sistemática de las rentas, que introduciría un sesgo en nuestro proceso de estimación de μ.

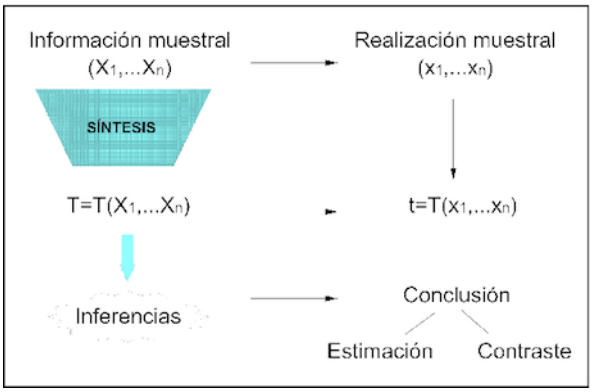
De modo complementario, debemos exigir a nuestro proceso de estimación que el riesgo de desviarnos del verdadero valor del parámetro sea moderado (esto es, que exista una baja probabilidad de obtener estimaciones muy alejadas de la verdadera renta esperada). Este requisito proporciona una especie de "garantía" para el proceso de estimación, al acotar la dispersión de nuestras estimaciones respecto al parámetro.

Por otra parte, parece necesario que los estimadores utilicen toda la información a su alcance, ya que de lo contrario estarían ignorando datos útiles en la aproximación del valor buscado. En este sentido, bastaría recordar que una medida de síntesis de las rentas observadas en la muestra debe abarcar todos y cada uno de los datos proporcionados por los hogares incluidos en la misma.

Cabe por último enfatizar la distinción entre los conceptos de *estimación* (resultado) y *proceso de estimación* (procedimiento). El primero dependerá de la información muestral disponible en cada caso, que conducirá a un valor concreto como aproximación (nunca exacta) del parámetro, mientras que el segundo describe nuestro método de trabajo. Como consecuencia, cabe esperar que, si el procedimiento empleado es correcto, la existencia de información adicional (nuevos datos sobre las rentas de los hogares) nos permita aproximarnos más adecuadamente a nuestro objetivo. En el límite, si la muestra creciese indefinidamente y llegásemos a disponer de información sobre todos los hogares de la población, cabría esperar que nuestra estimación se aproximase indefinidamente al parámetro μ.

El ejemplo propuesto pretende hacer explícita una lista de requisitos que resultan intuitivamente deseables en todo proceso de aproximación o estimación. Su cumplimiento no garantizará la bondad de resultados concretos pero sí en cambio del método

Figura 5.2.: Esquema del proceso inferencial



general.

Nuestro esquema de actuación, que aparece recogido en la figura 5.2, exige la presencia de instrumentos de síntesis denominados estadísticos que, por ser funciones de la muestra aleatoria, serán también aleatorios.

En efecto, consideremos una m.a.s. $(X_1, \ldots, X_n)$ extraída de una población X. Se trata de n v.a. independientes e idénticamente distribuidas (i.i.d.) y una vez que dicha muestra aleatoria se concrete en determinada observación muestral, podemos llevar a cabo una síntesis de su información mediante medidas descriptivas aplicadas a los valores obtenidos. Además, antes de que la muestra concreta haya sido seleccionada es posible también establecer expresiones matemáticas que son función de la muestra aleatoria y por tanto variables aleatorias. Dichas expresiones genéricas, que representaremos por $T = T(X_1, \ldots, X_n)$ se denominan estadísticos.

Definición 5.3. Sea $(X_1, \ldots, X_n)$ una muestra aleatoria de tamaño n de una variable X. Llamamos *estadístico* $T = T(X_1, \ldots, X_n)$ a cualquier función medible definida sobre las variables muestrales; esto es, una función observable del vector aleatorio.

Como ya hemos comentado en el capítulo segundo, toda función medible de v.a. es una v.a. Por lo tanto, un estadístico será una v.a., lo cual significa que llevará asociada una distribución de probabilidad y las correspondientes características: esperanza, varianza, etc.

A modo de ejemplo, a partir de una m.a.s. $(X_1, \ldots, X_n)$ definamos un estadístico $T = \sum_{i=1}^{n} X_i$. Se trata de una variable aleatoria cuyas características ya han sido obtenidas en temas anteriores; así, por ser la esperanza un operador lineal se tiene:

$$E(T) = E\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} E(X_i) = \sum_{i=1}^{n} \mu = n\mu$$

ya que las variables están idénticamente distribuidas y por tanto $E(X_i) = \mu\,, \forall i = 1, \ldots n$.

Para obtener la varianza se tiene en cuenta, además de la idéntica distribución, la independencia entre las variables $X_1, \ldots, X_n$:

$$Var(T) = Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i) = \sum_{i=1}^{n} \sigma^2 = n\sigma^2$$

La distribución de probabilidad del estadístico T vendría dada por la función:

$$F_T(t) = P(T \leq t) = P\left(\sum_{i=1}^{n} X_i \leq t\right)$$

en cuya expresión aparece la distribución de la suma, que en el capítulo 4 ha sido analizada para distintos supuestos.

En síntesis, considerábamos las siguientes situaciones:

- Si la población X se distribuye según un modelo reproductivo, es posible afirmar para cualquier tamaño muestral que la suma presenta el mismo modelo probabilístico, conociéndose además sus parámetros.
- Para tamaños elevados de muestra, el teorema central del límite garantiza -con independencia de la distribución de partida- la convergencia de la distribución de la suma a un modelo normal.
- Por último, para poblaciones desconocidas y tamaños pequeños de muestra, no es posible determinar la distribución de la suma por lo cual únicamente obtendríamos acotaciones de las probabilidades mediante la desigualdad de Chebyshev.

El estadístico es una función que puede tomar valores en una o más dimensiones, por lo que puede tratarse de una v.a. unidimensional o k-dimensional. Nos interesarán fundamentalmente los estadísticos de resumen, que a cada vector aleatorio $(X_1, \ldots, X_n)$ asocian un valor real $T(X_1, \ldots, X_n)$.

Generalmente las poblaciones investigadas dependen de ciertos parámetros desconocidos (μ, σ^2, p, ...) y la introducción de estadísticos suele tener como objetivo la aproximación de dichos parámetros a partir de muestras aleatorias. De ahí que a tales estadísticos se les denomine estimadores.

Definición 5.4. Consideremos una variable aleatoria X cuya distribución de probabilidad F depende de un parámetro (o vector de parámetros) θ perteneciente al espacio paramétrico Θ, esto es $F_X(x, \theta)$. Denominaremos *estimador* de θ a un estadístico $T(X_1, \ldots, X_n)$ que toma valores sólo en Θ.

Como puede apreciarse, la definición de estimador resulta más restrictiva que la de estadístico, ya que un estimador T de θ sólo podrá tomar valores en el espacio paramétrico Θ. Nuestro objetivo será seleccionar, de entre todas las expresiones que satisfacen este requisito, aquéllas cuyas propiedades proporcionen garantías en el proceso de estimación del parámetro.

En el caso de que la distribución de X dependa de k parámetros $\theta_1, \ldots, \theta_k$, podríamos plantearnos la utilización de k estimadores unidimensionales o bien la consideración de un estimador k-dimensional.

Como consecuencia de su definición, los estimadores $T(X_1, \ldots, X_n)$ serán variables aleatorias. Sin embargo, una vez seleccionada una muestra particular $(x_1, \ldots, x_n)$ se obtiene un valor concreto del estimador $t = T(x_1, \ldots, x_n)$ al que se denomina *estimación*.

Figura 5.3.: Interpretación de la función de verosimilitud

5.3.1. Función de verosimilitud

Dado que gran parte de la inferencia tiene como objetivo la aproximación de parámetros desconocidos a partir de información muestral, resulta necesario analizar la distribución probabilística de las muestras y de los estadísticos definidos a partir de las mismas.

La distribución de probabilidad de una muestra aparece conectada al concepto de *función de verosimilitud*. Se trata de una expresión que admite dos interpretaciones alternativas ilustradas en la figura 5.3 y que resulta de gran interés para examinar la idoneidad de las expresiones propuestas como estimadores.

Definición 5.5. Consideremos una población X cuya distribución depende de cierto parámetro desconocido $\theta \in \Theta$, esto es, $F_X(x, \theta)$. Si de esta población extraemos m.a.s. de tamaño n, $(X_1, \ldots, X_n)$, entonces la *distribución de probabilidad muestral* vendrá dada para cada realización $(x_1, \ldots, x_n)$, por la expresión:

$$F(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} F(x_i, \theta)$$

Esta expresión ha sido obtenida en el primer epígrafe de este tema cuando la distribución no dependía de ningún parámetro. En este caso el razonamiento sería análogo.

La consideración de F como función de parámetros poblacionales desconocidos θ conlleva nuevas interpretaciones para esta expresión, y más concretamente para su derivada que recibe la denominación de *función de verosimilitud* (f.v.) y se denota habitualmente por L (del término inglés *Likelihood*).

$$L = L(x_1, \ldots, x_n\theta) = \frac{\partial^n F(x_1, \ldots, x_n, \theta)}{\partial x_1 \cdots \partial x_n} = \prod_{i=1}^{n} \frac{\partial F(x_i, \theta)}{\partial x_i}$$

Según la tipología de la variable X, la f.v. puede ser expresada como:

- $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} p(x_i, \theta)$ para el caso discreto
- $L(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta)$ para el caso continuo

De este modo, si consideramos un valor fijo -aunque desconocido- del parámetro, que designamos por θ^*, la expresión $L(x_1, \ldots, x_n, \theta^*)$ representa la probabilidad de la muestra aleatoria $(x_1, \ldots, x_n)$.

De modo alternativo, si disponemos de una realización muestral concreta $(x_1^*, \ldots, x_n^*)$, la expresión $L(x_1^*, \ldots, x_n^*, \theta)$ dependerá únicamente del parámetro θ, respondiendo así a su denominación como *función de verosimilitud* (evalúa la verosimilitud o credibilidad de una observación muestral concreta en función del parámetro θ).

A modo de ilustración, consideremos dos ejemplos con los que trabajaremos a lo largo de este tema. El primero de ellos se corresponde con una variable aleatoria discreta X que recoge si un individuo activo se encuentra o no en paro, y por tanto sigue un modelo de Bernoulli de parámetro p, cuya distribución de probabilidad viene dada por la expresión:

$$p(x_i, p) = p^{x_i}(1-p)^{1-x_i}\ ,\ x_i = 0, 1$$

Por lo tanto, a partir de una muestra de tamaño n se obtendría la función de verosimilitud:

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} p(x_i, p) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

que, según las interpretaciones anteriormente comentadas, para valores fijos de p proporciona la distribución muestral, mientras que para cada muestra concreta evalúa su verosimilitud en función de la tasa de paro p.

Consideremos por otra parte una variable continua que recoge el gasto mensual de los hogares y se distribuye según un modelo normal en el que, para mayor operatividad, asumimos una dispersión unitaria. A partir de la distribución $X \approx \mathcal{N}(\mu, \sigma = 1)$ se obtiene:

$$L(x_1, \ldots, x_n, p) = \prod_{i=1}^{n} f(x_i, p) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x_i-\mu)^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

La interpretación de L como función de verosimilitud equivale a asumir una realización muestral concreta, es decir, determinados hogares para los que se ha examinando el gasto mensual. De este modo, no existe ya incertidumbre sobre la muestra, y $L(x_1, \ldots, x_n, \mu)$ mide su verosimilitud para cada potencial gasto esperado μ.

La función de verosimilitud genérica es $L(x_1, \ldots, x_n, \theta)$ con $\theta \in \Theta$, expresión que para valores concretos del parámetro proporciona resultados $L(x_1, \ldots, x_n, \theta) \in [0, 1]$, que evalúan el nivel de credibilidad o verosimilitud de nuestra realización muestral para cada valor potencial de θ.

Por tanto, si para dos posibles valores del parámetro θ_1 y θ_2 se tiene $L(x_1, \ldots, x_n, \theta_1) < L(x_1, \ldots, x_n, \theta_2)$ parece razonable afirmar que la probabilidad de haber obtenido la muestra $(x_1, \ldots, x_n)$ sería mayor con el valor θ_2 que con θ_1. Esta interpretación de la función de verosimilitud ha inspirado un método de obtención de estimadores que analizaremos en un apartado posterior.

En nuestro ejemplo de la tasa de paro, imaginemos que sobre una muestra de 20 activos hemos observado 5 parados. Si asumimos para mayor claridad que el parámetro p (tasa de paro) pertenece a un espacio paramétrico con sólo dos valores $\Theta = \{0,1; 0,3\}$, entonces se tiene:

$$L(x_1, \ldots, x_n, p = 0,1) = 0,1^5 0,9^{15}$$

$$L(x_1, \ldots, x_n, p = 0,3) = 0,3^5 0,7^{15}$$

verificándose $L(x_1, \ldots, x_n, p = 0,1) < L(x_1, \ldots, x_n, p = 0,3)$, con lo cual concluiríamos que $p = 0,3$ es el valor de la tasa de paro que hace más verosímil la muestra seleccionada.

5.4. Propiedades de los estimadores

A la vista del planteamiento general del apartado anterior, sería posible construir infinitos estadísticos con capacidad de resumir la información muestral. Sin embargo, los parámetros interesantes de una población son a menudo sus características más distintivas: la esperanza, la varianza, la proporción, ... y los estadísticos que analizaremos tendrán en general vocación de estimadores de estos parámetros.

Consideremos una v.a. X con distribución de probabilidad $F(x, \theta)$ cuyo parámetro θ pretendemos aproximar a partir de la información suministrada por una m.a.s. Existirían numerosas expresiones que proporcionan estimaciones del parámetro θ y, dado que en la práctica la utilización de una u otra no va a resultarnos indiferente, debemos enunciar los requisitos considerados deseables en un buen estimador, que servirán para discriminar entre expresiones alternativas.

5.4.1. Ausencia de sesgo

Un primer requisito deseable en un estimador es que su comportamiento sea "imparcial" o centrado, que no conduzca sistemáticamente a subvaloraciones o sobrevaloraciones del parámetro. Esta propiedad se conoce como *ausencia de sesgo*.

Al asumir como objetivo la aproximación de un parámetro desconocido θ, las limitaciones de información hacen inevitable la presencia de errores o desviaciones. De este modo, es posible definir:

Definición 5.6. Se denomina *error aleatorio* asociado a T que se genera como diferencia entre el estimador y el parámetro desconocido:

$$e_T = T - \theta$$

En el caso de que el origen de estos errores sea únicamente aleatorio, sin que se presente ninguna componente de tipo sistemático, se puede asumir fácilmente que éstos llegarán a compensarse, dando lugar a un error esperado nulo, $E(e_T) = 0$.

El requisito de ausencia de sesgo exige que no haya intencionalidad en los errores, esto es, que las desviaciones tengan carácter aleatorio y por tanto exista "neutralidad" en el proceso de estimación. Cuando una expresión T satisface esta condición de neutralidad recibe la denominación de estimador insesgado o centrado.

Definición 5.7. Se dice que $T(X_1, \ldots, X_n)$ es un *estimador insesgado* del parámetro $\theta \in \Theta$ si el valor esperado de su error aleatorio asociado existe y es nulo para cualquier valor posible del parámetro ($E(e_T) = 0$ o equivalentemente $E(T) = \theta$).

Cuando un estimador es centrado, su valor esperado coincide con el parámetro que pretende aproximar. De ahí que el requisito de ausencia de sesgo se exija habitualmente a los estimadores como garantía de su carácter objetivo.

En el caso de que un estimador T no sea centrado para estimar el parámetro ($E(e_T) \neq 0$) se denomina *sesgado*. Para cuantificar el sesgo o desviación sistemática inherente a un estimador comparamos su valor esperado con el parámetro que pretende aproximar.

Definición 5.8. El *sesgo* introducido por un estimador T para estimar un parámetro θ viene definido por la expresión:

$$B_T(\theta) = E(e_T) = E(T) - \theta$$

Dicho sesgo será una función del parámetro θ, que habitualmente se denota con la inicial del término inglés *Bias*.

Cuando un estimador T lleva asociado un sesgo positivo (esto es, un error esperado positivo) dicho estimador se desvía sistemáticamente "al alza" en su aproximación del parámetro desconocido, y lo contrario sucede para los estimadores con sesgo negativo (conducen a subestimaciones sistemáticas de θ). Por último, los estimadores que llevan asociados sesgos o errores esperados nulos han sido definidos anteriormente como centrados o insesgados.

Consideremos de nuevo la v.a. gasto mensual de los hogares, $X \approx \mathcal{N}(\mu, 1)$ de la que se ha extraído una m.a.s. de tamaño $n = 4$, definiendo las tres expresiones siguientes como estimadores del parámetro $\mu = E(X)$:

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4}{4} \; ; \; T_2 = \frac{2X_1 + X_4}{4} \; ; \; T_3 \frac{X_1 + X_2 + 2X_3 + X_4 + 50}{5}$$

Se comprueba fácilmente que sólo el primero de los estimadores propuestos es centrado o insesgado para estimar μ por ser el único que conduce a un error esperado nulo o, equivalentemente, el único que presenta esperanza coincidente con el parámetro μ:

$$E(T_1) = \mu \; ; \; E(T_2) = 0,75\mu \; ; \; E(T_3) = \mu + 10$$

[Compruébese]

Calculando ahora los sesgos de los estimadores anteriores $B_T(\mu) = E(T) - \mu$ se obtiene:

$$B_{T_1}(\mu) = 0 \; ; \; B_{T_2}(\mu) = -0,25\mu \; ; \; B_{T_3}(\mu) = 10$$

informando estos valores sobre la dirección y la cuantía del error sistemático cometido con cada estimador.

En el caso de T_1 el sesgo es nulo ya que dicho estimador es centrado y por tanto no introduce ningún error de carácter no aleatorio. En cambio no sucede lo mismo con T_2 y T_3.

Analizando las expresiones de los sesgos se aprecia que T_2 subestima el verdadero valor de μ, ya que $B_{T_2}(\mu) = -0,25\mu$ mientras que para T_3 se obtiene un sesgo positivo $B_{T_3}(\mu) = 10$ y por tanto una sobreestimación del parámetro.

Se observa además que en ciertos casos particulares (para T_3, por ejemplo) el sesgo adopta valor constante (en este caso 10 unidades). En general sin embargo el sesgo es función del parámetro desconocido (situación que se presenta por ejemplo para el estimador T_3, cuya expresión sobreestima sistemáticamente el parámetro μ).

Cuando el sesgo viene expresado en función de θ no puede ser calificado de "alto" o "bajo" ya que desconocemos la magnitud de θ. Puede entonces resultar útil definir el sesgo relativo, como cociente entre el sesgo y el parámetro desconocido:

$$B_T^R(\theta) = \frac{B_T(\theta)}{\theta}$$

A modo de ilustración, para el estimador T_2 se tenía un sesgo $B_{T_2}(\mu) = 0,25\mu$ que expresado en términos relativos proporciona el valor $B_{T_2}^R(\mu) = -0,25$. Este resultado permite afirmar que T_2 subestima el parámetro μ en un 25 % de su valor.

Los estimadores insesgados presentan propiedades interesantes:

Proposición 5.1. *Si T es un estimador insesgado de θ, entonces $aT + b$ es un estimador insesgado de $a\theta + b$, $\forall a, b \in \Re$.*

Demostración. La comprobación de esta propiedad es inmediata, con sólo tener presente que la esperanza es un operador lineal:

$$E(aT + b) = aE(T) + b = a\theta + b$$

□

Proposición 5.2. *Si T_1 y T_2 son dos estimadores insesgados de θ entonces cualquier combinación convexa de ellos es también un estimador insesgado.*

Demostración. En efecto, si T_1 y T_2 son insesgados, se tiene: $E(T_1) = \theta$ y $E(T_2) = \theta$. Dada ahora una combinación convexa: $T = \alpha T_1 + (1 - \alpha)T_2$, $0 < \alpha < 1$, su valor esperado viene dado por:

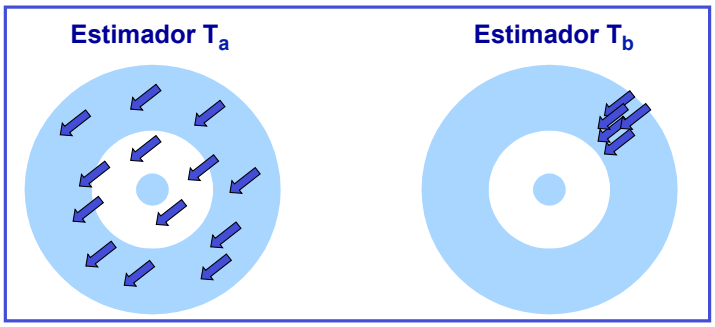
$$E(T) = E(\alpha T_1 + (1 - \alpha)T_2) = \alpha E(T_1) + (1 - \alpha)E(T_2) = \alpha\theta + (1 - \alpha)\theta = \theta$$

y por tanto T es insesgado. □

Así pues, si existen dos estimadores insesgados, entonces existen infinitos.

La figura 5.4 recoge una ilustración gráfica de la idea de estimador centrado o insesgado. En dicha figura representamos el proceso de estimación como lanzamientos sucesivos de dardos a una diana, cuyo punto central se corresponde con el parámetro

Figura 5.4.: Eficiencia



desconocido θ. Los lanzamientos representados son equivalentes a las estimaciones obtenidas al aplicar T a las realizaciones muestrales, y el error de estimación es en cada caso la distancia del dardo al centro de la diana. La figura a) representa una situación de ausencia de sesgo en la que los errores cometidos al estimar θ se compensan por no tener carácter sistemático. Los dardos lanzados se distribuyen aleatoriamente alrededor del centro de la diana o, en otras palabras, las estimaciones se distribuyen aleatoriamente en torno al parámetro θ por lo cual $E(T_a) = \theta$.

La figura 5.4b) representa una situación bastante distinta, en la que los dardos lanzados presentan un error sistemático o sesgo. En este caso los errores cometidos no parecen debidos únicamente al azar sino más bien a algún fallo del proceso (algún defecto de vista del propio tirador, corrientes de aire, un arco defectuoso, ...) por lo cual la esperanza ya no es el centro de la diana, sino que se sitúa en la parte superior derecha.

5.4.2. Eficiencia

El criterio de ausencia de sesgo exige que el valor esperado de los errores sea nulo; sin embargo, este requisito no ofrece garantías respecto al riesgo de obtener estimaciones muy alejadas del parámetro.

En efecto, si examinamos de nuevo los gráficos de la figura 5.4 vemos que pueden existir estimadores insesgados cuyo uso no resulta aconsejable. Esta sería la situación del estimador T_a, ya que su esperanza (resumen de los lanzamientos) coincide con el parámetro (centro de la diana) pero sin embargo puede conducir a estimaciones muy distantes del valor central (alta dispersión).

Por el contrario el estimador T_b presenta un sesgo pero, frente a este rasgo negativo, el estimador tiene a su favor la baja dispersión (los dardos se encuentran concentrados en un radio pequeño, lo que equivaldría a estimaciones muy próximas entre sí).

Supongamos que deseamos elegir un estimador de θ entre varias expresiones alternativas. Si estas expresiones resultan indiferentes respecto al sesgo (es decir, si todos son insesgados o bien presentan sesgos idénticos) parece claro que deberíamos seleccionar el estimador con menor riesgo o varianza.

Sin embargo, en general necesitaremos comparar estimadores con diferentes sesgos, por lo cual debemos considerar un criterio más amplio de selección, que tenga en cuenta tanto el sesgo como el nivel de riesgo asociados a un estimador. Surge así la idea de *eficiencia*.

Para comparar la eficiencia de varios estimadores alternativos de un parámetro θ estudiaremos los errores asociados a los mismos que, para evitar compensaciones de signo, elevamos al cuadrado. Este planteamiento conduce al concepto de *error cuadrático medio*.

Definición 5.9. El *error cuadrático medio* (ECM) asociado a un estimador T del parámetro θ es una medida de eficiencia definida como el valor esperado del error cuadrático de T:

$$ECM_T(\theta) = E(e_T^2) = E(T-\theta)^2$$

Dicha medida puede también ser formulada como:

$$ECM_T(\theta) = B_T^2(\theta) + Var(T)$$

expresión que se obtiene fácilmente a partir de la anterior y permite distinguir dentro del error cuadrático medio dos componentes: sesgo y varianza.

En efecto, a partir de la expresión de la varianza del error $Var(e_T) = E\left(e_T^2\right) - E^2(e_T)$, se obtiene:

$$E\left(e_T^2\right) = E^2(e_T) + Var(e_T) = B_T^2(\theta) + Var(T)$$

con sólo aplicar la definición de sesgo de T y tener en cuenta que se cumple $Var(e_T) = Var(T)$ [¿por qué ?].

Otra alternativa más habitual para llegar a esta expresión consiste en desarrollar el cuadrado del error:

$$\begin{aligned} ECM_T(\theta) = & \quad E\left(e_T^2\right) = E(T-\theta)^2 = E(T^2 - 2T\theta + \theta^2) = E(T^2) + \theta^2 - 2\theta E(T) = \\ = & \quad E(T^2) - E^2(T) + E^2(T) + \theta^2 - 2\theta E(T) = Var(T) + B_T^2(\theta) \end{aligned}$$

El error cuadrático medio es el criterio utilizado para comparar la eficiencia de varios estimadores de un parámetro.

Definición 5.10. Para cualesquiera T_1 y T_2 estimadores de θ se dirá que T_1 *es más eficiente que* T_2 si su error cuadrático medio es inferior:

$$ECM_{T_1}(\theta) < ECM_{T_2}(\theta)$$

Volviendo al ejemplo anteriormente considerado, podemos construir ahora los ECM de las tres expresiones propuestas como estimador del gasto esperado μ. Teniendo en cuenta los dos componentes sesgo y varianza, se obtienen los resultados recogidos a continuación:

Estimador	Sesgo	Varianza	ECM
$T_1 = \frac{X_1+X_2+X_3+X_4}{4}$	0	$\frac{1}{4}$	$\frac{1}{4}$
$T_2 = \frac{2X_1+X_4}{4}$	$-0,25\mu$	$\frac{5}{16}$	$0,0625\mu^2 + \frac{5}{16}$
$T_3 \frac{X_1+X_2+2X_3+X_4+50}{5}$	10	$\frac{7}{25}$	$100 + \frac{7}{25}$

que permiten concluir que, de las tres expresiones consideradas como estimadores del gasto mensual esperado, T_1 resulta ser la más eficiente.

El requisito de *eficiencia* permite formalizar las propiedades de precisión y exactitud o acuracidad a las que nos hemos referido en un apartado anterior. La *precisión* va referida a la concentración de la distribución de las estimaciones respecto al valor esperado, y por tanto aparece relacionada inversamente con la dispersión y su indicador, la varianza.

Por su parte, el requisito de *exactitud o acuracidad* es más estricto, por ir referido a la concentración de la distribución de estimaciones respecto al valor verdadero del parámetro y aparece por tanto inversamente relacionado con el error cuadrático medio.

Consideremos de nuevo la ilustración gráfica 5.4 de los estimadores T_a y T_b, planteándonos la pregunta ¿cuál de estos estimadores resultaría -en términos relativos- más eficiente?

La comparación de la eficiencia de ambos estimadores nos llevará a considerar simultáneamente los componentes de sesgo y riesgo a través del ECM.

Razonando en términos gráficos a partir de la figura 5.4, el estimador T_b tiene un sesgo representado por el segmento $B_T(\theta)$, mientras que su dispersión puede ser cuantificada a partir del radio de la circunferencia en la que se incluyen las estimaciones. Si a partir de ambos componentes construimos un triángulo, el cuadrado de su hipotenusa se correspondería con el ECM, y podría obtenerse como suma de los cuadrados de los catetos, es decir:

$$ECM_{T_b}(\theta) = B^2_{T_b}(\theta) + Var(T_b)$$

alcanzando esta expresión valores más reducidos cuanto más eficiente sea T_b.

Por su parte, el estimador T_a es insesgado, presentando por tanto un ECM coincidente con su varianza (aproximación -salvo constante- del área del círculo en el que se inscriben las estimaciones).

5.4.3. Mínima varianza

Hasta ahora hemos estudiado la eficiencia como criterio de selección entre varios estimadores. Sin embargo, teniendo en cuenta que podrían existir múltiples expresiones válidas como estimadores de cierto parámetro, cabe preguntarse cómo es posible seleccionar aquélla que resulte, en sentido absoluto, más eficiente.

Definición 5.11. Se dice que un estimador $T = T(X_1, \ldots, X_n)$ es *eficiente* cuando es insesgado y posee mínima varianza.

La condición de mínima varianza no resulta tan inmediata en su comprobación como la de ausencia de sesgo. Sin embargo es posible establecer una cota inferior para la varianza de cualquier estimador, de modo que si un estimador concreto alcanza esta cota podemos garantizar que no existe ningún otro estimador de varianza inferior.

Esta acotación, denominada *desigualdad de Frechet-Cramer-Rao*, permite una definición alternativa de la eficiencia:

Teorema 5.1. *Sea $\mathbf{x} = (x_1, \ldots, x_n)$ una muestra aleatoria extraída de una población X con distribución $F(\mathbf{x}, \theta)$ y denotemos por $L(\mathbf{x}, \theta)$ la función de verosimilitud asociada. Si T es un estimador cualquiera de θ, entonces la varianza de T verifica, bajo ciertas condiciones de regularidad, la desigualdad:*

$$Var(T) \geq \frac{\left(1 + \frac{\partial B_T(\theta)}{\partial \theta}\right)^2}{E\left(\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta}\right)^2}$$

Las condiciones de regularidad necesarias para la demostración de la desigualdad anterior hacen referencia a que el campo de variación de X no dependa del parámetro θ, que el espacio paramétrico Θ sea un intervalo que no se reduce a un punto, que existan derivadas de primero y segundo orden de la función de verosimilitud y que se puedan permutar las operaciones de derivabilidad e integrabilidad.

De estas condiciones nos interesa especialmente la primera, que iremos comprobando en las ilustraciones de este capítulo.

En la acotación de Frechet-Cramer-Rao (F-C-R) puede observarse que sólo el numerador depende del estimador considerado. En el caso particular de que dicho estimador sea insesgado, el numerador adopta valor unitario.

Por lo que se refiere al denominador de la cota de Frechet-Cramer-Rao, se trata de una expresión de gran interés, denotada habitualmente por I_n y denominada *cantidad de información de Fisher*. Esta medida es un indicador de la cantidad de información que la muestra contiene sobre el parámetro θ y para cada muestra aleatoria $\mathbf{x}$ viene dada por la expresión:

$$I_n(\mathbf{x}, \theta) = E\left(\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta}\right)^2$$

Entre los rasgos de esta medida de información destacan los siguientes:

Proposición 5.3. *La cantidad de información de Fisher puede ser también expresada de forma más operativa como:*

Proposición.

$$I_n(\mathbf{x}, \theta) = -E\left(\frac{\partial^2 \ln L(\mathbf{x}, \theta)}{\partial \theta^2}\right)$$

Demostración. Primero comprobamos:

$$\begin{aligned} E\left(\frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta}\right) = & \int \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathbf{x}, \theta) d\mathbf{x} = \int \left(\frac{\partial L(\mathbf{x}, \theta)}{\partial \theta} \frac{1}{L(\mathbf{x}, \theta)}\right) L(\mathbf{x}, \theta) d\mathbf{x} = \\ = & \frac{\partial}{\partial \theta} \int L(x, \theta) dx = \frac{\partial}{\partial \theta}(1) = 0 \end{aligned}$$

Con lo cual su derivada también es nula: $\frac{\partial}{\partial \theta}\left(\int \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathrm{x}, \theta) d\mathrm{x}\right) = 0$, y desarrollando esta derivada se tiene:

$$0 = \frac{\partial}{\partial \theta}\left(\int \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathrm{x}, \theta) d\mathrm{x}\right) = \int \left[\frac{\partial^2 \ln L(\mathbf{x}, \theta)}{\partial \theta^2} L(\mathrm{x}, \theta) + \frac{\partial \ln L(\mathrm{x}, \theta)}{\partial \theta} \underbrace{\frac{\partial L(\mathbf{x}, \theta)}{\partial \theta}}_{= \frac{\partial \ln L(\mathbf{x}, \theta)}{\partial \theta} L(\mathrm{X}, \theta)}\right] d\mathbf{x}$$

Por lo tanto:

$$0 = \int \left[\frac{\partial^2 \ln L(\mathbf{x},\theta)}{\partial \theta^2} L(\mathrm{x},\theta) + \underbrace{\frac{\partial \ln L(\mathrm{x},\theta)}{\partial \theta} \frac{\partial \ln L(\mathrm{x},\theta)}{\partial \theta}}_{\left(\frac{\partial \ln L(\mathrm{x},\theta)}{\partial \theta}\right)^2} L(\mathbf{x},\theta) \right] d\mathbf{x}$$

o lo que es lo mismo:

$$\int \left(\frac{\partial \ln L(\mathbf{x},\theta)}{\partial \theta} \right)^2 L(\mathrm{x},\theta) d\mathrm{x} = -\int \left(\frac{\partial^2 \ln L(\mathbf{x},\theta)}{\partial \theta^2} \right) L(\mathbf{x},\theta) d\mathbf{x}$$

que, por definición de esperanza, coincide con lo que queremos demostrar:

$$E\left(\frac{\partial \ln L(\mathbf{x},\theta)}{\partial \theta} \right)^2 = -E\left(\frac{\partial^2 \ln L(\mathbf{x},\theta)}{\partial \theta^2} \right)$$

□

Proposición 5.4. *La cantidad de información es una medida aditiva en el sentido de que la información de Fisher contenida en una m.a.s. de tamaño n coincide con la suma de la información contenida en n muestras de tamaño unitario:*

$$I_n(\mathbf{x},\theta) = nI_1(x,\theta)$$

Demostración. Partiendo de la definición de la cantidad de información, se tiene:

$$\begin{aligned} I_n(\mathbf{x},\theta) = &\quad E\left(\tfrac{\partial \ln L(\mathbf{x},\theta)}{\partial \theta} \right)^2 = E\left[\tfrac{\partial}{\partial \theta} \ln \left(\textstyle\prod_{i=1}^n f(x_i,\theta) \right) \right]^2 = \\ = &\quad E\left[\tfrac{\partial}{\partial \theta} \textstyle\sum_{i=1}^n \ln f(x_i,\theta) \right]^2 = E\left[\textstyle\sum_{i=1}^n \tfrac{\partial}{\partial \theta} \ln f(x_i,\theta) \right]^2 \end{aligned}$$

Desarrollando ahora el cuadrado de esta suma y teniendo en cuenta que los componentes de una m.a.s son independientes se llega al enunciado propuesto:

$$\begin{aligned} I_n(\mathbf{x},\theta) = &\quad E\left[\textstyle\sum_{i=1}^n \left(\tfrac{\partial}{\partial \theta} \ln f(x_i,\theta) \right)^2 \right] + E\left[\textstyle\sum_i \sum_{j \neq j} \left(\tfrac{\partial}{\partial \theta} \ln f(x_i,\theta) \right) \left(\tfrac{\partial}{\partial \theta} \ln f(x_j,\theta) \right) \right] = \\ = &\quad \textstyle\sum_{i=1}^n E\left(\tfrac{\partial}{\partial \theta} \ln f(x_i,\theta) \right)^2 + \underbrace{\sum_i \sum_{j \neq j} E\left(\frac{\partial}{\partial \theta} \ln f(x_i,\theta) \right) E\left(\frac{\partial}{\partial \theta} \ln f(x_j,\theta) \right)}_{=0 \text{ (propiedad anterior)}} = \\ = &\quad nE\left(\tfrac{\partial}{\partial \theta} \ln f(x,\theta) \right)^2 = nI_1(x,\theta) \end{aligned}$$

Con lo cual queda demostrada la propiedad enunciada.

□

Esta propiedad de aditividad permite obtener la cantidad de información asociada a una muestra global como suma de las cantidades de información correspondientes a las distintas submuestras que la integran. Es interesante señalar sin embargo que este rasgo de aditividad es susceptible de críticas, dado que no recoge la existencia de "rendimientos marginales decrecientes" en la información muestral.

Desde un punto de vista intuitivo, parece claro que la cantidad de información sobre θ aportada por el primer elemento de la muestra $I_1(x_1,\theta)$ supera la cantidad de información incorporada por el

elemento n-ésimo $I_1(x_n, \theta)$, ya que en el primer caso se parte de una situación de desconocimiento mientras que la información asociada al elemento n-ésimo $I_1(x_n, \theta)$ se incorpora a partir de un nivel determinado de información $I_{n-1}(x_1, \ldots, n_{n-1}, \theta)$.

A modo de ilustración, calculemos la cantidad de información y la acotación de Frechet-Cramer-Rao correspondientes a nuestro ejemplo del gasto normalmente distribuido con dispersión unitaria.

Debemos tener presente que dicha acotación sólo se cumple bajo ciertas condiciones de regularidad. En concreto, en nuestro ejemplo asumimos un modelo $\mathcal{N}(\mu, 1)$ por lo cual el recorrido de la variable es $(-\infty, +\infty)$ que no depende del parámetro, y el espacio paramétrico no se reduce a un punto por incluir todos los posibles valores de μ.

La función de verosimilitud, obtenida anteriormente, viene dada por la expresión:

$$L(\mathbf{x}, \mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

A partir de la cual se obtiene:

$$\ln L(\mathrm{x}, \mu) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2$$

Y derivando respecto al parámetro μ:

$$\frac{\partial}{\partial \mu}\ln L(\mathbf{x}, \mu) = \sum_{i=1}^{n}(x_i - \mu) = \sum_{i=1}^{n} x_i - n\mu$$

$$\frac{\partial^2}{\partial \mu^2}\ln L(\mathbf{x}, \mu) = -n$$

Se obtiene entonces para la cantidad de información de la muestra:

$$I_n(\mathrm{x}, \mu) = -E\left[\frac{\partial^2 \ln L(\mathbf{x}, \mu)}{\partial \mu^2}\right] = -E(-n) = n$$

es decir, la cantidad de información sobre μ contenida por una muestra coincide con su tamaño n.

Teniendo en cuenta que se cumplen las condiciones de regularidad, la acotación de Frechet-Cramer-Rao permite afirmar:

$$Var(T) > \frac{\left(1 + \frac{\partial B_T(\mu)}{\partial \mu}\right)^2}{n}$$

A partir de esta expresión es sencillo comprobar que la media muestral es un estimador eficiente, ya que su varianza sería:

$$Var(\bar{X}) = \frac{\sigma^2}{n} = \frac{1}{n}$$

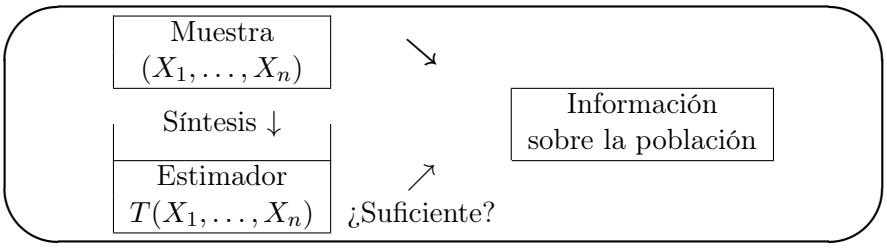
y la cota de F-C-R, cuyo numerador es unitario por ser el estimador insesgado, adopta el mismo valor:

$$\frac{\left(1 + \frac{\partial B_{\bar{X}}(\mu)}{\partial \mu}\right)^2}{n} = \frac{1}{n}$$

5.4.4. Suficiencia

Un estudio muestral no podrá ser calificado de adecuado si desperdicia o ignora parte de la información disponible. De ahí el concepto de *suficiencia*, entendido como capacidad de un estadístico para conservar toda la información que contiene una muestra.

Un estadístico suficiente deberá resultar igualmente útil -en cuanto al objetivo perseguido en cada caso- que la muestra inicial. Como recoge el esquema siguiente, la idea de suficiencia exige que toda la información de la muestra sea recogida o "atrapada" por T y en consecuencia, la distribución de la muestra una vez conocido T ya no dependerá del parámetro θ.



Como ilustra el esquema anterior, la utilización de estimadores supone en cierto sentido la aplicación de un filtro a nuestra información muestral. Desde un punto de vista conceptual, la propiedad de suficiencia resulta muy intuitiva, ya que se traducirá en que dicho filtro sea capaz de "asimilar" toda la información muestral disponible. No obstante, desde un punto de vista "técnico" la comprobación de la suficiencia no resulta sencilla.

El concepto de estadístico suficiente fue introducido por Fisher en 1922. Según dicho autor, un estadístico es suficiente para los objetivos de la inferencia estadística si contiene, en cierto sentido, toda la información acerca de la función de distribución a partir de la cual se ha generado la muestra.

Consideremos a modo de ejemplo la estimación del parámetro poblacional tasa de paro p. La situación de cada individuo activo se describe mediante una v.a. dicotómica X que adopta el valor 1 si el individuo se encuentra en paro y 0 en caso contrario, y a partir de muestras de 5 individuos activos se proponen dos estimadores alternativos de p:

$$T_1 = \frac{1}{5}\sum_{i=1}^{5} X_i \; ; \; T_2 = \frac{1}{5}(2X_1 + X_2 + X_5)$$

	Estimadores	
V.A.	T_1	T_2
(0,1,0,0,0)	$\frac{1}{5}$	$\frac{1}{5}$
(0,1,1,1,0)	$\frac{3}{5}$	$\frac{1}{5}$
(1,0,0,0,0)	$\frac{1}{5}$	$\frac{2}{5}$

Como se aprecia en el cuadro anterior, los estimadores presentan comportamientos distintos frente al requisito de suficiencia. Para estudiar este comportamiento, consideremos las tres muestras aleatorias representadas, que reflejan situaciones claramente distintas: en la primera y la tercera hay sólo un individuo parado, mientras que en la segunda el número de parados se eleva a 3.

¿Cómo recogen esta información muestral los dos estimadores propuestos para p? Puede verse que T_1 es capaz de diferenciar las situaciones muestrales registradas pero no sucede lo mismo con T_2, estimador para el que se aprecian dos tipos de contradicciones:

Por una parte, T_2 adopta el mismo valor ($\frac{1}{5}$) para dos situaciones muestrales distintas (la primera muestra con un sólo individuo en paro y la segunda con 3 parados).

Además, se observa que T_2 adopta valores distintos para dos situaciones muestrales que resultan indiferentes. En efecto, la tercera situación se obtiene como permutación de la primera, registrando ambas un sólo individuo parado; sin embargo los valores de T_2 asociados a dichas muestras son $\frac{1}{5}$ y $\frac{2}{5}$ respectivamente.

Definición 5.12. Se dice que T es un *estimador suficiente* de $\theta \in \Theta$ si y sólo si la distribución de una realización muestral $(x_1, \ldots, x_n)$ condicionada a un valor $T = t$ no depende del parámetro θ, esto es, si la expresión: $F(x_1, \ldots, x_n/T) = t)$ no depende de θ.

La definición anterior no resulta de fácil aplicación, ni permite conocer las modificaciones necesarias para transformar el estimador en otro suficiente. Debido a estas limitaciones, el método más habitual para comprobar la suficiencia de los estadísticos es el *teorema de factorización de Fisher-Neyman*:

Teorema 5.2. *Sea $(X_1, \ldots, X_n)$ una m.a.s. de una población X, con función de verosimilitud $L(x_1, \ldots, x_n, \theta)$ con $\theta \in \Theta$ y sea $T = T(X_1, \ldots, X_n)$ un estadístico para estimar θ. Entonces T es suficiente si y sólo si es posible la siguiente factorización:*

$$L(x_1, \ldots, x_n, \theta) = h(x_1, \ldots, x_n)t(t, \theta) \; ; \; \forall (x_1, \ldots, x_n) \in \Re^n$$

donde h es una función no negativa que sólo depende de la muestra $(x_1, \ldots, x_n)$ y g es una función no negativa que sólo depende de θ y del valor del estadístico t.

En este enunciado del teorema de factorización se asume que se cumplen las condiciones de regularidad exigidas por la acotación de Frechet-Cramer-Rao. En otro caso sería necesario que $g(t, \theta)$ coincidiera con la función de densidad del estimador considerado.

Bajo las condiciones de regularidad, un método alternativo para comprobar la suficiencia de un estimador se basa en la cantidad de información. Recordando el propio concepto de suficiencia, diremos que un estimador T de θ es suficiente si y sólo si la cantidad de información contenida en T coincide con la información de la muestra, es decir, si se cumple:

$$I_n(\mathbf{x}, \theta) = I(T, \theta) \; ; \; \forall \mathbf{x} = (x_1, \ldots, x_n) \in \Re^n$$

El concepto de suficiencia lleva asociadas varias propiedades de interés:

Proposición 5.5. *Toda función inyectiva de un estimador suficiente es también suficiente*

Proposición. *Si T_1 y T_2 son estimadores, el primero suficiente y el segundo con error cuadrático medio determinado, entonces es posible obtener otro estimador T_3, función del suficiente y con error cuadrático medio inferior al de T_2.*

La segunda propiedad permite limitar el número de estimadores a considerar en un problema: bastaría con elegir un estimador suficiente, estudiando sólo los que fuesen función de él ya que entre ellos estaría el óptimo, entendido como aquél que minimiza el error cuadrático medio. Como consecuencia de ambas propiedades es posible afirmar que la clase de estimadores eficientes es un subconjunto de la clase de estimadores suficientes.

La media muestral es un estimador suficiente de la esperanza poblacional μ. En concreto, para nuestro ejemplo del gasto normalmente distribuido $X \approx \mathcal{N}(\mu, 1)$, a partir de la función de verosimilitud:

$$L(\mathbf{x}, \mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

se obtiene, desarrollando el sumatorio del exponente:

$$L(\mathbf{x}, \mu) = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu+\bar{x}-\bar{x})^2} = \frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}n(\bar{x}-\mu)} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2}$$

expresión que cumple el criterio de factorización de Fisher-Neyman, ya que:

$$\frac{1}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}n(\bar{x}-\mu)} = g(\bar{x}, \mu) \text{ y } e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\bar{x})^2} = h(x_1, \ldots, x_n)$$

Siguiendo el método alternativo anteriormente comentado, para verificar la suficiencia de la media muestral basta comprobar que su cantidad de información coincide con la de la muestra, esto es:

$$I_n(\mathbf{x}, \mu) = I(\bar{x}, \mu) \; ; \; \forall \mathbf{x} = (x_1, \ldots, x_n) \in \Re^n$$

Para el primer término de la igualdad ya hemos obtenido $I_n(\mathbf{x}, \mu) = n$ y para calcular el segundo basta con tener presente que por ser $X \approx \mathcal{N}(\mu, 1)$, se tiene

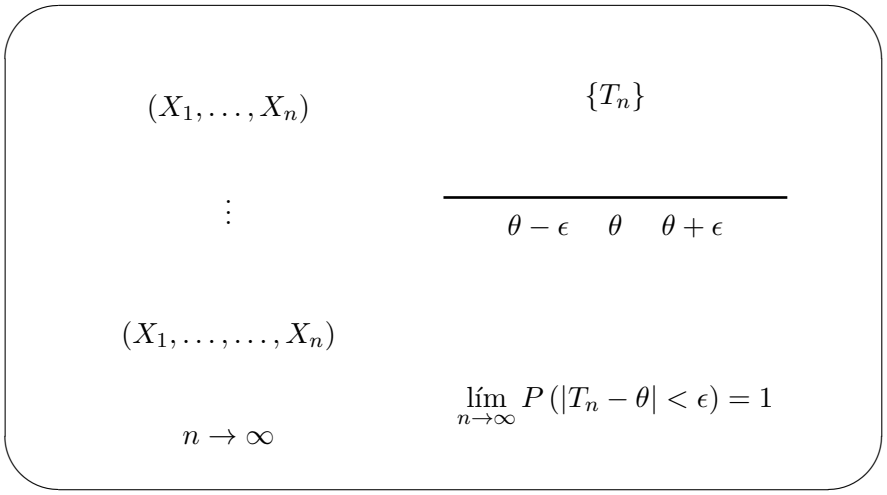
$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{1}{\sqrt{n}}\right)$$

con lo cual su distribución de probabilidad viene dada por:

$$f(\bar{x}, \mu) = \frac{1}{\frac{1}{\sqrt{n}}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\bar{x}-\mu}{\frac{1}{\sqrt{n}}}\right)^2}$$

obteniéndose a partir de ella:

Tabla 5.1.: Consistencia



$$I(\bar{x}, \mu) = -E\left[\frac{\partial^2 \ln f(\bar{x}, \mu)}{\partial \mu^2}\right] = n$$

[Compruébese].

5.4.5. Consistencia

Las propiedades que hemos examinado hasta ahora asumen como dado el tamaño muestral. Sin embargo, parece razonable que cuando la muestra aumente de tamaño se disponga de más información y tengamos una seguridad mayor de que las estimaciones se concentran en torno al verdadero valor del parámetro. Este requisito, denominado *consistencia*, se incluye también entre las propiedades exigidas a los estimadores.

El requisito de consistencia viene ilustrado en el esquema 5.1, donde los aumentos en el tamaño de la muestra se corresponden con estimadores que, cada vez con mayor probabilidad, adoptarán valores en determinado entorno de θ.

Definición 5.13. Si partimos de una muestra cuyo tamaño podemos aumentar indefinidamente ($n \to \infty$) y consideramos la sucesión de estimadores T_n de θ (cada uno de ellos asociado a un tamaño de muestra), se dice que esta sucesión es *consistente* si converge en probabilidad al valor del parámetro. Es decir, la sucesión estima consistentemente a θ si:

$$\forall \epsilon > 0, \forall \theta \in \Theta, \lim_{n\to\infty} P(|T_n - \theta| > \epsilon) = 0$$

Este enunciado puede ser también interpretado en términos del error ya que, a medida que el tamaño muestral aumenta, los errores $e_{T_n} = T_n - \theta$ convergen a 0.

Si consideramos el ejemplo de los gastos mensuales con el que venimos trabajando, para comprobar el requisito de consistencia bastaría con tener en cuenta que $\bar{X} \approx \mathcal{N}\left(\mu, \frac{1}{\sqrt{n}}\right)$, y por tanto $(\bar{X}-\mu)\sqrt{n} \approx \mathcal{N}(0,1)$, con lo cual se obtiene:

$$P\left(|\bar{X}-\mu| < \epsilon\right) = P\left(|\bar{X}-\mu|\sqrt{n} < \epsilon\sqrt{n}\right) = P\left(-\epsilon\sqrt{n} < (\bar{X}-\mu)\sqrt{n} < \epsilon\sqrt{n}\right) = 2F_X\left(\epsilon\sqrt{n}\right) - 1$$

y bastaría con tomar valores suficientemente elevados de n para que la probabilidad anterior se aproxime a 1 (y en consecuencia su complementaria a 0) tanto como queramos.

La aplicación general del criterio de consistencia al estimador media muestral puede ser efectuada gracias a la ley débil de los grandes números, cuyo postulado es:

$$\forall\epsilon,\ \lim_{n\to\infty} P\left(|\bar{X}-\mu| > \epsilon\right) = 0$$

y permite calificar a la media muestral de estimador consistente de la media poblacional.

Una formulación alternativa del criterio de consistencia, viene dada en los siguientes términos: dada una sucesión de estimadores T_n del parámetro θ, se dice que T_n es consistente para θ si se cumple:

$$\lim_{n\to\infty} E(T_n) = \theta\,,\ \lim_{n\to\infty} Var(T_n) = 0$$

Aunque el concepto de consistencia hace referencia a una sucesión de estimadores, habitualmente se presupone que todos sus miembros gozan de las mismas propiedades, hablando así de estimadores consistentes.

5.5. Métodos de obtención de estimadores

En el epígrafe anterior nos hemos ocupado de las propiedades que debe verificar un buen estimador, pero no hemos abordado el problema de cómo obtenerlos.

Aunque existen varios procedimientos alternativos, posiblemente el primer criterio al que acudiríamos para construir estimadores sería el de *analogía*.

Definición 5.14. Para estimar cierta característica poblacional, denominamos *estimador analógico* a la correspondiente expresión muestral.

Este método es altamente intuitivo, pero sin embargo resulta poco riguroso ya que no disponemos de herramientas para comprobar de modo general si los estimadores analógicos cumplen o no las propiedades consideradas deseables.

5.5.1. Método de la máxima verosimilitud

El procedimiento más empleado para la obtención de estimadores es el método de la *máxima verosimilitud*, debido a las buenas propiedades que presentan los estimadores que genera.

Tabla 5.2.: Estimación máximo verosímil

Zonas	Tasa de paro	Verosimilitud	EMV
Europa	12 %	L=0,0098	$\hat{\mathbf{p}} = \mathbf{12}\,\%$
EEUU	6,2 %	L=0,0032	
Japón	2,5 %	L=0,0007	

El planteamiento subyacente a este método es muy intuitivo, tal y como ilustramos en el siguiente ejemplo, y consiste en aprovechar la información muestral para obtener estimaciones verosímiles de los parámetros desconocidos.

Consideremos una nueva ilustración basada en nuestro ejemplo de la tasa de paro: disponemos de una muestra aleatoria de 5 trabajadores procedentes de una delegación cuyo origen desconocemos. Para mayor comodidad asumiremos que existen sólo las tres posibilidades siguientes:

Si la muestra fuese la recogida en el esquema de la figura 5.2 ¿cuál sería el origen más verosímil de los trabajadores? o, dicho de otro modo, ¿cuál sería la estimación máximo-verosímil de la tasa de paro p? Para responder a esta pregunta utilizaremos la información disponible, obteniendo la verosimilitud de la muestra de trabajadores para cada valor posible de p:

- $L(0,1,1,0,0,p_{EUROPA}) = 0,12^2(1-0,12)^3 = 0,0098$
- $L(0,1,1,0,0,p_{EEUU}) = 0,062^2(1-0,062)^3 = 0,0032$
- $L(0,1,1,0,0,p_{JAPÓN}) = 0,028^2(1-0,028)^3 = 0,0007$

A partir de estos resultados podemos calificar de más verosímil el primer supuesto (Europa, con tasa de paro del 12 %), ya que la muestra efectivamente extraída, con dos trabajadores en paro, resulta más verosímil o creible en ese caso.

Es fácil observar que en el método de máxima verosimilitud la muestra desempeña un papel central. En definitiva, el ejemplo anterior se limita a considerar que, si la muestra es representativa de la población, la muestra mantendrá la misma estructura de la población y por tanto resultará más probable bajo la composición correcta que bajo otra cualquiera.

Como hemos visto, el método de máxima verosimilitud consiste en elegir la estimación del parámetro que maximiza la función de verosimilitud muestral. Su idea es muy sencilla, ya que conduce a los valores del parámetro que hacen más probable la selección de la muestra realmente obtenida.

El principio de máxima verosimilitud puede ser descrito en los siguientes términos:

Definición 5.15. Se llama *estimación máximo verosímil* del parámetro θ al valor, si existe, $\hat{\theta} = T(x_1, \ldots, x_n)$ que maximiza la función de verosimilitud, esto es, un valor $\hat{\theta}$ tal que:

$$L(x_1, \ldots, x_n\hat{\theta}) = \sup_{\theta \in \Theta} L(x_1, \ldots, x_n)$$

Al estimador correspondiente se denomina *estimador máximo verosímil* de θ o abreviadamente $EMV(\theta)$.

Aunque la maximización de la función de verosimilitud suele conducir a un valor único $\hat{\theta}$, no puede garantizarse que esto sea cierto en general.

Además en ciertas ocasiones el valor $\hat{\theta}$ que hace máxima la función de verosimilitud no pertenece al espacio paramétrico Θ, concluyéndose entonces que el EMV de θ no existe.

La obtención práctica del estimador máximo verosímil parte de la pregunta ¿cuál es la probabilidad de selección de una muestra determinada? Para responderla es necesario construir la función de verosimilitud $L(x_1, \ldots, x_n, \theta)$ que depende de la distribución de la población y por tanto del parámetro desconocido.

Una vez construida la función L (producto de funciones de densidad o de probabilidad según la variable aleatoria sea continua o discreta) debemos buscar el valor del parámetro θ que maximice la probabilidad de la muestra, esto es:

$$\sup_{\theta} L(x_1, \ldots, x_n, \theta)$$

Dado que muchas de las funciones de verosimilitud presentan expresiones exponenciales, en la práctica resulta habitual -para aumentar la operatividad- trabajar con la función de verosimilitud transformada mediante logaritmos neperianos:

$$\sup_{\theta} \ln L(x_1, \ldots, x_n, \theta)$$

ya que dicha transformación (por ser el logaritmo una función monótona) linealiza las expresiones a maximizar, sin que ello afecte a sus puntos extremos.

La condición necesaria de extremo sería:

$$\frac{\partial \ln L(x_1 \ldots, x_n, \theta)}{\partial \theta} = 0$$

obteniéndose a partir de esta igualdad la EMV de θ, que denotamos por $\hat{\theta}$.

Este valor debe además verificar la condición suficiente de máximo:

$$\frac{\partial^2 \ln L(x_1, \ldots, x_n, \theta)}{\partial \theta^2} < 0$$

El método de máxima verosimilitud es consistente con la segunda de las interpretaciones contempladas al definir la función de verosimilitud L, esto es, asume como fijos los datos muestrales, dependiendo su valor del parámetro desconocido.

El método de máxima verosimilitud traslada importantes propiedades a sus estimadores, lo que es una buena garantía para utilizar EMV.

Proposición 5.6. *Entre las principales propiedades de los estimadores máximo verosímiles destacan las siguientes:*

- *Bajo condiciones generales, los estimadores máximo verosímiles son consistentes y además su distribución converge a una normal de esperanza θ y varianza $\frac{1}{I_n}$.*
- *El estimador máximo verosímil es invariante; es decir, si T es un estimador máximo verosímil de θ entonces $g(T)$ lo es de $g(\theta)$, siendo g una aplicación entre intervalos abiertos.*
- *Si existe un estimador suficiente, entonces el estimador máximo verosímil es función de él. Si además existe un estimador de mínima varianza, éste es de máxima verosimilitud.*
- *Los estimadores máximo verosímiles son asintóticamente eficientes.*

Estamos ahora en condiciones de deducir el estimador máximo verosímil de la tasa de paro p.

Siguiendo el método anteriormente expuesto, partimos de una realización muestral buscando el valor $\hat{p}$ que haga máxima la expresión:

$$L(\mathbf{x}, p) = p^{\sum_{i=1}^{n} x_i}(1-p)^{n-\sum_{i=1}^{n} x_i}$$

que, linealizada mediante logaritmos neperianos, conduce a:

$$\ln L(\mathbf{x}, p) = \sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)$$

La condición necesaria de extremo, $\frac{\partial \ln L(\mathbf{x},p)}{\partial p} = 0$, sería entonces:

$$\frac{\partial}{\partial p}\left[\sum_{i=1}^{n} x_i \ln p + \left(n - \sum_{i=1}^{n} x_i\right) \ln(1-p)\right] = 0 \Rightarrow \frac{1}{p}\sum_{i=1}^{n} x_i - \frac{1}{1-p}\left(n - \sum_{i=1}^{n} x_i\right) = 0$$

cuya solución conduce al valor:

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

[Compruébese que se cumple también la condición suficiente de máximo]

Con nuestra realización muestral anterior (0,1,1,0,0) se obtendría el EMV $\hat{p} = \frac{2}{5} = 0,4$; es decir, una tasa de paro del 40 % maximiza la verosimilitud de la muestra observada.

Obsérvese que el proceso seguido parte de una realización muestral concreta $(x_1, \ldots, x_n)$, por lo cual proporciona una estimación máximo verosímil (que será la solución de la ecuación a la que conduce la condición de extremo). El estimador máximo verosímil vendrá dado por la correspondiente expresión aleatoria, función de la muestra genérica $(X_1, \ldots, X_n)$, que en el ejemplo anterior sería:

$$EMV(p) = \frac{\sum_{i=1}^{n} X_i}{n}$$

De modo análogo, para deducir el EMV de μ en el ejemplo de los gastos mensuales, deberíamos partir de la función de verosimilitud dada por la expresión:

$$L(\mathbf{x}, \mu) = \frac{2}{(2\pi)^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu)^2}$$

que una vez linealizada da lugar a la expresión:

$$\ln L(\mathbf{x}, \mu) = \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 - \frac{n}{2}\ln(2\pi)$$

Si ahora derivamos respecto al parámetro e igualamos a cero, se tiene el EMV:

Tabla 5.3.: Método de los momentos

Población X Momentos Poblacionales		Muestra $(X_1, \ldots, X_n)$ Momentos muestrales
$E(X)$	$\longleftrightarrow$	$\frac{\sum_{i=1}^n X_i}{n}$
$E(X^2)$	$\longleftrightarrow$	$\frac{\sum_{i=1}^n X_i^2}{n}$
$\vdots$	$\longleftrightarrow$	$\vdots$
$E(X^k)$	$\longleftrightarrow$	$\frac{\sum_{i=1}^n X_i^k}{n}$

Sistema de k ecuaciones e incógnitas $\theta_1, \ldots \theta_k$

$$\frac{\partial \ln L(\mathbf{x}, \mu)}{\partial \mu} = 0 \Rightarrow \sum_{i=1}^{n}(x_i - \mu) = 0 \Rightarrow \hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

con lo cual el correspondiente estimador máximo verosímil vendría dado por la expresión:

$$EMV(\mu) = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

5.5.2. Método de los momentos

Otro procedimiento para la obtención de estimadores es el método de los momentos, basado en la conexión entre población y muestra.

Cuando nos planteamos aumentar indefinidamente el tamaño muestral, extendiendo la selección al conjunto poblacional parece lógico asumir que los resúmenes de la información muestral se aproximen a los parámetros desconocidos. De hecho, este es el planteamiento subyacente en el requisito de consistencia, que inspira también un nuevo procedimiento de obtención de estimadores llamado *método de los momentos*.

Definición 5.16. La estimación de k parámetros por el *método de los momentos* consiste en resolver el sistema de ecuaciones resultante de igualar los k primeros momentos poblacionales, si existen, $\alpha_1, \ldots, \alpha_k$ a los correspondientes momentos muestrales $a_1, \ldots, a_k$.

Dado que los momentos muestrales son estadísticos obtenidos a partir de la muestra y los momentos poblacionales dependen de la distribución probabilística de la variable aleatoria X y por tanto del correspondiente parámetro (o parámetros), la igualación de los momentos da lugar a un sistema de tantas ecuaciones como incógnitas (k parámetros), cuya resolución proporciona los valores considerados como estimaciones de los parámetros.

Denominando $m_1, \ldots, m_k$ a los momentos muestrales se llegaría mediante el procedimiento descrito a un sistema de k ecuaciones con incógnitas $\theta_1, \ldots, \theta_k$:

$$\begin{aligned}\theta_1 &= h_1(m_1, \ldots, m_k)\\ \theta_2 &= h_2(m_1, \ldots, m_k)\\ &\vdots\\ \theta_k &= h_k(m_1, \ldots, m_k)\end{aligned}$$

El método de los momentos resulta de aplicación más sencilla que el de máxima verosimilitud y conduce a estimadores consistentes. Sin embargo, su utilización es menos generalizada ya que los estimadores máximo verosímiles suelen resultar más eficientes.

5.5.3. Método de los mínimos cuadrados

Definición 5.17. El *método de los mínimos cuadrados* permite obtener estimadores basándose en minimizar la suma de las desviaciones cuadráticas entre las observaciones y sus valores esperados, esto es, la expresión:

$$\sum_{i=1}^{n}(X_i - E(X_i))^2$$

en la que $E(X_i)$ serán función de los parámetros desconocidos.

La aplicación de este método resulta operativa en el caso de que $E(X_i)$ sea una función lineal de los parámetros que deseamos estimar, pero sin embargo no proporciona estimadores para parámetros que no figuren en los valores esperados de las observaciones.

La obtención de *estimadores mínimo cuadráticos* (EMC) resulta habitual en los procesos de regresión, cuando una variable aleatoria Y presenta un valor esperado que es función lineal de una o varias características X:

$$E(Y) = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Los procedimientos alternativos que hemos examinado conducen en ocasiones a expresiones coincidentes para estimar determinado parámetro, llegando a menudo incluso a los estimadores analógicos que aconsejaba nuestra intuición inicial. Así la media poblacional μ sería estimada a través de la media muestral $\bar{X}$, la proporción poblacional p mediante la proporción muestral $\hat{p}$, etc.

Los métodos anteriormente recogidos, aunque habituales, no agotan las posibilidades para la construcción de estimadores. Así, en el caso de que las observaciones se agrupen en intervalos o bien sean frecuencias de sucesos disjuntos resulta aconsejable el *método de la chi-cuadrado mínima*, que consiste en minimizar la medida de discrepancia chi-cuadrado entre las frecuencias observadas y teóricas (estas últimas dadas en términos de la probabilidad de la variable y por tanto en función de los paráme-

tros). Una ventaja de este método es que los estimadores a los que conduce satisfacen el requisito de consistencia.

5.6. Algunos estimadores habituales

Las posibilidades inferenciales son ilimitadas como consecuencia de la diversidad de parámetros de los que puede depender la distribución de una magnitud aleatoria. Sin embargo, en la práctica los parámetros habitualmente investigados se corresponden con las principales características poblacionales: la esperanza μ, la varianza σ^2 y la proporción p.

Supongamos por ejemplo que nos interesa la v.a. X: "Consumo de electricidad de los hogares". Aunque por su propio carácter aleatorio no será posible llegar a anticipar el comportamiento de este consumo, si disponemos de información muestral podremos reducir nuestra incertidumbre inicial.

Más concretamente, si asumimos para X un modelo probabilístico determinado dependiente de uno o varios parámetros desconocidos, podríamos formular algunos interrogantes del tipo siguiente: ¿cuál es el consumo mensual esperado? ¿qué nivel de dispersión existe entre los consumos de los hogares? ¿qué proporción de hogares superan cierta cifra de consumo mensual?

Preguntas similares a las anteriores aparecen con gran frecuencia y una respuesta adecuada para ellas puede condicionar en gran medida el éxito de una investigación. Por tanto, resulta aconsejable examinar en detalle las expresiones idóneas para la estimación de las características poblacionales de interés.

En los apartados que siguen asumiremos que disponemos de m.a.s. de tamaño n $(X_1, \ldots, X_n)$ seleccionadas de la población considerada y analizaremos los estimadores adecuados para aproximar con la información muestral los parámetros objeto de estudio.

5.6.1. Parámetro media poblacional μ

Si nuestro primer objetivo es aproximar el consumo esperado μ, parece lógico resumir la información muestral calculando el promedio de consumo para los meses observados.

El estimador analógico *media muestral* resulta muy adecuado para llevar a cabo inferencias sobre la media poblacional $\mu = E(X)$. A partir de una m.a.s. $(X_1, \ldots, X_n)$ la media muestral es una nueva variable aleatoria definida por la expresión:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Por lo que respecta a las características de esta variable aleatoria, se obtiene fácilmente:

$$E\left(\bar{X}\right) = \mu \; ; \; Var\left(\bar{X}\right) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \; ; \; \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

[Efectuar la deducción de las características anteriores]

La media muestral es un estimador insesgado de la media poblacional y su riesgo viene dado en función de dos características: la dispersión poblacional y el tamaño de la muestra.

Para sucesivas realizaciones muestrales el estimador media muestral adoptaría valores $\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_k$ obteniéndose como valor esperado de todos ellos la esperanza poblacional μ.

La dispersión asociada a la media muestral viene recogida por su varianza $Var(\bar{X})$ o la correspondiente raíz cuadrada, denominada error estándar de la media.

Definición 5.18. El *error estándar de la media muestral* se obtiene como cociente entre la desviación típica poblacional y la raíz cuadrada del tamaño muestral:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Esta es la medida de dispersión habitualmente utilizada para la media muestral, ya que -a diferencia de la varianza- viene expresada en las mismas unidades que la variable aleatoria X. Como podemos apreciar en su expresión, el error estándar de la media aparece relacionado directamente con la dispersión poblacional e inversamente con el tamaño de la muestra.

Cuando seleccionamos una única observación de la variable aleatoria X la desviación estándar (que aproxima su dispersión respecto al valor esperado $E(X) = \mu$) viene dada por el parámetro σ. Si en cambio seleccionamos una muestra aleatoria simple de n elementos $(X_1, \ldots, X_n)$, el riesgo o dispersión respecto al valor esperado disminuirá a medida que aumenta el tamaño de la muestra.

Las expresiones anteriores han sido obtenidas para el supuesto de muestreo aleatorio simple en poblaciones infinitas o bien con reposición en el caso finito, que hemos adoptado como situación de referencia. Sin embargo, como ya hemos comentado, en la práctica resultan habituales otras técnicas de selección cuyas características serán analizadas con detalle en un capítulo específico dedicado al muestreo en poblaciones finitas.

En concreto, en la práctica es frecuente la selección de muestras aleatorias en las que cada elemento poblacional puede aparecer una vez a lo sumo, esto es, los *muestreos aleatorios sin reposición o sin reemplazamiento*. Las consecuencias de este cambio en el procedimiento de muestreo -que serán tratadas con detalle en un capítulo posterior- aparecen recogidas en la figura 5.4.

Como ya hemos visto, las condiciones de independencia no son necesarias para calcular

Tabla 5.4.: Media muestral en el muestreo aleatorio simple

	método de muestreo	
	con reposición	sin reposición
Esperanza	$E\left(\bar{X}\right)=\mu$	$E\left(\bar{X}\right)=\mu$
Varianza	$Var\left(\bar{X}\right)=\frac{\sigma^2}{n}$	$Var\left(\bar{X}\right)=\frac{N-n}{N-1}\frac{\sigma^2}{n}$
Error estándar	$\sigma_{\bar{X}}=\frac{\sigma}{\sqrt{n}}$	$\sigma_{\bar{X}}=\sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$

el valor esperado de una suma, por lo cual se sigue cumpliendo en este caso $E\left(\bar{X}\right)=\mu$. Sin embargo, se producirán cambios en las medidas de dispersión (varianza y error estándar), dado que el riesgo disminuye en el muestreo sin reposición, como consecuencia de la garantía de no poder observar más de una vez un mismo elemento.

La expresión de la varianza sería ahora:

$$Var\left(\bar{X}\right)=\frac{N-n}{N-1}\frac{\sigma^2}{n}$$

denominándose *factor de corrección* a la expresión $\frac{N-n}{N-1}$ que adoptará valores inferiores a la unidad siempre que $n>1$.

5.6.2. Parámetro varianza poblacional σ^2

Una vez que conocemos los consumos esperados, puede ser relevante investigar su dispersión poblacional, esto es, cuantificar en qué medida los hogares realizan consumos homogéneos o existen discrepancias entre las cifras de consumo de electricidad.

Si nuestro objetivo es efectuar inferencias sobre la varianza poblacional σ^2 una primera posibilidad podría ser partir del *estimador analógico de la varianza*, que vendría dado por la expresión aleatoria:

$$S_n^2=\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n}$$

No obstante, al examinar las características de esta expresión se comprueba que no resulta muy adecuada, por ser un estimador sesgado del parámetro σ^2.

En efecto, desarrollando la esperanza de S_n^2 se obtiene:

$$\begin{aligned}E(S_n^2)= & E\left[\frac{\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}{n}\right]=E\left(\frac{\sum_{i=1}^{n}X_i^2}{n}-\bar{X}^2\right)=\frac{1}{n}E\left(\sum_{i=1}^{n}X_i^2\right)-E\left(\bar{X}^2\right)= \\ = & E\left(X_i^2\right)-E\left(\bar{X}^2\right)=Var(X_i)+E^2(X_i)-Var\left(\bar{X}\right)-E^2\left(\bar{X}\right)= \\ = & \sigma^2+\mu^2-\frac{\sigma^2}{n}-\mu^2=\frac{n-1}{n}\sigma^2\end{aligned}$$

igualdad en la que hemos aplicado la definición de varianza para las variables X_i y para la media muestral.

El resultado anterior permite afirmar que el estimador analógico de la varianza poblacional subestima su verdadero valor, ya que conlleva un sesgo negativo:

$$B_{S_n^2}\left(\sigma^2\right) = \frac{-\sigma^2}{n}$$

[Compruébese]

Con el objetivo de solucionar las limitaciones que presenta en cuanto a la estimación insesgada la expresión muestral analógica de σ^2, definiremos la varianza muestral S^2 como la expresión:

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{n-1}$$

Dado que este estimador se diferencia del anterior S_n^2 sólo en el denominador, se verifica la relación

$$S^2 = \frac{n}{n-1}S_n^2$$

a partir de la cual se comprueba fácilmente que S^2 sí es un estimador insesgado de la varianza poblacional:

$$E(S^2) = \frac{n}{n-1}E\left(S_n^2\right) = \frac{n}{n-1}\cdot\frac{n-1}{n}\sigma^2 = \sigma^2$$

En estadística clásica resulta común la denominación de varianza muestral para el estimador analógico, designando al estimador insesgado S^2 cuasivarianza muestral. Sin embargo, hemos considerado más conveniente utilizar el término varianza muestral para la expresión S^2, que será la utilizada en todos los estudios inferenciales sobre la varianza.

Como hemos visto, la varianza de la media muestral y su error estándar dependen de la varianza poblacional σ^2. Por tanto la estimación de estas características se efectuará también a partir de la varianza muestral anteriormente definida. Así se obtiene:

$$S_{\bar{X}}^2 = \widehat{Var}\left(\bar{X}\right) = \frac{S^2}{n}\ ;\ S_{\bar{X}} = \frac{S}{\sqrt{n}}$$

La ausencia de sesgo es una propiedad de gran interés, que justifica la utilización de la varianza muestral S^2 en lugar de la expresión analógica S_n^2. Sin embargo, ello no permite concluir que el estimador S^2 sea "mejor", ya que sería necesario estudiar lo que ocurre con las restantes propiedades.

En concreto, si se admite el supuesto de normalidad para la población estudiada X, es posible demostrar que el estimador analógico S_n^2 presenta un menor error cuadrático medio (y por tanto una mayor eficiencia relativa) que la varianza muestral S^2.

También comprobaremos más adelante que en poblaciones finitas cuando las unidades observadas no se reponen a la población, el comportamiento de los estimadores difiere del obtenido aquí, puesto que la varianza muestral es un estimador sesgado.

5.6.3. Parámetro proporción poblacional p

La tercera pregunta planteada sobre nuestro ejemplo era ¿qué proporción de hogares realizan consumos de electricidad superiores a una cifra determinada? En este caso la situación consiste en describir una característica e investigar la proporción poblacional asociada a la misma.

Las inferencias sobre la proporción son frecuentes cuando trabajamos con características cualitativas (este es el caso de la tasa de paro, el porcentaje de votantes a favor de cierto candidato, la cuota de mercado de cierto producto, ...) y el estimador adecuado en estas situaciones es la proporción muestral que denominamos $\hat{p}$.

La *proporción muestral* se define como

$$\hat{p} = \frac{X}{n}$$

donde X es la v.a. que recoge el número de elementos de la muestra que presentan la característica analizada.

Así pues, se tiene ahora un caso particular de m.a.s. $(X_1, \ldots, X_n)$ donde las variables X_i son dicotómicas o Bernoulli:

$X_i = 1$	si se presenta la característica investigada	$P(X_i = 1) = p$
$X_i = 0$	en otro caso	$P(X_i = 0) = 1 - p$

siendo por tanto el numerador de la proporción muestral $X = \sum_{i=1}^{n} X_i$ una v.a. distribuida según un modelo binomial $\mathcal{B}(n, p)$.

Como consecuencia, se obtienen las siguientes características para la proporción muestral:

$$E(\hat{p}) = E\left(\frac{X}{n}\right) = \frac{1}{n}E(X) = \frac{np}{n} = p$$

$$Var(\hat{p}) = Var\left(\frac{X}{n}\right) = \frac{1}{n^2}Var(X) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

El estimador proporción muestral podría ser analizado como un caso particular de la media muestral para variables dicotómicas. No obstante, presenta como rasgo diferencial la presencia del parámetro p tanto en la esperanza como en la varianza del estimador, por lo cual resulta conveniente trabajar con estimaciones de la varianza, dadas por la expresión:

$$S^2(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n-1}$$

Puede comprobarse que esta expresión es insesgada para estimar $Var(\hat{p})$, ya que se obtiene:

$$\begin{aligned} E\left[S^2\left(\hat{p}\right)\right] = & \quad E\left[\frac{\hat{p}(1-\hat{p})}{n-1}\right] = \frac{1}{n-1}E\left(\hat{p}-\hat{p}^2\right) = \frac{1}{n-1}E\left(\hat{p}\right) - E\left(\hat{p}^2\right) = \\ = & \quad \frac{1}{n-1}\left[p - Var\left(\hat{p}\right) - E\left(\hat{p}^2\right)\right] = \frac{1}{n-1}\left[p(1-p) - Var\left(\hat{p}\right)\right] = \\ = & \quad \frac{1}{N-1}\left[nVar\left(\hat{p}\right) - Var\left(\hat{p}\right)\right] = \frac{(n-1)Var(\hat{p})}{n-1} = Var\left(\hat{p}\right) \end{aligned}$$

6. Herramientas inferenciales

Como hemos comentado en capítulos anteriores, la información muestral es el punto de partida para un amplio abanico de procesos inferenciales. Dichos procesos se basan en la información disponible y tienen como objetivo reducir la incertidumbre, que puede ir referida a parámetros concretos o a las poblaciones en su conjunto.

Cuando las inferencias que realizamos van referidas a características poblacionales concretas, es necesaria una etapa de diseño de estimadores que ya hemos abordado en el capítulo anterior. Una vez que dispongamos de estimadores adecuados para los parámetros de interés, debemos conectar sus expresiones con modelos probabilísticos conocidos, tarea de la que nos ocuparemos en este tema. En algunos casos será posible adaptar las expresiones a modelos empíricos ya estudiados, mientras que en otras situaciones las necesidades muestrales obligan a definir otra serie de distribuciones de carácter "artificial" cuya finalidad son precisamente los procesos inferenciales.

Cuando las inferencias son de carácter genérico (por ejemplo, si contrastamos hipótesis relativas al conjunto de la población) debemos aprovechar la información muestral, construyendo expresiones que permitan efectuar afirmaciones probabilísticas sobre nuestras conclusiones inferenciales.

6.1. Modelos probabilísticos asociados al muestreo

En este apartado analizamos las distribuciones de probabilidad usuales en los estudios inferenciales. Con excepción del modelo normal, que ya ha sido estudiado y ocupa un lugar central en los estudios empíricos, estas distribuciones muestrales pueden ser calificadas de "artificiales" por tratarse de modelos no observables en la realidad.

En efecto, las distribuciones probabilísticas asociadas a los procesos inferenciales no tienen por objeto describir el comportamiento de magnitudes aleatorias sino que se trata de construcciones "de laboratorio" que aparecen asociadas a ciertas expresiones muestrales bajo determinados supuestos. Sus distribuciones de probabilidad aparecen tabuladas y serán herramientas imprescindibles en los análisis inferenciales.

Como veremos en los apartados que siguen, estas distribuciones muestrales (chi-cuadrado, t de Student, F de Snedecor) se introducen asumiendo ciertos supuestos o hipótesis sobre la población de partida. Por tanto, resulta interesante conocer en qué medida se ven afectadas las distribuciones por la alteración de dichos supuestos, concepto que se conoce como robustez.

Una distribución, y los procesos inferenciales basados en la misma, se denominan robustos cuando no resultan muy sensibles a los cambios en los supuestos de partida, es decir, cuando no presentan

alteraciones graves ante el incumplimiento de las hipótesis poblacionales.

6.1.1. Distribución Normal

Al examinar los principales modelos probabilísticos útiles en el ámbito de la inferencia estadística debemos ocuparnos en primer lugar de la *distribución normal*, que a su importancia en la descripción de magnitudes económicas y como límite de agregados une ahora su interés desde una óptica inferencial.

Consideremos una m.a.s. $(X_1, \ldots, X_n)$ a partir de la cual es posible definir la v.a. media muestral:

$$\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

Como ya hemos visto en capítulos anteriores, existen diferentes situaciones en las que esta expresión seguirá una distribución normal:

- Siempre que la población de partida X se distribuya normalmente ($X \approx \mathcal{N}(\mu, \sigma)$), la propiedad de reproductividad garantiza para la media muestral:

$$\bar{X}_n \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

[Este resultado aparece recogido en un capítulo anterior, epígrafe 4.5]

- Aun cuando se desconozca el modelo poblacional de partida, los teoremas límites permiten afirmar que

$$\bar{X}_n \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

aproximación que suele realizarse para tamaños muestrales $n > 30$. Esta convergencia generaliza de modo considerable la aplicación del modelo normal como distribución de la media muestral.

Un caso particular de esta segunda situación se presenta cuando la muestra $(X_1, \ldots, X_n)$ está formada por variables dicotómicas o de Bernoulli. Se obtiene en este caso una suma distribuida según un modelo binomial $\mathcal{B}(n, p)$ que, gracias al Teorema de De Moivre puede ser aproximada para tamaños elevados por una distribución normal:

$$S_n \approx \mathcal{N}\left(np, \sqrt{np(1-p)}\right)$$

y en consecuencia

$$\bar{X}_n \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

6.1.2. Distribución chi-cuadrado

El modelo chi-cuadrado aparece conectado con la distribución normal al venir definido en los siguientes términos:

Definición 6.1. Dadas n v.a. $X_1, \ldots, X_n$ independientes y distribuidas según un modelo $\mathcal{N}(0,1)$, se define la v.a. *chi-cuadrado (o ji-cuadrado)* con n grados de libertad, que denotamos por χ^2_n, como:

$$\chi^2_n = \sum_{i=1}^{n} X_i^2$$

Consideremos una población normal estándar, $X \approx \mathcal{N}(0,1)$ y sea $(X_1, \ldots, X_n)$ una muestra aleatoria simple de esa población. Entonces la variable $\chi^2_n = \sum_{i=1}^{n} X_i^2$ sigue una distribución chi-cuadrado (o ji-cuadrado) con n grados de libertad (basta tener en cuenta que las componentes de una muestra genérica son independientes e idénticamente distribuidas).

Cuando partimos de una población normal genérica $Y \approx \mathcal{N}(\mu, \sigma)$, y consideramos una muestra aleatoria $(Y_1, \ldots, Y_n)$, entonces la suma de los cuadrados de las variables muestrales tipificadas se distribuye según un modelo chi-cuadrado, con n grados de libertad:

$$\sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right) \approx \chi^2_n$$

La justificación en este caso resulta sencilla con sólo llevar a cabo un proceso de tipificación sobre la muestra, definiendo variables $X_i = \left(\frac{Y_i - \mu}{\sigma}\right)$distribuidas según modelos $\mathcal{N}(0,1)$, a partir de las cuales se obtiene de forma inmediata la distribución chi-cuadrado anteriormente definida.

Como consecuencia de su definición, esta variable adopta valores no negativos, y su distribución de probabilidad viene caracterizada por el parámetro n, que recoge el número de sumandos que intervienen en su definición y se denomina *grados de libertad* (g.l.).

Los grados de libertad asociados a una expresión pueden ser interpretados como "número de valores que es posible fijar de modo arbitrario" y aparecen relacionados con el número de variables o tamaño muestral n. Una muestra de tamaño n tiene n grados de libertad, pues no establecemos ninguna restricción sobre los valores que pueden obtenerse en cada componente y éstos se eligen libremente. Por extensión, un estadístico definido a partir de esa muestra también tiene n grados de libertad, salvo que su expresión esté sometida a alguna restricción, en cuyo caso los niveles de holgura o libertad se reducen.

Para ilustrar el concepto de grados de libertad supongamos una población $\mathcal{N}(0,1)$ a partir de la cual extraemos una muestra aleatoria simple de tamaño $n = 3$ (X_1, X_2, X_3) y definimos el estadístico media aritmética. Tanto sobre la expresión muestral como sobre la media podemos seleccionar arbitrariamente 3 valores, por lo cual éste es el número de grados de libertad existentes.

Imaginemos por ejemplo, $x_1 = 4$, $x_2 = 2$ y $x_3 = 9$, con lo cual se obtendría la media $\bar{x} = 5$; hemos elegido 3 valores por lo cual el número de g.l. es $n = 3$ (podríamos también haber fijado dos valores y la media, con lo cual quedaría determinado el tercer valor; por tanto los g.l. siguen siendo 3).

Supongamos ahora que definimos la expresión:

$$\sum_{i=1}^{3}(X_i - \bar{X})^2$$

Resulta sencillo comprobar que en ella podemos seleccionar únicamente dos sumandos, ya que el tercero quedará automáticamente determinado. Así, a modo de ejemplo, con la muestra anterior se tendría $(x_1 - \bar{X}) = -1$, $(x_2 - \bar{X}) = -3$ y la tercera desviación deberá ser obligatoriamente $(x_3 - \bar{X}) = 4$ para que se cumpla la propiedad

$$\sum_{i=1}^{3}(X_i - \bar{X}) = 0$$

En definitiva, se aprecia que existe una restricción: $\sum_{i=1}^{3}(X_i - \bar{X}) = 0$, equivalente a la definición de la media

$$\bar{X} = \frac{\sum_{i=1}^{3} X_i}{3}$$

Como consecuencia, se reducen en uno los grados de libertad de la muestra, de modo que la expresión presenta en este caso 2 g.l.

En el caso de una muestra de tamaño n, la expresión $\sum_{i=1}^{n}(X_i - \bar{X})^2$ tendría $n - 1$ grados de libertad. De hecho, estos g.l. coinciden con el denominador del estimador insesgado varianza muestral

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n - 1}$$

A modo de resumen, la tabla siguiente recoge los grados de libertad asociados a expresiones genéricas con y sin restricción.

Expresión	Variables aleatorias	Restricciones	g.l.
$\sum_{i=1}^{n} X_i^2$	$X_1, \ldots, X_n$		n
$\sum_{i=1}^{n} (X_i - \bar{X})^2$	$X_1, \ldots, X_n$	$\frac{\sum_{i=1}^{n} X_i}{n} = \bar{X}$	n-1
	o bien		
	$X_1 - \bar{X}, \ldots, X_n - \bar{X}$	$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$	n-1

En general, para una muestra de tamaño n agrupada en k intervalos o clases, los grados de libertad serán $k - 1$ ya que, una vez especificadas $k - 1$ frecuencias, la frecuencia restante n_k vendrá determinada como $n - \sum_{i=1}^{k-1} n_i$.

Razonando de modo análogo, dada una muestra de tamaño n si se adoptan como constantes k funciones de los valores muestrales, el número de grados de libertad vendrá reducido en k.

La función de densidad del modelo chi-cuadrado para n g.l. viene dada por la expresión:

Tabla 6.1.: Modelo χ^2. Función de distribución

g.l. $n\rightarrow$ $\downarrow$Valores χ^2_n	5	10	20
5	0,5841	0,1088	0,0003
10	0,9248	0,5595	0,0318
20	0,9988	0,9709	0,5421

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}\ ,\ x > 0$$

donde $\Gamma\left(\frac{n}{2}\right)$ representa la función matemática gamma en el punto $\frac{n}{2}$.

La expresión de esta función de densidad puede obtenerse en dos etapas: en la primera, se parte de una variable $X_i \approx \mathcal{N}(0,1)$ efectuando sobre la misma el cambio de variable $Y_i = X_i^2$, con lo cual se obtiene para Y_i una función de densidad que corresponde a un modelo gamma de parámetros $p = \frac{1}{2}, a = \frac{1}{2}$. En la segunda etapa, teniendo en cuenta que las X_i son v.a. independientes e idénticamente distribuidas (i.i.d.), es posible aplicar la reproductividad del modelo gamma respecto al parámetro p; así se tiene:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} X_i^2 \approx \gamma\left(p = \frac{n}{2}, a = \frac{1}{2}\right)$$

La probabilidad de que χ^2_n tome valores en un intervalo $[a, b]$ sería la integral entre estos límites de la expresión anterior que sólo puede resolverse mediante métodos numéricos, por lo cual el modelo aparece tabulado para diferentes grados de libertad. A modo ilustrativo recogemos en la tabla 6.1 algunos valores de su función de distribución para ciertos g.l.

Sin embargo, conviene señalar que esta estructura de tablas resulta poco útil, dado que en las aplicaciones habituales de esta distribución nos interesa tener un amplio recorrido de g.l. y buscaremos el valor correspondiente a determinados centiles (esto es, valores cuya probabilidad acumulada se sitúa en el 0,1 %, 1 %, 5 %, etc). De ahí que una estructura más habitual sea la de la tabla 6.2:

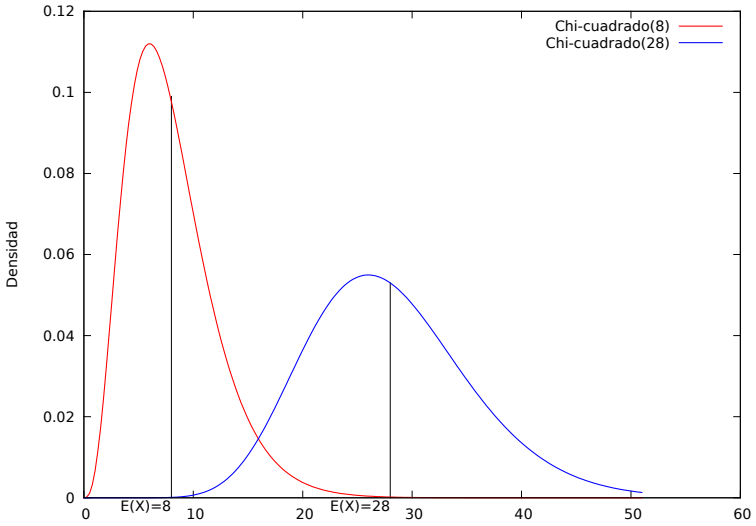
Como puede verse, en la primera columna se recogen los grados de libertad, en la primera fila el orden de los centiles indicados y en el interior de la tabla aparecen los distintos valores de la distribución χ^2_n.

En una aplicación usual de esta distribución, lo primero que conoceremos será el número de g.l., obtenido directamente a partir del tamaño muestral, en segundo lugar fijaremos el nivel de incertidumbre (1 %, 5 % o 10 % en general) o bien el nivel de confianza con el que deseamos trabajar (90 %, 95 % o 99 %) y luego buscaremos el valor de la χ^2_n correspondiente a esas restricciones.

Tabla 6.2.: Modelo χ^2_n. Valores x para $P(\chi^2_n \leq x)$

$n \backslash F$	0,01	0,025	0,05	0,1	0,9	0,95	0,975	0,99
1	0,0002	0,0010	0,0039	0,0158	2,7055	3,8415	5,0239	6,6349
2	0,0201	0,0506	0,1026	0,2107	4,6052	5,9915	7,3778	9,2103
3	0,1148	0,2158	0,3518	0,5844	6,2514	7,8147	9,3484	11,3449
4	0,2971	0,4844	0,7107	1,0636	7,7794	9,4877	11,1433	13,2767
5	0,5543	0,8312	1,1455	1,6103	9,2364	11,0705	12,8325	15,0863
6	0,8721	1,2373	1,6354	2,2041	10,6446	12,5916	14,4494	16,8119
7	1,2390	1,6899	2,1673	2,8331	12,0170	14,0671	16,0128	18,4753
8	1,6465	2,1797	2,7326	3,4895	13,3616	15,5073	17,5345	20,0902
9	2,0879	2,7004	3,3251	4,1682	14,6837	16,9190	19,0228	21,6660
10	2,5582	3,2470	3,9403	4,8652	15,9872	18,3070	20,4832	23,2093
11	3,0535	3,8157	4,5748	5,5778	17,2750	19,6751	21,9200	24,7250
12	3,5706	4,4038	5,2260	6,3038	18,5493	21,0261	23,3367	26,2170
13	4,1069	5,0088	5,8919	7,0415	19,8119	22,3620	24,7356	27,6882
14	4,6604	5,6287	6,5706	7,7895	21,0641	23,6848	26,1189	29,1412
15	5,2293	6,2621	7,2609	8,5468	22,3071	24,9958	27,4884	30,5779
16	5,8122	6,9077	7,9616	9,3122	23,5418	26,2962	28,8454	31,9999
17	6,4078	7,5642	8,6718	10,0852	24,7690	27,5871	30,1910	33,4087
18	7,0149	8,2307	9,3905	10,8649	25,9894	28,8693	31,5264	34,8053
19	7,6327	8,9065	10,1170	11,6509	27,2036	30,1435	32,8523	36,1909
20	8,2604	9,5908	10,8508	12,4426	28,4120	31,4104	34,1696	37,5662
21	8,8972	10,2829	11,5913	13,2396	29,6151	32,6706	35,4789	38,9322
22	9,5425	10,9823	12,3380	14,0415	30,8133	33,9244	36,7807	40,2894
23	10,1957	11,6886	13,0905	14,8480	32,0069	35,1725	38,0756	41,6384
24	10,8564	12,4012	13,8484	15,6587	33,1962	36,4150	39,3641	42,9798
25	11,5240	13,1197	14,6114	16,4734	34,3816	37,6525	40,6465	44,3141
26	12,1981	13,8439	15,3792	17,2919	35,5632	38,8851	41,9232	45,6417
27	12,8785	14,5734	16,1514	18,1139	36,7412	40,1133	43,1945	46,9629
28	13,5647	15,3079	16,9279	18,9392	37,9159	41,3371	44,4608	48,2782
29	14,2565	16,0471	17,7084	19,7677	39,0875	42,5570	45,7223	49,5879
30	14,9535	16,7908	18,4927	20,5992	40,2560	43,7730	46,9792	50,8922
40	22,1643	24,4330	26,5093	29,0505	51,8051	55,7585	59,3417	63,6907
50	29,7067	32,3574	34,7643	37,6886	63,1671	67,5048	71,4202	76,1539
60	37,4849	40,4817	43,1880	46,4589	74,3970	79,0819	83,2977	88,3794
70	45,4417	48,7576	51,7393	55,3289	85,5270	90,5312	95,0232	100,4252
80	53,5401	57,1532	60,3915	64,2778	96,5782	101,8795	106,6286	112,3288
90	61,7541	65,6466	69,1260	73,2911	107,5650	113,1453	118,1359	124,1163
100	70,0649	74,2219	77,9295	82,3581	118,4980	124,3421	129,5612	135,8067

Figura 6.1.: χ^2_n. Función de densidad



En la tabla 6.1 observamos cómo para determinados grados de libertad la probabilidad acumulada aumenta con el valor de la χ^2_n o bien para un valor fijo de ésta, la probabilidad disminuye conforme aumentan los grados de libertad. Siguiendo el mismo razonamiento, la tabla nos muestra cómo fijados los grados de libertad, los valores de χ^2_n aumentan con el valor de la probabilidad, mientras para una probabilidad acumulada fija estos valores aumentan con los grados de libertad.

Intuitivamente este comportamiento es muy razonable, teniendo en cuenta que $X_i \approx \mathcal{N}(0,1)$ y X_i^2 toma sólo valores positivos, con valor esperado la unidad. Por tanto, cuando definimos

$$\chi^2_n = \sum_{i=1}^{n} X_i^2$$

a medida que aumenta n se incrementa el valor esperado de la expresión y el punto donde se alcanza determinada probabilidad acumulada se desplaza a la derecha.

La representación gráfica del modelo chi-cuadrado aparece recogida en la figura 6.1. Esta función presenta una forma más simétrica a medida que aumentan sus grados de libertad n.

Las características del modelo chi-cuadrado dependen únicamente de sus grados de libertad: $E\left(\chi^2_n\right) = n$, $Var\left(\chi^2_n\right) = 2n$. Esta relación directa con n es razonable puesto que las variables cuya suma da lugar a la χ^2_n están normalizadas y por tanto no dependen de los valores de cada X_i. En la tabla 6.1 podemos observar que cuando el valor de la chi-cuadrado coincide con el número de grados de libertad, la probabilidad acumulada, aunque algo superior, se sitúa próxima a 0,5.

Proposición 6.1. *El modelo probabilístico chi-cuadrado es reproductivo respecto a los grados de libertad, esto es, dadas dos v.a. independientes X e Y con distribuciones respectivas χ^2_n y χ^2_m es posible afirmar que su suma $(X+Y)$ se distribuye según un modelo χ^2_{n+m}.*

Demostración. La comprobación de la reproductividad es inmediata a partir de la definición vista para el modelo chi-cuadrado, ya que se tendría:

$X = \sum_{i=1}^{n} X_i^2$, con $(X_1, \ldots, X_n)$ m.a.s. extraída de una población $X \approx \mathcal{N}(0,1)$

$Y = \sum_{i=1}^{m} Y_i^2$ con $(Y_1, \ldots, Y_m)$ m.a.s. extraída de una población $Y \approx \mathcal{N}(0,1)$

Si operamos el cambio $X_{n+1} = Y_1$, $X_{n+2} = Y_2$, ..., $X_{n+m} = Y_n$, entonces podemos escribir:

$$X + Y = \sum_{i=1}^{n+m} X_i^2 \tag{6.1}$$

Además $X_1, \ldots, X_n$ son independientes por tratarse de una m.a.s, $X_{n+1}, \ldots, X_{n+m}$ lo son por tratarse de otra muestra aleatoria y además, por ser las variables X e Y independientes, también lo son las muestras entre sí. Como consecuencia, el sumatorio 6.1 es por definición una chi-cuadrado con $n + m$ g.l. □

Desde un punto de vista intuitivo, la reproductividad va directamente asociada a la interpretación de los grados de libertad: si en las variables X e Y se tienen holguras de n y m valores respectivamente, la consideración conjunta de ambas características aleatorias nos proporcionará libertad para fijar un total de $n + m$ valores.

Proposición 6.2. *Para tamaños elevados de muestra la distribución chi-cuadrado puede ser aproximada por el modelo normal.*

Demostración. En efecto, aplicando el TCL a una sucesión de variables χ_n^2 independientes, su suma se distribuiría según un modelo normal cuya esperanza y varianza se obtienen como sumas de las correspondientes características de cada sumando.

Este resultado puede ser aplicado a la siguiente sucesión: $\chi_{1_1}^2 = X_1^2, \ldots, \chi_{1_n}^2 = X_n^2$ cuyos elementos presentan distribución chi-cuadrado con un grado de libertad y son variables independientes por serlo las componentes muestrales. Así se obtiene:

$$\chi_n^2 = \sum_{i=1}^{n} \chi_{1_i}^2 \to \mathcal{N}\left(n, \sqrt{2n}\right)$$

□

Sin embargo, para obtener una convergencia más rápida utilizamos la aproximación:

$$\sqrt{2\chi_n^2} - \sqrt{2n-1} \approx \mathcal{N}(0,1) \tag{6.2}$$

En la tabla que sigue aparecen calculadas por distintos métodos las probabilidades $P\left(\chi_n^2 \leq n\right)$, esto es, la probabilidad acumulada hasta el valor esperado para diferentes grados de libertad.

g.l. y valores	$P\left(\chi^2_n \leq x\right)$	Aprox. TCL	Aprox. rápida
n=30, x=30	0,5343	0,5	0,5258
n=50, x=50	0,5266	0,5	0,5200
n=100, x=100	0,5188	0,5	0,5141
n=500, x=500	0,5084	0,5	0,5063

En la segunda columna, que recoge los resultados de esta probabilidad calculada mediante el modelo chi-cuadrado, se observa que dicha probabilidad converge lentamente hacia 0,5 a medida que aumentan los tamaños muestrales. En cambio, la aproximación de estas probabilidades mediante la aplicación del TCL, que se recoge en la columna tercera, da siempre un resultado constante e igual a 0,5.

Por último, la aproximación que hemos denominado "rápida" (6.2) subvalora la verdadera probabilidad, aunque se aproxima considerablemente al valor verdadero a medida que n aumenta. De ahí que ésta será la aproximación utilizada siempre que dispongamos de tamaños muestrales suficientemente elevados.

[En realidad, habría que tener en cuenta que en todas las situaciones - incluida la que hemos llamado "verdadera probabilidad"- se utilizan algoritmos de cálculo numérico con lo cual se trata siempre de aproximaciones].

La distribución chi-cuadrado también aparece ligada a otros modelos de probabilidad. Así, dada una v.a. distribuida uniformemente en el intervalo $(0,1)$ y siendo $(X_1,\ldots,X_n)$ una m.a.s. de esa población, entonces la variable:

$$-\ln\left(\prod_{i=1}^{n} X_i^2\right) = -\sum_{i=1}^{n} \ln X_i^2$$

sigue una distribución χ^2_n.

El signo negativo de la expresión anterior se debe a que los valores de X_i son inferiores a la unidad y por tanto sus logaritmos resultan negativos.

El modelo chi-cuadrado desempeña un papel destacado en los procesos inferenciales. Concretamente, esta es la distribución de probabilidad que aparece asociada a las inferencias relativas a la dispersión poblacional, y además su utilización es muy frecuente en la inferencia no paramétrica (por ejemplo, cuando realizamos contrastes de independencia, de bondad de ajuste, de homogeneidad, ...).

Karl Pearson (1857-1936), considerado por algunos autores como el fundador de la ciencia estadística, fue el primero en introducir el modelo chi-cuadrado, en el año 1900, como expresión válida para contrastar la bondad del ajuste de una distribución teórica a la observada.

Pearson obtuvo también un sistema de curvas de frecuencias generalizadas basándose en una sola ecuación diferencial obteniendo los parámetros por el método de los momentos. Esta aportación convirtió al modelo chi-cuadrado en una herramienta básica del análisis estadístico, hecho que explica la mayor relevancia otorgada a Pearson que a Helmert, autor que le precedió cronológicamente obteniendo, en 1875, la distribución de la varianza muestral para una población con distribución normal.

Teorema de Fisher

La generalidad del modelo chi-cuadrado como distribución muestral se debe en gran medida al *Teorema de Fisher*, que garantiza la independencia entre los estadísticos media y varianza muestral, así como un modelo probabilístico relacionado con esta última.

Teorema 6.1. *Dada una m.a.s.* $(X_1, \ldots, X_n)$ *extraída de una población* $\mathcal{N}(\mu, \sigma)$*, se cumple:*

- *La media muestral* $\bar{X}$ *y la varianza muestral* S^2 *son variables aleatorias independientes.*
- *La expresión aleatoria* $\dfrac{(n-1)S^2}{\sigma^2}$ *se distribuye según un modelo chi-cuadrado con n-1 grados de libertad (* χ^2_{n-1}*).*

Demostración. El primero de los resultados del teorema de Fisher se basa en el hecho de que el vector $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$ es independiente de la media muestral por lo cual S^2, que es función del vector $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$, también será independiente de $\bar{X}$. Este resultado puede demostrarse construyendo la función generatriz de momentos n+1-dimensional del vector $(\bar{X}, X_1 - \bar{X}, \ldots, X_n - \bar{X})$ y viendo que dicha función se puede factorizar como un producto de dos f.g.m.: una correspondiente a $\bar{X}$ y la otra al vector $(X_1 - \bar{X}, \ldots, X_n - \bar{X})$, lo cual es una condición necesaria y suficiente (como hemos visto en el capítulo 4) para la independencia entre las dos variables.

Es posible garantizar que esta independencia entre las variables $\bar{X}$ y S^2 sólo se obtiene cuando la población de partida es normal.

Por lo que se refiere al segundo resultado del teorema de Fisher, éste se basa en la descomposición:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{\left(\bar{X} - \mu\right)^2}{\frac{\sigma^2}{n}}$$

en la que se cumple:

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \approx \chi^2_n \, ; \; \frac{\left(\bar{X} - \mu\right)^2}{\frac{\sigma^2}{n}} \approx \chi^2_1 \; ; \text{ [Justifíquese por qué]}$$

Además, gracias al primer resultado del teorema de Fisher podemos garantizar que los sumandos

$$\frac{(n-1)S^2}{\sigma^2} \qquad \text{y} \qquad \frac{\left(\bar{X} - \mu\right)^2}{\frac{\sigma^2}{n}}$$

son independientes, y por tanto la reproductividad de la distribución chi-cuadrado garantiza que la expresión $\dfrac{(n-1)S^2}{\sigma^2}$ se distribuirá según un modelo χ^2_{n-1}.

□

En el caso particular de que la distribución poblacional fuese N(0,1), entonces la comprobación $(n-1)S^2 \approx \chi^2_{n-1}$ podría hacerse de forma directa:

$$(n-1)S^2 = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 = n \sum_{i=1}^{n} \frac{\left(X_i - \bar{X}\right)^2}{n} = \sum_{i=1}^{n} X_i^2 - n\bar{X}^2 = \sum_{i=1}^{n} X_i^2 - \left(\sqrt{n}\bar{X}\right)^2$$

donde, teniendo en cuenta que

$$\bar{X} \approx \mathcal{N}\left(0, \frac{1}{\sqrt{n}}\right) \qquad \text{y} \qquad \sqrt{n}\bar{X} \approx \mathcal{N}(0,1)$$

obtenemos nuevamente la definición de χ^2_{n-1}.

El enunciado de Fisher garantiza un modelo probabilístico chi-cuadrado conectado a la dispersión muestral. Los grados de libertad de esta distribución serán $n-1$, debido a que en la expresión de S^2 aparece la restricción

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

que reduce en uno los niveles de libertad de la muestra.

Siguiendo este mismo planteamiento, cada restricción adicional que limite la posibilidad de elegir las componentes de la expresión supondría una nueva reducción en los grados de libertad.

6.1.3. Distribución t de Student

La necesidad de trabajar con muestras de pequeño tamaño limita la aplicabilidad del modelo normal, y justifica la utilización de la distribución denominada *t de Student*, cuya aparición histórica ilustra la interacción entre el desarrollo de las técnicas estadísticas y los problemas del mundo real.

Definición 6.2. Dadas dos variables aleatorias independientes $X \approx \mathcal{N}(0,1)$ e $Y \approx \chi^2_n$, la variable aleatoria

$$t = \frac{X}{\sqrt{\frac{Y}{n}}}$$

se distribuye según un modelo *t de Student* con n grados de libertad (t_n).

La derivación del estadístico t se debe a W. S. Gosset, empleado de las industrias cerveceras Guinness, quien se enfrentaba a la necesidad de estimar, a partir de muestras pequeñas, parámetros relativos a la fermentación de la cerveza. Dado que la compañía Guinness, para evitar el espionaje industrial, no autorizaba a sus empleados a publicar los resultados de sus investigaciones, Gosset (1908) utilizó el seudónimo Student que aún en la actualidad da nombre a la distribución t.

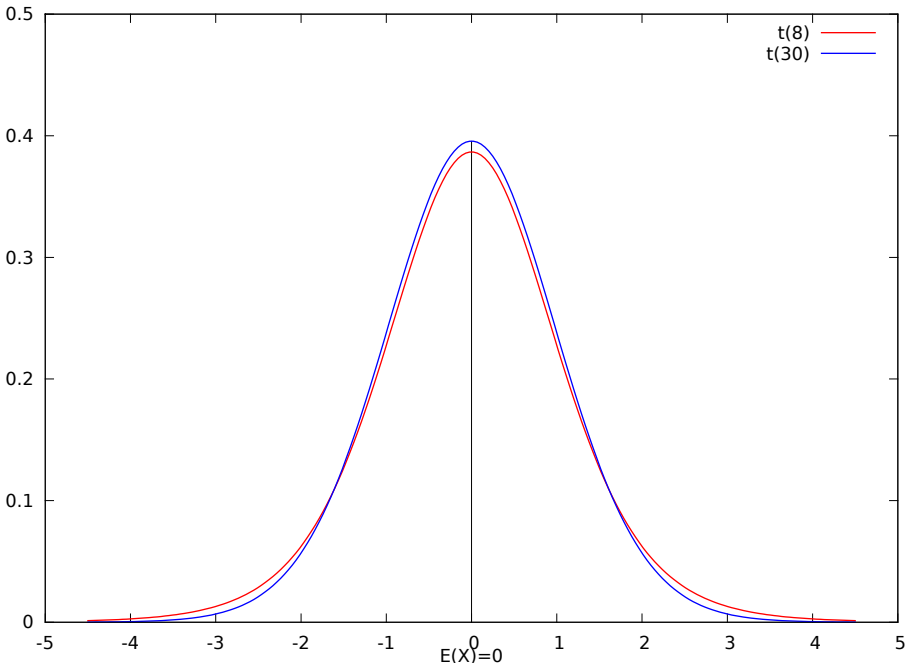
Si consideramos una población $X \approx \mathcal{N}(0,1)$ y una m.a.s. de la misma, $(X_1, \ldots, X_n)$, entonces la expresión:

$$t = \frac{X}{\sqrt{\frac{X_1^2 + \cdots + X_n^2}{n-1}}}$$

sigue una distribución t con n g.l. (t_n). Del mismo modo, teniendo en cuenta los comentarios del epígrafe anterior, podemos afirmar que:

$$t = \frac{X}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}}$$

Figura 6.2.: Modelo t de Student



sigue el modelo t con $n-1$ g.l. (t_{n-1}).

Si la población X sigue un modelo normal general, $X \approx \mathcal{N}(\mu, \sigma)$ y $(X_1, \ldots, X_n)$ es una m.a.s. de X, entonces considerando los cambios de variable:

$$Y = \frac{X-\mu}{\sigma} \qquad \text{y} \qquad Y_i = \frac{X_i-\mu}{\sigma}$$

para $i = 1, 2, \ldots, n$, y aplicando los estadísticos anteriores a las nuevas variables, se tiene:

$$t = \frac{X-\mu}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{n}}} \approx t_n \quad \text{y} \quad t = \frac{X-\mu}{\sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}} \approx t_{n-1}$$

El modelo t es en muchos sentidos similar al normal que aparece en su numerador, ya que se trata de una distribución unimodal simétrica y campaniforme. Este modelo viene caracterizado por el parámetro n, que representa sus grados de libertad, coincidentes con los de la distribución chi-cuadrado que aparece en su denominador.

La función de densidad del modelo t_n viene dada por la expresión:

$$f(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \qquad -\infty < x < \infty$$

La esperanza y la varianza de esta distribución son respectivamente:

$$E(t_n) = 0 \qquad \text{y} \qquad Var(t_n) = \frac{n}{n-2}$$

observándose que el riesgo del modelo disminuye a medida que aumentan los grados de libertad de la distribución t de Student (se cumple $\lim_{n\to\infty} Var(t_n) = 1$).

Por lo que se refiere a la utilización de las tablas (6.3), el procedimiento es similar al estudiado para la distribución χ^2_n y consiste en buscar en la tabla los valores correspondientes a determinados niveles de probabilidad y para ciertos grados de libertad.

Sin embargo, dado que la densidad de la distribución t de Student es simétrica, las tablas pueden ser utilizadas buscando probabilidades correspondientes a las áreas de una sola cola (indicadas en la fila superior) o de las dos colas (tal y como recoge la fila inferior). Obsérvese que para cualquier $t > 0$ se cumple: $P(|t_n| > t) = 2P(t_n > t)$.

Como ya hemos comentado, para valores elevados de n la distribución t se aproxima por el modelo $\mathcal{N}(0, 1)$. Sin embargo, como muestra la figura 6.3, se obtienen aproximaciones más adecuadas adoptando un modelo normal con la verdadera dispersión de la variable X (en la tabla se observa que, incluso para valores bajos de los grados de libertad n, las discrepancias aparecen sólo en la cuarta cifra decimal).

6.1.4. Distribución F de Snedecor

La distribución denominada F de Snedecor juega un importante papel en la comparación de la homogeneidad de varias poblaciones y viene definida en los siguientes términos:

Definición 6.3. Dadas dos v.a. independientes X e Y distribuidas según modelos de probabilidad χ^2_n y χ^2_m respectivamente, la expresión

$$F = \frac{\frac{X}{n}}{\frac{Y}{m}}$$

sigue un *modelo F de Snedecor* con n y m grados de libertad, que denotamos por $F_{n,m}$ o F^n_m (indicando así que tenemos n g.l. en el numerador y m en el denominador).

Debido a su construcción como cociente de dos modelos chi-cuadrado, la distribución F viene caracterizada por dos parámetros, que describen los grados de libertad en el numerador y el denominador respectivamente.

Sean X e Y dos poblaciones distribuidas según modelos $\mathcal{N}(0, 1)$. Dadas dos m.a.s. independientes extraídas de esas poblaciones $(X_1, \ldots, X_n)$ y $(Y_1, \ldots, Y_m)$, entonces la expresión:

$$F = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i^2}{\frac{1}{m}\sum_{i=1}^{m} Y_i^2}$$

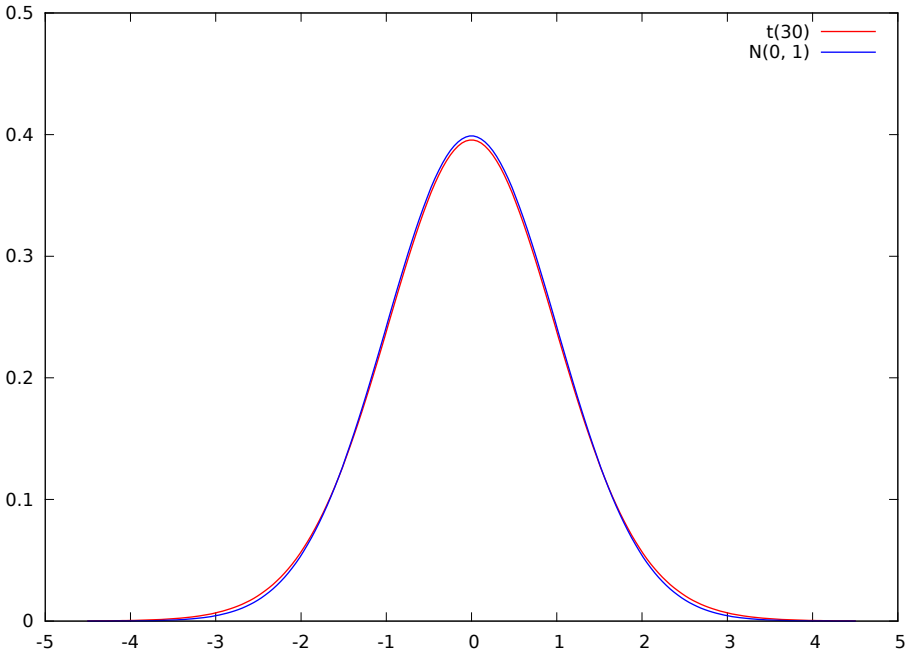
sigue una distribución F de Snedecor con grados de libertad n y m (F^n_m).

Si las variables poblacionales fuesen $\mathcal{N}(\mu_X, \sigma_X)$ y $\mathcal{N}(\mu_Y, \sigma_Y)$ respectivamente, entonces la expresión:

Tabla 6.3.: Modelo t de Student. Valores x para $p = P(|t_n| \geq x)$

n/p	0,001	0,0025	0,005	0,01	0,025	0,05	0,1	0,25
1	636,6192	254,6466	127,3213	63,6567	25,4517	12,7062	6,3138	2,4142
2	31,5991	19,9625	14,0890	9,9248	6,2053	4,3027	2,9200	1,6036
3	12,9240	9,4649	7,4533	5,8409	4,1765	3,1824	2,3534	1,4226
4	8,6103	6,7583	5,5976	4,6041	3,4954	2,7764	2,1318	1,3444
5	6,8688	5,6042	4,7733	4,0321	3,1634	2,5706	2,0150	1,3009
6	5,9588	4,9807	4,3168	3,7074	2,9687	2,4469	1,9432	1,2733
7	5,4079	4,5946	4,0293	3,4995	2,8412	2,3646	1,8946	1,2543
8	5,0413	4,3335	3,8325	3,3554	2,7515	2,3060	1,8595	1,2403
9	4,7809	4,1458	3,6897	3,2498	2,6850	2,2622	1,8331	1,2297
10	4,5869	4,0045	3,5814	3,1693	2,6338	2,2281	1,8125	1,2213
11	4,4370	3,8945	3,4966	3,1058	2,5931	2,2010	1,7959	1,2145
12	4,3178	3,8065	3,4284	3,0545	2,5600	2,1788	1,7823	1,2089
13	4,2208	3,7345	3,3725	3,0123	2,5326	2,1604	1,7709	1,2041
14	4,1405	3,6746	3,3257	2,9768	2,5096	2,1448	1,7613	1,2001
15	4,0728	3,6239	3,2860	2,9467	2,4899	2,1314	1,7531	1,1967
16	4,0150	3,5805	3,2520	2,9208	2,4729	2,1199	1,7459	1,1937
17	3,9651	3,5429	3,2224	2,8982	2,4581	2,1098	1,7396	1,1910
18	3,9216	3,5101	3,1966	2,8784	2,4450	2,1009	1,7341	1,1887
19	3,8834	3,4812	3,1737	2,8609	2,4334	2,0930	1,7291	1,1866
20	3,8495	3,4554	3,1534	2,8453	2,4231	2,0860	1,7247	1,1848
21	3,8193	3,4325	3,1352	2,8314	2,4138	2,0796	1,7207	1,1831
22	3,7921	3,4118	3,1188	2,8188	2,4055	2,0739	1,7171	1,1815
23	3,7676	3,3931	3,1040	2,8073	2,3979	2,0687	1,7139	1,1802
24	3,7454	3,3761	3,0905	2,7969	2,3909	2,0639	1,7109	1,1789
25	3,7251	3,3606	3,0782	2,7874	2,3846	2,0595	1,7081	1,1777
26	3,7066	3,3464	3,0669	2,7787	2,3788	2,0555	1,7056	1,1766
27	3,6896	3,3334	3,0565	2,7707	2,3734	2,0518	1,7033	1,1756
28	3,6739	3,3214	3,0469	2,7633	2,3685	2,0484	1,7011	1,1747
29	3,6594	3,3102	3,0380	2,7564	2,3638	2,0452	1,6991	1,1739
30	3,6460	3,2999	3,0298	2,7500	2,3596	2,0423	1,6973	1,1731
40	3,5510	3,2266	2,9712	2,7045	2,3289	2,0211	1,6839	1,1673
50	3,4960	3,1840	2,9370	2,6778	2,3109	2,0086	1,6759	1,1639
60	3,4602	3,1562	2,9146	2,6603	2,2990	2,0003	1,6706	1,1616
70	3,4350	3,1366	2,8987	2,6479	2,2906	1,9944	1,6669	1,1600
80	3,4163	3,1220	2,8870	2,6387	2,2844	1,9901	1,6641	1,1588
90	3,4019	3,1108	2,8779	2,6316	2,2795	1,9867	1,6620	1,1578
100	3,3905	3,1018	2,8707	2,6259	2,2757	1,9840	1,6602	1,1571

Figura 6.3.: Modelo t. Aproximación normal



g.l. y valores	$P(t_n \leq x)$	Aprox. $\mathcal{N}(0,1)$	Aprox. $\mathcal{N}\left(0, \sqrt{\frac{n}{n-2}}\right)$
$n = 10, x = 1,96$	0,9608	0,9750	0,9602
$n = 30, x = 1,96$	0,9703	0,9750	0,9709
$n = 50, x = 1,96$	0,9722	0,9750	0,9726
$n = 100, x = 1,96$	0,9736	0,9750	0,9738

$$F = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \mu_X)^2}{\frac{1}{m}\sum_{i=1}^{m}(Y_i - \mu_Y)^2}$$

sigue también un modelo F_m^n.

R. A. Fisher (1890-1962) fue el primero en estudiar la distribución del cociente de varianzas. Estos estudios fueron proseguidos por G. W. Snedecor (1881-1974), autor de la obra Statistical Methods (1937) quien denominó F a la distribución de la razón de varianzas en honor de Fisher.

Existe también una distribución denominada *z de Fisher* que se obtiene mediante una transformación de la F de Snedecor:

$$z = \frac{1}{2}\ln F$$

que resulta de utilidad para llevar a cabo inferencias relativas a la correlación entre variables.

La función de densidad de la distribución F_m^n viene dada por la expresión:

$$f(x) = n^{\frac{n}{2}} m^{\frac{m}{2}} \frac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)} \frac{x^{\frac{n}{2}-1}}{(nx+m)^{\frac{n+m}{2}}} \; ; \qquad x > 0$$

Las características del modelo F de Snedecor aparecen relacionadas con sus grados de libertad. Así se obtiene:

$$E\left(F_m^n\right) = \frac{n}{n-2} \qquad \text{con } n > 2$$

y

$$Var\left(F_m^n\right) = \frac{2n^2(n+m-2)}{m(n-2)^2(n-4)} \qquad \text{con } n > 4$$

Por lo que se refiere a la representación gráfica, esta distribución presenta una forma similar a la del modelo chi-cuadrado, tal y como puede apreciarse en la figura 6.4.

Para tabular las probabilidades de este modelo es necesario recoger los grados de libertad tanto del numerador (n) como del denominador (m), por lo cual cada tabla contiene valores de la distribución que llevan asociada una probabilidad fija. En la tabla 6.4 recogemos una de las situaciones más habituales, con probabilidades en la cola derecha del 5 % (esto es $P\left(F_m^n > x\right) = 0,05$ y por tanto $P\left(F_m^n \leq x\right) = 0,95$).

En general, utilizaremos las tablas del modelo F cuando disponemos de información sobre los tamaños muestrales y fijamos alguna probabilidad para la cola derecha de la distribución.

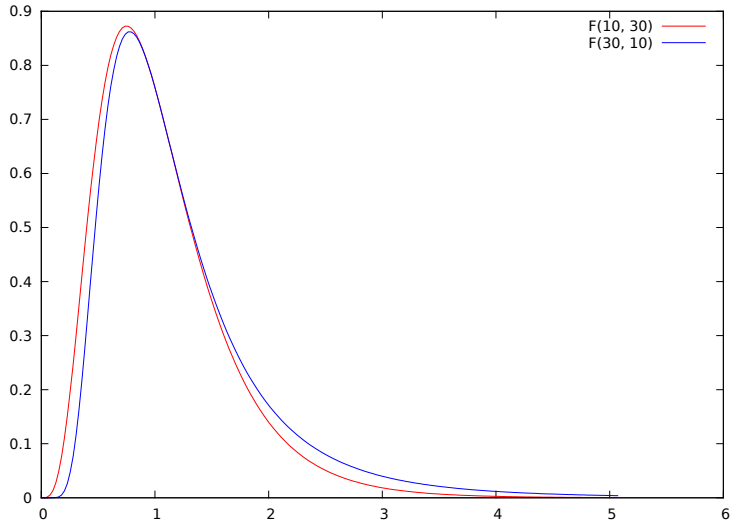
La intersección en las tablas entre la columna y la fila asociadas a los g.l. del numerador y del denominador proporciona el valor de la distribución que deja a su derecha la probabilidad fijada.

La utilización práctica del modelo F de Snedecor se beneficia en gran medida de la propiedad de inversión.

Tabla 6.4.: Modelo F de Snedecor. Valores x para $P(F_m^n \geq x) = 0,05$

$gl\, d/n$	1	2	3	4	5	6	7	8	9	10
1	161,448	199,500	215,707	224,583	230,162	233,986	236,768	238,883	240,543	241,882
2	18,513	19,000	19,164	19,247	19,296	19,330	19,353	19,371	19,385	19,396
3	10,128	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,812	8,786
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,999	5,964
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,772	4,735
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,099	4,060
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,677	3,637
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,388	3,347
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,179	3,137
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	3,020	2,978
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,896	2,854
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,796	2,753
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,714	2,671
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,646	2,602
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,588	2,544
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,538	2,494
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,494	2,450
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,456	2,412
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,423	2,378
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,393	2,348
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,366	2,321
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,342	2,297
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,320	2,275
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,300	2,255
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,282	2,236
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,265	2,220
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,250	2,204
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,236	2,190
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,223	2,177
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,211	2,165
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,124	2,077
50	4,034	3,183	2,790	2,557	2,400	2,286	2,199	2,130	2,073	2,026
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	2,040	1,993
70	3,978	3,128	2,736	2,503	2,346	2,231	2,143	2,074	2,017	1,969
80	3,960	3,111	2,719	2,486	2,329	2,214	2,126	2,056	1,999	1,951
90	3,947	3,098	2,706	2,473	2,316	2,201	2,113	2,043	1,986	1,938
100	3,936	3,087	2,696	2,463	2,305	2,191	2,103	2,032	1,975	1,927

Figura 6.4.: Modelo F de Snedecor. Función de densidad



Proposición 6.3. *Si una variable X se distribuye según un modelo F_m^n entonces su inversa $\frac{1}{X}$ aparece también distribuida según este modelo, invertido el orden de sus grados de libertad (F_n^m).*

Demostración. En efecto, si una variable X sigue distribución F con grados de libertad n y m, entonces puede ser expresada como:

$$X = \frac{\frac{\chi_n^2}{n}}{\frac{\chi_m^2}{m}} \approx F_m^n$$

y calculando su inversa se obtiene

$$\frac{1}{X} = \frac{\frac{\chi_m^2}{m}}{\frac{\chi_n^2}{n}} \approx F_n^m$$

□

Esta propiedad de inversión rentabiliza el uso de las tablas de la F, ya que permite limitar la información contemplando un mayor recorrido de los grados de libertad en el numerador o en el denominador, de forma que si la probabilidad buscada no aparece en las tablas podemos llevar a cabo la transformación:

$$F\left(F_m^n \leq x\right) = P\left(\frac{1}{F_m^n} \geq \frac{1}{x}\right) = P\left(F_n^m \geq \frac{1}{x}\right)$$

De esta forma, combinando inversos y complementarios podemos resolver buena parte de las lagunas que presentan estas tablas.

Otra propiedad interesante de la distribución F de Snedecor es su conexión con el modelo t de Student:

Proposición 6.4. *Si X es una variable con distribución t de Student de m grados de libertad ($X \approx t_m$) entonces la variable X^2 sigue un modelo F_m^1.*

Demostración. En efecto, por definición de la distribución t de Student, la variable X es de la forma:

$$X = \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi_m^2}{m}}}$$

con lo cual para su cuadrado se tiene:

$$X^2 = \frac{(N(0,1))^2}{\frac{\chi_m^2}{m}} = \frac{\frac{\chi_1^2}{1}}{\frac{\chi_m^2}{m}} = F_m^1$$

□

6.2. Procesos inferenciales y distribuciones asociadas

Un proceso inferencial consiste en utilizar de forma coordinada gran parte de las técnicas que hemos analizado en capítulos anteriores. De este modo, una vez decidido el objetivo de nuestros estudios será posible determinar las expresiones muestrales adecuadas para cada caso y conectar dichas expresiones con algún modelo probabilístico conocido.

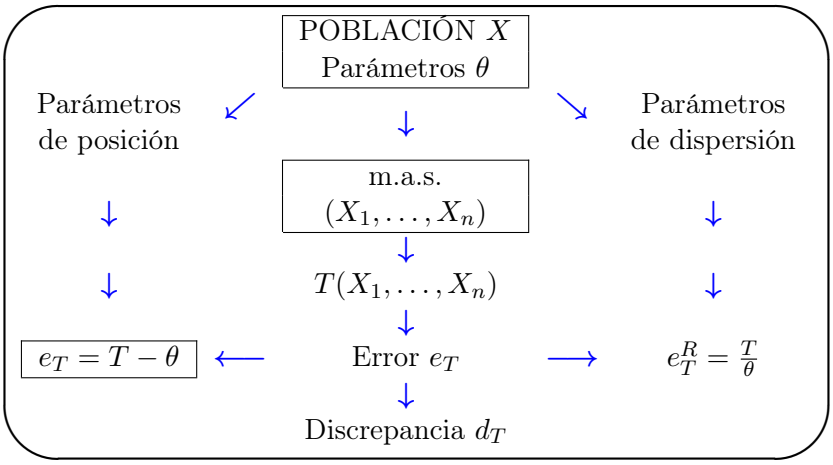
Este procedimiento será a su vez el punto de partida para los capítulos posteriores, en los que abordaremos diversos objetivos inferenciales mediante dos técnicas diferenciadas: la estimación y el contraste de hipótesis.

El objetivo de la inferencia será en cualquier caso la investigación de una o varias poblaciones que identificamos con variables aleatorias y sobre las que existe incertidumbre. En ocasiones nos centraremos en ciertos parámetros poblacionales que denotamos genéricamente por θ, mientras que en otros casos abordaremos características más genéricas como la aleatoriedad, la distribución probabilística de la población o la independencia entre dos poblaciones.

Conviene destacar la importancia de estas características genéricas, que a menudo coinciden con los supuestos asumidos sobre una o varias poblaciones (normalidad, independencia,...). De ahí la necesidad de ser rigurosos en nuestros procesos inferenciales, en el sentido de especificar (y contrastar) los supuestos que en cada caso estamos admitiendo como válidos.

Entre dichos supuestos merece una mención especial la hipótesis de normalidad, que ocupa un papel central en la inferencia estadística: por una parte, es frecuente asumir que la población de partida se distribuye normalmente y por otra, aun partiendo de poblaciones desconocidas, siempre que analicemos muestras de tamaño suficientemente elevado y se cumplan los supuestos necesarios,

Figura 6.5.: Esquema de trabajo en inferencia



podremos aplicar el teorema central del límite que garantiza la convergencia de agregados o promedios a una distribución normal.

6.2.1. Inferencias relativas a parámetros

Supongamos que deseamos llevar a cabo inferencias sobre algún parámetro poblacional que denominamos θ. En este caso el esquema de trabajo que seguiremos aparece descrito en la figura 6.5 y consiste en aprovechar toda la información muestral, llegando a resumir dicha información mediante expresiones que nos permitan realizar afirmaciones probabilísticas.

Como puede apreciarse en el esquema, el punto de partida es la población X cuya distribución probabilística viene dada por $F_X(x, \theta)$ que depende de ciertos parámetros desconocidos θ.

La primera etapa del proceso inferencial consistirá en seleccionar una muestra aleatoria que podemos identificar con una v.a. n-dimensional, cuya distribución probabilística dependerá también de los parámetros θ.

A partir de esta muestra se definen estimadores T que, por ser funciones de la muestra, serán también variables aleatorias cuya distribución de probabilidad denotamos por $F_T(t, \theta)$.

Como hemos visto en el capítulo anterior, el estimador T no coincidirá con el valor del parámetro desconocido θ, por lo cual definimos los correspondientes errores aleatorios. Cuando el parámetro investigado sea una característica de posición (esperanza, proporción) este error se define como diferencia $e_T = T - \theta$, mientras que si θ es una característica de dispersión debemos definir errores relativos, que vienen dados por el cociente $e_T^R = \frac{T}{\theta}$.

Las características de estos errores aleatorios aparecen conectadas con las del esti-

mador T, tal y como muestra la siguiente tabla:

Error	Esperanza	Varianza
$e_T = T - \theta$	$E(e_T) = E(T) - \theta$	$Var(e_T) = Var(T)$
$e_T^R = \frac{T}{\theta}$	$E\left(e_T^R\right) = \frac{E(T)}{\theta}$	$Var\left(e_T^R\right) = \frac{Var(T)}{\theta^2}$

Como ya hemos visto, las propiedades de ausencia de sesgo y eficiencia pueden ser formuladas indistintamente sobre los errores o sobre el estimador. Así, si T es insesgado se cumple $E(T) = \theta$ o equivalentemente $E(e_T) = 0$ y en el caso del error relativo $E(e_T^R) = 1$.

De modo similar, se observa que si T es un estimador de mínima varianza, entonces también se hace mínima la varianza del error.

Si T es un estimador sesgado del parámetro investigado θ se tiene:

$$E(e_T) = E(T) - \theta = B_T(\theta) \quad ; \quad E\left(e_T^R\right) = 1 + \frac{B_T(\theta)}{\theta} = 1 + B_T^R(\theta)$$

[Compruébese]

Una vez conocidas las características de los errores, nos interesará llevar a cabo un proceso de transformación de los mismos, efectuando ciertos ajustes sobre las expresiones de e_T y e_T^R hasta llegar a obtener *discrepancias tipificadas o estandarizadas* que denotaremos por d_T.

La definición de estas discrepancias abarca dos etapas:

1. La primera etapa consiste en reducir los errores a su expresión de referencia o "estándar".

En el caso de los parámetros de posición, este objetivo se consigue mediante una tipificación de los errores que conduce a expresiones:

$$\frac{e_T - E(e_T)}{\sigma_{e_T}} = \frac{T - E(T)}{\sigma_T}$$

que presentan esperanza nula y varianza unitaria.

Para los parámetros de dispersión el procedimiento es distinto como consecuencia de su carácter multiplicativo. En este caso los errores relativos deben presentar esperanza unitaria, para lo cual -si es necesario- se efectúa el ajuste:

$$\frac{e_T^R}{E\left(e_T^R\right)}$$

2. La segunda etapa abarca los ajustes necesarios en las expresiones anteriores hasta obtener modelos de probabilidad conocida, que no dependan de ningún parámetro desconocido.

En general, las discrepancias tipificadas se adaptarán a los modelos probabilísticos analizados en el epígrafe anterior (normal, chi-cuadrado, t de Student y F de Snede-

cor).

En los apartados que siguen estudiamos la construcción de las discrepancias utilizadas en las inferencias sobre los parámetros de interés.

Es conveniente observar que, aunque este apartado tiene por objetivo las inferencias relativas a parámetros, ello no implica que todas las inferencias examinadas sean de tipo paramétrico. La situación más habitual -y más conveniente para nuestros objetivos- será que la población X investigada sea conocida, limitándose la ausencia de información a los parámetros con lo que el estudio sería de inferencia paramétrica.

Sin embargo, es posible también que nos enfrentemos a poblaciones completamente desconocidas, en cuyo caso nos situaríamos en el ámbito de la inferencia no paramétrica. En estos casos, a menudo deberemos renunciar a la segunda etapa de la construcción de discrepancias, ya que no resulta posible garantizar su conexión con modelos probabilísticos conocidos.

6.2.2. Inferencias sobre la media

Como hemos justificado en temas anteriores, cuando deseamos llevar a cabo inferencias sobre la esperanza poblacional μ, resulta adecuado el estimador media muestral que viene dado por la expresión

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

En concreto, sus características esperanza y varianza vienen dadas por:

$$E\left(\bar{X}\right) = \mu \quad ; \quad Var\left(\bar{X}\right) = \frac{\sigma^2}{n}$$

El error cometido con este estimador será una variable aleatoria, que se obtiene por diferencia entre la media muestral y la media poblacional: $e_{\bar{X}} = \bar{X} - \mu$.

Siguiendo el esquema anteriormente descrito, debemos analizar las características del error, para el que se obtiene un valor esperado nulo:

$$E(e_{\bar{X}}) = E\left(\bar{X}\right) - \mu = 0$$

y una dispersión dada por las expresiones:

$$Var(e_{\bar{X}}) = Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \quad ; \quad \sigma_{e_{\bar{X}}} = \frac{\sigma}{\sqrt{n}}$$

Así pues, es posible en una primera etapa llevar a cabo una tipificación o ajuste del error aleatorio, llegando a una expresión:

$$\frac{e_{\bar{X}} - E(e_{\bar{X}})}{\sigma_{e_{\bar{X}}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

que presenta esperanza nula y dispersión unitaria.

Por lo que se refiere a la segunda etapa, la distribución de probabilidad de esta expresión dependerá de los supuestos asumidos sobre la población de partida X. Así, en el caso de que asumamos que X se distribuye normalmente con varianza conocida, se obtendría la *discrepancia tipificada para la media*:

$$d_{\bar{X}} = \frac{e_{\bar{X}}}{\sigma_{e_{\bar{X}}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0,1)$$

No obstante, si la varianza σ^2 fuese desconocida la expresión anterior no sería válida. Podríamos en este caso proponer una discrepancia dada por:

$$d_{\bar{X}} = \frac{e_{\bar{X}}}{S_{e_{\bar{X}}}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \approx t_{n-1}$$

expresión en la que hemos sustituido σ^2 por su estimador S^2.

La deducción de estos modelos probabilísticos resulta sencilla aplicando resultados anteriores. En efecto, siempre que la población investigada sea normal, $X \approx \mathcal{N}(\mu, \sigma)$ y por estar todas las variables muestrales idénticamente distribuidas se tiene:

$$\frac{X_i}{n} \approx \mathcal{N}\left(\frac{\mu}{n}, \frac{\sigma}{n}\right)$$

Dado que dichas variables son independientes, podemos aplicar la propiedad de reproductividad del modelo normal, con lo cual obtenemos:

$$\bar{X} = \sum_{i=1}^{n} \frac{X_i}{n} \approx \mathcal{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Para el error aleatorio se tiene por tanto:

$$e_{\bar{X}} = \bar{X} - \mu \approx \mathcal{N}\left(0, \frac{\sigma}{\sqrt{n}}\right)$$

y en consecuencia para la discrepancia tipificada:

$$d_{\bar{X}} = \frac{e_{\bar{X}}}{\frac{\sigma}{\sqrt{n}}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0,1)$$

Así pues, la discrepancia es función de la dispersión poblacional y si ésta es conocida podemos acudir a las tablas del modelo normal estándar para obtener la probabilidad de que esta expresión se encuentre en cualquier intervalo $[a, b]$.

Si por el contrario la varianza poblacional resulta desconocida entonces no podríamos calcular las probabilidades anteriores y por tanto tendríamos que construir una expresión alternativa. Para ello acudimos a las distribuciones definidas en el apartado anterior, construyendo una distribución t de Student (t_n) como cociente entre una distribución $\mathcal{N}(0,1)$ y la raíz cuadrada de una χ^2_n entre sus grados de libertad.

En efecto, la discrepancia tipificada anterior

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0,1)$$

nos proporciona el numerador, mientras que para el denominador se tiene, gracias al teorema de Fisher:

$$\frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$$

Teniendo en cuenta que -también por el teorema de Fisher- ambas expresiones son v.a. independientes, se obtiene:

$$d_{\bar{X}} = \frac{\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2(n-1)}}} \approx t_{n-1}$$

expresión que puede también ser formulada como:

$$d_{\bar{X}} = \frac{\bar{X}-\mu}{S/\sqrt{n}}$$

Hasta la introducción de la t de Student por parte de Gosset (1908), el procedimiento habitual en los estudios inferenciales sobre la media consistía en calcular $\bar{X}$ y S^2 procediendo como si la media muestral siguiese un modelo

$$\bar{X} \approx \mathcal{N}\left(\mu, \frac{S}{\sqrt{n}}\right)$$

No obstante, teniendo en cuenta que este supuesto no necesariamente era correcto (al ignorar el riesgo asociado a la estimación de σ por S) Gosset intuía que este procedimiento aplicado a muestras pequeñas daría una falsa idea de exactitud. Dicho inconveniente se soluciona al introducir el modelo probabilístico t, que en tamaños pequeños de muestra viene representado por curvas campaniformes con menor apuntamiento que la normal.

Parece lógico pensar, y así lo confirmaremos en capítulos posteriores, que si utilizamos la información muestral para estimar los dos parámetros μ y σ^2 perderemos fiabilidad, de modo que para conseguir un mismo nivel de probabilidad, la distribución t de Student exige un mayor recorrido que el modelo normal.

Las dos distribuciones obtenidas (normal y t de Student) se basan en el supuesto de normalidad de X. Resulta interesante examinar cómo se verían afectados estos resultados si la población de partida no es normal.

El efecto de la ausencia de normalidad sobre las distribuciones anteriores dependerá de la información poblacional. Así, si la población X no es normal pero su varianza es conocida y el tamaño de muestra n es elevado, entonces la distribución de $d_{\bar{X}}$ puede ser aproximada por una $\mathcal{N}(0,1)$.

En efecto, en el desarrollo anterior hemos obtenido $\bar{X}$ como un agregado de v.a. independientes e idénticamente distribuidas. Por tanto, para tamaños suficientemente elevados, el TCL garantiza la convergencia a un modelo normal.

Cuando la varianza poblacional es conocida y el tamaño de muestra elevado, el procedimiento descrito resulta válido con hipótesis de normalidad o sin ella. Sin embargo, la situación cambia cuando la varianza es desconocida, ya que en ese caso el incumplimiento del supuesto de normalidad invalida el procedimiento desarrollado anteriormente para construir $d_{\bar{X}} \approx t_{n-1}$.

Obsérvese que en esta situación podemos aplicar el TCL, con lo cual el numerador convergería a una distribución $\mathcal{N}(0,1)$. Sin embargo, la no normalidad nos impide aplicar el teorema de Fisher, con lo cual no tenemos garantizada la independencia entre numerador y denominador ni tampoco la distribución χ^2_n para el denominador.

No obstante, es interesante señalar que aunque Gosset derivó la expresión de la t a partir del supuesto de poblaciones normales, se ha comprobado que las poblaciones no normales aproximadamente simétricas proporcionan expresiones que se aproximan mucho a la distribución t.

Este rasgo permite calificar a la distribución t de *robusta* y constituye una garantía de estabilidad para los procesos inferenciales. En general, la distribución t de Student asociada a la discrepancia de la media $d_{\bar{X}}$ resulta insensible a la hipótesis de normalidad cuando $n > 15$ y se ha comprobado que para tamaños pequeños de muestra dicha distribución se ve más afectada por la asimetría que por la no normalidad.

6.2.3. Inferencias sobre la varianza

Las inferencias sobre la varianza poblacional se basan en la varianza muestral S^2 cuya expresión se compara con σ^2 mediante el error relativo

$$e^R_{S^2} = \frac{S^2}{\sigma^2}$$

que adoptará valores superiores a la unidad si la varianza muestral sobreestima la poblacional y será inferior a 1 en caso contrario.

Obsérvese que en este caso la comparación del estimador con el parámetro se lleva a cabo por cociente y no por diferencia como se hacía para la esperanza. Este hecho se debe al propio concepto de dispersión, que tiene carácter multiplicativo (en este sentido, basta recordar que la dispersión no viene afectada por el origen sino únicamente por la escala y también que en el procedimiento de tipificación de variables aleatorias se eliminan la esperanza y la dispersión, pero mientras la primera se elimina mediante diferencias, en cambio la dispersión se elimina por cociente, dividiendo entre la desviación típica).

El error relativo $\frac{S^2}{\sigma^2}$ es una v.a. definida en función de la varianza muestral, cuya esperanza es unitaria por cumplirse $E(S^2) = \sigma^2$.

Por lo que respecta a la segunda etapa, si asumimos que la población de partida es normal, la expresión anterior puede ser ajustada a un modelo conocido con sólo multiplicar por los grados de libertad de la varianza muestral (n-1).

Definimos entonces la *discrepancia tipificada de la varianza*:

$$d_{S^2} = \frac{(n-1)S^2}{\sigma^2}$$

expresión que, gracias al teorema de Fisher, sigue un modelo chi-cuadrado con n-1 grados de libertad siempre que X sea normal.

A diferencia de los procesos inferenciales referidos a la media, los procesos asociados a la varianza resultan poco robustos, en el sentido de que el incumplimiento del supuesto de normalidad para la población de partida invalida la obtención de una distribución chi-cuadrado asociada a la varianza muestral.

6.2.4. Inferencias sobre proporciones

La proporción poblacional puede ser analizada como caso particular de la esperanza, cuando la población investigada es dicotómica o de Bernoulli $\mathcal{B}(p)$.

Como consecuencia, en esta situación la m.a.s. sería del tipo $(X_1, \ldots, X_n)$ donde cada X_i está definida como sigue:

$X_i = 0$	si el elemento i presenta el rasgo estudiado	$P(X_i = 1) = p$
$X_i = 1$	si el elemento i no presenta el rasgo estudiado	$P(X_i = 0) = 1 - p$

y el estimador analógico proporción muestral viene definido por la expresión:

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

siendo $\sum_{i=1}^{n} X_i$ una v.a. que sigue un modelo binomial $\mathcal{B}(n, p)$. Definimos el error como la diferencia entre la proporción muestral y la poblacional $e_{\hat{p}} = \hat{p} - p$, que presenta las características:

$$E\left(e_{\hat{p}}\right) = 0 \quad ; \quad Var\left(e_{\hat{p}}\right) = \frac{p(1-p)}{n}$$

La deducción de la varianza puede ser efectuada en los siguientes términos:

$$Var\left(e_{\hat{p}}\right) = Var\left(\hat{p}\right) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} p(1-p) = \frac{p(1-p)}{n}$$

La *discrepancia tipificada para la proporción* se obtiene entonces como:

$$d_{\hat{p}} = \frac{e_{\hat{p}}}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$$

y para tamaños muestrales suficientemente elevados es posible aplicar la convergencia del modelo binomial al normal, con lo cual se tiene: $d_{\hat{p}} \approx \mathcal{N}(0, 1)$.

En efecto, el teorema de De Moivre nos indica que:

$$\sum_{i=1}^{n} X_i \xrightarrow{L} N\left(np, \sqrt{np(1-p)}\right)$$

con lo cual para tamaños elevados de n asumimos

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i \approx N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

y por tanto se obtienen también aproximaciones normales para el error y la discrepancia tipificada:

$$e_{\hat{p}} \approx \mathcal{N}\left(0, \sqrt{\frac{p(1-p)}{n}}\right) \quad ; \quad d_{\hat{p}} \approx \mathcal{N}(0,1)$$

Puede observarse que el error estándar o desviación típica del estimador, $\sqrt{\frac{p(1-p)}{n}}$, depende de la proporción p y por tanto, si esta desviación típica fuese conocida podríamos despejar de forma exacta el valor de la proporción, sin necesidad de utilizar la información muestral para su estimación.

No obstante, en el caso de la proporción esta afirmación no es enteramente correcta. En muchos supuestos asumiremos como conocido el valor de la dispersión (a veces en problemas reales aprovechamos cierta información obtenida de algún censo anterior o relativo a una actividad afín a la que estamos estudiando, con lo cual estamos asumiendo que no cambia el riesgo). Sin embargo, la información sobre la dispersión no es válida para establecer el valor concreto del parámetro p, ya que los hábitos, el entorno y otras características influirán en alteraciones de la proporción poblacional, aun cuando la banda de riesgo que acompaña a estas estimaciones en períodos no muy grandes de tiempo se mantenga constante.

Si la varianza del error es desconocida, ésta podría aproximarse por su estimador insesgado $\frac{\hat{p}(1-\hat{p})}{n-1}$.

En efecto, en el capítulo anterior hemos comprobado que se cumple:

$$E\left[\frac{\hat{p}(1-\hat{p})}{n-1}\right] = \frac{p(1-p)}{n} = Var\left(\hat{p}\right)$$

La aproximación normal para la discrepancia tipificada de la proporción exige tamaños de muestra elevados ya que se basa en los teoremas límites. Debemos tener presente que la variable aleatoria X que aparece en el numerador de la proporción muestral es discreta y sin embargo las transformaciones sucesivas que operamos sobre ella (cálculo de errores y tipificación de los mismos) desembocan en un modelo aproximadamente normal; por tanto, cuando nos interese aproximar probabilidades para valores concretos de X conviene llevar a cabo la corrección de continuidad estudiada en el capítulo 4.

Los desarrollos precedentes son también aplicables si nuestro objetivo inferencial es una combinación de proporciones (generalmente diferencias de proporciones referidas a varias poblaciones, ...).

6.2.5. Inferencias sobre la diferencia de medias

En el ámbito económico-empresarial resultan frecuentes las situaciones en las que se deben realizar inferencias relativas a las esperanzas de dos poblaciones. Este sería el caso si deseamos estimar la diferencia de ingresos medios entre dos regiones, contrastar si ha aumentado la esperanza de vida a lo largo del tiempo o si se ha producido ganancia salarial en una población tras la aplicación de cierta medida económica.

En estos casos, aun siendo perfectamente válido el planteamiento visto en los apartados anteriores, existe una mayor complejidad en el estudio como consecuencia de la diversidad de situaciones que pueden presentarse. Así, debemos considerar si las

muestras a partir de las cuales extraemos información son o no independientes, si los parámetros poblacionales son conocidos, etc.

Las distintas situaciones posibles presentan como objetivo común llevar a cabo inferencias sobre el parámetro $\mu_X - \mu_Y = E(X) - E(Y)$, pero la casuística que puede aparecer es amplia, tal y como se describe en los apartados siguientes.

6.2.5.1. Diferencia de medias con datos pareados

Comencemos por considerar aquellas situaciones en las que extraemos la información de dos muestras aleatorias que no son independientes. Más concretamente, asumiremos que disponemos de dos *muestras dependientes con datos pareados*, esto es, que sobre cada integrante de la muestra estudiamos dos características dependientes.

Ilustraciones de esta situación podrían ser las siguientes: muestras de individuos sobre los que observamos renta y gasto, muestras de empresas sobre las que estudiamos los beneficios antes y después de impuestos, muestras de artículos sobre los que analizamos precio de adquisición y de venta, etc.

Si denotamos por X e Y las características aleatorias analizadas y por $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_n)$ las muestras respectivas [¿por qué tienen el mismo tamaño?], resultaría posible definir la v.a. diferencia $D = X - Y$, pasando así de las dos muestras anteriores a una muestra única $(D_1, ..., D_n)$.

Nuestro objetivo será realizar inferencias sobre $E(D) = E(X) - E(Y)$ para lo cual debemos basarnos en la información muestral $(D_1, ..., D_n)$. Por tanto el problema es idéntico al ya visto con anterioridad para una media poblacional, y conducirá en general a una distribución t de Student con $n - 1$ grados de libertad.

Dado que nuestro objetivo inferencial es $E(D)$, si consideramos como estimador la diferencia media, el error correspondiente puede ser definido como: $e_{\bar{D}} = \bar{D} - E(D)$.

Resulta sencillo comprobar que la esperanza de este error es nula y su varianza viene dada por la expresión:

$$Var\left(e_{\bar{D}}\right) = Var\left(\bar{D}\right) = \frac{Var(D)}{n}$$

que sólo podrá ser determinada si conocemos las varianza de D o bien las varianzas de las variables originales y su covarianza.

En general, la varianza anterior resulta desconocida, por lo cual debería ser estimada a partir de la muestra mediante la expresión:

$$S^2_{e_{\bar{D}}} = \frac{S^2_D}{n} = \frac{1}{n}\left[\frac{\sum_{i=1}^{n}\left(D_i - \bar{D}\right)^2}{n-1}\right] = \frac{1}{n}\left(\frac{\sum_{i=1}^{n}\left[(X_i - Y_i) - \left(\bar{X} - \bar{Y}\right)\right]^2}{n-1}\right)$$

Por tanto, bajo el supuesto de normalidad para la variable D la expresión de la discrepancia tipificada vendrá dada por:

$$d_{\bar{D}} = \frac{e_{\bar{D}}}{\frac{S_D}{\sqrt{n}}} = \frac{\bar{D} - E(D)}{\frac{S_D}{\sqrt{n}}} \approx t_{n-1}$$

Obsérvese que en el supuesto poco frecuente de que la varianza de D resultase conocida se tendría

$$d_{\bar{D}} = \frac{e_{\bar{D}}}{\sqrt{Var\left(e_{\bar{D}}\right)}} \approx \mathcal{N}(0,1)$$

La distribución normal resultaría también aplicable cuando la variable D no fuera normal, siempre que el tamaño de muestra fuese elevado y la varianza conocida. En tal situación, la aplicación del TCL conduciría a la discrepancia tipificada

$$d_{\bar{D}} = \frac{e_{\bar{D}}}{\sqrt{Var\left(e_{\bar{D}}\right)}} \xrightarrow{L} \mathcal{N}(0,1)$$

6.2.5.2. Diferencia de medias con muestras independientes

Supongamos ahora que a partir de las poblaciones X e Y hemos extraído muestras aleatorias independientes: $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$.

Si denotamos por μ_X y μ_Y las medias poblacionales y por σ_X^2 y σ_Y^2 las varianzas de X e Y respectivamente, definimos el estimador de la diferencia de las medias muestrales $T = \bar{X} - \bar{Y}$, a partir del cual obtenemos el error aleatorio $e_{\bar{X}-\bar{Y}} = (\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)$, cuyas características son:

$$E\left(e_{\bar{X}-\bar{Y}}\right) = 0 \quad ; \quad Var\left(e_{\bar{X}-\bar{Y}}\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

En efecto, se tiene:

$$Var\left(e_{\bar{X}-\bar{Y}}\right) = Var\left(\bar{X} - \bar{Y}\right) = Var\left(\bar{X}\right) + Var\left(\bar{Y}\right) - 2Cov\left(\bar{X}, \bar{Y}\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}$$

puesto que al ser las muestras independientes $Cov(X,Y) = 0$ y en consecuencia $Cov(\bar{X}, \bar{Y}) = 0$ [¿Por qué?]

Así pues, adoptamos como discrepancia en este caso la expresión:

$$d_{\bar{X}-\bar{Y}} = \frac{e_{\bar{X}-\bar{Y}}}{\sqrt{Var\left(e_{\bar{X}-\bar{Y}}\right)}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{Var\left(\bar{X} - \bar{Y}\right)}}$$

que podemos ajustar a modelos probabilísticos conocidos en función de las hipótesis de partida.

Supongamos que las poblaciones X e Y se distribuyen normalmente:

$$X \approx \mathcal{N}(\mu_X, \sigma_X)\ ,\ Y \approx \mathcal{N}(\mu_y, \sigma_y)$$

Siempre que las varianzas poblacionales de X e Y sean conocidas, también lo será la varianza de $e_{\bar{X}-\bar{Y}}$. En consecuencia, la expresión

$$d_{\bar{X}-\bar{Y}} = \frac{e_{\bar{X}-\bar{Y}}}{\sqrt{Var\left(e_{\bar{X}-\bar{Y}}\right)}} = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}}$$

se distribuye según un modelo $\mathcal{N}(0,1)$.

En efecto, según hemos visto al estudiar la distribución de la media,

$$\bar{X} \approx \mathcal{N}\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \text{ e } \bar{Y} \approx \mathcal{N}\left(\mu_Y, \frac{\sigma_Y}{\sqrt{m}}\right)$$

Se trata de muestras independientes y aplicando la propiedad de reproductividad del modelo normal se tiene

$$\bar{X}-\bar{Y} \approx N\left(\mu_X\text{-}\mu_Y, \sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)$$

Así pues, se obtienen las siguientes distribuciones para el error aleatorio y para la discrepancia tipificada:

$$e_{\bar{X}-\bar{Y}} = \bar{X}-\bar{Y}-(\mu_X\text{-}\mu_Y) \approx \mathcal{N}\left(0, \sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)$$

$$d_{\bar{X}-\bar{Y}} = \frac{e_{\bar{X}-\bar{Y}}}{\sqrt{Var\left(e_{\bar{X}-\bar{Y}}\right)}} = \frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}} \approx \mathcal{N}(0,1)$$

Si no se verifica la hipótesis de normalidad pero los tamaños de las dos muestras, n y m, son suficientemente elevados, entonces -gracias al teorema central del límite- obtenemos una distribución de la discrepancia que es aproximadamente $\mathcal{N}(0,1)$.

La comprobación de este caso se limita a aplicar el teorema central del límite, ya que si n y m son elevados, entonces obtenemos:

$$\bar{X} \approx \mathcal{N}\left(\mu_X, \frac{\sigma_X}{\sqrt{n}}\right) \text{ e } \bar{Y} \approx \mathcal{N}\left(\mu_Y, \frac{\sigma_Y}{\sqrt{m}}\right)$$

siendo válido el resto del razonamiento sin más que tener presente que trabajamos con distribuciones aproximadas.

Obsérvese que si una de las dos muestras tuviese un tamaño inferior a 30 ya no podríamos aplicar este desarrollo.

En el caso de que las variables se distribuyan normalmente pero las varianzas poblacionales sean desconocidas, la expresión anterior de la discrepancia no resulta válida. No obstante, este problema podrá ser solventado si las varianzas, aunque desconocidas, son coincidentes: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$.

En esta situación, la varianza poblacional σ^2 podrá ser estimada utilizando la información que proporcionan las dos muestras, mediante la expresión:

$$S^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

que como vemos es una media ponderada de las varianzas muestrales, adoptando como pesos los grados de libertad.

Utilizando dicha estimación se llegaría a una discrepancia dada por la expresión

$$d_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} \approx t_{n+m-2}$$

que se distribuye según un modelo t de Student con $n+m-2$ grados de libertad.

La deducción de la expresión anterior se realiza asumiendo los siguiente supuestos:

- Normalidad de las poblaciones: $X \approx \mathcal{N}(\mu_X, \sigma_X)$, $Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$
- Igualdad de varianzas: $\sigma_X^2 = \sigma_Y^2 = \sigma^2$
- Muestras $(X_1, \ldots, X_n)$ e $(Y_1, \ldots, Y_n)$ independientes

Bajo estas hipótesis, el error asociado a la diferencia de medias de dos muestras independientes: $e_{\bar{X}-\bar{Y}} = (\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)$, se distribuye normalmente con las siguientes características:

$$E\left(e_{\bar{X}-\bar{Y}}\right) = 0 \; ; \; Var\left(e_{\bar{X}-\bar{Y}}\right) = \frac{\sigma_X^2}{n} + \frac{\sigma_X^2}{m} = \sigma^2\left(\frac{1}{n}+\frac{1}{m}\right)$$

Así pues, la discrepancia tipificada vendría dada inicialmente por:

$$\frac{e_{\bar{X}-\bar{Y}}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}} \approx \mathcal{N}(0,1)$$

pero, dado que en esta expresión aparece el parámetro desconocido σ, resulta necesario acudir a la información proporcionada por las dispersiones muestrales.

Gracias al supuesto de normalidad, es posible aplicar a cada muestra el teorema de Fisher, según el cual se verifica:

$$(n-1)\frac{S_X^2}{\sigma^2} \approx \chi_{n-1}^2 \; ; \; (m-1)\frac{S_Y^2}{\sigma^2} \approx \chi_{m-1}^2$$

Teniendo en cuenta que ambas variables son independientes por serlo las muestras y aplicando la reproductividad de la distribución chi-cuadrado se obtiene:

$$\frac{(n-1)S_X^2 + (m-1)S_Y^2}{\sigma^2} \approx \chi_{n-1}^2 + \chi_{m-1}^2 = \chi_{n+m-2}^2$$

El teorema de Fisher garantiza además que esta distribución chi-cuadrado y la expresión normal anterior son independientes, por lo cual podemos definir a partir de ambas una nueva variable distribuida según un modelo t de Student:

$$\frac{\frac{e_{\bar{X}-\bar{Y}}}{\sigma\sqrt{\frac{1}{n}+\frac{1}{m}}}}{\frac{1}{\sigma}\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}}\sqrt{\frac{1}{n}+\frac{1}{m}}} \approx t_{n+m-2}$$

Resulta interesante analizar cómo se ve afectada esta expresión por la alteración de los supuestos de partida.

Comenzando por la hipótesis de normalidad de las poblaciones investigadas, ya hemos comentado en epígrafes anteriores que la distribución t de Student resulta ser muy robusta, es decir, poco sensible a la no normalidad.

En concreto, los estudios efectuados por Barlett (1935), Gayen (1949,1951) y Boneau (1960) ponen de manifiesto que, siempre que las muestras investigadas tengan tamaños coincidentes, la distribución de la expresión no se ve alterada por la no normalidad (incluso cuando las poblaciones de partida sean muy asimétricas). De modo similar, si las distribuciones de partida son aproximadamente simétricas, la expresión resulta robusta aun cuando las muestras tengan tamaños distintos.

La alteración del supuesto de igualdad de varianzas invalida la deducción efectuada para la t de Student, dando lugar al problema conocido como de Behrens-Fisher, ampliamente tratado por varios autores sin que exista una solución universalmente aceptada.

En general, las propuestas para solucionar este problema parten de la consideración de las varianzas muestrales, estimadores consistentes de las varianzas poblacionales, que conducen a la expresión:

$$d_{\bar{X}-\bar{Y}} = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}}$$

que sigue aproximadamente una distribución t de Student con grados de libertad:

$$g.l. = \frac{(n-1)\frac{S_X^2}{n} + (m-1)\frac{S_Y^2}{m}}{\frac{S_X^2}{n} + \frac{S_Y^2}{m}}$$

Esta aproximación, que se debe a Cochran (1964), conduce a un número de g.l. que, en general, no será entero, por lo cual cuando consultemos en las tablas tendremos que buscar el número de g.l. más próximo o bien interpolar.

Por último, el supuesto de independencia entre las muestras resulta de gran importancia ya que, si esta hipótesis se incumple, las expresiones anteriormente deducidas pierden su validez. De ahí el interés de distinguir las inferencias sobre diferencia de medias con muestras independientes de las correspondientes a datos pareados, analizada con anterioridad.

El planteamiento recogido en este apartado para la diferencia de medias es susceptible de ser generalizado a la suma o a cualquier combinación lineal de esperanzas del tipo $\alpha\mu_X + \beta\mu_Y$.

Aunque no recogemos aquí estos desarrollos, en el supuesto más sencillo de normalidad y varianzas conocidas, la discrepancia normalizada viene dada por la expresión:

$$d_{\alpha\bar{X}-\beta\bar{Y}} = \frac{\alpha\bar{X} + \beta\bar{Y} - (\alpha\mu_X + \beta\mu_Y)}{\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}} \approx \mathcal{N}(0,1)$$

[Llevar a cabo la deducción de esta expresión]

Otra posible generalización consiste en llevar a cabo comparaciones de las esperanzas en más de dos poblaciones. Este planteamiento conduce al análisis de varianza, que será analizado en un capítulo posterior.

6.2.6. Inferencias sobre la razón de varianzas

Como hemos visto en el apartado anterior, al analizar la diferencia de medias poblacionales resulta de gran importancia conocer si las varianzas de las poblaciones investigadas son coincidentes. En consecuencia, estaremos interesados en llevar a cabo inferencias sobre el parámetro $\frac{\sigma_X^2}{\sigma_Y^2}$, o más concretamente, contrastar si esta expresión es unitaria.

Supongamos dos poblaciones X e Y normales:

$$X \approx \mathcal{N}(\mu_X, \sigma_X)\ ,\ Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$$

y consideremos dos muestras independientes de cada población $(X_1, \ldots, X_n), Y_1, \ldots, Y_n)$.

El estimador analógico del parámetro investigado será

$$T = \frac{S_X^2}{S_Y^2}$$

y definiremos un error relativo que -como ya hemos justificado en las inferencias sobre la varianza- resulta más adecuado para las características de dispersión. Dicho error

$$e^R_{S_X^2/S_Y^2} = \frac{\frac{S_X^2}{S_Y^2}}{\frac{\sigma_X^2}{\sigma_Y^2}} = \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2}$$

presenta esperanza unitaria y distribución conocida, por lo cual no es necesario efectuar ningún ajuste sobre el mismo. Se define así la discrepancia tipificada:

$$d_{\frac{S_X^2}{S_Y^2}} = \frac{\frac{S_X^2}{S_Y^2}}{\frac{\sigma_X^2}{\sigma_Y^2}} = \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2}$$

expresión que se adapta a un modelo F de Snedecor con $n-1$ g.l. en el numerador y m-1 en el denominador.

Aplicando el teorema de Fisher a las dos muestras, se tiene

$$(n-1)\frac{S_X^2}{\sigma_X^2} \approx \chi^2_{n-1}\ ;\ (m-1)\frac{S_Y^2}{\sigma_Y^2} \approx \chi^2_{m-1}$$

Además ambas variables son independientes por serlo las muestras, luego el cociente de dos variables chi-cuadrado divididas por sus g.l. define una variable F de Snedecor. Así:

$$\frac{\frac{(n-1)S_X^2/\sigma_X^2}{n-1}}{\frac{(m-1)S_Y^2/\sigma_Y^2}{m-1}} = \frac{\frac{S_X^2}{\sigma_X^2}}{\frac{S_Y^2}{\sigma_Y^2}} = \frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2} \approx F^{n-1}_{m-1}$$

Gracias a la propiedad de inversión de la distribución F, es sencillo comprobar que en el caso de que las inferencias fuesen referidas a $\frac{\sigma_Y^2}{\sigma_X^2}$ se llegaría a un modelo F_{n-1}^{m-1} [Compruébese]

Este proceso resulta poco robusto, ya que la distribución F se verá muy afectada por posibles alteraciones en el supuesto de normalidad de las poblaciones investigadas.

6.2.7. Inferencias sobre otras características

Es evidente que las situaciones señaladas no agotan toda la casuística inferencial. Sin embargo, sí cubren los parámetros de mayor interés práctico.

En ciertas ocasiones podemos estar interesados en parámetros que aparecen relacionados mediante alguna expresión con las esperanzas o las varianzas poblacionales, por lo cual resultarían aplicables los procedimientos anteriormente descritos, utilizando como punto de partida los estimadores adecuados en cada caso y explicitando los correspondientes supuestos.

A modo de ilustración, podríamos estar interesados en llevar a cabo inferencias sobre el parámetro b en un modelo uniforme $\mathcal{U}[0, b]$ (obsérvese que en este caso $b = 2\mu$) o sobre la desigualdad en un modelo de Pareto (es decir, la expresión $\frac{\alpha}{\alpha-1}$, que coincide con el ratio $\frac{\mu}{x_0}$).

Por otra parte, existen situaciones en las que, por el propio carácter del parámetro investigado, la metodología inferencial cambia considerablemente. Este será el caso de las inferencias relativas a la mediana o, más en general, a cualquier característica de posición no central (cuantiles).

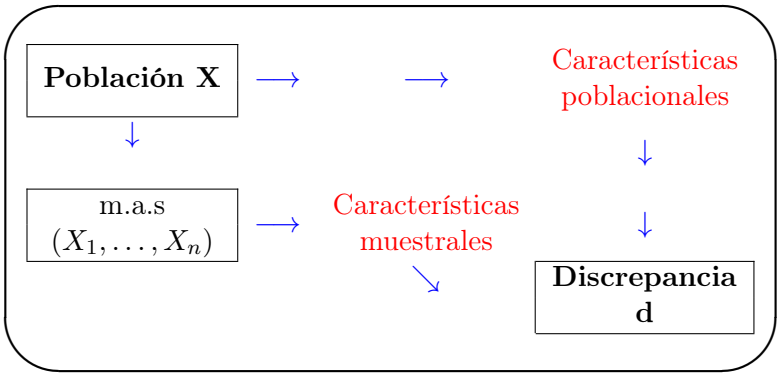
Las inferencias relativas a la mediana Me se abordan desde un marco no paramétrico, esto es, sin explicitar supuestos sobre la población de partida X. (Obsérvese que bajo la hipótesis de normalidad se cumple $Me = \mu$, por lo cual nos remitiríamos a las deducciones ya estudiadas para la media).

La característica que identifica al parámetro Me es por definición su probabilidad acumulada $F_X(Me) = P(X \leq Me) = 0,5$. Así pues, a partir de una m.a.s. $(X_1, \ldots, X_n)$ podemos también garantizar para una X_i cualquiera: $P(X_i \leq Me) = 0,5$.

Dado que nuestro objetivo es llegar a un modelo probabilístico conocido que utilice las informaciones muestral y poblacional, definiremos ahora una v.a. Z que recoge el número de observaciones muestrales inferiores o iguales a Me. Dicha variable seguirá un modelo $\mathcal{B}(n, 0,5)$ y en consecuencia podremos calcular cualquier probabilidad asociada a valores concretos de Z, y a partir de ellas llevar a cabo inferencias (estimaciones o contrastes) del parámetro Me.

La utilización de la mediana presenta como ventaja su robustez, pero en cambio supone pérdidas de eficiencia con respecto a otros procesos inferenciales. Para solucionar este inconveniente, en ocasiones se defiende la utilización de una *media ajustada* (*trimmed mean*) obtenida como promedio de una muestra de la que se han eliminado las observaciones extremas (por exceso y por defecto).

Figura 6.6.: Inferencias genéricas



La media ajustada de nivel k para una muestra n se obtiene como promedio de sus $n-2k$ observaciones centrales. Puede comprobarse que la mediana se corresponde con el caso particular de nivel $\frac{n-1}{2}$ para n impar y $\frac{n-2}{2}$ para n par.

De modo similar, para cualquier cuantil Q se tiene una probabilidad asociada p_Q, por lo cual a partir de la muestra garantizamos $P(X_i \leq Q) = p_Q$, definiendo ahora la variable Z: "número de observaciones muestrales inferiores a Q" que sigue un modelo $\mathcal{B}(n, p_Q)$.

6.2.8. Inferencias genéricas sobre poblaciones

Cuando los procesos inferenciales no tienen como objetivo una característica concreta, el planteamiento cambia ligeramente respecto al visto en apartados anteriores (figura 6.6).

En efecto, en este caso no estamos interesados en parámetros concretos sino en características más globales de la población, tales como su distribución de probabilidad. Es habitual también, en el caso de que nos interesen varias poblaciones, que investiguemos la independencia entre ambas o su homogeneidad.

En este tipo de inferencias resulta de gran interés la *distribución chi-cuadrado*, que surge, siguiendo el esquema anterior, en los siguientes términos: para cada elemento de la muestra comparamos la información muestral (C_i^m) con la correspondiente información teórica poblacional (C_i^p), construyendo de este modo unos errores $e = (C_i^m - C_i^p)$ que, por depender de la información muestral, son aleatorios.

Dado que estos errores deben ser sintetizados y reducidos a un modelo probabilístico teórico, el procedimiento de tipificación consiste en este caso en elevar los errores al cuadrado (evitando así la compensación de signos), y dividir entre la característica teórica.

Una vez llevado a cabo este procedimiento para todos los componentes de la muestra, se obtiene la expresión de la discrepancia:

Tabla 6.5.: Inferencias sobre características genéricas

Característica investigada	C_i^m	C_i^p	Error e	Expresión Ajustada	Grados de Libertad
Distribución de probabilidad	n_i	np_i	$n_i - np_i$	$\sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$	$r-1-k$ r=nº de clases k=nº de parámetros estimados
Independencia	n_{ij}	$\frac{n_{i\cdot}n_{\cdot j}}{n}$	$n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}$	$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$	$(r-1)(s-1)$ r,s=nº de clases de las dos características investigadas
Homogeneidad	n_{ij}	$\frac{n_{i\cdot}n_{\cdot j}}{n}$	$n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}$	$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$	$(r-1)(s-1)$ r=nº de poblaciones s=modalidades de la característica investigada

$$d = \sum_{i=1}^{n} \frac{(C_i^m - C_i^p)^2}{C_i^p}$$

que se distribuye, siempre que se garanticen ciertos tamaños muestrales mínimos, según un modelo chi-cuadrado con un número de grados de libertad igual a n (tamaño de muestra) menos k (número de restricciones).

El procedimiento descrito es aplicable a varias situaciones diferenciadas, que aparecen sintetizadas en la figura 6.5.

En general estas expresiones son aplicadas al contraste de ciertas hipótesis (contrastes de bondad de ajuste, contrastes de independencia y contrastes de homogeneidad) que analizaremos con detalle en un capítulo posterior, por lo cual nos limitamos aquí a comentar brevemente sus rasgos más destacables.

Las inferencias basadas en la distribución chi-cuadrado se llevan a cabo agrupando la información muestral en intervalos o clases y comparando las frecuencias observadas en la muestra con las frecuencias esperadas (esto es, las asociadas a la característica poblacional investigada).

En concreto, cuando llevamos a cabo *inferencias sobre la distribución de probabilidad*, denotamos por n_i la frecuencia observada en el intervalo i-ésimo y por np_i su frecuencia esperada (calculada asumiendo que la variable sigue determinado modelo probabilístico, que nos proporciona para cada intervalo su probabilidad p_i).

Repitiendo el proceso para todos los intervalos se llega a la discrepancia tipificada:

$$\sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \approx \chi^2_{r-1-k}$$

cuyos grados de libertad se obtienen como diferencia entre el tamaño muestral (r clases o intervalos) y el número de restricciones (al menos existe una:

$$\sum_{i=1}^{r} np_i = \sum_{i=1}^{r} n_i = n$$

pero podría haber nuevas restricciones, k en general, si el modelo probabilístico investigado depende de k parámetros que debemos estimar).

Las condiciones de convergencia hacia la distribución chi-cuadrado exigen que ninguna de las frecuencias esperadas np_i adopte un valor bajo, considerándose habitualmente 5 como frecuencia mínima para un intervalo.

Si nuestro objetivo es investigar (en general contrastar) la *independencia entre dos poblaciones o características*, debemos partir de la información de una muestra conjunta, agrupada en clases, para las dos variables investigadas. Para construir la discrepancia en este caso se comparan las frecuencias conjuntas observadas (n_{ij}) con las esperadas, que serán $\frac{n_{i\cdot}n_{\cdot j}}{n}$ bajo la condición de independencia, llegando así a la expresión:

$$\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}} \approx \chi_{(r-1)(s-1)} \tag{6.3}$$

cuyos grados de libertad se obtienen como producto de los asociados a cada una de las características investigadas (para la primera característica se tienen $r-1$ g.l., ya que existen r clases sometidas a la restricción $\sum_{i=1}^{r} n_{i\cdot} = n$; de modo análogo para la segunda se tienen $s-1$ g.l., ya que las s clases deben cumplir $\sum_{j=1}^{s} n_{\cdot j} = n$).

La independencia puede ser investigada sobre características tanto cuantitativas como cualitativas, siendo aplicable en ambas situaciones la distribución chi-cuadrado deducida, en la que tan sólo intervienen las frecuencias. Para que este modelo probabilístico quede garantizado es necesario que ninguna de las frecuencias esperadas sea excesivamente pequeña.

Por último, la distribución chi-cuadrado es aplicable al estudio de la homogeneidad de varios colectivos o poblaciones respecto a una característica. En este caso, aunque se produce un cambio de planteamiento, las expresiones resultan muy similares a las estudiadas anteriormente para la independencia (ecuación 6.3), si bien en este caso r representa el número de colectivos o poblaciones investigados, n_i recoge el tamaño de muestra correspondiente al colectivo r y s indica las modalidades excluyentes que presenta la característica respecto a la cual se investiga la homogeneidad.

7. Estimación

Cada día, los medios de comunicación difunden noticias basadas en estimaciones: la subida media de los precios durante el último mes, la audiencia que ha tenido una retransmisión deportiva en televisión, la proporción de votos que obtendría un partido político en las próximas elecciones, ... En todos estos ejemplos las noticias se basan en información parcial, no exhaustiva, y por tanto los datos publicados no serán exactos, pero sin embargo resultarán de gran utilidad.

Así, en el caso de la subida media de precios, la información proporcionada por el IPC (Indice de Precios de Consumo) no puede medir los precios de todos los bienes y servicios consumidos y por tanto las noticias que se publican cada mes en los medios de comunicación corresponden a una estimación realizada por el Instituto Nacional de Estadística (INE) a través de un muestreo muy completo. De modo similar, las audiencias de programas televisivos se estiman a partir de encuestas entre los espectadores y la intención de voto se estima mediante encuestas o sondeos electorales.

Es importante tener presente que el hecho de que las estimaciones no tengan carácter exacto no afecta a su veracidad ni a su utilidad: en el contexto socioeconómico nos interesará disponer de aproximaciones fiables de la subida de precios, la renta percápita, la tasa de paro, ... y la calidad de las estimaciones dependerá básicamente de dos factores que ya hemos analizado en los temas precedentes: en primera instancia la información muestral disponible (que será la "materia prima" en la que se fundamenta cualquier estudio inferencial) y en segundo lugar la "calidad" de las técnicas aplicadas (término que abarca tanto las expresiones de los estimadores como el método de estimación utilizado).

Como consecuencia de los condicionantes señalados, a menudo encontramos diferentes estimaciones para los mismos parámetros. Así, distintos organismos e instituciones publican sus estimaciones y predicciones referidas al crecimiento del PIB, el IPC o la tasa de paro, con resultados no coincidentes.

Analizando los comentarios anteriores llegamos a la conclusión de que disponemos de algunas "garantías" para los resultados de nuestras estimaciones. Tal y como indica la Figura 7.1 la primera de estas garantías será el diseño muestral, ya que trabajamos con muestras aleatorias y a partir de ellas inferimos resultados para el conjunto poblacional.

Además, conocemos los requisitos exigibles a los estimadores por lo cual podremos saber hasta qué punto estamos trabajando con instrumental adecuado. Como consecuencia de la propia definición vista para estas propiedades, si el estimador es insesgado garantiza que no introduce errores sistemáticos y si es eficiente consigue una dispersión mínima respecto al parámetro.

Figura 7.1.: Garantías del proceso de estimación

- Muestras **aleatorias** garantizan que la selección es probabilística
- **Estimadores** y sus propiedades:
  - **Insesgados**: garantizan que no hay errores sistemáticos
  - **Eficientes**: garantizan mínimo riesgo respecto al parámetro
  - **Suficientes**: garantizan aprovechamiento de la información
  - **Consistentes**: garantizan la convergencia al parámetro
- **Discrepancia** asociada al estimador con distribución conocida: garantiza cierta fiabilidad en la estimación.

La suficiencia por su parte es una garantía de aprovechamiento de la totalidad de la información muestral mientras que la consistencia asegura la convergencia al valor verdadero.

Por otra parte, en el capítulo anterior hemos visto que es posible especificar una distribución probabilística asociada a la discrepancia del estimador respecto al parámetro. Basándonos en estas distribuciones podremos efectuar afirmaciones probabilísticas, garantizando que cierta proporción de las discrepancias se sitúa entre ciertos márgenes.

Este último tipo de garantía aparece directamente relacionado con el procedimiento de estimación, para el que son posibles dos modalidades que estudiaremos con detalle en los epígrafes que siguen: la primera de ellas -denominada estimación puntual- proporciona una aproximación concreta del parámetro desconocido mientras la segunda -estimación por intervalos- consiste en estimar un intervalo o banda de confianza.

7.1. Estimación puntual y por intervalos

Un proceso global de estimación consiste en utilizar de forma coordinada gran parte de las técnicas que hemos analizado en capítulos anteriores. Gracias a ellas estamos ya en condiciones de responder a preguntas del tipo ¿qué característica de la población tratamos de aproximar?, ¿cuál es la expresión más adecuada como estimador?, ¿qué modelo probabilístico podemos llegar a obtener para la discrepancia asociada a ese estimador?

Nos preocuparemos ahora de estudiar la estimación de la característica investigada a la que llegaremos con la información muestral disponible, y de evaluar la fiabilidad asociada a dicha estimación.

Consideramos como punto de partida del proceso de estimación una muestra aleatoria simple $(X_1, \ldots, X_n)$ extraída de la población investigada, a partir de la cual definimos un estimador $T = T(X_1, \ldots, X_n)$ que será también una v.a.

Para cada muestra concreta $(x_1, \ldots, x_n)$, el estimador T proporciona una estimación puntual determinada $t = T(x_1, \ldots, x_n)$ que aproxima el parámetro θ desconocido y por tanto conlleva un error concreto $e = t - \theta$.

¿Será aceptable esta estimación de θ? Para responder a esta pregunta sería necesario conocer la magnitud del error cometido, objetivo que no resulta factible en la práctica por ser dicho error función del parámetro θ desconocido.

Debemos tener presente que las propiedades estudiadas para los estimadores garantizan un buen comportamiento probabilístico de los mismos pero no permiten efectuar ninguna afirmación sobre las estimaciones particulares. De este modo es perfectamente posible que, aun utilizando un estimador centrado, eficiente, suficiente y consistente cometamos errores de gran magnitud al estimar el parámetro.

En definitiva, las propiedades de los estimadores avalan el instrumento utilizado pero no cada resultado particular obtenido con éste. De hecho, a partir de la expresión única de un estimador, cada muestra concreta nos conducirá a estimaciones diferentes, que llevan asociados los correspondientes errores.

En el capítulo anterior hemos analizado las distribuciones probabilísticas asociadas a los procesos inferenciales relativos a diversos parámetros. No obstante, por tratarse habitualmente de modelos continuos (normal, chi-cuadrado, t de Student o F de Snedecor) es imposible cuantificar probabilidades para estimaciones concretas de un parámetro desconocido θ.

Consideremos a modo de ejemplo que nuestro objetivo es estimar la renta esperada en determinada población. A partir de una m.a.s. de tamaño n y dado que la media muestral -como ya hemos visto con anterioridad- es un estimador adecuado de μ podríamos enunciar afirmaciones del tipo:

$$E\left(\bar{X}\right) = \mu \; ; \; Var\left(\bar{X}\right) = \frac{\sigma^2}{n} \; ; \; \lim_{n\to\infty} P\left(\left|\bar{X} - \mu\right| > k\sigma_{\bar{X}}\right) = 0$$

o bien, en términos de las correspondientes discrepancias tipificadas:

$$E\left(d_{\bar{X}}\right) = 0 \; ; \; Var\left(d_{\bar{X}}\right) = 1 \; ; \; \lim_{n\to\infty} P\left(|d_{\bar{X}}| > k\right) = 0$$

Sin embargo, una vez que partimos de una muestra concreta no tiene sentido plantearse ese tipo de afirmaciones, (habremos obtenido, por ejemplo, $\bar{x} = 50$, y no resultará posible asignar una probabilidad al error asociado a este valor concreto). En realidad para cada estimación puntual sólo cabrían dos grandes posibilidades excluyentes: haber estimado de forma exacta el valor de μ o bien (alternativa más habitual) haber cometido algún error en dicha estimación.

Si queremos llegar a efectuar afirmaciones probabilísticas sobre un resultado inferencial, cabe la posibilidad de ampliar la óptica de la estimación pasando a la construcción de *intervalos o bandas de confianza*.

Este procedimiento, que aparece conectado de modo natural con la estimación pun-

tual, resulta sin embargo más general. A grandes rasgos, la aportación de la estimación por intervalos respecto a la puntual consiste en hacer explícito el error o discrepancia inherente al proceso de estimación, incorporando márgenes -por exceso y por defecto- respecto a la estimación puntual.

Siguiendo con el ejemplo anterior, la estimación de la renta esperada se efectuaría ahora adoptando como punto de partida la media muestral a la que incorporamos márgenes de error ϵ, cuya determinación estudiaremos más adelante.

Al igual que ocurría en la estimación puntual, cada muestra concreta llevará a unos valores particulares de $\bar{x} - \epsilon$ y $\bar{x} + \epsilon$ (47 y 53, por ejemplo) que determinan un intervalo en el que no podemos asegurar que esté contenido el parámetro μ. Sin embargo, la gran novedad es que ahora resulta posible efectuar afirmaciones probabilísticas referidas a los intervalos genéricos $\left(\bar{X} - \epsilon, \bar{X} + \epsilon\right)$.

Para clarificar las conexiones y diferencias entre la estimación puntual y la estimación por intervalos, supongamos que nuestro objetivo es determinar la tasa de paro en una población. Como ya hemos visto en capítulos anteriores se trata de un caso particular de estimación de la proporción p, problema que resulta muy habitual en el ámbito económico y del que por tanto pueden encontrarse otros muchos ejemplos (la cuota de mercado de un producto, la participación femenina en el mercado laboral, el peso relativo de un sector económico, ...).

A partir de una m.a.s. el estimador analógico de la proporción poblacional sería la proporción muestral $\hat{p} = \frac{X}{n}$ donde n es el tamaño muestral (número de activos investigados para nuestro estudio) y X es la variable aleatoria que recoge el número de activos que se encuentran en paro.

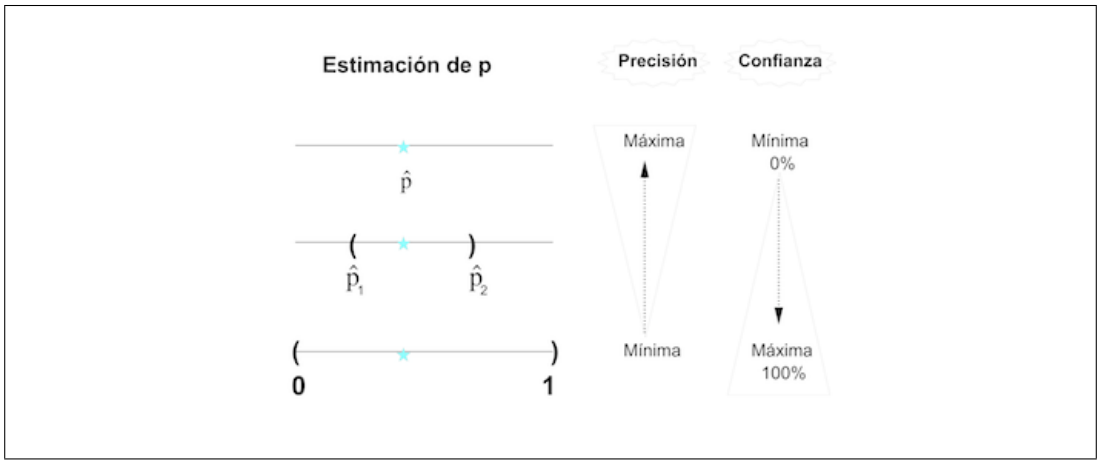
La proporción muestral nos permite estimar la tasa de paro p tanto puntualmente como por intervalos. En el primero de estos casos, el estimador nos proporciona un valor único, que queda determinado a partir de la muestra, mientras que la incorporación de márgenes de confianza -tal y como se indica en la figura 7.2- nos permitiría garantizar cierta probabilidad de que la tasa p se encuentre comprendida entre dos valores aleatorios.

En principio podría parecer deseable maximizar esta probabilidad o confianza del intervalo, pero ello nos llevaría a aumentar su tamaño, cosa que no resulta deseable. En realidad, al llevar a cabo este tipo de estimación debemos cuestionarnos en un doble sentido: ¿qué probabilidad podemos garantizar? pero también ¿qué precisión tiene el intervalo que facilitamos?

Ambas características, que denominaremos *confianza* y *precisión* respectivamente, resultan de gran interés, y puede percibirse fácilmente la relación de sustitución entre las mismas. Como muestra la figura, un afán por garantizar la máxima confianza (probabilidad del 100 %) conduce inevitablemente al intervalo $[0, 1]$ para la tasa de paro. Dicho resultado no es en absoluto informativo, hasta el punto de que coincide con el espacio paramétrico inicial, ya que por definición se tiene $p \in [0, 1]$.

Frente a este caso extremo de máxima confianza con mínima precisión se encontraría la situación opuesta, en la que consideraríamos prioritaria la obtención de intervalos

Figura 7.2.: Estimación puntual y por intervalos



precisos. La maximización de la precisión o minimización de la amplitud nos conduciría a intervalos con margen de error nulo, que coinciden con la estimación puntual y a los que resulta imposible asignar un nivel de confianza o probabilidad (por tanto el nivel de confianza adoptaría su valor mínimo, 0 %).

Entre los dos casos extremos descritos existen infinitas situaciones intermedias en las que se presentarían distintas opciones de las características de precisión y confianza.

A la vista de los comentarios anteriores cabría preguntarse si una estimación por intervalos resulta siempre preferible a la estimación puntual. La respuesta es negativa puesto que, si bien la estimación por intervalos presenta claras ventajas sobre la puntual, la elección de uno u otro procedimiento dependerá de los objetivos que persigamos en cada investigación.

A modo de ilustración, una empresa no estará demasiado interesada en una cifra concreta de ventas esperadas sino que preferirá conocer con ciertas garantías un intervalo en el que éstas se encuentren situadas. Este mismo razonamiento podría ser aplicable a la evolución de precios, pero deja de ser válido si lo que se pretende es obtener un índice que posteriormente pueda ser utilizado como deflactor, función ésta que no puede desempeñar un intervalo.

En definitiva, aunque la estimación por intervalos presenta la ventaja de garantizar una probabilidad o nivel de confianza, no siempre es éste nuestro objetivo prioritario. En muchas ocasiones resulta imprescindible disponer de un dato único y así la propia estadística oficial muestra numerosos ejemplos en los que las estimaciones se llevan a cabo puntualmente (IPC, tasa de paro, PIB, ...).

Esta opción permite disponer de cifras precisas -no avaladas por ninguna probabilidad directa pero sí por su método de obtención- que en ocasiones serán utilizadas como deflactores o indicadores económicos. Además, en la estadística oficial estos resultados aparecen a menudo complementados con información relativa a su margen de error, que permite la construcción de intervalos en el caso de que ésta se considere

conveniente.

7.2. Intervalos de confianza. Construcción y características

Una vez examinadas las ventajas de la estimación por intervalos, nos ocuparemos de su determinación, estudiando también los factores que afectan a su precisión y su nivel de confianza.

Definición 7.1. Llamamos *intervalo de confianza* a un intervalo del espacio paramétrico Θ limitado por dos valores aleatorios T_1 y T_2 entre los que, con cierta probabilidad se halle comprendido el verdadero valor del parámetro desconocido θ.

Llamamos *nivel de confianza*, que denotamos por $1-\alpha$, a la probabilidad o confianza de que el parámetro se encuentre entre los límites anteriores:

$$1 - \alpha = P\left(T_1 \leq \theta \leq T_2\right)$$

7.2.1. Construcción de intervalos de confianza

Se trata por tanto de determinar los valores que delimitan el intervalo, y que dependerán tanto de la información muestral $(X_1, \ldots, X_n)$ como del estimador T.

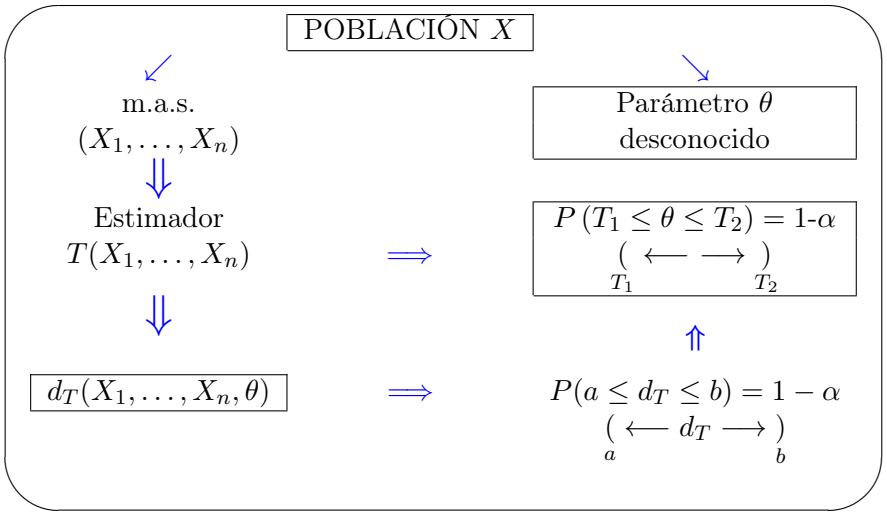
Dado que sería difícil determinar los valores extremos que podría presentar la muestra, la construcción de intervalos se lleva a cabo mediante la consideración de valores extremos para el estimador. Así pues, el método utilizado para la determinación de intervalos consiste en incorporar ciertos márgenes de error al estimador T, hasta llegar a obtener un recorrido aleatorio (T_1, T_2) cuya amplitud denotamos por $A = T_2 - T_1$. Cuanto más reducida sea esta amplitud, calificaremos al intervalo de más preciso.

En ocasiones el intervalo de confianza se obtiene añadiendo un mismo margen a derecha e izquierda de la estimación puntual, con lo cual el intervalo es simétrico respecto al punto, y su amplitud coincide con el doble del margen. No obstante, esta situación no resulta aplicable a todos los parámetros de interés.

Consideremos una población que identificamos con una variable aleatoria X cuya distribución de probabilidad depende de cierto parámetro desconocido θ. El procedimiento general que seguiremos para la construcción de intervalos de confianza para θ aparece descrito en la figura 7.3.

La primera etapa, ya conocida, consiste en resumir la información muestral mediante un estimador $T(X_1, \ldots, X_n)$ y construir la discrepancia tipificada asociada a este estimador. Para ello seguiremos la metodología estudiada en el capítulo anterior, llegando a expresiones d_T que son función tanto de la m.a.s. $(X_1, \ldots, X_n)$ como del parámetro investigado θ: $d_T = d_T(X_1, \ldots, X_n, \theta)$ y que, en general, seguirán un mo-

Figura 7.3.: Esquema de Estimación por intervalos



delo probabilístico conocido.

Las discrepancias tipificadas d_T que hemos introducido en el capítulo anterior son expresiones aleatorias construidas por comparación de T y θ, cuyo modelo probabilístico es habitualmente conocido y no depende del parámetro θ.

Así, dada una v.a. X con distribución $F_X(x)$, hemos estudiado las distintas expresiones tipificadas $d_T(X_1, \ldots . X_n, \theta)$, cuya distribución de probabilidad es conocida (Normal, t de Student, chi-cuadrado o F de Snedecor) y no depende de ningún parámetro desconocido.

La distribución de d_T aparecerá en general tabulada. Por tanto, una vez fijado el *nivel de confianza* $1 - \alpha$ que deseamos garantizar para nuestra estimación, es posible determinar con ayuda de las correspondientes tablas un par de valores a y b tan próximos como sea posible tales que se cumpla:

$$P(a \leq d_T(X_1, \ldots, X_n, \theta) \leq b) = 1 - \alpha$$

Dado que d_T es una expresión aleatoria que resume la discrepancia entre T y θ convenientemente tipificada, mediante la igualdad anterior garantizamos una probabilidad de que la discrepancia se encuentre en cierto recorrido $[a, b]$.

Los niveles de confianza más habituales son el 90 %, 95 % y 99 %. Una vez fijado un nivel de confianza determinado $1 - \alpha$, en principio existen infinitas posibilidades para determinar los valores a y b que encierran esa probabilidad. Sin embargo, nuestro objetivo es obtener intervalos precisos por lo cual intentaremos que a y b se encuentren lo más próximos posible. En concreto, para expresiones d_T distribuidas simétricamente (caso de los modelos Normal o t de Student) el recorrido óptimo -en el sentido de máxima precisión- se obtiene para valores opuestos, esto es, con $a = -b$.

Mediante las etapas descritas hasta ahora hemos llegado a obtener intervalos constantes $[a, b]$ para la variable aleatoria d_T. Se trata de un paso intermedio hacia nuestro objetivo, que es la construcción de intervalos aleatorios para el parámetro θ.

Por tanto, debemos ocuparnos ahora de la etapa final de la figura 7.3, consistente en pasar del intervalo constante $[a, b]$ que incluye un $(1-\alpha)\,\%$ de la probabilidad de d_T, a otro intervalo con límites aleatorios T_1 y T_2 entre los que, con probabilidad $1-\alpha$, se encontrará el parámetro θ.

Dado que d_T es una función continua e inyectiva de θ, a partir de su expresión $d_T(X_1, \ldots, X_n, \theta)$ es posible obtener (igualando d_T a los extremos constantes a y b) un par de funciones de la muestra $T_1(X_1, \ldots, X_n)$ y $T_2(X_1, \ldots, X_n)$ tales que se cumpla:

$$P(T_1 \leq \theta \leq T_2) = 1 - \alpha$$

El proceso de obtención de $[T_1, T_2]$ a partir de $[a, b]$ se basa en el siguiente razonamiento: al igualar la discrepancia d_T a su valor mínimo a, estamos asumiendo el máximo error por defecto (subestimación) y como consecuencia, debemos corregir T al alza para compensar esa subestimación de θ, llegando así al extremo superior del intervalo T_2.

Este razonamiento se aplicaría de modo análogo a la situación opuesta en la que la discrepancia d_T adopta su valor máximo b, por lo cual al estimar θ corregiremos el valor de T a la baja hasta llegar a T_1. Así pues, se tiene:

$$d_T(X_1, \ldots, X_n, \theta) = a \Rightarrow \hat{\theta} = T_2$$

$$d_T(X_1, \ldots, X_n, \theta) = b \Rightarrow \hat{\theta} = T_1$$

y con la obtención de los límites aleatorios T_1 y T_2 hemos concluido la construcción de un intervalo de confianza (IC) para el parámetro θ.

En apartados posteriores deduciremos las expresiones de los intervalos de confianza correspondientes a los parámetros de interés. Sin embargo, para ilustrar el procedimiento anteriormente descrito, recogemos aquí la construcción de un intervalo de confianza para la esperanza μ.

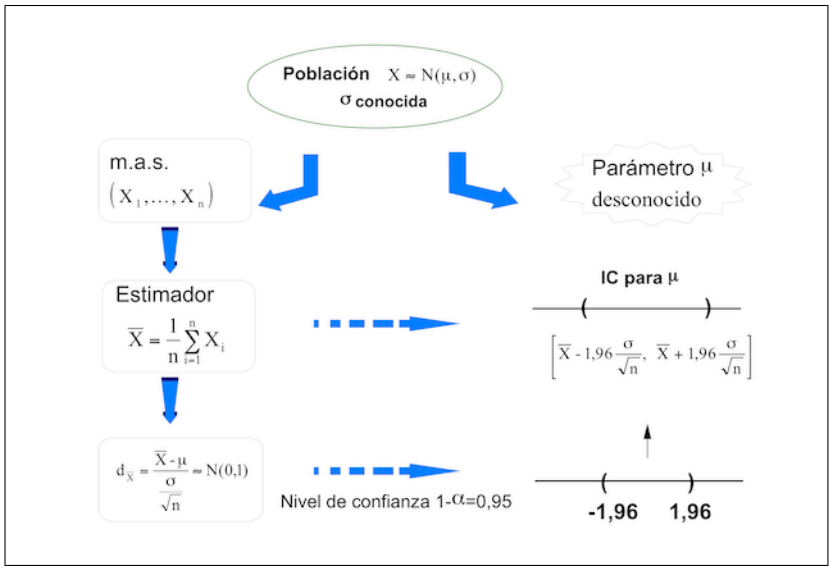
Consideremos una población $X \approx \mathcal{N}(\mu, \sigma)$ con σ conocida y supongamos que deseamos obtener un intervalo para μ con nivel de confianza $1-\alpha = 0,95$. En tal situación, las etapas a seguir aparecen recogidas en el esquema de la figura 7.4.

A partir de la media muestral se construye la discrepancia tipificada (que en este caso sigue una distribución normal estándar) y se determina el intervalo $[-1,961, 96]$, que es el mínimo recorrido de $d_{\bar{X}}$ que encierra la probabilidad pedida $(1-\alpha = 0,95)$.

Una vez obtenido este recorrido constante, basta despejar el parámetro μ para llegar a un intervalo aleatorio

$$\left[\bar{X} - 1,96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1,96\frac{\sigma}{\sqrt{n}}\right]$$

Figura 7.4.: Etapas en la estimación por intervalos para μ



que, con probabilidad del 95 %, incluye el verdadero valor esperado.

7.2.2. Precisión de los intervalos

Una vez que hemos comentado el procedimiento general de obtención de intervalos que garanticen cierto nivel de confianza 1-α , pasamos a analizar su *precisión*, característica que, como ya hemos comentado, evaluamos a través de la amplitud $A = T_2 - T_1$.

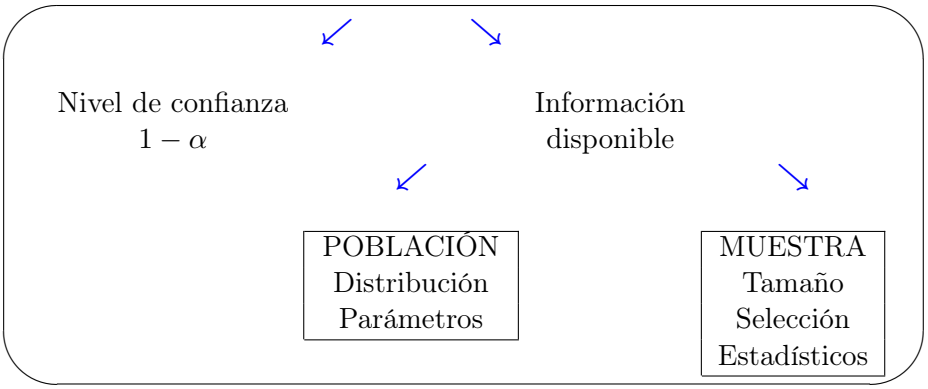
Los factores que determinan la precisión de un intervalo, recogidos en la figura 7.5, son de dos tipos: por una parte, el *nivel de confianza* que queramos garantizar en nuestra estimación, y por otra la *información disponible*, que -en sentido amplio- hace referencia tanto a la población como a la muestra.

Ya hemos justificado que el nivel de confianza de un intervalo aparece inversamente relacionado con su precisión, siendo imposible optimizar simultáneamente ambas características. En la práctica, la construcción de intervalos se lleva a cabo prefijando un nivel de confianza $1 - \alpha$ y buscando a continuación la expresión del intervalo que, garantizando dicho nivel de confianza, optimiza la precisión.

Como hemos visto en el apartado anterior, una vez fijado el nivel de confianza $1 - \alpha$ se determinan las constantes a y b que, en el mínimo recorrido posible, encierran esa probabilidad. Resulta evidente que a y b estarán más distanciados cuanto mayor haya sido la probabilidad $1 - \alpha$ fijada; por tanto, para una situación dada, el intervalo óptimo al 99 % será siempre más amplio (menos preciso) que el intervalo óptimo al 90 %.

Por lo que se refiere a la información disponible, podemos afirmar que, para un nivel de confianza dado, el intervalo será más preciso (es decir, presentará una menor am-

Figura 7.5.: Precisión del intervalo



plitud) a medida que mejoremos nuestra información, tanto poblacional (distribución de X y parámetros característicos) como muestral (tamaño, métodos de selección y estadísticos utilizados).

7.2.2.1. Información sobre la población

En primera instancia, nos interesa la información referida a la población investigada, esto es, su distribución de probabilidad y sus parámetros. En la práctica, gran parte de los procesos inferenciales se llevan a cabo partiendo del supuesto de normalidad, pero resulta interesante estudiar la construcción de intervalos en poblaciones desconocidas.

Consideremos una población X, desconocida tanto en lo que respecta a su distribución $F_X(x)$ como al parámetro θ. En esta situación no es posible garantizar modelos probabilísticos para las discrepancias d_T, por lo cual acudiremos a la desigualdad de Chebyshev, que da lugar a la siguiente expresión:

$$P\left(|d_T - E\left(d_T\right)| \leq k\sigma_{d_T}\right) \geq 1 - \frac{1}{k^2}\ ;\ \forall k > 0$$

Obsérvese que el enunciado de Chebyshev iría en realidad referido a la probabilidad de que d_T se desvíe de su valor esperado en menos de un cierto margen. No obstante, teniendo en cuenta que d_T es una función continua de T, la desigualdad estricta es equivalente a la condición "menor o igual".

Aunque hasta ahora utilizábamos la desigualdad de Chebyshev para aproximar probabilidades, en este caso nos interesará el planteamiento opuesto, ya que deseamos garantizar cierta probabilidad (coincidente con el nivel de confianza $1-\alpha$) pero desconocemos el margen que debemos incorporar a la discrepancia.

Así pues, igualaremos la probabilidad que deseamos garantizar para nuestro intervalo con la cota de Chebyshev:

$$1 - \alpha = 1 - \frac{1}{k}^2 \Rightarrow k = \sqrt{\frac{1}{\alpha}}$$

obteniendo la constante k en función del nivel de confianza fijado.

Una vez determinado, este margen k proporciona un par de valores a y b entre los cuales se encuentra d_T con probabilidad de al menos 1-α:

$$P\left(|d_T - E\left(d_T\right)| \leq k\sigma_{d_T}\right) \geq 1-\alpha \Rightarrow P\left(a \leq d_T \leq b\right) \geq 1-\alpha$$

y a partir de la expresión $d_T(X_1, \ldots, X_n, \theta)$, podremos despejar un intervalo aleatorio (T_1, T_2) para θ, tal que:

$$P\left(T_1 \leq \theta \leq T_2\right) \geq 1-\alpha$$

Estos intervalos basados en la desigualdad de Chebyshev serán -para niveles de confianza dados- menos precisos que los obtenidos a partir de distribuciones probabilísticas conocidas, puesto que la ausencia de información inicial es un inconveniente que conlleva un "coste" en términos de precisión. Así pues, la aplicación de Chebyshev únicamente es aconsejable cuando no existe otra alternativa, ya que cualquier información adicional sobre X proporcionará información sobre d_T y en consecuencia mejorará la precisión de nuestro intervalo.

La construcción de intervalos basados en la desigualdad de Chebyshev sería un caso de estimación no paramétrica, en el sentido de que los intervalos no se basan en ningún modelo probabilístico conocido.

Como veremos en apartados posteriores, en estas situaciones, la amplitud del intervalo de confianza dependerá del nivel de confianza fijado y de la eficiencia de nuestros estimadores.

Si consideramos ahora que la población X presenta una distribución conocida (habitualmente normal), podremos obtener intervalos de confianza basados en los modelos probabilísticos asociados a las discrepancias d_T.

Dado que las características poblacionales afectan a las distribuciones de las d_T, también condicionarán la precisión de los intervalos. En general, obtendremos intervalos más precisos cuanto más homogéneas sean las poblaciones investigadas.

En un apartado posterior analizaremos la importancia que tiene en la estimación de diversos parámetros la información sobre la varianza poblacional. Así, ya hemos visto que las inferencias sobre la esperanza con σ^2 conocida conducen a un modelo normal mientras que si σ^2 es desconocida debemos utilizar su estimación S^2, que conduce a una t de Student y, en general, a estimaciones menos precisas.

A su vez, en las situaciones con σ^2 conocida, veremos que la amplitud del intervalo aumentará con el valor de la varianza.

7.2.2.2. Información muestral

La información muestral abarca factores relevantes como el tamaño de muestra (n), los métodos de selección empleados y las expresiones utilizadas como estadísticos muestrales.

El *tamaño de la muestra* resulta de gran importancia en los procesos de estimación. En primer lugar, ya hemos visto que los tamaños elevados de muestra permiten aplicar los teoremas límites, garantizando así la convergencia al modelo normal.

Además, la expresión final de los intervalos de confianza para θ:

$$P\left(T_1 \leq \theta \leq T_2\right) = 1 - \alpha$$

conduce a dos límites aleatorios T_1 y T_2 que en general son funciones de n. En apartados posteriores analizaremos la relación entre la amplitud de un intervalo $A = T_2 - T_1$ y el tamaño de muestra.

La *selección muestral* puede colaborar en gran medida a mejorar la precisión de los intervalos. Nos referiremos aquí únicamente al muestreo aleatorio simple que es la técnica de selección más sencilla y se adopta habitualmente como referencia, pero los diseños muestrales más elaborados contribuyen a mejorar la precisión de las estimaciones.

Por último, debemos tener presente que la información muestral debe ser resumida mediante *estadísticos o estimadores*, por lo cual su definición será también de importancia en la determinación de los intervalos.

Las expresiones $T(X_1, \ldots, X_n)$ utilizadas como estimadores serán las que resulten en cada caso más adecuadas según los criterios estudiados en capítulos anteriores (ausencia de sesgo, eficiencia, suficiencia, consistencia) ya que estas propiedades deseables se trasladan a las discrepancias tipificadas $d_T(X_1, \ldots, X_n, \theta)$ y también a la precisión de los intervalos de confianza para θ.

7.2.3. Nivel de confianza: Interpretación

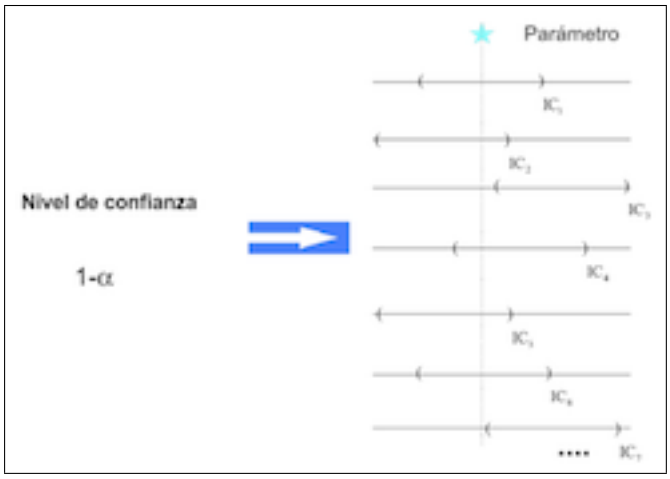
Como hemos visto, el proceso de construcción de intervalos concluye con la determinación de un recorrido aleatorio al que, con cierto nivel de confianza $(1 - \alpha)$, pertenecerá el parámetro desconocido. Es importante tener presente que esta interpretación probabilística, válida para la expresión $[T_1, T_2]$ deja de serlo cuando se obtiene una m.a.s. concreta $(x_1, \ldots, x_n)$ y a partir de la misma las estimaciones $t_1(x_1, \ldots, x_n)$ y $t_2(x_1, \ldots, x_n)$ que ya no tienen carácter aleatorio.

En efecto, cada concreción de un intervalo de confianza proporciona un recorrido numérico concreto sobre el que no puede efectuarse ninguna afirmación probabilística. Sin embargo, el método de construcción de los intervalos de confianza garantiza que si efectuásemos numerosos estudios muestrales, que se concretarían en los correspondientes intervalos IC_1, IC_2, etc, entonces una proporción (1-α) % de ellos contendría el verdadero valor del parámetro θ.

Esta interpretación aparece ilustrada en la figura 7.6 donde representamos el parámetro θ que pretendemos estimar y los intervalos de confianza obtenidos a partir de diferentes realizaciones muestrales.

Como puede apreciarse, la mayoría de estos intervalos abarcan en su recorrido a θ pero existen algunos (en la ilustración, el tercero y el séptimo) en los que no se

Figura 7.6.: Nivel de confianza



encuentra el parámetro. Así pues, la concepción frecuencialista de la probabilidad nos llevaría a afirmar que la proporción de intervalos estimados que contienen en su recorrido el verdadero θ se aproxima al nivel de confianza $1-\alpha$ utilizado para su construcción.

Obsérvese que esta afirmación probabilística debe ser efectuada en relación al intervalo aleatorio y no al parámetro. Así, sería incorrecto decir que "θ tiene una probabilidad $1-\alpha$ de pertenecer al intervalo" ya que el parámetro, aunque desconocido, es un valor constante al que por tanto no podemos asignar ninguna probabilidad.

7.3. Algunos intervalos de confianza particulares

El esquema expuesto anteriormente (figura 7.4) es aplicable a la construcción de intervalos de confianza para los parámetros de interés, esto es, las características poblacionales relativas a valores esperados, proporciones y varianzas.

7.3.1. Intervalos de confianza para la esperanza

Las inferencias sobre el parámetro μ se llevan a cabo habitualmente partiendo del supuesto de normalidad que, además de ser adecuado desde el punto de vista empírico, proporciona distribuciones conocidas para las discrepancias normalizadas.

Dada una v.a. $X \approx \mathcal{N}(\mu, \sigma)$ y partiendo de la información proporcionada por una m.a.s. adoptaremos como punto de partida el estimador analógico media muestral, que viene avalado por sus propiedades, a partir del cual construimos las discrepancias estudiadas en el capítulo anterior:

$$d_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0,1) \ , \ \text{si} \ \sigma^2 \ \text{es conocida}$$

$$d_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \approx t_{n-1} \ , \ \text{si} \ \sigma^2 \ \text{es desconocida}$$

En el supuesto más sencillo, con σ^2 conocida, se parte de la expresión tipificada con distribución normal estándar a partir de la cual, una vez fijado el nivel de confianza $1-\alpha$, es posible determinar el valor k tal que:

$$P\left(-k \leq d_{\bar{X}} \leq k\right) = 1 - \alpha$$

Puede apreciarse que en este caso particular los valores que encierran la probabilidad $1-\alpha$ son opuestos ya que, como consecuencia de la simetría del modelo normal, ésta es la opción que, fijado el nivel de confianza, conduce al intervalo óptimo (de mínima amplitud o máxima precisión).

A modo de ilustración, los valores k para los niveles de confianza habituales son:

Nivel de confianza	k
0,9	1,645
0,95	1,96
0,99	2,576

Teniendo en cuenta que en este caso la discrepancia viene dada por la expresión

$$d_{\bar{X}} = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

es posible despejar el parámetro μ, llegando a la expresión:

$$P\left(\bar{X} - k\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + k\frac{\sigma}{\sqrt{n}}\right)$$

que proporciona un intervalo aleatorio para μ con nivel de confianza $1-\alpha$.

$$\left[\bar{X} - k\frac{\sigma}{\sqrt{n}}, \bar{X} + k\frac{\sigma}{\sqrt{n}}\right]$$

El procedimiento seguido hasta la obtención de este intervalo final consiste en igualar la discrepancia a cada uno de sus valores límites $-k$ y $+k$. De este modo, si $d_{\bar{X}}$ adoptase el valor k entonces se obtendría la máxima discrepancia -por exceso- de la media muestral respecto a μ; por tanto, el límite inferior del intervalo se obtiene al corregir el estimador media muestral en el máximo error por exceso

$$\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} = k \Rightarrow \hat{\mu} = \bar{X} - k\frac{\sigma}{\sqrt{n}}$$

Con el razonamiento exactamente simétrico, se llegaría al estimador considerado límite superior:

$$\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} = -k \Rightarrow \hat{\mu} = \bar{X} + k\frac{\sigma}{\sqrt{n}}$$

A partir de la expresión obtenida para el intervalo de confianza de μ se observa que éste se encuentra centrado en $\bar{X}$ y presenta amplitud

$$A = 2k\frac{\sigma}{\sqrt{n}}$$

en la que pueden apreciarse tres factores: el nivel de confianza (que determina k), la dispersión poblacional (σ) y el tamaño muestral (n).

Estudiemos ahora cómo se vería alterado el procedimiento descrito si la varianza poblacional se desconoce, esto es, si la discrepancia tipificada viene dada por

$$d_{\bar{X}} = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \approx t_{n-1}$$

A partir de esta expresión, una vez fijado el nivel de confianza $1-\alpha$ podríamos obtener en las tablas de la distribución t un par de valores simétricos $-k$ y k tales que:

$$P(-k \leq d_{\bar{X}} \leq k) = 1-\alpha$$

obteniéndose a partir de la expresión anterior:

$$P\left(\bar{X} - k\frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} + k\frac{S}{\sqrt{n}}\right)$$

que conduce al siguiente intervalo de confianza para μ:

$$\left[\bar{X} - k\frac{S}{\sqrt{n}}, \bar{X} + k\frac{S}{\sqrt{n}}\right]$$

Dicho intervalo sigue estando centrado en la media muestral pero presenta ahora amplitud variable

$$A = 2k\frac{S}{\sqrt{n}}$$

dependiente de la dispersión muestral.

Al comparar esta amplitud con la asociada al IC para μ con σ conocida se aprecian dos cambios. En primer lugar, el valor k aumenta, ya que para un mismo nivel de confianza el valor obtenido en las tablas t de Student será superior al del modelo normal (si bien estas diferencias se atenúan al aumentar n).

Por otra parte, la amplitud pasa de ser constante a variable, por lo cual no es posible comparar la precisión de ambos tipos de intervalos.

Por último, en ausencia de información sobre la distribución poblacional y si los tamaños de muestra no son suficientemente elevados para aplicar los teoremas límites, tampoco conoceríamos la distribución de la discrepancia, por lo cual deberíamos acudir a la desigualdad de Chebyshev:

$$P\left(|d_{\bar{X}} - E\left(d_{\bar{X}}\right)| \leq k\sigma\right) \geq 1 - \frac{1}{k^2}$$

La igualación de la cota con el nivel de confianza exigido proporciona el resultado

$$k = \frac{1}{\sqrt{\alpha}}$$

y, dado que la discrepancia $d_{\bar{X}}$ tiene esperanza nula y dispersión unitaria, se obtiene el intervalo:

$$\left[\bar{X} - \frac{1}{\sqrt{\alpha}}\frac{\sigma}{\sqrt{n}}, \bar{X} + \frac{1}{\sqrt{\alpha}}\frac{\sigma}{\sqrt{n}}\right]$$

que, como consecuencia de su obtención, tiene un nivel de confianza de al menos $1-\alpha$.

Este intervalo presenta una mayor amplitud que los anteriormente vistos y para situaciones con varianza desconocida podría ser aproximado mediante la dispersión muestral.

7.3.2. Intervalos de confianza para la varianza

En ocasiones nuestro objetivo es aproximar la dispersión poblacional, obteniendo *intervalos de confianza para la varianza* σ^2. En tales casos, partiendo de poblaciones normales, la discrepancia se construye como un cociente entre la varianza muestral S^2 y la varianza poblacional, ajustado por los grados de libertad $(n-1)$ hasta obtener la expresión

$$d_{S^2} = \frac{(n-1)S^2}{\sigma^2} \approx \chi^2_{n-1}$$

Una vez fijado el nivel de confianza $1-\alpha$ sería necesario obtener un par de valores k_1 y k_2 tales que: $P\left(k_1 \leq d_{S^2} \leq k_2\right) = 1-\alpha$

Como consecuencia de la asimetría del modelo chi-cuadrado, el método de determinación de estos valores no coincide con el visto para los intervalos de μ. En concreto, la opción más habitual consiste en determinar recorridos $[k_1, k_2]$ que dejan a su izquierda y a su derecha colas con idéntica probabilidad $\frac{\alpha}{2}$.

A partir de esos valores se llega a intervalos de confianza para σ^2 dados por la expresión:

$$\left[\frac{(n-1)S^2}{k_2}, \frac{(n-1)S^2}{k_1}\right]$$

que se encuentran próximos al óptimo.

Los extremos de este intervalo se obtienen al igualar la discrepancia normalizada a las constantes k_1 y k_2 obtenidas anteriormente. En el primer caso se hace coincidir d_{S^2} con su valor mínimo k_1, por lo cual la varianza muestral se corrige al alza multiplicando por el índice $\frac{n-1}{k_1} > 1$, llegando al extremo superior $\frac{(n-1)S^2}{k_1}$; la situación contraria aparece al igualar la discrepancia a k_2 y conduce al límite inferior $\frac{(n-1)S^2}{k_2}$.

Otra posibilidad sería adoptar soluciones unilaterales que, aunque en ciertos casos resultan interesantes, en general son poco informativas ya que proporcionan un sólo extremo para el intervalo.

Así, cuando el nivel de confianza se acumula a la izquierda $P\left(d_{S^2} \leq k\right) = 1 - \alpha$, se obtiene $[0, k]$ con lo cual el intervalo de confianza para σ^2 es $\left[\frac{(n-1)S^2}{k}, +\infty\right)$ que no tiene cota superior.

Si en cambio se obtiene el valor k tal que $P\left(d_{S^2} \geq k\right) = 1 - \alpha$, entonces el intervalo para σ^2 resulta ser $\left(0, \frac{(n-1)S^2}{k}\right]$.

En cualquiera de las situaciones comentadas, los intervalos de confianza para la varianza poblacional presentan un rasgo que los diferencia de los construidos para la media. Se trata de la incorporación de un coeficiente o margen de carácter multiplicativo, que sustituye a los márgenes aditivos considerados hasta ahora.

7.3.3. Intervalos de confianza para la proporción

Cuando investigamos características cualitativas, aparece un caso particular de la media poblacional: se trata de la proporción, p, parámetro que -como ya hemos comentado- presenta gran interés en análisis inferenciales y se aproxima por la proporción muestral.

En el capítulo anterior hemos visto que para tamaños muestrales suficientemente elevados es posible -gracias al teorema de De Moivre- construir discrepancias tipificadas

$$d_{\hat{p}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \approx \mathcal{N}(0,1)$$

Aplicando ahora el procedimiento descrito para la construcción de intervalos se obtiene la expresión:

$$\left[\hat{p} - k\sqrt{\frac{p(1-p)}{n}}, \hat{p} + k\sqrt{\frac{p(1-p)}{n}}\right]$$

que no puede ser determinada en la práctica ya que sus extremos dependen del parámetro desconocido p. Para solucionar este inconveniente, se suele sustituir la varianza poblacional de p por su estimador insesgado, con lo cual se obtiene el intervalo:

$$\left[\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}, \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right]$$

donde k ha sido calculado mediante las tablas de la distribución normal para el nivel de confianza $1-\alpha$ fijado.

El razonamiento anterior no resultará sin embargo aplicable para muestras de tamaño pequeño. En estas situaciones, únicamente es posible afirmar que X (numerador de la proporción muestral) sigue una distribución binomial $\mathcal{B}(n,p)$.

Así pues, conocida la proporción muestral $\hat{p}$ y dado un nivel de confianza $1-\alpha$ se buscan dos valores de probabilidad p_1 y p_2 tales que:

$$P\left(X < n\hat{p}/p_2\right) = \frac{\alpha}{2}\ ;\ \mathrm{P}(X > n\hat{p}/p_1) = \frac{\alpha}{2}$$

con lo cual se obtiene directamente el intervalo de confianza $[p_1, p_2]$ para p.

La estimación de p puede también llevarse a cabo mediante la utilización de *bandas gráficas de confianza*. El método consiste en seleccionar en el eje de abscisas el valor muestral $\hat{p}$, para el cual se obtienen en ordenadas los límites inferior y superior para el parámetro p con un cierto nivel de confianza y determinado tamaño de muestra.

C. J. Clopper y E. S. Pearson (1934) elaboraron bandas gráficas para la proporción con niveles de confianza del 95 % y el 99 % y para ciertos tamaños de muestra. Posteriormente autores como Clark (1953) y Pachares (1960) construyeron mediante el mismo procedimiento bandas asociadas a nuevos niveles de confianza y tamaños muestrales.

Partiendo del valor $\hat{p}$ observado en una muestra, estas gráficas proporcionan bandas de confianza que en general no son simétricas respecto a $\hat{p}$, ya que para proporciones observadas bajas se incorpora un mayor margen a la derecha y viceversa si la proporción observada es cercana a 1. (A modo de ilustración, para $n = 20$ con $\hat{p} = 0,8$ se llegaría a una banda $[0,55\ ,\ 0,95]$ mientras que para el mismo tamaño con $\hat{p} = 0,1$ se obtiene el intervalo $[0,01\ ,\ 0,33]$).

7.3.4. Intervalos de confianza para combinaciones lineales de medias

A menudo nos interesa llevar a cabo inferencias sobre la diferencia, la suma u otra combinación lineal de esperanzas de dos poblaciones. Para mayor generalidad, consideremos el parámetro $\alpha\mu_X + \beta\mu_Y$ que deseamos aproximar mediante intervalos de confianza.

Para la construcción de estos intervalos debemos tener presentes las diversas situaciones descritas en el capítulo anterior para la diferencia de medias (muestras dependientes e independientes, poblaciones normales y desconocidas, varianzas conocidas y desconocidas, etc.).

Si las muestras son dependientes se obtienen datos pareados, que se reducen a una muestra única sobre la cual resulta aplicable el procedimiento descrito para la construcción de IC para la esperanza poblacional.

Por otra parte, en el supuesto de independencia, partiendo de muestras de tamaños n y m la construcción de intervalos para $\alpha\mu_X + \beta\mu_Y$ se basa en el estimador insesgado $\alpha\bar{X} + \beta\bar{Y}$ que conduce a la discrepancia tipificada:

$$d_{\alpha\bar{X}+\beta\bar{Y}} = \frac{\alpha\bar{X} + \beta\bar{Y} - (\alpha\mu_X + \beta\mu_Y)}{\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}}$$

cuya distribución de probabilidad depende de la información disponible sobre las poblaciones.

En el supuesto más sencillo, con $X \approx \mathcal{N}(\mu_X, \sigma_X)$, $Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$ y varianzas conocidas, los intervalos de confianza para la combinación de esperanzas vienen dados por la expresión:

$$\left[\alpha\bar{X} + \beta\bar{Y} - k\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}\,,\, \alpha\bar{X} + \beta\bar{Y} + k\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}\right]$$

donde k se determina en las tablas del modelo normal.

[Dedúzcase la expresión anterior] [¿Cuál sería el intervalo de confianza si las varianzas fuesen desconocidas y coincidentes?].

Para poblaciones desconocidas la solución consiste en aplicar la desigualdad de Chebyshev, cuya cota igualamos al nivel de confianza deseado. Se llega entonces a la expresión:

$$\left[\alpha\bar{X} + \beta\bar{Y} - \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}\,,\, \alpha\bar{X} + \beta\bar{Y} + \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\alpha^2\sigma_X^2}{n} + \frac{\beta^2\sigma_Y^2}{m}}\right]$$

[Compruébese]

7.3.5. Intervalos de confianza para la razón de varianzas

Como hemos visto, en la obtención de los intervalos para combinaciones de medias resulta importante conocer si las varianzas poblacionales son iguales. De ahí que la razón de varianzas sea también un objetivo inferencial.

Dadas dos poblaciones normales $X \approx \mathcal{N}(\mu_X, \sigma_X)$ e $Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$ supongamos que deseamos obtener un *intervalo de confianza para la razón de varianzas* $\frac{\sigma_X^2}{\sigma_Y^2}$.

Partiremos de la información suministrada por dos muestras aleatorias independientes de tamaños n y m respectivamente, a partir de las cuales es posible construir la discrepancia $\frac{S_X^2\sigma_Y^2}{S_Y^2\sigma_X^2}$ que sigue una distribución F de Snedecor con $(n-1)$ y $(m-1)$ grados de libertad.

Siguiendo el mismo procedimiento visto para la varianza, podemos construir intervalos bilaterales o unilaterales. En el primer caso, buscaríamos en las tablas de la F dos valores k_1 y k_2 tales que:

$$P\left(d_{\frac{S_X^2}{S_Y^2}} < k_1\right) = \frac{\alpha}{2}\ ;\ P\left(d_{\frac{S_X^2}{S_Y^2}} > k_2\right) = \frac{\alpha}{2}$$

con lo cual se obtiene la expresión del intervalo bilateral para la razón de varianzas $\frac{\sigma_X^2}{\sigma_Y^2}$:

$$\left[\frac{S_X^2}{S_Y^2 k_2}\,,\,\frac{S_X^2}{S_Y^2 k_1}\right]$$

[¿Cómo se obtendrían los intervalos unilaterales?] [¿y los intervalos para $\frac{\sigma_Y^2}{\sigma_X^2}$]

7.3.6. Intervalos de confianza para la mediana

La construcción de *intervalos para la mediana* de una población es un problema de inferencia no paramétrica, dado que no es necesario en este caso asumir ningún supuesto sobre la población X de partida.

Teniendo en cuenta que la mediana (Me) es el parámetro poblacional que garantiza $F_X(Me) = 0,5$, entonces para todo elemento X_i de una m.a.s. $(X_1, \ldots, X_n)$ se cumple $P(X_i \leq Me) = 0,5$.

Aprovechando la información muestral definiremos un intervalo aleatorio $[X_a\,,\,X_b]$ que, con el nivel de confianza deseado, contenga al valor mediano. En esta construcción resultará útil la variable Z: "Número de observaciones muestrales menores o iguales a Me" distribuida según un modelo $\mathcal{B}(n, 0,5)$, ya que se cumple:

$$P\left(X_a \leq Me \leq X_b\right) = P(a \leq Z \leq b) = \sum_{k=a}^{b} \binom{n}{k} 0,5^n$$

y en consecuencia para cada nivel de confianza $1-\alpha$ determinaremos con las tablas del modelo binomial (o con ayuda de un ordenador) dos cantidades a y b que posteriormente conducen a un intervalo numérico $[x_a\,,\,x_b]$ para la mediana.

Entre las distintas posibilidades para determinar los valores a y b de la variable binomial, optaremos por aquélla que proporcione a y b más próximos. Sin embargo, esta opción no garantiza que el intervalo $[x_a\,,\,x_b]$ al que conduce sea óptimo en el sentido de maximizar la precisión.

El procedimiento descrito puede ser aplicado a la construcción de intervalos para cualquier cuantil Q, con sólo definir la variable aleatoria Z: "número de observaciones muestrales inferiores al cuantil Q" que sigue un modelo $\mathcal{B}(n, p_Q)$.

7.4. Determinación del tamaño muestral

Hasta ahora nos hemos ocupado de determinar intervalos de confianza para los parámetros de interés, analizando diversas situaciones con distintos niveles de información. Consideramos ahora otra posibilidad habitual, consistente en determinar el tamaño muestral que permita obtener cierto intervalo.

Este planteamiento resulta de interés en el ámbito económico, ya que son frecuentes las situaciones en las que el investigador debe determinar el tamaño de muestra necesario para que un intervalo cumpla ciertos requisitos (precisión y nivel de confianza).

Obsérvese que el tamaño de muestra es determinante para conocer el presupuesto de una investigación. De ahí el interés de optimizar, buscando el mínimo tamaño que garantice las condiciones de precisión y confianza que se consideran necesarias en el intervalo buscado.

7.4.1. Tamaño de muestra en intervalos para la esperanza

Consideremos una población normal con varianza conocida para la que deseamos estimar el valor esperado μ con un cierto nivel de confianza $1-\alpha$. En ocasiones podemos estar interesados en obtener el tamaño de muestra n necesario para garantizar determinada precisión en nuestras estimaciones.

Dado que los intervalos para la media son simétricos, su precisión puede ser cuantificada indistintamente mediante la amplitud A o mediante su margen de error $\epsilon = \frac{A}{2}$. Así pues, se obtiene:

$$A = 2k\frac{\sigma}{\sqrt{n}} \Rightarrow \epsilon = k\frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{k\sigma}{\epsilon}\right)^2$$

En esta expresión se observa que el tamaño muestral n aumenta con la dispersión poblacional (σ), con el nivel de confianza (que determina el valor k) y con la precisión (inverso del margen ϵ).

Puede verse además que este tamaño aumenta cuando nos enfrentamos a mayor incertidumbre sobre la población, tal y como se recoge en la tabla comparativa que sigue:

Situación	Margen de error(ϵ)	Tamaño muestral
$X \approx \mathcal{N}(\mu,\sigma)$ con σ conocida	$\epsilon = k\frac{\sigma}{\sqrt{n}}$	$n = \left(\frac{k\sigma}{\epsilon}\right)^2$
X desconocida con σ conocida	$\epsilon = \frac{\sigma}{\sqrt{n\alpha}}$	$n = \left(\frac{\sigma}{\epsilon\sqrt{\alpha}}\right)^2$

El tamaño aumenta al enfrentarnos a poblaciones desconocidas, ya que para un nivel de confianza dado se obtiene $k < \frac{1}{\sqrt{\alpha}}$

$1-\alpha$	90 %	95 %	99 %
k	1,645	1,96	2,576
$\frac{1}{\sqrt{\alpha}}$	3,1623	4,4721	10

En las expresiones deducidas para el tamaño muestral n aparece la dispersión poblacional, característica que en la práctica puede ser desconocida. Para solucionar este inconveniente, la práctica más habitual consiste en obtener una estimación de la dispersión mediante una muestra piloto, considerando a continuación este parámetro como conocido.

Debemos tener en cuenta que la realización de un estudio piloto para estimar la dispersión poblacional aumentará el presupuesto necesario para nuestra investigación. Sin embargo, también presenta ciertas ventajas, ya que en la medida en que detectemos errores en este estudio previo podremos mejorar el diseño de la encuesta definitiva.

7.4.2. Tamaño de muestra en intervalos para la proporción

Supongamos que deseamos estimar la proporción asociada a determinada característica poblacional con cierto nivel de confianza $1-\alpha$. Como hemos visto en un apartado anterior, la expresión del intervalo viene dada por:

$$\left[\hat{p} - k\sqrt{\frac{p(1-p)}{n}}, \hat{p} + k\sqrt{\frac{p(1-p)}{n}}\right]$$

con lo cual podemos obtener el tamaño muestral necesario para garantizar un nivel de confianza $(1-\alpha)$ y una precisión (ϵ) concretos:

$$\epsilon = k\sqrt{\frac{p(1-p)}{n}} \Rightarrow n = \frac{k^2 p(1-p)}{\epsilon^2}$$

Puede verse que el tamaño obtenido aumenta con las exigencias de confianza y precisión para nuestro intervalo. Además, esta expresión depende de la proporción p desconocida, problema que puede ser solucionado de dos formas:

- Seleccionar una muestra piloto que proporcione una primera estimación de p.
- Sustituir el valor desconocido $p(1-p)$ por su cota máxima, que es 0,25.

Esta segunda alternativa es la más habitual y, como consecuencia de asumir la dispersión máxima, conduce a valores de n que siempre proporcionan una precisión superior a la inicialmente fijada.

Puede comprobarse fácilmente que la expresión de la dispersión $p(1-p)$ alcanza su valor máximo para $p = 0,5$. Dado que ésta es la situación más desfavorable, en ocasiones podríamos disponer de información para acotar p y en consecuencia la dispersión.

8. Contraste de hipótesis

Nuestra vida cotidiana está repleta de decisiones y actuaciones basadas en hipótesis. Si estos supuestos de partida son adecuados aumentarán nuestras posibilidades de éxito mientras que, si partimos de hipótesis o supuestos inadecuados, nuestras decisiones pueden llegar a tener consecuencias contrarias a las deseadas. De ahí la importancia de aprovechar al máximo la información estadística disponible, llevando a cabo contrastes en los que nuestras hipótesis de partida se enfrentarán a la realidad, para analizar si ambas informaciones son coherentes o contradictorias.

Dada la trascendencia del contraste de hipótesis, es importante prestar atención a todo el proceso, que incluye el enunciado de los supuestos de partida, el tratamiento de la información muestral, la eleccion del estadístico de contraste y la conclusión final, que consistirá a decidir si debemos o no rechazar la hipótesis planteada.

8.1. Conceptos básicos

El primer aspecto que nos planteamos en un problema de contraste es el relativo a la formulación de una hipótesis, que debería recoger un postulado o supuesto de trabajo, elaborado a partir de teorías, experiencias previas, e incluso nuestras propias convicciones.

Puede verse por tanto que el origen de las hipótesis es diverso, al igual que lo es su naturaleza. A modo de ejemplo, nuestros supuestos podrían ir referidos a características de tipo técnico ("existencia de rendimientos a escala constantes"), de comportamiento ("la propensión marginal al consumo es del 80 %"), comerciales ("se espera una demanda de al menos 30.000 unidades de cierto producto"), políticos ("un 75 % de ciudadanos está a favor de la ampliación de la UE"), ... pero también podrían ser afirmaciones genéricas sobre ciertos colectivos ("el peso de cierto producto se distribuye según un modelo normal") o las relaciones entre varias características ("el tipo impositivo medio es independiente del nivel de inflación", ...).

En un problema de contraste de hipótesis existen distintos tipos de información: la *información básica*, que consideraremos segura y no está sujeta a contraste (por ejemplo, considerar que el modelo de distribución de renta es conocido), la *información adicional, a priori o contrastable* que nos permite establecer la hipótesis de trabajo (todas las ilustraciones anteriores entrarían en esta categoría) y finalmente para contrastar la hipótesis se utiliza la *información muestral*.

En función de la información básica disponible cabe hacer una primera distinción entre los contrastes:

Definición 8.1. Se dice que un contraste es *paramétrico* si existe una información

básica que nos permite garantizar el modelo de probabilidad de la población que se va a contrastar.

Un contraste se dice *no paramétrico* cuando no hay información básica y todos los supuestos se incluyen en la hipótesis de trabajo

Los tests no paramétricos son enunciados más globales relativos a una o varias distribuciones poblacionales.

Esta distinción, que hemos propuesto de forma genérica para todos los procesos inferenciales, se basa en el grado de desconocimiento sobre la población, que es parcial para la inferencia paramétrica y total para la no paramétrica.

Conviene tener presente este criterio de clasificación ya que, como hemos visto, existen algunos procedimientos inferenciales que, pese a ir referidos a parámetros tienen carácter no paramétrico (por ejemplo, este es el caso de la estimación por intervalos con desconocimiento del modelo poblacional mediante la desigualdad de Chebyshev).

Una vez investigada su naturaleza ¿cómo se decide si una hipótesis debe o no ser rechazada? Algunas veces tenemos pruebas inequívocas sobre la validez de un supuesto y entonces éste se incluye entre la información básica o núcleo no contrastable. Sin embargo, en la mayor parte de los supuestos asumimos cierto riesgo o incertidumbre probabilística, dado que en general las observaciones se hallan expuestas a variaciones y por tanto podrían haber sido generadas bajo la hipótesis enunciada pero también bajo su complementaria o alternativa.

La filosofía del contraste de hipótesis se basa en recopilar información muestral que nos permita decidir si las desviaciones observadas con respecto a la hipótesis teórica son demasiado elevadas o "significativas" como para poder atribuirlas al azar. En este caso, la información muestral contradice claramente nuestro supuesto y debemos rechazar nuestra hipótesis de partida.

En definitiva, las decisiones relativas a la hipótesis se basan en la información muestral disponible. Como consecuencia, se trata de un proceso de inferencia estadística, que lleva inherente el correspondiente riesgo inferencial.

8.1.1. Contraste de hipótesis e intervalos de confianza

Con los rasgos descritos hasta ahora ¿qué analogías y diferencias existen entre los procedimientos de contraste y los métodos de estimación? Para responder a este interrogante, la figura 8.1 recoge un esquema en el que se analizan paralelamente ambas técnicas.

En principio, dado que tanto la estimación como el contraste de hipótesis son procesos inferenciales, existen abundantes rasgos comunes a los dos métodos. En realidad, ambos procedimientos son similares en cuanto a la utilización de la información muestral y a los instrumentos o expresiones estadísticas que comparan dicha información con las características poblacionales.

Tabla 8.1.: Comparación: estimación y contraste

	Estimación	Contraste
Objetivo	Aproximar características poblacionales desconocidas	Contrastar supuestos sobre la población
Información	Básica Muestral	Básica A priori o contrastable Muestral
Herramienta	Discrepancia estimador-parámetro	Discrepancia muestra-hipótesis
Resultado	Estimación puntual o intervalo de confianza	Conclusión: Rechazar o no rechazar
Garantías	Nivel de confianza	Nivel de significación Nivel crítico

Las expresiones de las discrepancias tipificadas estudiadas en capítulos anteriores y sus correspondientes distribuciones probabilísticas siguen siendo válidas para la realización de contrastes estadísticos, con la única diferencia de que ahora evaluaremos dichas discrepancias bajo ciertas hipótesis que deseamos contrastar.

Como consecuencia de su carácter inferencial, los resultados a los que lleguemos tanto en estimación como en contraste, serán afirmaciones probabilísticas basadas en información parcial. De ahí el interés de conocer sus "garantías" que vendrán medidas en términos de probabilidad (en el caso de los intervalos conocemos su *nivel de confianza* y para los contrastes introduciremos el *nivel de significación* y el *nivel crítico*).

A pesar de las coincidencias señaladas, existen también diferencias notables entre los métodos de estimación y contraste. Una de ellas es la información presente en cada uno de los procesos ya que, además de la información básica y la muestral, un contraste viene caracterizado por la existencia de *información a priori o contrastable* que da lugar a un supuesto o hipótesis inicial.

Como consecuencia, en los contrastes de hipótesis el investigador se ve más involucrado en el problema ya que, además de conocer el instrumental estadístico, para establecer la información básica necesita tener un conocimiento teórico y empírico del marco de referencia.

Además, los métodos de estimación y de contraste son diferentes en cuanto a su objetivo: en estimación se pretende aproximar un parámetro desconocido (mediante un valor único o un intervalo) mientras que en el contraste de hipótesis nuestro objetivo final es llegar a tomar una decisión (rechazar o no rechazar) sobre la hipótesis inicial.

Para ilustrar la conexión entre estimación y contraste consideremos a modo de ejemplo la producción mensual de cierto mineral (X, datos expresados en miles de Tm.), variable aleatoria que

se distribuye según un modelo normal y cuyo valor esperado, según la hipótesis de trabajo de la empresa, es de 410 miles de Tm./mes.

Analizando esta información se aprecia que la hipótesis de normalidad se asume como información básica (no entra en este caso en el contraste) mientras el supuesto relativo a la producción mensual esperada ($\mu = 410$) es una *información a priori* que nos interesa contrastar.

Si disponemos de información muestral, podemos comenzar por analizar la validez del supuesto utilizando el planteamiento ya conocido de la estimación.

La estimación puntual no resulta de gran ayuda ya que, aunque la población sea normal con media $\mu = 410$, no cabe exigir que la media muestral coincida exactamente con este valor. Así pues, sería más adecuado construir un intervalo de confianza para el parámetro μ, utilizando las expresiones estudiadas en el capítulo anterior.

Supongamos por ejemplo que, para un nivel de confianza del 95 %, el intervalo obtenido con nuestra información muestral es [350, 390]. Teniendo en cuenta que el 95 % de los intervalos contendrían al verdadero valor de la esperanza, en principio pensaríamos que nuestro intervalo particular se encuentra en esa proporción.

Podemos observar sin embargo que el recorrido estimado [350, 390] no incluye el valor hipotético de la producción esperada ($\mu = 410$), hecho que nos llevaría a pensar que el verdadero valor de μ se sitúa por debajo de nuestro supuesto inicial y por tanto a rechazar la hipótesis $\mu = 410$.

Siguiendo el mismo procedimiento, ¿cuál sería la conclusión si hubiéramos obtenido el intervalo [380, 420]? Parece claro que esta estimación no contradice el supuesto de partida (el recorrido contiene el valor hipotético 410), por lo cual no conllevaría un rechazo de la hipótesis.

Hemos visto que los intervalos de confianza pueden conducirnos a una decisión sobre el rechazo de una hipótesis. Sin embargo, conviene observar que en el proceso de construcción de intervalos no hemos tenido en cuenta el supuesto que sometemos a contraste (los IC anteriores no cambiarían si la hipótesis inicial hubiera sido $\mu = 400$ o cualquier otra), hecho que permite apreciar hasta qué punto estamos menospreciando el supuesto de partida.

Como consecuencia, la estimación no es un método recomendable cuando se pretende contrastar una hipótesis: aunque desde un punto de vista instrumental pueda conducir a resultados válidos, no sucede lo mismo desde una óptica conceptual, dado que ignora por completo el supuesto planteado o *información a priori*.

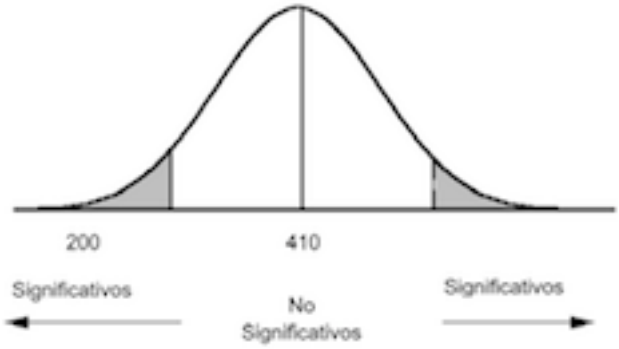
8.1.2. Contrastes de significación

Una vez justificado que los procedimientos de estimación no están diseñados para contrastar una hipótesis, estudiaremos nuevas opciones que, utilizando adecuadamente la información muestral, incorporen además explícitamente el supuesto formulado.

Una opción en este sentido consiste en estudiar la discrepancia entre muestra e hipótesis, aplicando el siguiente razonamiento: debido a las variaciones aleatorias, asumimos como aceptables ligeras desviaciones entre las observaciones muestrales y las hipótesis poblacionales. Sin embargo, cuando estas discrepancias sean considerables, su presencia ya no es atribuible únicamente al azar. Así pues, la muestra resulta poco verosímil bajo la hipótesis inicial y ello nos llevará a pensar que hemos partido de una hipótesis falsa.

Evidentemente, necesitaremos algún criterio estadístico que nos permita decidir si las discrepancias observadas son suficientemente elevadas para rechazar nuestra hipótesis. Este criterio no es único ya que se han desarrollado varios métodos para el contraste de hipótesis.

Figura 8.1.: Contraste de significación



El procedimiento más habitual y el primero históricamente es el de los *contrastes de significación*, ampliamente utilizados y cuya descripción puede resumirse en las siguientes etapas:

- Establecer la hipótesis
- Definir la expresión de la discrepancia tipificada en la que se basa el contraste
- Decidir, a partir de información muestral, rechazar o no rechazar la hipótesis

Los primeros desarrollos de los contrastes de significación fueron llevados a cabo por Karl Pearson, sobre 1900; sin embargo, la sistematización y desarrollo de este método se deben a R.A. Fisher, hacia el año 1920.

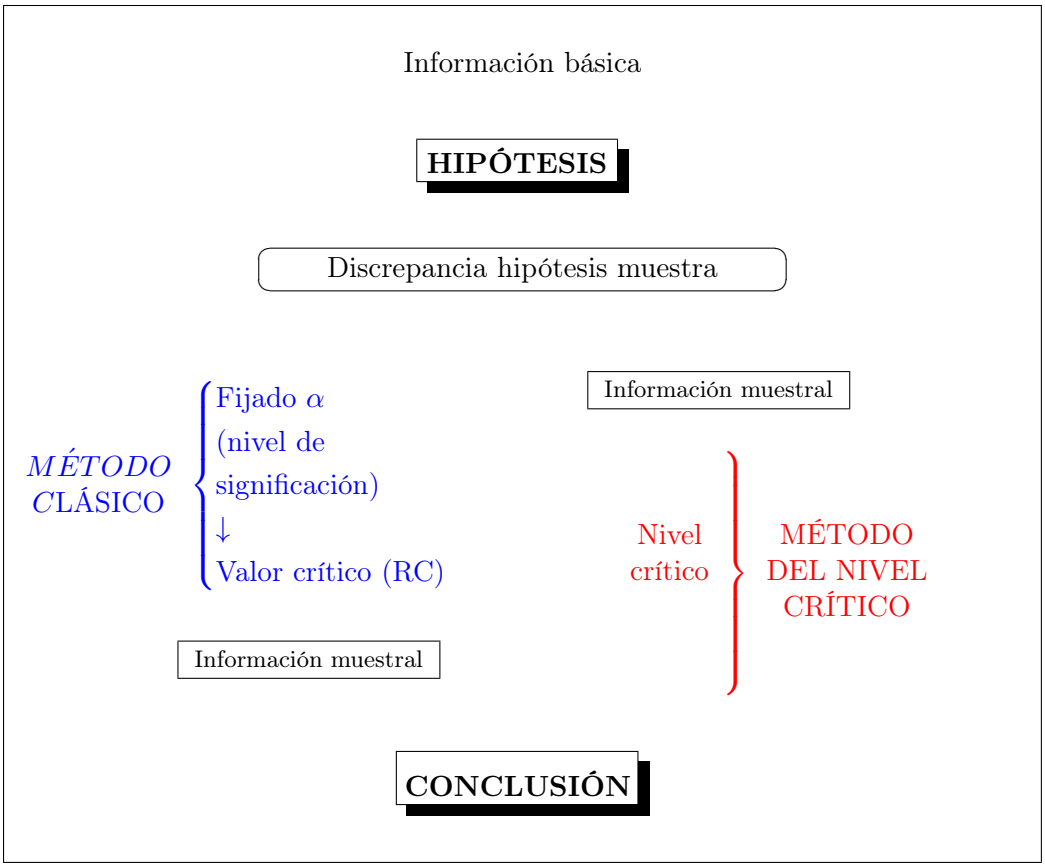
A lo largo de las etapas señaladas, los contrastes de significación estudian las discrepancias entre la información muestral y nuestra hipótesis hasta decidir si éstas son *significativas para rechazar*. El criterio de decisión será probabilístico: diremos que las discrepancias son significativas cuando éstas resultan muy poco probables bajo el supuesto de partida, y en caso contrario las calificaremos de *no significativas*.

En el ejemplo considerado con hipótesis $\mu = 410$, una media muestral de valor $\bar{x} = 415$ puede resultar coherente con el supuesto de partida, ya que conduce a un error de magnitud 5 que, una vez tipificado, proporciona una discrepancia de valor moderado. Por tanto, existirán probabilidades razonablemente elevadas de que, con una producción esperada de 410 miles de Tm/mes, se obtengan muestras como la observada.

Ahora bien, si en la muestra se registrase una producción media $\bar{x} = 200$ nuestra conclusión cambiaría ya que bajo el supuesto $\mu = 410$ resultaría muy poco probable extraer muestras con producciones medias tan bajas. Así pues, este resultado nos hace dudar del valor $\mu = 410$ o, en otras palabras, es significativo para rechazar el supuesto planteado.

Como puede apreciarse en la figura 8.1, los valores significativos se presentan cuando la muestra adopta valores muy alejados del supuesto inicial, esto es, en las colas sombreadas a derecha e izquier-

Figura 8.2.: Esquema del contraste de hipótesis



da. Por el contrario, los valores centrales se corresponderían con informaciones muestrales compatibles con la hipótesis.

Existen dos técnicas alternativas para resolver los contrastes, esquematizadas en la figura 8.2, que siguen procedimientos distintos para elaborar una conclusión de rechazo o no rechazo de las hipótesis.

a) El procedimiento que denominamos *tradicional o clásico* se basa en determinar una alta banda de tolerancia y admitir esas discrepancias como atribuibles al azar o no significativas (por ejemplo el 95 % de las posibles), considerando las restantes (sólo un 5 % de las mayores) como valores suficientemente atípicos o significativos para rechazar la hipótesis.

La proporción de las observaciones que vamos a considerar significativas debe ser fijada de antemano en función del problema que estemos considerando y se denomina *nivel de significación*, que denotamos por α. Esta probabilidad nos permite determinar un valor de la discrepancia, que denominamos *valor crítico* y marca el límite o separación entre valores significativos y no significativos.

El nivel de significación representará la probabilidad de que, bajo la hipótesis de partida, se presenten discrepancias superiores a las que marca el valor crítico. Dicha probabilidad α será muy baja (usualmente del 1 % o el 5 %) de modo que, si nuestra hipótesis de trabajo es cierta, al seleccionar muchas muestras, aproximadamente el $(1-\alpha)$ % de las veces aparecerían discrepancias admisibles o no significativas.

Por tanto, si al seleccionar una muestra concreta la discrepancia supera al valor crítico entonces sería poco verosímil justificar este resultado como fruto de la mala suerte y asignarlo a ese α % de casos: lo más razonable en cambio sería pensar que la discrepancia no se debe al azar y lo que nos ha fallado es la hipótesis de partida.

b) El otro procedimiento para llegar a conclusiones sobre nuestra hipótesis se conoce como *método del nivel crítico* p, y a diferencia del anterior no impone ninguna restricción a priori sobre las discrepancias admisibles. En este caso se optimiza la información muestral en el sentido de obtener la probabilidad asociada al valor observado de la discrepancia.

De este modo, definiremos el nivel crítico como la probabilidad de obtener, bajo la hipótesis establecida, discrepancias iguales o mayores a la observada. Cuando este nivel crítico p adopta valores muy bajos indica que nuestros resultados muestrales resultan poco verosímiles bajo la hipótesis de partida, luego lo más razonable sería dudar sobre la validez de dicha hipótesis.

Si por el contrario el valor p es alto, indica que la muestra está muy identificada con la hipótesis. Por tanto no sería lógico rechazar, ya que estaríamos aplicando un criterio muy rígido que no superarían la mayoría de las muestras (o, dicho en otras palabras, rechazaríamos con pocos argumentos).

Así, en el ejemplo anterior de la hipótesis $\mu = 410$ para la producción esperada, siguiendo el procedimiento tradicional debemos fijar un cierto nivel de significación α que nos permitirá delimitar regiones de rechazo y de aceptación (más o menos amplias según el valor de α) y conducirá a reglas de decisión del tipo:

- Rechazar siempre que $\bar{X} < 390$ o $\bar{X} > 430$
- No rechazar si $390 \leq \bar{X} \leq 430$

Se dispone así de un esquema general que delimita dos regiones complementarias en las que clasificamos la información muestral y según en cuál de ellas nos situemos decidiremos rechazar o no rechazar la hipótesis inicial. Si ahora introducimos la información muestral en el problema, calculamos $\bar{x} = 415$ y por lo tanto decidimos no rechazar la hipótesis.

En cambio, siguiendo el segundo procedimiento, la información muestral se considera en una etapa previa a la regla de decisión. Así si la hipótesis es $\mu = 410$ y en la muestra obtenemos $\bar{x} = 415$, debemos calcular la probabilidad de que se presenten errores de al menos 5 miles de Tm/mes: $P\left(\left|\bar{X} - \mu\right| \geq 5\right)$. Este valor de p sería en nuestro caso elevado (por ejemplo $p = 0,6$) y a partir de él llegaríamos a la decisión de no rechazar el supuesto de partida.

[¿Cuál sería el razonamiento si la información muestral proporcionase un resultado =200?]

Cabe preguntarse hasta dónde los resultados de p pueden ser calificados de "moderados" y a partir de qué valor pasan a ser suficientemente bajos para rechazar. Evidentemente no existen respuestas exactas a estos interrogantes y en esta característica reside precisamente una de las ventajas de este método: el investigador, a la

vista del nivel crítico obtenido decidirá si rechaza o no la hipótesis pero además, al proporcionar el valor de p, da una idea del nivel de "fuerza" de su conclusión (así, una hipótesis puede ser rechazada con $p = 0,05$, esto es, con un resultado significativo al $5\,\%$, pero evidentemente la conclusión de rechazar parece mucho más sólida si obtenemos un nivel crítico $p = 0,0001$).

Hemos llamado al primer método tradicional o clásico porque ha sido el usual durante muchas décadas y permitió resolver el problema del contraste mediante el uso de las tablas estadísticas disponibles. Estas tablas están calculadas para determinados niveles de significación y resulta sumamente complicado realizar interpolaciones o extrapolaciones (no lineales) de sus correspondientes funciones de distribución para obtener el nivel crítico asociado al resumen muestral observado.

Sin embargo, en épocas recientes este problema fue superado gracias al uso masivo del ordenador, que realiza los cálculos anteriores en un tiempo casi despreciable. Así, hoy en día prácticamente todos los programas informáticos de estadística proporcionan tanto el valor muestral de la discrepancia asociada al contraste como su nivel crítico (p).

Si podemos disponer de las dos alternativas parece claro que el método basado en el nivel crítico sería preferible al clásico, ya que en éste último la información muestral no se aprovecha completamente, sino que se utiliza más bien en un sentido cualitativo de rechazar o no rechazar la hipótesis.

De este modo, si para un contraste determinado dos investigadores seleccionan muestras, resultando una de ellas muy coherente con la hipótesis y la otra con una discrepancia en el límite de las significativas, ambos estudios conducirían a la misma conclusión (no rechazar). Sin embargo, el método del nivel crítico pondría de manifiesto que en el primer caso existe más evidencia para la conclusión que en el segundo (valor más alto de p), y sin embargo esta diferencia entre las dos situaciones no se detecta en el método tradicional.

Existen otros enfoques sobre los procedimientos de contraste de hipótesis estadísticas, cuya metodología resulta más compleja y que estudiaremos en el anexo a este capítulo. Se trata del enfoque introducido por Neyman y Pearson y el método de la razón de verosimilitudes.

J. Neyman y E. S. Pearson (1928, 1933) propusieron una metodología para seleccionar contrastes óptimos. Esta propuesta introduce un cambio en el planteamiento del contraste, ya que la hipótesis de trabajo (que en este enfoque se denomina hipótesis nula) se enfrenta a una alternativa (o hipótesis alternativa), de modo que la elección del mejor test no depende sólo de la hipótesis sino también de la alternativa.

Así, cuando comparamos varios tests para contrastar hipótesis no debemos tener en cuenta sólo el nivel de significación α (que mide probabilidad de equivocarnos al rechazar la hipótesis cuando sea cierta) sino también la probabilidad del error contrario (equivocarnos al aceptar la hipótesis nula cuando la correcta sea la alternativa). Al complementario de esta segunda probabilidad se la denomina *potencia del test*, y el criterio de optimización de Neyman y Pearson consiste en elegir, entre todos los tests que tienen un mismo nivel de significación, aquél que tenga una menor probabilidad del segundo tipo de error (o bien que presente una potencia mayor).

Figura 8.3.: Contraste de hipótesis

Etapas del contraste de hipótesis

- **Enunciado:** Hipótesis (H_0) y alternativa (H_1)
- **Desarrollo:**
  - Método tradicional
  - Método del nivel crítico
- **Conclusión:** Rechazar o no rechazar

8.2. Metodología del contraste de hipótesis

La secuencia seguida para el contraste de una hipótesis puede estructurarse en varias etapas que, de forma simplificada (figura 8.3), denominamos enunciado o formulación, desarrollo y conclusión.

8.2.1. Enunciado

El enunciado de la hipótesis es, sin duda, un punto clave ya que traduce el supuesto que deseamos someter a verificación o contraste. Es importante insistir en que las hipótesis se corresponden con nuestros postulados de trabajo, considerados válidos a priori, que sometemos al control de un test estadístico del que pueden salir refutados.

Como justificaremos más adelante, los contrastes pueden llevarnos a rechazar un supuesto o hipótesis pero nunca a aceptarlo. Ello se debe a que estamos utilizando técnicas estadísticas, que pueden conducirnos a la conclusión de que cierto supuesto es inadecuado (en este sentido "rechazar" sería equivalente a detectar contraejemplos) pero en cambio nunca servirán para demostrar la validez general de un supuesto. De ahí que evitemos el uso del término "aceptar", utilizando en su lugar "no rechazar".

El enunciado de una hipótesis debe reflejar fielmente un supuesto del investigador. En la investigación económica, a menudo se enuncian como hipótesis ciertos postulados teóricos relativos al comportamiento de los agentes económicos (por ejemplo, el principio de la utilidad marginal decreciente, la ley de consumo keynesiana, ...).

Lógicamente nos interesan tan sólo las hipótesis que puedan ser contrastadas empíricamente mediante la aplicación de técnicas estadísticas a la información muestral, que se denominan *hipótesis estadísticas*. En los problemas paramétricos, esto es, para poblaciones conocidas, dichas hipótesis van habitualmente referidas a algún parámetro o característica poblacional (la esperanza μ, la varianza σ^2, la razón de dos varianzas

$\frac{\sigma_X^2}{\sigma_Y^2}$, la proporción p, ...) mientras que en los problemas no paramétricos las hipótesis suelen ser más amplias (supuestos referidos al modelo de probabilidad de la población, la independencia entre v.a., la homogeneidad, ...) y, en general, los procedimientos son menos eficientes.

El supuesto o hipótesis que queremos contrastar se denomina habitualmente *hipótesis nula*, H_0 y enfrentamos a él las restantes posibilidades, que aglutinamos en la *hipótesis alternativa*, H_1. A modo de ejemplo, consideremos el siguiente contraste paramétrico: dada $X \approx \mathcal{N}(\mu, 10)$ queremos contrastar cierto valor esperado concreto ($\mu = 410$), por lo cual incluimos en la alternativa todas las posibles esperanzas diferentes a ese valor:

$$H_0 : \mu = 410$$
$$H_1 : \mu \neq 410$$

Por su parte, en situaciones de inferencia no paramétrica, podríamos someter a contraste un modelo determinado, por ejemplo $\mathcal{N}(\mu = 410, \sigma = 10)$, frente a la alternativa de otra distribución cualquiera. En este caso la formulación sería:

$$H_0 : F(x, \mu, \sigma) = F_0(x)$$
$$H_1 : F(x, \mu, \sigma) \neq F_0(x)$$

donde $F_0(x) \approx \mathcal{N}(\mu = 410, \sigma = 10)$.

Es importante tener presente que la *hipótesis nula* será siempre un supuesto avalado por la información a priori que en principio suponemos verdadero, designando como alternativa a su complementario. Esta distinción resulta de gran trascendencia, ya que el papel de ambas hipótesis en el desarrollo de los contrastes no es en absoluto simétrico.

A menudo se ilustra el papel de las hipótesis nula y alternativa acudiendo a una comparación con los procesos judiciales: una vez enunciada nuestra hipótesis de trabajo, adoptaríamos como principio su "inocencia" o validez, de modo que "una hipótesis será inocente o válida mientras la muestra no demuestre lo contrario".

Siguiendo con este paralelismo, la metodología del contraste de significación consiste en evaluar en qué medida la muestra pone en evidencia la "culpabilidad" o falsedad de H_0.

A modo de ilustración, si deseamos llevar a cabo un análisis inferencial sobre cierta función de demanda $D = \beta_1 + \beta_2 P$, la teoría económica nos llevaría a enunciar como hipótesis nula $H_0 : \beta_2 \leq 0$ frente a la alternativa $H_1 : \beta_2 > 0$ y no viceversa. De igual modo, si la información a priori nos lleva a defender el supuesto de un aumento en el ahorro medio de un período A respecto a otro B, el contraste deberá ser formulado como $H_0 : \mu_A \geq \mu_B$ frente a $H_1 : \mu_A < \mu_B$.

Tanto la hipótesis nula como la alternativa pueden ser clasificadas en simples o compuestas. Una hipótesis (o su alternativa) se dice *simple* cuando de ser cierta especifica plenamente la población; por el contrario las hipótesis son *compuestas* cuando, incluso siendo ciertas, no determinan a la población investigada.

En realidad, en el caso de contrastes genéricos las definiciones anteriores deben ser consideradas de una forma más amplia, no limitándonos a una determinación del modelo sino a la nitidez de la hipótesis. El ejemplo propuesto sobre la normalidad encaja perfectamente en la definición establecida, pero si consideramos el contraste: H_0 : X e Y son poblaciones independientes frente a la alternativa H_1 : existe relación entre X e Y, la hipótesis nula de ser cierta es nítida y por tanto sería simple, aunque no especifique el modelo probabilístico de ninguna de las dos poblaciones.

Consideremos nuevamente el ejemplo de la producción mensual de mineral X, que se distribuye según un modelo normal con varianza determinada.

- Supongamos que la información básica establece únicamente dos posibilidades para la producción media: $\mu = 410$ o $\mu = 350$ y la empresa, con información adicional, defiende el supuesto $\mu = 410$. Entonces el enunciado del contraste sería:

$$H_0 : \mu = 410$$
$$H_1 : \mu = 350$$

tratándose en ambos casos de hipótesis simples.

- Si no existe información básica sobre la producción media y la información adicional nos lleva a defender que ésta será de al menos 410 miles de Tm/mes, el enunciado sería:

$$H_0 : \mu \geq 410$$
$$H_1 : \mu < 410$$

siendo ambas hipótesis compuestas.

[Clasificar las siguientes hipótesis y enunciar las correspondientes alternativas: $\sigma_X^2 \leq \sigma_Y^2$; $\mu_X \neq \mu_Y$; $p \geq 0,2$; $p_X = p_Y$]

Podemos llevar a cabo la siguiente formalización del problema de contraste de hipótesis:

• **Hipótesis referidas a parámetros** Supongamos una población X cuya función de distribución $F_X(x, \theta)$, depende de uno o más parámetros. En lo que sigue nos referiremos a un parámetro aunque el razonamiento sería válido para un vector paramétrico $\boldsymbol{\theta}$.

La información básica de nuestra investigación debe proporcionarnos los posibles valores del parámetro o su recorrido. Este conjunto de posibles valores de θ se denomina *espacio paramétrico* y se denota por Θ.

En la segunda etapa de la formulación, la información contrastable nos permite establecer el valor o los valores teóricos que debemos asumir para el parámetro. Estos valores que constituyen nuestra hipótesis nula forman un subconjunto Θ_0 del espacio paramétrico y por tanto la hipótesis alternativa estará formada por el subconjunto complementario $\Theta_1 = \Theta - \Theta_0$.

Así pues las hipótesis nos conducen a una partición del espacio paramétrico y el contraste puede enunciarse en los siguientes términos

$$H_0 : \theta \in \Theta_o \quad \text{frente a} \quad H_1 : \theta \in \Theta_1 \quad (\Theta = \Theta_0 \cup \Theta_1)$$

Si Θ_0 consta de un solo elemento entonces la hipótesis es simple (de ser cierta queda determinado el valor del parámetro y en consecuencia también la distribución de la población), y si por el contrario Θ_0 consta de dos o más elementos entonces se dice que la hipótesis es compuesta. La clasificación de la hipótesis alternativa se hace en los mismos términos.

A partir de la información muestral pretendemos contrastar si rechazamos que el parámetro θ se sitúe en el subespacio Θ_0 o por el contrario no tenemos razones para este rechazo.

• **Hipótesis genéricas** Cuando las hipótesis tienen carácter genérico admiten una mayor diversidad de posibilidades, por lo que no pueden ser formalizadas con tanta concreción como en el caso anterior. Sin embargo, sustituyendo el espacio paramétrico por otro tipo de espacios la idea de establecer una partición del mismo y su formulación se mantiene.

En esta situación, Θ recogerá todas las especificaciones asociadas al contraste que se plantea, incluyéndose en Θ_0 las favorables a la hipótesis de partida. Así, si queremos contrastar que una población se distribuye según un modelo de Poisson de parámetro λ frente a la alternativa de que se trata de cualquier otro modelo, entonces el espacio paramétrico es sustituido por el de todos los posibles modelos de probabilidad (con los infinitos parámetros admisibles) y el subconjunto correspondiente a la hipótesis nula está constituido por un único punto.

Si por ejemplo estamos realizando un contraste de independencia entre dos poblaciones, el espacio paramétrico será sustituido por el conjunto de todas las posibles relaciones entre esas variables (relación lineal, hiperbólica, independencia, etc.) y el subconjunto correspondiente a la hipótesis constará de un único elemento {independencia}, mientras las restantes posibilidades se incluyen en la alternativa.

La hipótesis nula de un contraste se corresponde habitualmente con un supuesto más concreto que su alternativa. La concreción máxima tiene lugar en el caso de hipótesis nulas simples, como por ejemplo $H_0 : \theta = \theta_0$ que serán rechazadas cuando la información muestral se desvía mucho del supuesto.

A menudo esta desviación puede presentarse en cualquiera de las dos direcciones, con lo cual se rechaza en las colas derecha e izquierda. En tales situaciones los contrastes se denominan *bilaterales o de dos colas* (por ejemplo, frente a la hipótesis anterior podríamos plantear la alternativa bilateral $H_1 : \theta \neq \theta_0$).

Por lo que se refiere a las hipótesis compuestas, éstas podrían estar formadas por un número finito de elementos, aunque en la práctica resulta usual que dichas hipótesis hagan referencia a subintervalos del tipo $H_0 : \theta \geq \theta_0$ o $H_0 : \theta \leq \theta_0$.

En tales casos, la formulación de la hipótesis nula suele incluir en su enunciado el signo igual, mientras que la alternativa aparece con desigualdad estricta. El rechazo de la hipótesis tiene lugar cuando nos desviamos considerablemente en la dirección de la alternativa, por lo cual los contrastes se denominan *unilaterales o de una cola*.

8.2.2. Desarrollo

Como hemos justificado en el epígrafe anterior, los contrastes de significación se basan exclusivamente en la hipótesis nula, ya que la alternativa se introduce como contrapunto para efectuar un contraste estadístico pero no influye en nuestras conclusiones.

Una vez formulada la hipótesis nula, su contraste se basa en la *información muestral* suministrada por una o varias muestras. Generalmente los contrastes se centran en una población X y para ello seleccionamos una muestra aleatoria de tamaño n $(X_1, \ldots, X_n)$ de esa población, pero si queremos llevar a cabo contrastes referidos a dos poblaciones X e Y (igualdad de medias, independencia, homogeneidad, ...) entonces seleccionaremos muestras aleatorias de tamaños n y m, $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_m)$ de las respectivas poblaciones.

Para contrastar la hipótesis poblacional con la información muestral utilizaremos las expresiones de las discrepancias tipificadas (que en general denotamos por d) deducidas en el capítulo 6 tanto para el caso de inferencias paramétricas como no paramétricas. Las discrepancias tipificadas serán expresiones aleatorias que, bajo la hipótesis H_0, suelen seguir modelos probabilísticos conocidos.

A modo de ejemplo, recogemos dos contrastes habituales y sus correspondientes discrepancias tipificadas:

- $H_0 : \mu = \mu_0$

$$d_{\bar{X}/H_0} = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \approx \mathcal{N}(0,1)$$

- H_0 : Independencia de X e Y

$$d_{IND/H_0} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}\, n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}\, n_{\cdot j}}{n}} \approx \chi^2_{(r-1)(s-1)}$$

Como vemos, la aleatoriedad de estas discrepancias depende exclusivamente de la intervención del azar en la selección de la muestra. Dicha aleatoriedad debe ser repartida entre la tolerancia para aceptar la hipótesis nula y la que asignamos a su rechazo y dado que la hipótesis nula es asumida inicialmente como válida este reparto será asimétrico a favor de la hipótesis contrastada.

Llegados a este punto tendremos que diferenciar las metodologías según que el procedimiento utilizado sea el clásico o el basado en el nivel crítico. En el método del nivel crítico la fase de desarrollo concluye con la construcción de la discrepancia, mientras que en el método tradicional el paso siguiente será fijar el nivel de significación α, en función de la seguridad que nos merece nuestra hipótesis.

Supongamos que somos totalmente rígidos y para aceptar la hipótesis exigimos discrepancia nula. Como esta variable es continua la probabilidad de un punto es nula y por tanto en términos de probabilidad nunca aceptaríamos la hipótesis. Parece claro que esta forma de proceder no sería en absoluto recomendable.

Imaginemos ahora que queremos ser imparciales y asignamos el 50 % de esa componente de azar para rechazar y el mismo porcentaje para no rechazar. En este caso si la hipótesis nula es cierta tendremos una probabilidad del 50 % de confundirnos y rechazarla, es decir, la misma que de actuar correctamente (no rechazar una hipótesis que es cierta). Este razonamiento equivale a ignorar la información adicional que nos llevó a enunciar la hipótesis, cosa que no parece lógica si tenemos cierta convicción en nuestra información a priori (basada en teoría, experiencia, etc.).

La pregunta clave sería entonces ¿en cuánto estaríamos dispuestos a valorar la fiabilidad de nuestra hipótesis? Con un planteamiento neutral (valoración nula) llegaríamos al 50 % anterior, pero si tenemos una alta seguridad en nuestro postulado inicial, el margen a favor de mantener nuestra hipótesis podría situarse en el 95 % o el 99 %, niveles de confianza habituales.

En estos casos estamos manteniendo un nivel de significación α del 5 % o el 1 %, lo cual significa que estamos muy seguros de nuestra hipótesis y que, aunque el azar intervenga en la selección de la muestra, vamos a ser muy tolerantes con él ya que, de ser cierto el supuesto inicial, sólo asumimos un riesgo del 1 % de equivocarnos y rechazarlo.

Obsérvese sin embargo que un nivel α tan bajo tiene también sus implicaciones negativas, ya que podemos estar considerando como azar desviaciones que pudieran ser atribuibles a otras circunstancias. Dicho de otro modo, nuestra elevada tolerancia podría llevarnos a no rechazar hipótesis incluso cuando éstas son falsas.

A la vista de estos comentarios, parece claro que el nivel de significación debería depender de nuestra seguridad en el supuesto planteado y por lo tanto asumimos un tratamiento asimétrico para las decisiones de rechazar y no rechazar. Cuando realizamos un experimento no debemos esperar que pueda conducirnos a cualquier decisión, sino que el resultado natural sería no rechazar, y solamente si aparecen discrepancias demasiado elevadas para ser asumidas decidiremos rechazar.

En definitiva, la metodología del contraste de hipótesis estadísticas está diseñada de modo que cuando tomemos la decisión de rechazar, ésta se encuentre estadísticamente avalada. Sin embargo, cuando no rechazamos ello no significa que la estadística esté avalando la decisión de aceptar, sino únicamente que la estadística se abstiene y es la seguridad con la que hemos planteado nuestra hipótesis (esto es, la información adicional o complementaria) la que garantizará nuestra decisión. En definitiva, aceptamos por un conocimiento profundo del problema económico y no porque la distribución de las discrepancias así nos lo aconseje.

Siguiendo con el mismo razonamiento, parece claro que si queremos equilibrar más las responsabilidades deberemos elevar el nivel crítico.

Una vez fijado el nivel de significación α (en general 5 % o 1 %) es posible determinar reglas de decisión del test, con las que concluye el desarrollo del método tradicional o clásico.

A partir de la distribución de la discrepancia tipificada d y fijado un nivel de significación α, podemos determinar las constantes k que recogen los *valores críticos* a partir de los cuales rechazaremos la hipótesis:

- Contraste Bilateral: $P(d > k_2/H_0) = P(d < k_1/H_0) = \frac{\alpha}{2}$
  - $P(|d| > k/H_0) = \alpha$ si la distribución es simétrica
- Contraste Unilateral: $P(d > k/H_0) = \alpha$ si rechazamos en la cola derecha y $P(d < k/H_0) = \alpha$ si rechazamos en la cola izquierda

La expresión $P(d > k/H_0)$ representa la probabilidad de que la discrepancia supere determinado valor cuando la hipótesis nula es cierta. Obsérvese sin embargo que esta notación, habitual en el contraste de hipótesis estadísticas, no es completamente correcta, ya que las probabilidades anteriores no son realmente probabilidades condicionadas, al no tener la hipótesis carácter aleatorio.

A través de este proceso hemos pasado de una partición del espacio paramétrico a otra partición sobre la recta real. En los problemas de contraste paramétrico será muy útil esta transformación, que permite expresar las reglas de decisión en términos de la muestra.

En principio, la región crítica o de rechazo sería un subconjunto del espacio muestral, integrado por todas las muestras $(x_1, \ldots, x_n)$ de $\Re^n$ que conducen al rechazo de H_0. Sin embargo, teniendo en cuenta que la información muestral aparece resumida mediante las discrepancias, resultará más operativo definir la región crítica como un cierto recorrido de la recta real asociado a las discrepancias o bien, si el contraste va referido a parámetros, a los correspondientes estimadores.

Los valores críticos delimitan la *región crítica* (RC) en la que se producirá el rechazo de nuestra hipótesis y su complementaria, la *región de aceptación* (RA). Además, en

el caso de que las hipótesis vayan referidas a parámetros es posible obtener a partir de estos valores críticos (k) unas nuevas constantes C que definen las regiones crítica y de aceptación en relación al estimador T.

8.2.3. Conclusión

La conclusión es la última etapa del procedimiento y abarca los pasos que se realizan desde la selección de una muestra particular hasta la decisión de rechazar o no la hipótesis.

Una vez que se selecciona una muestra concreta $(x_1, \ldots, x_n)$, sobre ella la discrepancia tipificada adoptará un valor determinado d^*, cuyo papel es distinto según que sigamos el método clásico o el del nivel crítico.

En el método clásico la conclusión se reduce a comprobar si este valor d^* de la discrepancia se sitúa dentro de la región crítica. En caso afirmativo, la decisión final será rechazar la hipótesis al nivel de significación establecido (el resultado es significativo a ese nivel) mientras que en el supuesto contrario no existe evidencia para rechazar la hipótesis.

En el método del nivel crítico no hemos introducido ningún valor de α y por tanto no es posible determinar una región de rechazo. El procedimiento consiste entonces en utilizar el valor muestral de la discrepancia d^*, con el cual podemos calcular el *nivel crítico* p:

- Contraste Bilateral: $p = P(|d| > |d^*|\,/H_0)$ si la distribución es simétrica y $p = 2P(d > d^*/H_0)$ o $p = 2P(d < d^*/H_0)$ en otro caso
- Contraste Unilateral: $p = P(d > d^*/H_0)$ o $p = P(d < d^*/H_0)$, según el enunciado de la hipótesis

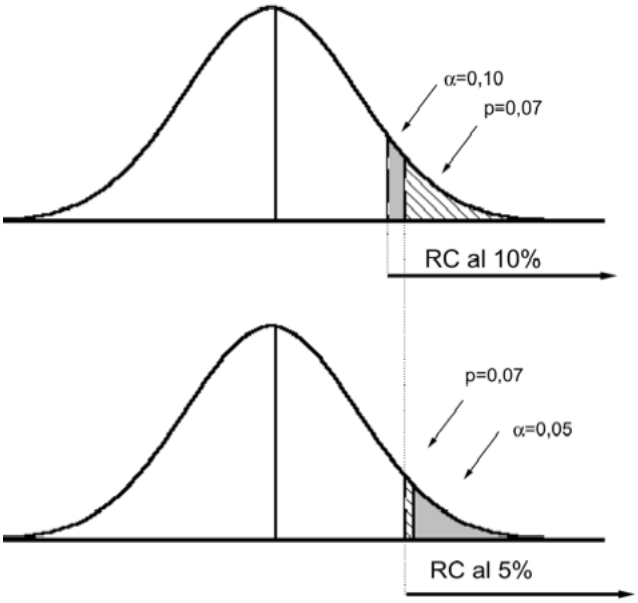
La regla de decisión se establecerá a partir de los datos observados, de modo que un mayor nivel crítico indicará una mayor conformidad de la evidencia empírica (muestra) con la hipótesis de trabajo, y en cambio los niveles críticos bajos van asociados a grandes discrepancias. Esto es:

- Si la información muestral no discrepa significativamente del supuesto de partida no existen razones para rechazar la hipótesis (valor de p alto).
- Si la información muestral no es compatible con el supuesto de partida, los resultados pueden ser calificados de "significativos" para rechazar la hipótesis (valor de p bajo).

Los comentarios anteriores ponen de manifiesto el distinto papel de la información muestral en los dos procedimientos de contraste. En el método clásico esta información se utiliza tan sólo de forma cualitativa: una vez determinadas dos categorías correspondientes a la región crítica (RC) y la región de aceptación (RA), nuestra conclusión se limitará a observar a cuál de ellas nos conduce la muestra observada.

Por el contrario, en el método del nivel crítico la información muestral es el punto de partida para llegar a una decisión, es decir, para evaluar si el resultado debe ser considerado significativo.

Figura 8.4.: Regiones críticas



Es posible establecer una conexión entre el método del nivel crítico y el método clásico en los siguientes términos: dado un nivel crítico p, la conclusión sería rechazar para valores de significación α superiores a él. En cambio, para niveles de significación inferiores a p, la hipótesis no se rechazaría.

Así, si nuestra muestra conduce a un nivel crítico $p = 0,07$ la conclusión debería ser rechazar la hipótesis para cualquier nivel de significación superior (10 %, por ejemplo). Ello se debe a que el valor muestral que lleva asociado el nivel crítico obtenido se encuentra necesariamente dentro de la región crítica fijada al 10 %.

Sin embargo no rechazaríamos para niveles α inferiores (5 %, 1 %) ya que éstos llevan asociadas regiones críticas más pequeñas, en las que no se encontraría situada nuestra información muestral. (figura 8.4)

Obsérvese que hemos utilizado tres términos: *valor crítico*, *nivel crítico* y *nivel de significación*, que -aunque puedan resultar similares- tienen significados distintos. Así, cuando hablamos de niveles criticos o niveles de significación nos estamos refiriendo a probabilidades, mientras que los valores críticos son valores de la discrepancia que delimitan la región crítica de un contraste.

El nivel de significación es una probabilidad asociada a la región crítica del contraste, esto es, delimita la región de los valores significativos para rechazar. Por el contrario el nivel crítico es una probabilidad asociada a la muestra, a partir de la cual el investigador deberá llevar a cabo su decisión.

Los procedimientos descritos serán aplicados a los contrastes habituales en los apartados que siguen. No obstante, conviene llevar a cabo una reflexión general sobre la importancia del supuesto o hipótesis como punto de partida del proceso.

En efecto, hemos visto que los contrastes comienzan con un enunciado teórico que sometemos a un test y finalizan con una conclusión basada en la información mues-

tral. Desde un punto de vista práctico, el investigador podría plantearse "espiar los datos", es decir, examinar la información muestral antes de establecer su supuesto, pero esta posibilidad invalidaría el planteamiento ya que la hipótesis o información a priori quedaría desvirtuada.

En épocas recientes algunos paquetes estadísticos han puesto de moda una técnica llamada "análisis exploratorio de datos" que podríamos resumir de modo simple mediante el interrogante "dado este conjunto de datos ¿qué podemos afirmar?". Este tipo de análisis puede ser útil para llevar a cabo una síntesis de la información muestral, pero desde el punto de vista del método científico resulta peligroso, ya que puede inducir a confusión entre las informaciones a priori y a posteriori, con lo cual el investigador plantearía directamente "hipótesis para rechazar" o "hipótesis para validar".

8.3. Contrastes de hipótesis básicas

En todos los desarrollos inferenciales que hemos estudiado nos basamos en una serie de hipótesis acerca de la muestra y de la población: hemos supuesto la utilización de muestras aleatorias simples (m.a.s.) que llevan implícitas las hipótesis de *aleatoriedad*, *independencia* entre componentes e *idéntica distribución* y en muchos casos también partíamos de que la población se distribuía según un *modelo normal*.

Estos supuestos o *hipótesis estructurales* de trabajo suelen incluirse entre la información básica en los estudios inferenciales. En este apartado estudiamos estas hipótesis desde una doble vertiente: en primer lugar cómo podemos contrastar empíricamente estos supuestos, y en segundo lugar analizar las consecuencias que se derivan de la no verificación de las *hipótesis básicas*.

8.3.1. Hipótesis de m.a.s.

Hasta ahora hemos basado los desarrollos inferenciales en m.a.s., esto es, un proceso de selección aleatorio con variables muestrales independientes e idénticamente distribuidas.

Estos tres supuestos aparecen estrechamente relacionados, de forma que cuando la población es infinita o el muestreo es con reposición, la no verificación de cualquiera de ellos nos lleva a la invalidación de los otros.

Naturalmente cuando trabajamos con poblaciones finitas y métodos de muestreo sin reposición, el proceso de selección de la muestra puede ser aleatorio, pero en cambio las variables muestrales no son independientes (los valores que pueden tomar dependen de las observaciones anteriores) ni están idénticamente distribuidas (el sistema de probabilidades irá cambiando en cada selección).

Hemos aclarado algunas veces el concepto "aleatorio" identificándolo con "estocástico" y nos hemos planteado cómo seleccionar una muestra aleatoria. La respuesta consiste en utilizar tablas de números aleatorios, una vez indexadas las unidades (valores) poblacionales, ya que la aleatoriedad de la tabla garantiza la de la muestra.

Sin embargo, hasta ahora no nos hemos ocupado de medir el nivel de aleatoriedad de la muestra o contrastar si es asumible o no que los datos muestrales son aleatorios o, equivalentemente, que se trata de una m.a.s.

Para llevar a cabo este contraste enunciamos la hipótesis nula H_0: *los datos constituyen una m.a.s.* frente a la que podrían formularse diferentes alternativas. Así, la hipótesis H_1 podrá venir dada por violaciones de hipótesis concretas (no aleatoriedad, no independencia, distribuciones no idénticas) o bien ser explicitada en términos amplios (*los datos no constituyen una m.a.s.*).

8.3.1.1. Test de rachas

El *test de rachas* se construye asociando a una población dos categorías alternativas y clasificando según este criterio los elementos que forman parte de la muestra seleccionada. De este modo se obtendrá una secuencia de categorías ordenadas según las observaciones muestrales.

Cada vez que se produce un cambio de categoría decimos que hay una nueva racha; posteriormente observamos las rachas que se presentan, esperando que si la muestra es aleatoria éstas sean un número moderado. Si por el contrario el número de rachas es excesivamente elevado o muy pequeño rechazamos la hipótesis de que las muestras hayan sido seleccionadas al azar.

A modo de ilustración supongamos que se lanza diez veces una moneda. Si todos los resultados fueran caras o todos cruces parece poco creible que el comportamiento de la moneda fuese aleatorio, y se diría lo mismo si se observan dos rachas (k caras y $n-k$ cruces consecutivas). Además, aunque nos resulte llamativo, tampoco parece justificado por el azar un número excesivo de rachas (por ejemplo, si se presentasen diez rachas esto significaría que se irían alternando sistemáticamente caras y cruces, que sería un efecto contrario al comportamiento azaroso).

Por el contrario, si obtuviésemos un número intermedio de rachas (cinco o seis rachas de diferente longitud) entonces sí podríamos asumir la hipótesis de aleatoriedad.

¿Cómo se puede trasladar este proceso a una variable aleatoria cualquiera? Se trata de definir sobre dicha variable dos categorías y observar las rachas que se producen con las mismas.

En el caso de que trabajásemos con variables dicotómicas la asignación de estas categorías sería automática, al presentarse sólo dos posibilidades. Si en cambio partimos de una variable aleatoria arbitraria, dado un conjunto de observaciones de la misma podemos calcular la mediana de esos datos y establecer las categorías "menor que la mediana" y "mayor que la mediana".

Teniendo en cuenta que la mediana es el punto central de la distribución tendríamos tantos elementos inferiores como superiores con lo cual se presentarían secuencias del tipo $ABAABABBBA\ldots$, $01101001110\ldots$, $+-++--+----+--\ldots$, etc. La forma de denotar las categorías sería secundaria y por comodidad vamos a considerar el valor 0 si $x_i < Me$ y 1 si $x_i > Me$ (no incluimos los valores que tengan una coincidencia exacta con la mediana).

Denotemos por n_0 el número de ceros observado en la secuencia, por n_1 los correspondientes unos, y $n = n_0 + n_1$. A partir de nuestra secuencia de unos y ceros

observaremos las rachas, definidas como secuencias de observaciones consecutivas de la misma categoría hasta la aparición de una categoría distinta.

Cabe preguntarse ¿qué sucedería si las muestras con las que trabajamos no fuesen m.a.s.? Podrían presentarse muestras con cierta tendencia, esto es, muchos ceros al principio y unos al final o viceversa), en cuyo caso aparecerían pocas rachas. También podrían obtenerse muestras en las que se presentasen alternativamente ceros y unos, con lo cual el número de rachas sería elevado.

Si en vez de plantear este esquema sobre una muestra concreta lo hacemos sobre una muestra genérica, el número de rachas será una v.a. R. Este razonamiento conduce a una regla de decisión basada en el número de rachas observado r^* , y nos llevará a rechazar la hipótesis nula si r^* sobrepasa cierto valor r_2 o bien es inferior a r_1, siendo estos valores tales que $P(R \notin [r_1, r_2]/H_0) \leq \alpha$. Se trata de un contraste bilateral, puesto que podemos rechazar la hipótesis tanto por un número excesivamente bajo como alto de rachas.

La función de probabilidad de R (y por tanto la determinación de los límites de la región crítica r_1 y r_2) depende del tamaño muestral y viene dada por la siguiente expresión:

$$P(R=r) = \begin{cases} \dfrac{2\binom{n_1-1}{m-1}\binom{n_0-1}{m-1}}{\binom{n}{n_1}} & \text{si r es par } (r=2m) \\ \dfrac{\binom{n_1-1}{m}\binom{n_0-1}{m-1}+\binom{n_1-1}{m-1}\binom{n_0-1}{m}}{\binom{n}{n_1}} & \text{si r es impar } (r=2m+1) \end{cases}$$

con $E(R) = \dfrac{2n_0n_1}{n_0+n_1} + 1$ y $Var(R) = \dfrac{2n_0n_1(2n_0n_1-n_0-n_1)}{(n_0+n_1)^2(n_0+n_1-1)}$

Los límites r_1 y r_2 pueden ser obtenidos como: $P(R < r_1/H_0) = P(R > r_2/H_0) = \frac{\alpha}{2}$.

Si resolvemos el problema por el método del nivel crítico, calcularíamos $p = P(R \leq r/H_0)$, y si esta probabilidad (o su complementaria) fuesen muy bajas estarian indicando que, bajo el supuesto de aleatoriedad de la muestra, la presencia de un número tan reducido (o tan elevado interpretando el complementario) de rachas no es atribuible al azar, y en consecuencia rechazamos la hipótesis. Así pues, el nivel crítico viene dado en este caso por el doble de la probabilidad correspondiente a la cola más pequeña, es decir: $p = 2P(R \leq r^*/H_0)$ o $p = 2P(R \geq r^*/H_0)$ según los casos.

En ciertas situaciones el test de rachas puede plantearse como un contraste unilateral. Este es el caso cuando la hipótesis alternativa indica la existencia de una tendencia en algún sentido, que iría asociada a una sola cola.

En el caso de que el tamaño muestral n sea elevado, la proporción de unos (p) y de ceros $(1-p)$ puede considerarse como un valor fijo. Wald y Woldfowitz demostraron que la distribución de R es asintóticamente normal con características $E(R) = 2np(1-p)$ y $Var(R) = 4np^2(1-p)^2$, con lo cual se obtiene la discrepancia tipificada asociada al estimador de rachas:

$$d_R = \frac{R - 2np(1-p)}{2\sqrt{n}p1-p)} \approx \mathcal{N}(0,1)$$

La regla de decisión del test es análoga al caso de la distribución exacta.

8.3.1.2. Test de rangos

Un planteamiento alternativo para contrastar la hipótesis nula H_0: los datos constituyen una m.a.s. es el *test de rangos*, que consiste en comparar la muestra observada $(x_1, \ldots, x_n)$ con la misma muestra ordenada. En concreto, para cada observación x_i de la muestra compararemos la posición que ocupa en el proceso de selección, i, con la posición que ocupa en la ordenación, que denominamos rango r_i[1].

¿Qué comportamiento cabría esperar bajo la hipótesis nula de aleatoriedad? Parece claro que en ese supuesto la posición y el rango serían independientes, mientras que la existencia de relaciones entre i y r_i iría en contra de la hipótesis nula.

Para estudiar la situación trabajamos con las diferencias cuadráticas $(i - r_i)^2$, que se resumen bajo la expresión $r = 1 - \dfrac{6}{n(n^2-1)} \sum_{i=1}^{n} (i - r_i)^2$, denominada *coeficiente de correlación de Spearman*.

Es interesante destacar dos características de esta expresión:

$$r = \frac{\sum_{i=1}^{n} (i - \bar{i})\,(r_i - \bar{r})}{\sqrt{\sum_{i=1}^{n} (i - \bar{i})} \sqrt{\sum_{i=1}^{n} (r_i - \bar{r})}} \; ; \; -1 \leq r \leq 1$$

en primer lugar, su coincidencia con el coeficiente de correlación muestral: y en segundo lugar que, para tamaños muestrales elevados, dicha expresión converge bajo la hipótesis nula a un modelo normal

$$\mathcal{N}\left(0, \frac{1}{\sqrt{n-1}}\right)$$

Definimos en este caso la discrepancia tipificada asociada al test de los rangos como $d_r = r\sqrt{n-1} \approx \mathcal{N}(0,1)$. Así, se observará el valor muestral d_r^* que, si H_0 es cierta, debería adoptar valores cercanos a 0.

El nivel crítico vendrá dado para cada valor d_r^* por la probabilidad $P(|d_r| > |d_r^*|/H_0)$ que, si es suficientemente baja, permitirá calificar nuestro resultado de significativo para rechazar la hipótesis de m.a.s.

Si el tamaño muestral es inferior a 30 la aproximación normal no resulta adecuada pero es posible recurrir a la transformación $\sqrt{n-2}\dfrac{r}{1-r^2} \approx t_{n-2}$.

Este contraste resulta de gran interés en el ámbito de las series temporales, cuando deseamos estudiar si una serie presenta o no tendencia. En tal situación, la hipótesis nula sería H_0: *la serie no tiene tendencia*, equivalente al supuesto de m.a.s.

[1] En el caso de que varias observaciones sean coincidentes y por tanto ocupen el mismo lugar en la ordenación, la solución más frecuente consiste en asignar a todas estas observaciones el rango promedio.

Los tests considerados no agotan todas las posibilidades para contrastar la aleatoriedad de una muestra. Un método que se utiliza habitualmente cuando los datos se generan en un soporte temporal consiste en estudiar la correlación entre las observaciones. Si las observaciones fuesen independientes y consideramos los conjuntos de datos $(x_1, \ldots, x_{n-1})$ y $(x_2, \ldots, x_n)$ éstos presentarían una correlación nula. Podríamos considerar cualesquiera otros subconjuntos de esas observaciones (por ejemplo los $\frac{n}{2}$ primeros datos y los $\frac{n}{2}$ últimos) de modo que, si entre alguno de estos subconjuntos encontramos correlación (en este caso se denomina autocorrelación) entonces el proceso generador de la muestra no sería aleatorio sino que presentaría un patrón determinista o sistemático.

La distribución de la discrepancia tipificada asociada al coeficiente de autocorrelación sería análoga a la desarrollada en los párrafos anteriores para el coeficiente de Spearman: para valores pequeños de n se realiza el ajuste a la distribución t_{n-2} y para tamaños grandes de n se utiliza la aproximación normal.

8.3.1.3. Consecuencias del incumplimiento del supuesto de m.a.s.

Si rechazamos la hipótesis de muestreo aleatorio simple (m.a.s.) gran parte de los procesos inferenciales descritos tendrán que ser revisados.

En primer lugar no podríamos obtener la función de verosimilitud como producto de las funciones de densidad marginales, por lo que deberemos revisar la expresión de esta función. En consecuencia, si estamos utilizando un EMV éste debe ser recalculado.

La no verificación de la independencia entre los componentes muestrales afectará a la varianza del estimador y en consecuencia a las propiedades de mínima varianza y eficiencia.

Para obtener las distribuciones de las discrepancias no podremos basarnos en la propiedad de reproductividad. Las aproximaciones de Chebyshev y TCL deben ser revisadas, pues debemos tener en cuenta la covarianza entre las variables muestrales.

Además, para garantizar que la discrepancia tipificada asociada a la varianza muestral sigue un modelo χ^2_{n-1} necesitamos que las v.a. muestrales sean independientes.

Hemos enumerado sólo algunas de las consecuencias directas de la no verificación del supuesto de m.a.s.; no pretendemos ser exhaustivos sino solamente señalar cómo el rechazo de esta hipótesis nos conduciría a una revisión profunda de las técnicas inferenciales.

8.3.2. Contrastes de bondad de ajuste. Test de normalidad

Existen varios contrastes de carácter general relativos al modelo probabilístico de la población, que se conocen como *contrastes de bondad de ajuste* , y se refieren a la aproximación de los datos a un modelo probabilístico teórico. Además, dada la

importancia de la distribución normal en los procesos inferenciales, consideraremos en este apartado varios contrastes específicos para este modelo.

En los contrastes de bondad de ajuste no existe información básica por lo que se trata de contrastes en un contexto no paramétrico. La información complementaria nos llevará a establecer la hipótesis nula de que el modelo de probabilidad de la población X es uno determinado $F_0(x)$, es decir $H_0 : F(x) = F_0(x)$, y la hipótesis complementaria o alternativa será que el modelo es otro diferente $H_1 : F(x) \neq F_0(x)$.

Supongamos que la información muestral consta de un conjunto de datos $(x_1, \ldots, x_n)$ que proceden de la observación de una m.a.s. de tamaño n de la población X, con los que se pretende contrastar la hipótesis.

Existen diversos tipos de contrastes que en síntesis persiguen evaluar la discrepancia entre la distribución de frecuencias observadas y la distribución teórica. Cuando esta discrepancia es muy elevada entonces rechazaremos la hipótesis de que el modelo es el especificado y de lo contrario diremos que los datos no son significativos para rechazar la función de distribución especificada.

Este tipo de tests suelen basarse en el hecho de que la función de distribución muestral converge en probabilidad a la poblacional (distribución origen que da lugar a la muestra). Este resultado, aunque admite diversos enunciados, suele conocerse como lema de Glivenko-Cantelli.

8.3.2.1. Test de Bondad de Ajuste

El test de la χ^2 se construye a partir de las discrepancias entre el histograma de la distribución de frecuencias de la muestra y el que se obtendría calculando la frecuencia teórica que correspondería a esos mismos valores bajo el supuesto de que la hipótesis nula es correcta.

Karl Pearson (1857-1936) se planteó la necesidad de un criterio para evaluar hasta qué punto una curva ajustada a un conjunto de observaciones era válida. Así ideó la medida chi-cuadrado (1900), estudiando además su distribución que posteriormente se reveló como un instrumento de gran utilidad en diversas aplicaciones estadísticas.

Supongamos que se toma una m.a.s. $(X_1, \ldots, X_n)$ cuyos valores agrupamos en r clases o intervalos $I_1, \ldots, I_r$ y representemos las frecuencias absolutas de esos intervalos por $n_1, \ldots, n_r$ $(\sum_{i=1}^{r} n_i = n\ )$.

En este contraste es necesario tener presente que, cuando la distribución es continua, los intervalos deben abarcar todo el recorrido de la variable.

Cada uno de esos intervalos será del tipo $I_i = (a_i\,,\, b_i]$ por lo que, de ser cierta la hipótesis nula, la probabilidad de que una observación pertenezca a ese intervalo será:

$$p_i = P(X \in I_i) = P(a_i < X \leq b_i) = F_0(b_i) - F_0(a_i)$$

Como la muestra consta de n observaciones, el número de éxitos en el intervalo I_i seguiría un modelo multinomial $\mathcal{M}(n, p_i)$, y en consecuencia su esperanza será np_i. Para cada intervalo, podemos evaluar el error aleatorio asociado al mismo como diferencia entre la frecuencia observada y la teórica que postula nuestra hipótesis: $e_i = n_i - np_i$.

Tal y como ya hemos descrito en el capítulo 6, la discrepancia tipificada asociada a las frecuencias n_i sigue una distribución que, para valores elevados de n, es aproximadamente χ^2 con r-1 g.l.

$$d_n = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} \to \chi^2_{r-1}$$

Hasta aquí estamos suponiendo una m.a.s. por lo que las frecuencias de cada intervalo son aleatorias y en consecuencia también lo será la discrepancia tipificada asociada a este estadístico.

Si la especificación de la distribución incluye parámetros desconocidos, entonces $p_i = p_i(\theta)$. La discrepancia anterior sigue siendo válida siempre que los parámetros se estimen a partir de la muestra por el método de máxima verosimilitud; los EMV serán

$$\hat{p}_i = \frac{n_i}{n}$$

y cada parámetro actúa como una restricción que limita el número de g.l., por lo que si estimamos h parámetros la discrepancia anterior seguirá aproximadamente una distribución χ^2_{r-1-h}.

Si partimos ahora de una muestra concreta $(x_1, \ldots, x_n)$ debemos calcular sobre esos datos el valor de la discrepancia:

$$d^*_n = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

si la discrepancia entre las frecuencias observadas y las teóricas es tolerable no rechazaremos la hipótesis, pero si por el contrario es elevada la rechazaremos. El límite de tolerancia d_0 puede ser obtenido a partir de las tablas de la χ^2 para un cierto nivel de significación α ($P(d_n > d_0/H_0) = \alpha$).

Sin embargo, la práctica más habitual para construir la regla de decisión del test consiste en calcular el nivel crítico asociado al contraste, es decir, obtener la probabilidad asociada al valor muestral d^*: $p = P\left(d_n > d^*/H_0\right)$; cuando esta probabilidad sea baja significará que la discrepancia observada es muy alta y por tanto las frecuencias observadas discrepan mucho de las correspondientes al modelo teórico postulado, por lo que tendríamos evidencia significativa para rechazar la hipótesis.

Cuanto mayor sea el número de intervalos en los que agrupemos los valores muestrales, tanto más fino será el contraste. Sin embargo, el proceso de convergencia a la χ^2 exige que la frecuencia esperada de cada intervalo sea mayor o igual que 5, con lo cual si para algún i el producto $np_i < 5$ debemos

proceder a reagrupar los intervalos.

A modo de ilustración, supongamos que deseamos contrastar sobre la v.a. X la hipótesis $H_0 : X \approx \mathcal{U}(0, 100)$, a partir de la siguiente información muestral agrupada en intervalos:

Intervalo	Frecuencia observada n_i
$(0, 10]$	9
$(10, 20]$	11
$(20, 50]$	12
$(50, 100]$	20

Para contrastar el supuesto de uniformidad de X deberíamos comparar las frecuencias observadas (n_i) con las teóricas (np_i) teniendo en cuenta que bajo la hipótesis nula la probabilidad de cada intervalo se obtiene como

$$p_i = P(a_i < X \leq b_i) = \frac{b_i - a_i}{100}$$

Intervalo	n_i	np_i
$(0, 10]$	9	5,2
$(10, 20]$	11	5,2
$(20, 50]$	12	15,6
$(50, 100]$	20	26
Total	52	52

A partir de la información anterior se obtiene la discrepancia

$$d_n^* = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i} = 11,4615$$

que lleva asociado un nivel crítico

$$p = P(\chi^2_{4-1} > 11,4615) = 0,00947$$

que es suficientemente bajo para calificar el resultado de significativo y rechazar la hipótesis de uniformidad propuesta para X.

El contraste de normalidad de una población podría tratarse como un caso particular del test anterior, donde la hipótesis nula sería $H_0 : F(x) \approx \mathcal{N}(\mu, \sigma)$. Cuando los parámetros μ y σ se estiman a partir de la muestra la distribución chi-cuadrado presenta $n - 3$ g.l.

Una de las ventajas del contraste de bondad de ajuste mediante la chi-cuadrado es que, en el caso de que la conclusión sea el rechazo de la hipótesis, este test permite detectar las observaciones causantes de dicho rechazo. Para ello bastaría examinar las discrepancias individuales que aparecen agregadas en la expresión final chi-cuadrado y de este modo podremos saber si dichas discrepancias son homogéneas o bien existe un único valor extremo, que incluso podría deberse a errores en la muestra.

Así, en el ejemplo anterior se observa que la mayor discrepancia corresponde al intervalo (10,20] en el que se han registrado 11 observaciones muestrales cuando la frecuencia teórica sería aproximadamente la mitad [Compruébese que para este intervalo se obtiene $np_i = (52)(0,1) = 5,2$].

8.3.2.2. Test de Kolmogorov-Smirnov

El *test de Kolmogorov-Smirnov* (K-S) basa la discrepancia entre la muestra y la población en la función de distribución en vez de la función de probabilidad del test chi-cuadrado.

Consideremos una población X para la cual establecemos la hipótesis de que su distribución es una determinada: $H_0 : F(x) = F_0(x)$.

Para contrastar este supuesto disponemos de una muestra particular de esa población $(x_1, \ldots, x_n)$ para la cual construimos la f.d. de la muestra que comparamos con nuestra hipótesis nula.

Vamos a establecer la hipótesis adicional de continuidad para la variable, de forma que la probabilidad de cualquier punto es nula y por tanto la probabilidad de que dos valores muestrales coincidan también será cero.

Podemos suponer que cada valor muestral se repite una única vez, y disponemos estos valores ordenados de forma creciente: $x_1 < \cdots < x_n$.

Denotemos por $S_n(x)$ la f.d. muestral que vendrá dada por:

$$S_n(x) = \begin{cases} 0 & si\, x < x_1 \\ \frac{i}{n} & si\, x_i \leq x < x_{i+1} \\ 1 & si\, x_n \leq x \end{cases}$$

Definimos el error de estimación de la distribución poblacional a partir de la muestral como el supremo de las diferencias en todo el recorrido:

$$D_n^* = \sup_{-\infty<x<+\infty} |S_n(x) - F_0(x)|$$

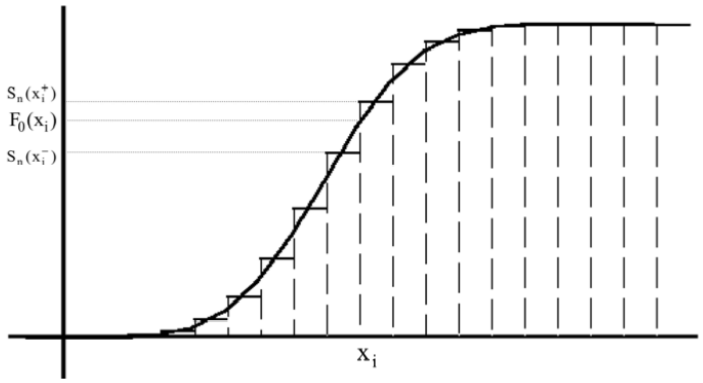
Cuando esta diferencia máxima es pequeña quiere decir que ambas f.d. se aproximan, por lo cual no podríamos rechazar la hipótesis nula. Si por el contrario la diferencia no fuese admisible rechazaríamos la hipótesis.

En el caso de una muestra genérica, el supremo anterior será una v.a. cuya distribución de probabilidad exacta para tamaños pequeños de n, bajo el supuesto de que la hipótesis nula es cierta, fue obtenida por Massey (1952). Para tamaños elevados de muestra la probabilidad de que D_n sea mayor que el valor observado se aproxima mediante la expresión:

$$\lim_{n\to\infty} P\left(D_n > D_n^*/H_0\right) = 2\sum_{i=1}^{\infty}(-1)^{i-1}e^{-2i^2\frac{D_n^{*2}}{n}}$$

En cualquiera de los dos casos podemos obtener el nivel crítico que nos indica si debemos o no rechazar la hipótesis de que la población sigue un modelo determinado.

Figura 8.5.: Test de Kolmogorov-Smirnov



Podemos observar que esta expresión depende (además del valor) únicamente de n (número de observaciones) y no de la distribución teórica supuesta. Esto significa que el límite de tolerancia admisible para rechazar una hipótesis de población exponencial, gamma o normal es la misma.

Para determinar D_n^* debemos calcular todas las diferencias entre la distribución muestral y la teórica. Las mayores de estas diferencias se encontrarán en los puntos de salto de la distribución muestral por lo que nos bastará con observar las desviaciones en los puntos $(x_1, \ldots, x_n)$. Sin embargo, debemos tener en cuenta que $S_n(x_i^-) \neq S_n(x_i^+)$ (dado que la f.d. muestral es escalonada) por lo que para calcular la desviación suprema es necesario contemplar las $2n$ diferencias:

$$\left|S_n\left(x_i^-\right) - F_0(x_i)\right| \ , \ \left|S_n\left(x_i^+\right) - F_0(x_i)\right| \ , \ \forall i = 1, \ldots, n$$

A.N. Kolmogorov introdujo en 1933 el estadístico D_n para el que elaboró las primeras tablas de probabilidad. Por su parte, N.V Smirnov (1939) publicó tablas más precisas y posteriormente (1944) llegó a acotar las probabilidades del estadístico de Kolmogorov.

El test K-S establece ciertas restricciones al modelo supuesto para la población. En concreto, hemos hecho explícito que debe tratarse de un modelo continuo, pero además los parámetros de la distribución supuesta deben ser conocidos para poder calcular la distribución anterior.

Sin embargo, dada la importancia de la distribución normal existe una *corrección de la distribución de K-S*, debida a Lilliefors (1967), que nos permite aplicar este test cuando se estiman los parámetros a partir de la muestra.

Como puede apreciarse en la tabla que sigue, dados un tamaño muestral n y un nivel de significación α, las tablas de Lilliefors proporcionan valores críticos inferiores a los de Kolmogorov-Smirnov. Este hecho se debe a que el desconocimiento de los parámetros poblacionales debe ser compensado siendo más estrictos en el contraste, es decir, admitiendo menores desviaciones.

Tamaño muestral	K-S ($\alpha = 0,05$)	Lilliefors ($\alpha = 0,05$)	K-S ($\alpha = 0,01$)	Lilliefors ($\alpha = 0,01$)
5	0,565	0,337	0,669	0,405
10	0,41	0,258	0,49	0,294
15	0,338	0,22	0,404	0,257

8.3.2.3. Test de normalidad de Jarque-Bera

El contraste de normalidad desarrollado por C.M. Jarque y A.K. Bera (1980) se basa en el estudio de la forma de la distribución, examinando sus discrepancias respecto a la curva campaniforme característica del modelo normal.

Estas discrepancias respecto a la normalidad deben ser evaluadas mediante dos características de forma: la simetría y la kurtosis o apuntamiento. Para ello suelen emplearse los coeficientes g_1 y g_2 que se definen e interpretan como sigue[2]:

Característica	Simetría	Kurtosis
	$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^3}{S^3}$	$g_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^4}{S^4} - 3$
g=0	Distribuciones simétricas	Distribuciones mesocúrticas
g>0	Distr. con asimetría positiva	Distribuciones leptocúrticas
g<0	Distr. con asimetría negativa	Distribuciones platicúrticas

El *contraste de normalidad de Jarque-Bera* se basa en una discrepancia definida por la expresión:

$$d_{JB} = \frac{n}{6}\left(g_1^2 + \frac{1}{4}g_2^2\right)$$

en la que aparece el tamaño muestral n y los coeficientes muestrales g_1 y g_2 de Fisher elevados al cuadrado. Bajo la hipótesis nula de normalidad esta expresión se distribuye según un modelo chi-cuadrado con 2 grados de libertad.

Para tamaños elevados de muestra la distribución de g_1 es aproximadamente normal con

$$E(g_1) = 0 \text{ y } Var(g_1) = \frac{6}{n}$$

Por su parte, la medida de apuntamiento g_2 es asintóticamente normal con esperanza nula y

[2] En el capítulo 2 hemos definido los coeficientes de asimetría y apuntamiento γ_1 y γ_2 de Fisher, que para una variable aleatoria X vienen dados por las expresiones

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \text{ y } \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

Para el contraste de normalidad examinamos las correspondientes características muestrales:

$$g_1 = \frac{m_3}{S^3} \text{ y } g_2 = \frac{m_4}{S^4} - 3$$

$$Var(g_2) = \frac{24}{n}$$

y de ahí que se combinen ambas medidas en un contraste conjunto:

$$\frac{n}{6}g_1^2 + \frac{n}{24}g_2^2 \approx \chi_2^2$$

La discrepancia d_{JB} resume las características de forma de la muestra, y adoptará valores bajos si la distribución observada es aproximadamente simétrica y mesocúrtica. En otro caso, a medida que se detectan asimetrías (positivas y negativas) o desviaciones en la kurtosis (distribuciones platicúrticas o leptocúrticas) la discrepancia aumenta de valor. Para llegar a una conclusión del contraste, bastaría con calcular el nivel crítico asociado al resultado muestral: $p = P\left(d_{JB} > d_{JB}^*/H_0\right)$, que si es suficientemente bajo conduce al rechazo de la hipótesis de normalidad.

Los métodos anteriores no agotan las posibilidades para contrastar un modelo probabilístico. Así, la normalidad puede ser contrastada mediante el método de Shapiro y Wilks, que estudia si una muestra representada en papel probabilístico normal puede ser ajustada adecuadamente a una recta.

8.4. Algunos contrastes paramétricos

Cuando la población investigada sigue un modelo probabilístico conocido las inferencias sobre dicha población son de tipo paramétrico. Dichos contrastes suelen ir referidos a los parámetros poblacionales y podrán ser resueltos mediante cualquiera de los dos procedimientos descritos anteriormente: el tradicional o clásico y el del nivel crítico. Ambos métodos difieren como hemos visto en su desarrollo, pero coinciden sin embargo en la herramienta utilizada para analizar las discrepancias entre hipótesis y muestra.

Siguiendo el esquema general, deberemos comenzar por enunciar la hipótesis que deseamos contrastar sobre el parámetro de interés, con su correspondiente alternativa.

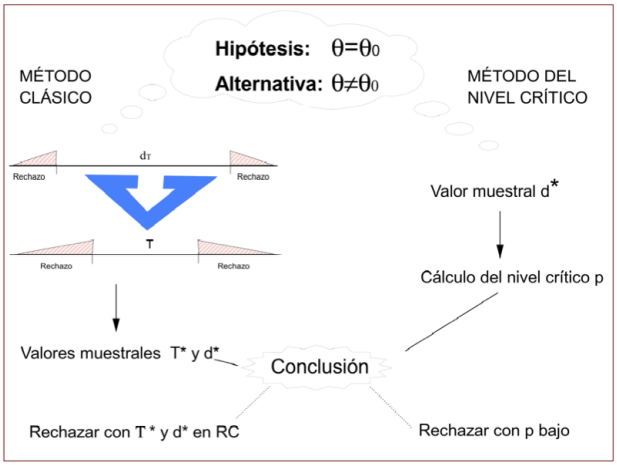
Como ya hemos visto, en el desarrollo de un contraste de hipótesis nos interesa distinguir entre contrastes unilaterales y bilaterales. Así pues, para un parámetro genérico θ, podríamos plantear las siguientes situaciones:

$$H_0 : \theta = \theta_0 \qquad H_0 : \theta \geq \theta_0 \qquad H_0 : \theta \leq \theta_0$$

$$H_1 : \theta \neq \theta_0 \qquad H_1 : \theta < \theta_0 \qquad H_1 : \theta > \theta_0$$

La primera posibilidad se corresponde con un contraste bilateral o de dos colas, en el que la hipótesis alternativa se sitúa tanto a izquierda como a derecha de la hipótesis planteada H_0. Por el contrario, los otros contrastes son unilaterales ya que ahora la alternativa se correspondería con una sola cola (la izquierda en el primer caso y la derecha en el segundo).

Figura 8.6.: Métodos de contraste de hipótesis



Es evidente que las tres situaciones comentadas no agotan la casuística de los contrastes. Sin embargo, son suficientes para describir la metodología de los contrastes de significación, ya que otros posibles enunciados se resolverían de modo similar.

Así, si planteamos un contraste de hipótesis nula simple frente a alternativa simple:
$H_0 : \theta = \theta_0$
$H_1 : \theta = \theta_1$
el espacio paramétrico tendría dos regiones (cada una de ellas con un valor único) por lo cual el planteamiento sería de una sola cola (a la derecha si $\theta_1 < \theta_0$ y a la izquierda en caso contrario).

Obsérvese que en todos los casos se incluye en la hipótesis nula el valor concreto del parámetro θ_0 que marca su límite con la alternativa.

En los apartados que siguen desarrollamos los contrastes referidos a los parámetros más habituales, tanto por el método clásico como por el del nivel crítico. En síntesis, para un contraste bilateral el esquema de trabajo es el ilustrado en la figura 8.6.

Por su parte, los contrastes de hipótesis unilaterales se resolverían según el mismo esquema de trabajo, con la salvedad de que consideraríamos una sola cola, tanto en la región crítica como en la probabilidad calculada como nivel crítico.

Como podemos apreciar en los esquemas anteriores, el método del nivel crítico empezaría respondiendo a la pregunta ¿qué dice la muestra? Para evaluar si la muestra dista mucho del valor hipotético se calcula la discrepancia tipificada d^* y su correspondiente nivel crítico p.

Por el contrario, el método clásico proporciona reglas de decisión, estableciendo regiones críticas o de rechazo de una hipótesis. Sólo al final del proceso se aplica dicha regla genérica a la muestra concreta para obtener una conclusión relativa a la hipótesis planteada.

8.4.1. Contrastes sobre la media

Consideremos una población normal $X \approx \mathcal{N}(\mu, \sigma)$ en la que deseamos contrastar algún supuesto inicial sobre el valor esperado μ. Bastaría con aplicar el planteamiento general recogido en los esquemas anteriores, distinguiendo en cada caso si se trata de contrastes bilaterales o unilaterales.

Las situaciones más habituales son las descritas a continuación:

Enunciado I:

$$\boxed{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{array}}$$

Al contrastar la hipótesis $H_0 : \mu = \mu_0$ frente a la alternativa $H_1 : \mu \neq \mu_0$, si fijamos un nivel de significación α, es posible obtener un valor k de la discrepancia tal que se cumpla:

$$P\left(|d_{\bar{X}}| > k / H_0\right) = \alpha$$

Como ya hemos justificado en temas anteriores, el cálculo de k se efectuaría en las tablas de la distribución $\mathcal{N}(0,1)$ siempre que σ fuese conocido. En otro caso, para pequeños tamaños de muestra deberíamos acudir a las tablas t de Student con $n-1$ g.l. Obsérvese que, en cualquiera de los dos casos, el valor k es más elevado cuanto menor sea nuestro nivel de significación α.

A partir de este valor k se obtiene la región crítica para la discrepancia $d_{\bar{X}}$:

$$(-\infty, -k) \cup (k, +\infty)$$

y equivalentemente la región crítica para el estimador sería:

$$\left(-\infty, \mu_0 - k\frac{\sigma}{\sqrt{n}}\right) \bigcup \left(\mu_0 + k\frac{\sigma}{\sqrt{n}}, +\infty\right) \quad \text{si } \sigma \text{ es conocido}$$

$$\left(-\infty, \mu_0 - k\frac{S}{\sqrt{n}}\right) \bigcup \left(\mu_0 + k\frac{S}{\sqrt{n}}, +\infty\right) \quad \text{si } \sigma \text{ es desconocido}$$

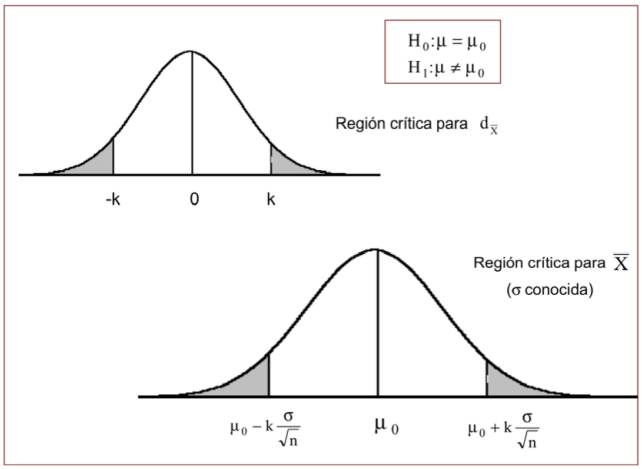
Como consecuencia, si con nuestra información disponible se obtienen una discrepancia $d^*_{\bar{X}}$ y una media muestral $\bar{X}^*$ incluidas en la correspondiente región crítica, la conclusión sería el rechazo de la hipótesis, y viceversa en el caso contrario.

Cuando el contraste se resuelve por el procedimiento del nivel crítico bastará con evaluar el nivel p, esto es, la probabilidad de que, siendo la hipótesis cierta, se presenten discrepancias respecto a ella tanto o más elevadas que las observadas en la muestra:

$$p = P\left(|d_{\bar{X}}| > \left|d^*_{\bar{X}}\right| / H_0\right)$$

Cuando el valor d^* de la discrepancia sea elevado, la probabilidad p resultará baja y en consecuencia rechazaremos la hipótesis $\mu = \mu_0$. Tales situaciones se presentan cuando la media muestral observada

Figura 8.7.: Región crítica para contrastes sobre la media



es muy distinta de la hipotética μ_0, y de ahí que el resultado sea calificado de "significativo para rechazar".

Si por el contrario el valor de la media muestral fuese muy similar al hipotético, se obtendría una discrepancia cercana a 0, a la que se asocia un nivel crítico elevado. En este caso es perfectamente admisible que las desviaciones observadas se deban al azar, y no existen razones fundadas para rechazar la hipótesis.

Consideremos de nuevo el ejemplo inicial donde la hipótesis de producción esperada era $H_0 : \mu = 410$ frente a $H_1 : \mu \neq 410$. Si asumimos que la población es normal con $\sigma = 20$, ¿cuál sería la conclusión adoptada si en una muestra de 16 observaciones se obtiene $\bar{x} = 430$?

Plantearemos en primer lugar la obtención de la región crítica al nivel de significación $\alpha = 0,05$: puede comprobarse fácilmente que el valor crítico de la discrepancia es $k = 1,96$, ya que se cumple $P\left(|d_{\bar{X}}| > 1,96/H_0\right) = 0,05$.

Como consecuencia la región crítica puede ser definida sobre la media muestral:

$$\left(-\infty, 410 - 1,96\frac{20}{\sqrt{16}}\right) \bigcup \left(410 + 1,96\frac{20}{\sqrt{16}}, +\infty\right)$$

es decir

$$(-\infty, 400,2) \cup (419,8, +\infty)$$

con lo cual el valor observado $\bar{x} = 430$ pertenece a la región crítica y conduce al rechazo de la hipótesis.

Mediante el enfoque del nivel crítico llegaríamos a la misma conclusión ya que, asumiendo como cierta la hipótesis nula, se obtendría:

$$P\left(|d_{\bar{X}}| > |d^*_{\bar{X}}| / H_0\right) = P\left(\left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| > \frac{430 - 410}{\frac{20}{\sqrt{16}}}\right) = P\left(\left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| > 4\right) = 0,0000$$

y dicho resultado nos informa de que, asumiendo una producción esperada de 410 miles de Tm/mes, sería inverosímil una muestra como la observada (la probabilidad p es muy baja); por tanto el resultado es claramente "significativo" para rechazar.

[Estudiar cuál sería la conclusión en el caso de que en la muestra se hubiera observado $\bar{x} = 415$]

Enunciado II:

$$\boxed{\begin{array}{c} H_0 : \mu \geq \mu_0 \\ H_1 : \mu < \mu_0 \end{array}}$$

Si planteamos ahora un contraste unilateral del tipo $H_0 : \mu \geq \mu_0$ frente a $H_1 : \mu < \mu_0$ el método clásico nos llevaría, una vez fijado el nivel de significación α, a buscar un valor de k de tal que:

$$P\left(d_{\bar{X}} < k/H_0\right) = \alpha \Rightarrow d_{\bar{X}} \in (-\infty, k)$$

Como muestra la condición anterior, en este caso nos preocuparán únicamente las discrepancias que apoyen la hipótesis alternativa, esto es, los valores del estimador muy inferiores a los hipotéticos que pertenecen a la región crítica $(-\infty, k)$. En efecto, adoptando como representante de H_0 su valor menos favorable (μ_0, que es el límite inferior supuesto para la esperanza), se obtendría la correspondiente región crítica para la media muestral $\bar{X}$ representada por la cola de la izquierda:

$$\left(\mu_0 + k\frac{\sigma}{\sqrt{n}}, +\infty\right) \quad \text{si } \sigma \text{ es conocido}$$

$$\left(\mu_0 + k\frac{S}{\sqrt{n}}, +\infty\right) \quad \text{si } \sigma \text{ es desconocido}$$

Obsérvese que para los niveles de significación habituales (1 %, 5 % o 10 %) el valor k (obtenido en las tablas de la Normal o la t, según los casos) es negativo.

Las consideraciones anteriores son también válidas en el caso de que el contraste se lleve a cabo por el método del nivel crítico. Este evaluará la probabilidad de que, siendo cierta la hipótesis (esto es, siendo la esperanza al menos μ_0), las desviaciones por defecto respecto a ella sean tan notables como las observadas, esto es: $p = P\left(d_{\bar{X}} < d^*_{\bar{X}}/H_0\right)$.

Como puede verse, la expresión de cálculo del nivel crítico depende de la formulación de las hipótesis, pues pretende evaluar la probabilidad de distanciarnos tanto o más que la muestra de la hipótesis nula y, según cuáles sean los enunciados, se traducirá en la probabilidad de dos colas, de la cola a la derecha o de la cola a la izquierda.

Enunciado III:

$$\boxed{\begin{array}{c} H_0 : \mu \leq \mu_0 \\ H_1 : \mu > \mu_0 \end{array}}$$

Este contraste es también unilateral y se plantea en términos simétricos al enunciado anteriormente visto, por lo cual conduce a una región crítica y un nivel crítico asociados a la cola de la derecha [Indicar cómo se resolvería este contraste por los métodos clásico y del nivel crítico]

Enunciado IV:

$$\boxed{\begin{array}{l} H_0 : \mu = \mu_0 \\ H_1 : \mu = \mu_1 \end{array}}$$

Para el contraste de una hipótesis simple frente a una alternativa también simple, nos remitimos a la descripciones anteriores, ya que cuando $\mu_1 < \mu_0$ el planteamiento coincidiría con el visto para el enunciado II y en caso contrario ($\mu_1 > \mu_0$) con el III.

8.4.1.1. Extensión a poblaciones desconocidas

Los contrastes anteriores sobre un valor esperado asumían la normalidad para la población de partida. Además, en el caso de que el tamaño muestral sea suficientemente elevado, el teorema central del límite garantiza un comportamiento aproximadamente normal para la media muestral y para la discrepancia $d_{\bar{X}}$ aun partiendo de poblaciones desconocidas.

Por otra parte, en temas anteriores hemos contemplado la posibilidad de llevar a cabo inferencias sobre μ a partir de poblaciones desconocidas mediante la desigualdad de Chebyshev, planteamiento que resulta también aplicable al contraste de hipótesis, que sería en este caso de tipo no paramétrico.

Consideremos por ejemplo el primero de los enunciados anteriores donde la hipótesis nula es $H_0 : \mu = \mu_0$ frente a la alternativa $H_1 : \mu \neq \mu_0$. Una vez fijado el nivel de significación α, se trata de buscar un valor k tal que:

$$P\left(|d_{\bar{X}}| > k/H_0\right) = \alpha$$

y la desigualdad de Chebyshev aplicada a las discrepancias $d_{\bar{X}}$ garantiza que para cualquier $k > 0$:

$$P\left(|d_{\bar{X}}| \geq k/H_0\right) \leq \frac{1}{k^2}$$

con lo cual, para garantizar que nuestro nivel de significación no excederá el α fijado debemos considerar un valor $k = \frac{1}{\sqrt{\alpha}}$ a partir del cual se obtiene la región crítica para $\bar{X}$:

$$\left(-\infty, \mu_0 - k\frac{\sigma}{\sqrt{n}}\right) \bigcup \left(\mu_0 + k\frac{\sigma}{\sqrt{n}}, +\infty\right)$$

que, en caso de que σ^2 fuese desconocido, podría ser aproximada mediante la correspondiente estimación muestral con S^2.

De modo similar, la desigualdad de Chebyshev permitiría llegar a obtener cotas superiores para el nivel crítico:

$$p = P\left(|d_{\bar{X}}| > \left|d^*_{\bar{X}}\right| / H_0\right) \leq \frac{1}{\left(d^*_{\bar{X}}\right)^2}$$

Puede observarse que en ambos métodos estamos "costeando" la ausencia de infor-

mación poblacional ya que, para rechazar una hipótesis al mismo nivel de significación, la evidencia muestral debe ser ahora más fuerte. Este hecho se debe a que con distribución desconocida serán mayores los errores debidos al azar, con lo cual admitiremos mayores discrepancias, llegando por tanto a regiones críticas menores. Alternativamente, si optamos por el método del nivel crítico solamente podríamos llegar a una cota superior para la probabilidad p.

A modo de ilustración, analicemos cómo afectaría la no normalidad al contraste anteriormente desarrollado $H_0 : \mu = 410$ frente a la alternativa $H_1 : \mu \neq 410$, donde seguimos asumiendo $\sigma = 20$ y una observación muestral $\bar{x} = 430$ obtenida a partir de 16 observaciones.

Tal y como hemos visto, la región crítica a un nivel de significación no superior a $\alpha = 0,05$ vendría dada ahora para $\bar{X}$ por:

$$\left(-\infty, 410 - 4,4721\frac{20}{\sqrt{16}}\right) \bigcup \left(410 + 4,4721\frac{20}{\sqrt{16}}, +\infty\right)$$

es decir , $(-\infty, 387,64) \cup (432,36, +\infty)$ donde $k = 4,4721$ ha sido obtenida como $\frac{1}{\sqrt{0,05}}$ y, conscientes de no poder garantizar ninguna distribución para $d_{\bar{X}}$, hemos asumido como válida una mayor discrepancia debida al azar de modo que el valor muestral $\bar{x} = 430$ no pertenece a la región crítica y por tanto no conduce al rechazo de la hipótesis.

Siguiendo el enfoque del nivel crítico llegaríamos a la misma conclusión ya que, asumiendo como cierta la hipótesis nula, se obtendría:

$$P\left(|d_{\bar{X}}| > |d^*_{\bar{X}}| / H_0\right) = P\left(\left|\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}\right| > 4\right) \leq \frac{1}{4^2} = 0,0625$$

es decir, podemos garantizar que el nivel crítico no supera el 6,25 % sin que este resultado pueda ser calificado de "significativo" para rechazar a los niveles habituales (5 % y 1 %).

A la vista de los resultados anteriores podemos concluir que un valor medio de 430 en la muestra es suficientemente significativo para rechazar la hipótesis $\mu = 410$ cuando la población es normal pero no así en el caso de que estemos trabajando con una población desconocida.

8.4.2. Contrastes sobre la varianza

Asumiendo el supuesto de normalidad para la población X y siguiendo la misma metodología vista para μ, planteemos algunos contrastes habituales para la varianza:

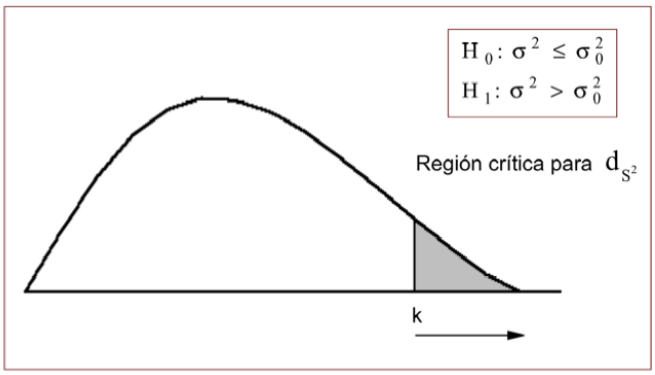
Enunciados unilaterales: Supongamos que deseamos llevar a cabo un contraste unilateral sobre la varianza en los siguientes términos:

$$\boxed{\begin{array}{l} H_0 : \sigma^2 \leq \sigma_0^2 \\ H_1 : \sigma^2 > \sigma_0^2 \end{array}}$$

Si seguimos el método clásico, una vez fijado el nivel de significación α deberíamos buscar el valor k que garantice: $P\left(d_{S^2} > k / H_0\right) = \alpha$. Teniendo en cuenta que bajo H_0 la discrepancia es en este caso

$$d_{S^2/H_0} = \frac{(n-1)S^2}{\sigma_0^2} \approx \chi^2_{n-1}$$

Figura 8.8.: Región crítica para contrastes sobre la varianza



dicho valor se determinará en las tablas chi-cuadrado con n-1 g.l.

Se llega así a las siguientes regiones críticas para la discrepancia y para la varianza muestral:

- RC para d_{S^2}: $(k, +\infty)$
- RC para S^2: $\left(\frac{k\sigma_0^2}{n-1}, +\infty\right)$

que, como consecuencia del enunciado de la hipótesis, se corresponden con las colas de la derecha (figura 8.8)

De modo similar, si llevamos a cabo el contraste por el método del nivel crítico, bastaría con calcular la probabilidad asociada a la cola derecha del valor observado d^*: $p = P\left(d_{S^2} > d^*_{S^2}/H_0\right)$ ya que en este caso nos preocupan únicamente las discrepancias por exceso, que son las que apoyan la alternativa.

Este razonamiento resulta aplicable a otros contrastes unilaterales, tanto de hipótesis simples como compuestas.

[¿Cómo se resolvería el contraste de la hipótesis $\sigma^2 = 36$ frente a la alternativa $\sigma^2 = 60$?]

Enunciado bilateral:

$$\boxed{\begin{array}{l} H_0 : \sigma^2 = \sigma_0^2 \\ H_1 : \sigma^2 \neq \sigma_0^2 \end{array}}$$

La resolución por el método clásico se lleva a cabo buscando en las tablas de la distribución chi-cuadrado con $n-1$ g.l dos valores k_1 y k_2 de la discrepancia tales que:

$$P\left(d_{S^2} < k_1/H_0\right) = P\left(d_{S^2} > k_2/H_0\right) = \frac{\alpha}{2}$$

La región crítica será en este caso bilateral:

- RC para d_{S^2}: $(0, k_1) \cup (k_2, +\infty)$

- RC para S^2: $\left(0, \frac{k_1\sigma_0^2}{n-1}\right) \bigcup \left(\frac{k_2\sigma_0^2}{n-1}, +\infty\right)$

En el tema de estimación hemos visto que la construcción de intervalos de confianza para σ^2 se llevaba a cabo multiplicando su estimador S^2 por los índices

$$\frac{n-1}{k_1} > 1 \text{ y } \frac{n-1}{k_2} < 1$$

Ahora la determinación de la región crítica para S^2 se realiza incorporando al valor hipotético σ_0^2 los inversos de ambas expresiones

$$\frac{k_1}{n-1} < 1 \text{ y } \frac{k_2}{n-1} > 1$$

¿Cómo se resolvería este contraste por el método del nivel crítico? Por tratarse de un contraste bilateral debemos tener presentes las discrepancias en los dos sentidos, con lo cual se obtienen niveles críticos dados por:

- $P\left(d_{S^2} < d^*_{S^2}/H_0\right)$ si la menor probabilidad se encuentra en la cola de la izquierda
- $P\left(d_{S^2} > d^*_{S^2}/H_0\right)$ si la menor probabilidad se encuentra en la cola derecha

Obsérvese que en realidad este planteamiento coincide con el de los contrastes bilaterales para μ, si bien en aquel caso el cálculo resultaba más sencillo al tratarse de una distribución simétrica, que permitía la utilización de discrepancias en valor absoluto.

8.4.3. Contrastes sobre la proporción

En el ámbito económico son habituales los supuestos sobre proporciones poblacionales (tasas de actividad, participaciones sectoriales,...) en los que interesará conocer si son bilaterales o unilaterales. A modo de ilustración nos centraremos en un contraste bilateral, pero la descripción resulta fácilmente trasladable a los contrastes de una sola cola.

Enunciado bilateral:

$$\boxed{\begin{array}{c} H_0 : p = p_0 \\ H_1 : p \neq p_0 \end{array}}$$

El desarrollo de este contraste será distinto según el tamaño de la muestra en la que nos basemos. Comenzando por las muestras de tamaño elevado, el planteamiento sería muy similar al visto para la esperanza poblacional μ, ya que en este caso la discrepancia es

$$d_{\hat{p}/H_0} = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \approx \mathcal{N}(0,1)$$

y en consecuencia bastaría con buscar un valor k tal que $P\left(|d_{\hat{p}}| > k/H_0\right) = \alpha$ donde α es el nivel de significación.

Una vez determinada la constante k, las regiones críticas para la discrepancia y para el estimador se obtienen como sigue:

- RC para $d_{\hat{p}}$: $(-\infty, -k) \cup (k, +\infty)$
- RC para $\hat{p}$: $\left(-\infty, p_0 - k\sqrt{\frac{p_0(1-p_0)}{n}}\right) \bigcup \left(p_0 + k\sqrt{\frac{p_0(1-p_0)}{n}}, +\infty\right)$

Obsérvese que en este caso no es necesario estimar la varianza de la proporción muestral, dado que ésta quedará completamente determinada bajo la hipótesis nula $p = p_0$.

Del mismo modo, si optásemos por el método del nivel crítico, éste se calcularía mediante la expresión:

$$p = P\left(|d_{\hat{p}}| > \left|d^*_{\hat{p}}\right| / H_0\right)$$

cuyo resultado permite decidir si se debe o no rechazar H_0.

¿Qué sucedería si el tamaño de muestra n resultase insuficiente para aplicar los teoremas límites? En estas situaciones, los contrastes deben ser planteados aprovechando únicamente el hecho de que, bajo la hipótesis nula, el numerador de la proporción muestral es $X \approx \mathcal{B}(n, p_0)$.

Así, se buscan dos valores x_1 y x_2 tales que:

$$P\left(X < x_1 / H_0\right) = P\left(X > x_2 / H_0\right) = \frac{\alpha}{2}$$

Dichos valores determinan directamente la región crítica para $\hat{p}$:

$$\left(0, \frac{x_1}{n}\right) \bigcup \left(\frac{x_2}{n}, 1\right)$$

8.4.4. Contrastes sobre medias de dos poblaciones

Cuando investigamos conjuntamente dos poblaciones, a menudo resultará interesante comparar sus valores esperados. Se trata en estos casos de contrastar hipótesis relativas a la diferencia de medias, para lo cual -como hemos analizado en el capítulo 6- deben contemplarse diferentes supuestos.

Así, una primera posibilidad sería aquella en la que las muestras $(X_1, \ldots, X_n)$ y $(Y_1, \ldots, Y_n)$ aparecen pareadas, con lo cual podemos definir una nueva variable $D = X - Y$ con esperanza $\mu_D = \mu_X - \mu_Y$ y trabajar sobre la muestra $(D_1, \ldots, D_n)$.

A partir de aquí, el objetivo consistiría en contrastar hipótesis para la esperanza de una población (μ_D), por lo cual resultan aplicables todas las consideraciones vistas para dicho parámetro.

Sin embargo, en la práctica más habitual, los contrastes sobre diferencia de medias asumen como válidos dos supuestos básicos: la normalidad de las poblaciones X e Y y la independencia entre las muestras aleatorias $(X_1, \ldots, X_n)$ y $(Y_1, \ldots, Y_m)$ extraídas de las mismas.

Tabla 8.2.: Cuadro resumen de los contrastes de la diferencia de medias

HIPOTESIS	REGION CRITICA PARA $\bar{X}-\bar{Y}$	NIVEL CRITICO
$H_0:\mu_X-\mu_Y=0$ $H_1:\mu_X-\mu_Y\neq 0$	$\left(-\infty,-k\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)\bigcup\left(k\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},+\infty\right)$ con $P\left(\lvert d_{\bar{X}-\bar{Y}}\rvert>k/H_0\right)=\alpha$	$p=P\left(\lvert d_{\bar{X}-\bar{Y}}\rvert>\lvert d^*_{\bar{X}-\bar{Y}}\rvert/H_0\right)$
$H_0:\mu_X-\mu_Y\geq 0$ $H_1:\mu_X-\mu_Y<0$	$\left(-\infty,k\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}}\right)$ con $P\left(d_{\bar{X}-\bar{Y}}<k/H_0\right)=\alpha$	$p=P\left(d_{\bar{X}-\bar{Y}}<d^*_{\bar{X}-\bar{Y}}/H_0\right)$
$H_0:\mu_X-\mu_Y\leq 0$ $H_1:\mu_X-\mu_Y>0$	$\left(k\sqrt{\frac{\sigma_X^2}{n}+\frac{\sigma_Y^2}{m}},+\infty\right)$ con $P\left(d_{\bar{X}-\bar{Y}}>k/H_0\right)=\alpha$	$p=P\left(d_{\bar{X}-\bar{Y}}>d^*_{\bar{X}-\bar{Y}}/H_0\right)$

La resolución de estos contrastes seguirá procedimientos similares a los vistos para una sola media, con sólo tener presente si el enunciado es de una o dos colas. A modo de resumen, la tabla 8.2 recoge el esquema de estos contrastes por los métodos clásico y del nivel crítico para la situación más sencilla, esto es, con varianzas poblacionales conocidas:

En la construcción de la región crítica hemos sustituido la diferencia de medias por el valor 0 correspondiente a la hipótesis nula. Por lo que se refiere a la determinación de la constante k, ésta se obtendría con las tablas del modelo $\mathcal{N}(0,1)$ dado que hemos asumido que las varianzas son conocidas.

En otro caso, ya hemos visto que sería posible -bajo condiciones de proporcionalidad entre las varianzas poblacionales- utilizar las varianzas muestrales para la obtención de discrepancias que conduzcan a las correspondientes regiones críticas.

Así, en el caso de varianzas desconocidas pero coincidentes, la región crítica asociada al contraste bilateral sería del tipo:

$$\left(-\infty,-k\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)}\right)\bigcup\left(k\sqrt{\frac{(n-1)S_X^2+(m-1)S_Y^2}{n+m-2}\left(\frac{1}{n}+\frac{1}{m}\right)},+\infty\right)$$

donde k se obtendría en tablas de la t de Student con $n+m-2$ g.l.

El planteamiento anterior podría ser generalizado a los contrastes relativos a la suma o la combinación lineal de esperanzas. En términos generales, podríamos plantear contrastes del tipo:

$$\boxed{\begin{array}{l}H_0:\alpha\mu_X+\beta\mu_Y=c\\H_1:\alpha\mu_X+\beta\mu_Y\neq c\end{array}}$$

con α, β y c constantes, cuya resolución se llevaría a cabo de modo similar a los contrastes de diferencias de medias.

8.4.5. Contrastes sobre varianzas de dos poblaciones

A menudo estamos interesados en contrastar algún supuesto relativo a la dispersión de dos poblaciones. De hecho, en el apartado anterior hemos visto que la relación entre las varianzas poblacionales resulta relevante a la hora de conocer la distribución de la diferencia de medias.

En este tipo de situaciones, podemos plantear el contraste:

$$\boxed{\begin{array}{l} H_0 : \sigma_X^2 = \sigma_Y^2 \\ H_1 : \sigma_X^2 \neq \sigma_Y^2 \end{array}}$$

Consideremos dos poblaciones normales $X \approx \mathcal{N}(\mu_X, \sigma_X)$ e $Y \approx \mathcal{N}(\mu_Y, \sigma_Y)$ de las que se han extraído independientemente las muestras aleatorias $(X_1, \ldots, X_n)$, $(Y_1, \ldots, Y_m)$. En esta situación el contraste del supuesto de igualdad de varianzas puede llevarse a cabo a partir de la discrepancia tipificada que, bajo la hipótesis nula, adopta la expresión

$$d_{\frac{S_X^2}{S_Y^2}/H_0} = \frac{S_X^2}{S_Y^2} \approx F_{m-1}^{n-1}$$

[¿Por qué?]

Dado que el contraste planteado es bilateral, la región crítica vendría determinada por los valores inferiores a k_1 o superiores a k_2. De modo análogo, si se sigue el método del nivel crítico la probabilidad correspondiente al nivel crítico sería el área encerrada en las dos colas de la distribución F de Snedecor.

Obsérvese que en este contraste la discrepancia coincide con la razón de varianzas muestrales, por lo cual la región crítica es en ambos casos $(0, k_1) \cup (k_2, +\infty)$.

El planteamiento expuesto resulta también aplicable cuando nuestra información *a priori* nos lleva a establecer que una población es más homogénea que otra ($\sigma_X^2 \geq \sigma_Y^2$ o bien $\sigma_X^2 \leq \sigma_Y^2$). A modo de resumen, recogemos las diferentes situaciones en la tabla siguiente:

HIPÓTESIS	REGIÓN CRÍTICA PARA $\frac{S_X^2}{S_Y^2}$	NIVEL CRÍTICO
$H_0 : \sigma_X^2 = \sigma_Y^2$ $H_0 : \sigma_X^2 \neq \sigma_Y^2$	$(0, k_1) \cup (k_2, +\infty)$ con $P\left(\frac{S_X^2}{S_Y^2} > k_2/H_0\right) = P\left(\frac{S_X^2}{S_Y^2} < k_1/H_0\right) = \frac{\alpha}{2}$	$p = 2P\,(d > d^*/H_0)$ o $p = 2P\,(d < d^*/H_0)$
$H_0 : \sigma_X^2 \geq \sigma_Y^2$ $H_0 : \sigma_X^2 < \sigma_Y^2$	$(0, k)$ con $P\left(\frac{S_X^2}{S_Y^2} < k/H_0\right) = \alpha$	$p = P\,(d < d^*/H_0)$
$H_0 : \sigma_X^2 \leq \sigma_Y^2$ $H_0 : \sigma_X^2 > \sigma_Y^2$	$(k, +\infty)$ con $P\left(\frac{S_X^2}{S_Y^2} > k/H_0\right) = \alpha$	$p = P\,(d > d^*/H_0)$

Cabe por último señalar que este planteamiento puede ampliarse a cualquier contraste de proporcionalidad de varianzas, en los que la hipótesis nula sería:

$$H_0 : \sigma_X^2 = c\sigma_Y^2 \ , \ H_0 : \sigma_X^2 \geq c\sigma_Y^2 \ o \ H_0 : \sigma_X^2 \leq c\sigma_Y^2$$

8.5. Algunos contrastes no paramétricos

En un apartado previo hemos planteado el contraste de los supuestos básicos, que se incluyen en el ámbito de los tests no paramétricos. También en el apartado anterior, al estudiar los contrastes sobre ciertos parámetros, hemos considerado algunas situaciones particulares con modelo probabilístico desconocido, que en general se resuelven mediante la desigualdad de Chebyshev.

En la práctica pueden interesarnos muy diversos tests de tipo no paramétrico referentes a poblaciones independientes, datos homogéneos, medidas de posición, etc.

Los contrastes no paramétricos constituyen un conjunto muy extenso y útil de herramientas estadísticas que, sin suponer información básica sobre la distribución de partida, permiten obtener pruebas con cierta eficiencia para contrastes muy variados.

Uno de los primeros problemas que nos encontramos es el de clasificar los diferentes tests no paramétricos para sistematizar su estudio y aplicación. Podemos clasificar estas pruebas en función de la información empírica, según que se refiera a una muestra, a dos muestras (relacionadas en distintos períodos de tiempo o independientes), o a k muestras.

También podemos agrupar los contrastes no paramétricos atendiendo al tipo de test utilizado; así tendríamos pruebas basadas en rachas, en rangos, en estadísticos de orden, en distancias chi-cuadrado, ...

Otra forma de clasificar estos contrastes sería en función de las hipótesis: tests de localización, de independencia de poblaciones, de homogeneidad, ... Este es el esquema que vamos a seguir en este epígrafe en el que no pretendemos realizar un desarrollo exhaustivo sino que describimos algunas de las principales técnicas no paramétricas.

Dos aspectos que debemos tener en cuenta para una mejor comprensión de todo lo que sigue son los siguientes:

- Ligeras modificaciones en el enfoque de un mismo test permiten que éste pueda ser aplicado a contrastes, en principio, diferentes. Por este motivo no debe extrañarnos que para diferentes hipótesis poblacionales aparezca la misma prueba.
- Prácticamente en todos los casos tendremos que ir distinguiendo si se trata de muestras pequeñas o si pueden aplicarse condiciones de convergencia.

8.5.1. Contrastes del modelo poblacional

Este tipo de contrastes se conocen como de *bondad de ajuste* y fueron descritos en un apartado anterior. A modo de síntesis, recogemos en la tabla 8.3 sus rasgos más

Tabla 8.3.: Test de bondad de ajuste

Test	Criterio	Discrepancia	Distribución y condiciones	Nivel crítico
Chi-cuadrado	Compara frecuencias teóricas y observadas en los k intervalos	$d_n = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$	χ^2_n (Aprox.)	$P\left(d_n > d_n^*/H_0\right)$
Kolmogorov-Smirnov	Compara frecuencias acumuladas muestrales y teóricas	$D_n^* = \sup_x \lvert S_n(x) - F_0(x)\rvert$	Tabulada (v.a. continuas, parámetros dados)	$P(D_n > D_n^*/H_0)$
Jarque-Bera	Compara características de forma con las del modelo normal	$d_{JB} = \frac{n}{6}\left(g_1^2 + \frac{1}{4}g_2^2\right)$ g_1, g_2: Medidas de asimetría y kurtosis	χ^2_2 (Aprox.)	$P(d_{JB} > d_{JB}^*/H_0)$

destacables:

Como vemos, estos contrastes presentan diferencias en cuanto a la discrepancia considerada y a las condiciones de aplicación.

El test chi-cuadrado introduce una arbitrariedad inicial al tener que agrupar los datos muestrales en intervalos. La clasificación que se haga depende del criterio de la persona que realice el contraste y el valor de la discrepancia es distinto según la agrupación realizada.

Por otra parte, tanto el test chi-cuadrado como el de Jarque-Bera conducen sólo a una distribución aproximada de la discrepancia, mientras el test de K-S proporciona una distribución exacta.

Tanto el test chi-cuadrado como el propuesto por Jarque y Bera para la normalidad admiten una mayor holgura que el de K-S. En este sentido siempre que sea posible aplicar el test de K-S (distribuciones continuas y con parámetros conocidos), este contraste resultará preferible.

8.5.2. Contrastes de independencia de dos poblaciones

Consideremos ahora un colectivo sobre el cual se observa una v.a. bidimensional (X, Y); supongamos que las variables unidimensionales son susceptibles de clasificación en r y s categorías $A_1, \ldots, A_r$ para X y $B_1, \ldots, B_s$ para Y, respectivamente. Se desea contrastar si las dos poblaciones X e Y son independientes para lo cual debemos seleccionar una muestra aleatoria de tamaño n de esa v.a. bidimensional: (X_1, Y_1), ...,(X_n, Y_n).

En ocasiones, las categorías de clasificación se corresponden con intervalos: $L_0 - L_1$, $L_1 - L_2$, ..., $L_{r-1} - L_r$ y $M_0 - M_1$, $M_1 - M_2$, ..., $M_{s-1} - M_s$ que cubren el recorrido de las variables X e Y respectivamente.

Podemos resumir la información de los valores muestrales en la siguiente tabla de doble entrada:

Y/X	A_1	$\cdots$	A_r	$n_{\cdot j}$
B_1	n_{11}	$\cdots$	n_{r1}	$n_{\cdot 1}$
$\vdots$	$\ddots$	$\ddots$	$\ddots$	$\vdots$
B_s	n_{1s}	$\cdots$	n_{rs}	$n_{\cdot s}$
$n_{i\cdot}$	$n_{1\cdot}$	$\cdots$	$n_{r\cdot}$	$n_{\cdot\cdot}$

Dado que estamos trabajando con una muestra genérica, las frecuencias absolutas n_{ij} serán v.a. La notación es la usual en Estadística Descriptiva donde

$$n_{i\cdot} = \sum_{j=1}^{s} n_{ij} \ , \ n_{\cdot j} = \sum_{i=1}^{r} n_{ij} \ , \ n_{\cdot\cdot} = \sum_{i=1}^{r} n_{i\cdot} = \sum_{j=1}^{s} n_{\cdot j}$$

Las frecuencias marginales $n_{i\cdot}$ y $n_{\cdot j}$ son también magnitudes aleatorias y se cumple $n = n_{\cdot\cdot}$.

La hipótesis de independencia entre las variables X e Y puede ser expresada como

$$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j} \ , \ \forall i = 1, \ldots, r \ , \ \forall j = 1, \ldots, s$$

la hipótesis alternativa será que para algún par (i, j) no se verifique la igualdad anterior. El contraste planteado puede ser equivalente a uno de bondad en el que ajustamos cada elemento de la tabla bidimensional al comportamiento teórico de independencia.

La probabilidad conjunta o frecuencia observada puede calcularse como $p_{ij} = \frac{n_{ij}}{n}$ y las probabilidades marginales $p_{i\cdot}$ y $p_{\cdot j}$ (correspondientes a los valores poblacionales $p_{i\cdot} = P(X \in A_i)$, $p_{\cdot j} = P(Y \in B_j)$) serán desconocidas por lo que debemos proceder a su estimación a partir de los datos muestrales. Los estimadores MV de estas probabilidades son:

$$\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n} \text{ y } \hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n} \ , \ \forall i = 1, \ldots, r \, \forall j = 1, \ldots s$$

Así pues, sustituyendo podemos expresar la hipótesis nula como:

$$H_0 : n_{ij} = \frac{n_{i\cdot}n_{\cdot j}}{n_{\cdot\cdot}} \ , \ \forall i = 1, \ldots, r \, \forall j = 1, \ldots s$$

La discrepancia tipificada de este contraste (que ya ha sido introducida en el capítulo 6), será:

$$d_{IND} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

que, en el supuesto de independencia, converge a una χ^2 con $(r-1)(s-1)$ g.l.

Para determinar los g.l. debemos tener en cuenta que en principio tenemos rs frecuencias observadas y partimos de una restricción inicial:

$$n = \sum_{i=1}^{r}\sum_{j=1}^{s} n_{ij}$$

con lo cual el número de g.l. serían $rs - 1 -$(nº de parámetros que estimamos).

El número de parámetros estimados para X sería $r - 1$ ($\hat{p}_1, \ldots, \hat{p}_{r-1}$, ya que $\hat{p}_r$ se obtiene como $\hat{p}_r = 1 - \sum_{i=1}^{r-1} \hat{p}_i$). De la misma forma para Y el número de parámetros estimados serán $s - 1$. Por tanto los grados de libertad serán:

$$rs - 1 - (r-1) - (s-1) = rs - r - s + 1 = (r-1)(s-1)$$

A partir de la información proporcionada por una muestra concreta, la decisión sobre la hipótesis de independencia se tomará a partir del valor de la discrepancia observada:

$$d^*_{IND} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_{i\cdot}n_{\cdot j}}{n}\right)^2}{\frac{n_{i\cdot}n_{\cdot j}}{n}}$$

Esta discrepancia nos conduce a un nivel crítico $p = P\left(d_{IND} > d^*_{IND}/H_0\right)$, en función del cual rechazaremos o no la hipótesis formulada.

Para que podamos obtener una buena aproximación de esta discrepancia al modelo χ^2 se exige que la frecuencia esperada en cada casilla sea no inferior a 5. Por tanto, si para una clasificación determinada no se verificase este supuesto, deberíamos proceder a una reclasificación de las categorías hasta conseguir el objetivo anterior.

8.5.3. Contrastes de homogeneidad de poblaciones clasificadas según varias categorías

Presentamos aquí un caso particular de los contrastes de homogeneidad entre poblaciones, cuando se tienen variables que se pueden clasificar en categorías. Analizamos en primer lugar la prueba exacta de Fisher para el caso en que se consideren sólo dos categorías y dos muestras; posteriormente extenderemos este contraste a un caso arbitrario con r muestras y s clasificaciones.

8.5.3.1. Prueba exacta de Fisher

Este test es aplicable cuando disponemos de dos poblaciones que pueden ser clasificadas con arreglo a dos categorías excluyentes. Esto es, puede referirse a atributos que presentan sólo dos modalidades, o bien a variables cuantitativas en las que definimos las categorías "menor que b" y "mayor o igual a b".

Contrastar la identidad de las poblaciones respecto a esta clasificación dicotómica será equivalente a contrastar la identidad entre las proporciones que cada muestra presenta en las dos categorías.

Para realizar el contraste tomamos una m.a.s. de cada población y clasificamos ambas muestras con arreglo a las categorías (A, B). De esta forma elaboramos una tabla 2×2:

	Categoría A	Categoría B	Sumas
Muestra 1	n_{1A}	n_{1B}	$n_1 = n_{1A} + n_{1B}$
Muestra 2	n_{2A}	n_{2B}	$n_2 = n_{2A} + n_{2B}$
Sumas	$n_A = n_{1A} + n_{2A}$	$n_B = n_{1B} + n_{2B}$	$n_m = n_1 + n_2 = n_A + n_B$

Si encontramos diferencias significativas entre $\frac{n_{1A}}{n_1}$ y $\frac{n_{2A}}{n_2}$ entonces podríamos garantizar que las poblaciones de partida no coinciden.

La probabilidad exacta de obtener esta distribución de frecuencias al clasificar un total de n unidades se obtiene a partir del modelo hipergeométrico. Para obtener este resultado limitemos el total de unidades al conjunto de las dos muestras, identifiquemos la muestra o unidades observadas con la muestra 1 y el número de unidades poblacionales favorables (por ejemplo a la categoría A) n_A. En este caso tendríamos una distribución hipergeométrica $\mathcal{H}(N = n_m, M = n_A, n = n_1)$.

El número de unidades favorables sobre la muestra sería $x = n_{1A}$; por tanto la probabilidad exacta vendría dada por:

$$p = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} = \frac{\binom{n_A}{n_{1A}}\binom{n_B}{n_{1B}}}{\binom{n_m}{n_1}}$$

Ahora bien, para obtener el nivel crítico asociado a este test tendremos que calcular la probabilidad de obtener esta distribución de frecuencias o cualquier otra más extrema que ella, esto es, cualquier otra distribución obtenida a partir de la actual que presente mayor disparidad entre las proporciones de las categorías.

Obsérvese que esta expresión podría ser también obtenida razonando de forma simétrica, esto es, considerando como casos posibles todas las combinaciones de n_A elementos a partir del total n_m y como casos favorables las combinaciones de tamaño n_{1A} a partir de la muestra n_1 y de tamaño n_{2A} a partir de la muestra n_2.

Se tendría entonces la expresión:

$$p = \frac{\binom{n_1}{n_{1A}}\binom{n_2}{n_{2A}}}{\binom{n_m}{n_A}}$$

que resulta coincidente con la anteriormente vista

$$p = \frac{\binom{n_A}{n_{1A}}\binom{n_B}{n_{1B}}}{\binom{n_m}{n_1}}$$

[Compruébese]

Consideremos por ejemplo la siguiente tabla de frecuencias:

	A	B	Sumas
Muestra 1	6	2	8
Muestra 2	1	2	3
Sumas	7	4	11

La hipótesis nula es que las poblaciones no difieren en cuanto a la clasificación en las categorías anteriores o en otras palabras, que la proporción de unidades clasificadas en la categoría A es la misma en las dos poblaciones.

La probabilidad asociada a esta tabla será:

$$p = \frac{\binom{7}{6}\binom{4}{2}}{\binom{11}{8}} = 0,25454$$

Para obtener tablas más extremas que la actual, debemos tener en cuenta que el total de unidades ($N = 11$) debe mantenerse, al igual que el total de unidades a favor ($M = 7$) y el de unidades observadas ($n = 8$); es decir, debemos considerar las combinaciones que se pueden hacer con esta tabla que conduzcan a una distribución con más desequilibrio, pero manteniendo los totales marginales de la tabla anterior. En esta situación, una tabla más extrema que la inicial sería:

	A	B	Sumas
Muestra 1	7	1	8
Muestra 2	0	3	3
Sumas	7	4	11

Y su probabilidad correspondiente viene dada por:

$$p = \frac{\binom{7}{4}\binom{4}{1}}{\binom{11}{8}} = 0,1212$$

Al obtener un 0 en una de las frecuencias de la tabla, ya hemos llegado a la situación extrema. No podemos obtener tablas con mayores desequilibrios que ésta, ya que cualquier transformación que hagamos manteniendo los totales marginales sería una vuelta a la situación anterior.

Por tanto el nivel crítico asociado a esta distribución de frecuencias sería:

$$p = 0,25454 + 0,1212 = 0,37574$$

si el contraste se considera de una sola cola o bien el doble de esta probabilidad si se plantease como bilateral.

Parece claro que cuando la frecuencia más pequeña de la tabla es elevada, el número de tablas transformadas será alto y por tanto el proceso se complica. Sin embargo, algunos programas informáticos nos resuelven este problema mediante sencillos algoritmos.

La regla de decisión del *contraste de Fisher* se establece en función del nivel crítico (probabilidad exacta).

Cuando las frecuencias son elevadas debemos tener en cuenta que la distribución hipergeométrica se aproxima por otros tipos de modelos, por lo cual sería preferible utilizar tests alternativos.

Una variante de este test es la prueba de la mediana. En este caso las categorías A y B representan los sucesos $(-\infty, Me)$ y $[Me, +\infty)$ respectivamente, y la prueba contrasta si las poblaciones son homogéneas o no en cuanto a su tendencia central.

8.5.3.2. Contraste χ^2 de homogeneidad entre poblaciones

Este test permite una generalización, en cuanto a objetivos, de la prueba de Fisher. Podemos contrastar la homogeneidad de dos poblaciones clasificadas en dos catego-

rías o, de forma más general, r poblaciones, $(X_1, \ldots, X_r)$, clasificadas en s categorías $A_1, \ldots, A_s$ mutuamente excluyentes. En definitiva, se trata de contrastar si el comportamiento de las componentes X_i es homogéneo sobre esas categorías.

Estas categorías suelen corresponderse con intervalos disjuntos $L_0 - L_1$, $L_1 - L_2$, ..., $L_{s-1} - L_s$ que cubren el recorrido de las variables estudiadas.

Si denotamos por p_{ij} la probabilidad de que la v.a. X_i tome valores en la categoría A_j $(p_{ij} = P(X_i \in A_j))$, la hipótesis de *homogeneidad* se expresaría:

$$H_0 : p_{i1} = p_{h1}, \ldots, p_{is} = p_{hs}\ ,\ \forall i, h = 1, \ldots, r$$

es decir, la probabilidad de cada categoría es la misma en todas las variables y por tanto tiene que coincidir con la marginal $(p_j = P(A_j) = p_{ij}, \forall i = 1, \ldots, r)$; así pues, para toda categoría j, debe verificarse: $H_0 : p_{ij} = p_j\ ,\ \forall i = 1, \ldots, r$.

Para contrastar esta hipótesis se toma una m.a.s. de cada una de las variables $X_1, \ldots, X_r$ con tamaños respectivos $n_1, \ldots n_r$. A partir de estas muestras se pretende contrastar la hipótesis anterior. Una vez clasificadas las muestras con arreglo a las s categorías anteriores, podemos resumir su información, en términos de frecuencias, en la siguiente tabla:

A_j/X_i	Muestra X_1	...	Muestra X_r	$n_{.j}$
A_1	n_{11}	$\cdots$	n_{r1}	$n_{.1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
A_S	n_{1s}	$\cdots$	n_{rs}	$n_{.s}$
Tam.muestra n_i	n_1	$\cdots$	n_r	n

con las notaciones usuales empleadas en el test χ^2 de independencia.

El supuesto de homogeneidad $H_0 : p_{ij} = p_j$ puede ser también enunciado como: $H_0 : n_{ij} = n_i p_j$, para cada categoría $j = 1, \ldots, s$ y cada muestra $i = 1, \ldots, r$. Se trata de nuevo de un test de bondad de ajuste donde en cada una de las $r \times s$ casillas dispondremos de una frecuencia que puede ser observada n_{ij} y una frecuencia esperada $n_i p_j$,

La medida de discrepancia χ^2 viene definida por:

$$d_{HOMOG} = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(n_{ij} - n_i p_j)^2}{n_i p_j}$$

Si la probabilidad de cada categoría es conocida, bajo la hipótesis nula, esta discrepancia puede aproximarse por una distribución χ^2 con $(r-1)s$ g.l.

Observemos que aquí el número de g.l. no es $rs - 1$, puesto que esta cantidad tiene algunas restricciones más. Ello se debe a que los tamaños de las r muestras, n_i, vienen dados y también su suma,

n.

Como en general esas probabilidades resultarán desconocidas y será necesario estimarlas a partir de la información muestral, debemos recurrir a los EMV: $\hat{p}_j = \frac{n_{\cdot j}}{n}$ y entonces la discrepancia tipificada asociada al test sería:

$$d_{HOMOG} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_i n_{\cdot j}}{n}\right)^2}{\frac{n_i n_{\cdot j}}{n}}$$

que se distribuirá asintóticamente según un modelo χ^2 con g.l. $(r-1)(s-1)$. [Justifíquese el número de estos g.l.].

A partir de la información proporcionada por muestras concretas, podemos calcular la discrepancia observada:

$$d^*_{HOMOG} = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(n_{ij} - \frac{n_i n_{\cdot j}}{n}\right)^2}{\frac{n_i n_{\cdot j}}{n}}$$

que nos conduce a un nivel crítico $p = P(d > d^*/H_0)$, en función del cual rechazaremos o no la hipótesis formulada.

Como se señaló en el contraste de independencia, si para alguna casilla la frecuencia esperada fuese inferior a 5 debemos reagrupar categorías hasta conseguir que todas las frecuencias esperadas rebasen este límite.

8.5.4. Contrastes de identidad de la población a partir de muestras independientes

En este epígrafe presentamos en primer lugar pruebas para contrastar la identidad de la población de la cual proceden dos m.a.s. independientes (Mann-Whitney, Wald-Wolfowitz y Kolmogorov-Smirnov) y posteriormente describimos tests para r muestras (Kruskal-Wallis).

Supongamos que disponemos de dos muestras particulares $(x_1, \ldots, x_n)$ e $(y_1, \ldots, y_m)$, y estamos interesados en contrastar, a partir de esos datos, que las muestras provienen de la misma población, o lo que es equivalente, si denotamos por X e Y las poblaciones originales y por $F(x)$ y $G(y)$ sus respectivas f.d., se tiene: $H_0 : F(x) = G(y)$.

Los supuestos básicos de estos contrastes son que las poblaciones de partida X e Y son continuas y que las muestras son independientes.

8.5.4.1. Test de Mann-Whitney (M-W)

El *test de Mann-Whitney* permite contrastar si dos muestras independientes proceden de la misma población. El procedimiento consiste en agrupar las dos muestras y estudiar el rango de cada observación en el total. Si las poblaciones de origen fuesen distintas entonces se esperarían diferencias en la ordenación, de forma que los valores

de una de las muestras serían sistemáticamente más elevados que los de la otra, y por tanto se situarían en las colas de la serie ordenada por rangos.

El estadístico de contraste utilizado es:

$$U_X = nm + \frac{n(n+1)}{2} - R_X$$

donde R_X denota la suma de rangos de la muestra $(x_1, \ldots, x_n)$. Simétricamente, el test podría plantearse en términos de la muestra $(y_1, \ldots, y_m)$.

Si la hipótesis nula fuera cierta, las muestras estarían mezcladas en rango y por tanto R_X, única v.a. en la expresión anterior, sería la suma de n números naturales en el conjunto $\{1, 2, \ldots, n+m\}$, con lo cual podríamos calcular su valor esperado y su varianza y en consecuencia los correspondientes al estadístico U_X.

$$E\left(U_X/H_0\right) = \frac{nm}{2} \; ; \; Var\left(U_X/H_0\right) = \frac{nm(n+m+1)}{12}$$

Bajo la hipótesis de coincidencia de poblaciones, los resultados del estadístico U_X serán en general cercanos a su esperanza; por tanto, rechazaríamos la hipótesis nula cuando el valor U_X fuese excesivamente alto o bajo.

Esta distribución se encuentra tabulada para ciertos niveles de significación en función de los cuales establecemos la regla de decisión del test.

Cuando los tamaños muestrales son elevados (al menos 10) y la hipótesis nula es cierta, podemos utilizar una aproximación normal con los parámetros señalados. Si por el contrario los tamaños muestrales son bajos, entonces la distribución exacta de U se encuentra tabulada, y a partir de ella podemos obtener para determinados niveles de significación los valores críticos para realizar el contraste.

A modo de ilustración, supongamos que disponemos de dos muestras relativas a niveles de contaminación diarios (partículas en suspensión, $ug/m^3N.$) y deseamos contrastar si proceden de la misma población.

X	12	25	17	32	19
Y	18	13	16	15	24

Siguiendo el procedimiento de Mann-Whitney debemos estudiar conjuntamente ambas muestras, examinando los rangos de cada observación en el total:

Rangos X	1	9	5	10	7
Rangos Y	6	2	4	3	8

A partir de estos resultados se obtiene la suma de rangos $R_X = 32$ que conduce al valor

$$U_X^* = 5 \times 5 + \frac{5 \times 6}{2} - 32 = 8$$

Este resultado no difiere mucho del valor esperado, $E\left(U_X/H_0\right) = 12,5$ y por tanto no conduce al rechazo de la hipótesis nula (en concreto, dado el pequeño tamaño muestral, para obtener el nivel crítico asociado a este valor deberíamos acudir a la distribución tabulada de la U, donde aparece $P(U_X \leq 8) = 0,21$ y por tratarse de un contraste bilateral se obtendría $p = 0,42$).

8.5.4.2. Test de Wald-Wolfowitz

El proceso seguido en el *test de Wald-Wolfowitz* es una mezcla de los utilizados en el de Mann-Whitney (M-W) y el de rachas.

Dadas dos muestras aleatorias independientes de las poblaciones X e Y, sus observaciones se agrupan y se clasifican por rangos, definiendo las categorías "pertenecer a la muestra $(x_1, \ldots, x_n)$" que representamos por 0 y "pertenecer a la muestra $(y_1, \ldots, y_m)$" que denotamos por 1. A continuación observamos las rachas (R) de 0 y 1, cuyo número oscilará entre 2 y $n + m$.

Si la población de partida fuera la misma, los valores muestrales se encontrarían mezclados y las rachas serían mayores que un cierto límite r_1. En caso contrario, si las muestras procedieran de poblaciones diferentes sería de esperar que el número de rachas fuera reducido, observándose bloques de valores consecutivos de la misma muestra.

Para determinar la regla de decisión del contraste y la distribución de probabilidad de las rachas remitimos al lector al contraste de rachas desarrollado en el epígrafe 8.3.

8.5.4.3. Test de Kolmogorov-Smirnov para dos muestras

Otra posibilidad para realizar el contraste de identidad de poblaciones es adaptar a este caso el *test de Kolmogorov-Smirnov (K-S)* . La hipótesis nula sería la enunciada y ahora calcularíamos las funciones de distribución muestrales para ambas muestras, de modo que si la diferencia máxima entre ambas distribuciones es excesiva rechazaríamos la hipótesis de que las muestras procedan de la misma población, y si por el contrario las desviaciones son tolerables entonces podemos asumir la hipótesis de que no hay discrepancias entre las poblaciones de partida.

De una forma más operativa, ordenamos la primera muestra: $x_1 < \cdots < x_n$ y obtenemos su distribución muestral $S_{n,X}(x)$; a continuación ordenamos la segunda muestra y de nuevo calculamos su f.d. muestral $S_{m,Y}(x)$ (denotamos por x el recorrido de ambas variables). Definimos entonces el estadístico de K-S:

$$D^*_{n,m} = \sup_{-\infty<x<+\infty} |S_{n,X}(x) - S_{m,Y}(x)|$$

$D^*_{n,m}$sería la mayor de las $2(n + m)$ discrepancias que obtendríamos calculando las desviaciones por la izquierda y por la derecha en cada uno de los n puntos de discontinuidad de $S_{n,X}$ y los m de $S_{m,Y}$. La distribución del estadístico y la regla de decisión del contraste coinciden con las desarrolladas para el test de K-S para una muestra, que ha sido recogido en el apartado 8.3.

8.5.4.4. Prueba de Kruskal-Wallis para r muestras

La *prueba de Kruskal-Wallis* es una extensión del test de rangos y se utiliza para contrastar la hipótesis nula de que las poblaciones de las que proceden r muestras seleccionadas son idénticas.

Supongamos que se tienen r muestras $\left(X_{i_1}, \ldots, X_{i_{n_i}}\right)$, $i = 1, \ldots r$, con tamaños respectivos n_i siendo $\sum_{i=1}^{r} n_i = n$. Si reunimos todos los elementos muestrales en un solo conjunto y ordenamos sus valores en forma creciente, podemos asignar a cada uno de los elementos la posición o rango que ocupa en el conjunto total.

Denotemos por R_j la suma de los rangos correspondientes a la muestra j-ésima. Si las muestras procediesen de poblaciones idénticas (o si se quiere de la misma población) sería de esperar que todos los rangos tuviesen valores similares y en caso contrario el resultado sería significativo para rechazar la hipótesis de partida.

La discrepancia asociada a este test viene definida como:

$$d_{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{r} \frac{R_j^2}{n_j} - 3(n+1)$$

expresión que bajo la hipótesis nula sigue aproximadamente una distribución χ^2 con r-1 g.l.

Sobre muestras concretas calculamos el valor de la discrepancia observada, d^*, que permite determinar el nivel crítico $p = P\left(d_{KW} > d^*_{KW}/H_0\right)$ en función del cual se establece la decisión del contraste.

En este contraste tenemos las mismas restricciones que en otras aplicaciones de la χ^2: se exige que el tamaño de cada muestra n_j sea mayor o igual que 5 ya que en otro caso habría que desarrollar un modelo de probabilidad específico.

Entre las aplicaciones del contraste de Kruskal-Wallis resulta especialmente interesante, en el caso de series temporales, el *contraste de estacionalidad*. En esta situación la hipótesis nula es la no existencia de estacionalidad, o equivalentemente, la identidad de las poblaciones correspondientes a los diferentes subperíodos considerados (meses, trimestres, etc.).

8.5.5. Contrastes de cambios sucesivos sobre una población

El contraste que ahora planteamos resulta muy útil en problemas socioeconómicos donde se trata de estudiar si se han producido cambios en la distribución de una población a lo largo del tiempo (por ejemplo, después de que se hayan adoptado algunas medidas de política económica, o bien cuando trabajamos con datos de panel para estudiar cambios en la intención de voto del electorado o reacciones ante ciertos acontecimientos).

8.5.5.1. Test de McNemar

Supongamos una población que se puede clasificar en dos categorías A y B e imaginemos que llevamos a cabo un seguimiento de una misma muestra en dos períodos de tiempo t_1 y t_2. Podemos clasificar la información muestral en una tabla del tipo:

Periodo $\downarrow t_1/t_2 \rightarrow$	Categoría A	Categoría B
Categoría A	n_A	n_{AB}
Categoría B	n_{BA}	n_B

donde n_A representa las unidades que en el período t_1 se encontraban en la categoría A y en el período posterior t_2 mantienen la misma categoría (no cambiaron de opinión); n_B indica la misma estabilidad respecto a la categoría B, y en cambio n_{AB} y n_{BA} representan el número de unidades que han cambiado de situación entre estos períodos.

Si la hipótesis que deseamos contrastar es la de estabilidad en la población entre estos períodos respecto a las categorías A y B, entonces $n_{AB} + n_{BA}$ debería ser una cifra relativamente pequeña; si por el contrario la hipótesis que estamos contrastando no es la estabilidad, sino la aleatoriedad en los cambios (es decir, que estos no respondan a ninguna tendencia especial), entonces las frecuencias n_{AB} y n_{BA} deberían ser magnitudes muy próximas.

El *test de McNemar* da respuesta a este segundo planteamiento. En este caso disponemos de unas frecuencias observadas de los cambios, y otras frecuencias teóricas que, por asumir que los cambios no tienen influencia, vienen dadas por $\frac{n_{AB}+n_{BA}}{2}$. Comparando ambos tipos de frecuencias podemos proponer una discrepancia χ^2:

$$d_M = \frac{\left(n_{AB} - \frac{n_{AB}+n_{BA}}{2}\right)^2}{\frac{n_{AB}+n_{BA}}{2}} + \frac{\left(n_{BA} - \frac{n_{AB}+n_{BA}}{2}\right)^2}{\frac{n_{AB}+n_{BA}}{2}} = \frac{(n_{AB}-n_{BA})^2}{n_{AB}+n_{BA}}$$

que, bajo la hipótesis nula, se distribuye aproximadamente como una χ^2 con 1 g.l.

Para una buena aproximación la frecuencia esperada debería ser mayor o igual que 5 y además se recomienda introducir la corrección de continuidad:

$$d_M = \frac{(|n_{AB} - n_{BA}| - 1)^2}{n_{AB} + n_{BA}}$$

El resto del contraste consiste en construir la discrepancia observada d^* y rechazar o no la hipótesis nula en función el nivel crítico $p = P(d_M > d_M^*/H_0)$.

8.5.5.2. Prueba Q de Cochran

El *test Q de Cochran* es una extensión de la prueba anterior para el caso en que consideremos r muestras. En esta situación observamos N elementos o grupos durante r períodos de tiempo $t_1, \ldots, t_r$, y queremos comprobar si los cambios se han debido únicamente al azar o por el contrario dichos cambios pueden atribuirse a otras circunstancias.

La información muestral recoge el número de éxitos, esto es, las veces que se ha presentado cierta característica sobre los elementos investigados. La hipótesis nula establecería que las proporciones de éxitos son idénticas en la distintas muestras, presentándose únicamente variaciones aleatorias.

Si designamos por n_j el número de éxitos en la muestra j, por $\bar{n}_j$ su promedio

$$\bar{n}_j = \frac{\sum_{j=1}^{r} n_j}{r}$$

y por n_i el número total de éxitos correspondientes al elemento o grupo i-ésimo, la discrepancia viene dada por la expresión:

$$d_Q = \frac{r(r-1)\sum_{j=1}^{r}(n_j - \bar{n}_j)^2}{r\sum_{i=1}^{N} n_i - \sum_{i=1}^{N} n_i^2}$$

que bajo la hipótesis nula sigue aproximadamente una distribución χ^2 con $r-1$ g.l.

Consideremos a modo de ilustración que observamos una muestra de 10 jóvenes activos en tres períodos de tiempo distintos, con la intención de analizar si se han producido cambios significativos en su situación laboral, que clasificamos en las categorías "ocupado" y "parado".

La información muestral aparece representada en la tabla siguiente, donde denotamos por 1 el éxito (joven ocupado) y por 0 su complementario:

Elementos→ ↓ Muestras	1	2	3	4	5	6	7	8	9	10	n_j
Muestra 1	1	1	1	0	1	1	1	1	1	1	0
Muestra 2	1	0	0	1	1	1	1	0	1	0	6
Muestra 3	1	1	1	0	0	1	1	1	1	0	7
n_i	3	2	2	1	2	3	3	2	3	1	22
n_i^2	9	4	4	1	4	9	9	4	9	1	54

La hipótesis nula sería que la proporción de jóvenes ocupados no ha cambiado en los tres períodos de tiempo considerados y para su contraste utilizamos la expresión de la discrepancia de Cochran, cuyo resultado es en este caso:

$$d_Q^* = \frac{3(3-1)\left[(9-7,\hat{3})^2 + (6-7,\hat{3})^2 + (7-7,\hat{3})^2\right]}{(3)\,(22) - 54} = 2,\hat{3}$$

valor que lleva asociado un nivel crítico $p = P\left(\chi_2^2 > 2,\hat{3}/H_0\right) = 0,3114$ y por tanto no resulta significativo para rechazar la hipótesis planteada.

8.6. Anexo: Diseño de contrastes óptimos

A lo largo de este capítulo nos hemos ocupado de los contrastes de hipótesis estadísticas desde una óptica eminentemente práctica. Sin embargo ¿podemos estar seguros de que los contrastes aplicados son adecuados? o, en otros términos, ¿es posible llegar a diseñar tests óptimos para contrastar un supuesto determinado?

Este interrogante no tiene una respuesta única, ya que ésta dependerá de nuestras condiciones de trabajo (hipótesis simples o compuestas, información poblacional disponible, ...). En cualquier caso, el diseño de contrastes adecuados debe tener en cuenta no sólo la hipótesis nula sino también su alternativa.

Para ilustrar hasta qué punto la consideración de la alternativa afecta a nuestras conclusiones, imaginemos que un investigador enuncia una hipótesis relativa a la proporción de familias que atraviesan dificultades económicas: $H_0 : p = 0,2$.

Si al seleccionar una muestra aleatoria simple de 100 familias se observa que 31 declaran tener dificultades económicas para llegar a fin de mes, la proporción muestral sería $\hat{p} = 0,31$ y el contraste de la hipótesis nula se llevaría a cabo calculando la discrepancia tipificada y el correspondiente nivel crítico:

$$d^*_{\hat{p}}/H_0 = \frac{0,31 - 0,2}{\sqrt{\frac{(0,2)(0,8)}{100}}} \;\; ; \;\; p = P\left(\left| \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \right| > 2,75/H_0 \right) = 0,0059$$

Según el planteamiento visto en el tema, esta probabilidad adopta un valor suficientemente bajo, que permite calificar el resultado de "significativo para rechazar". Sin embargo, hasta ahora la hipótesis ha sido contrastada por sí misma, sin tener en cuenta cuál sería la alternativa al supuesto planteado.

Supongamos ahora que el investigador considera como hipótesis alternativa, $H_1 : p = 0,15$ ¿nos llevaría el rechazo de la hipótesis nula $p = 0,2$ a admitir como válida la alternativa $p = 0,15$? Parece claro que la respuesta es negativa, ya que la discrepancia de la muestra respecto a la alternativa es aún mayor que la correspondiente a la hipótesis nula (se obtiene $d^*_{\hat{p}}/H_1 = 4,481$, que lleva asociada una probabilidad prácticamente nula: $p = 0,0000074$; es decir, una muestra con un 31 % de familias atravesando dificultades económicas es poco verosímil si la proporción poblacional es del 20 %, pero aún lo es menos bajo la alternativa $p = 15\,\%$).

Así pues, en situaciones como la descrita es importante tener presente que la hipótesis nula no se contrasta por sí misma sino enfrentada a una alternativa. Por tanto, sólo en la medida en que la hipótesis alternativa explique la realidad observada con un mayor grado de probabilidad tendríamos motivos para rechazar la hipótesis nula.

Aplicando este razonamiento, se obtendría ahora el nivel crítico correspondiente a la cola de la izquierda (en la dirección de la alternativa): $p = P\left(d_{\hat{p}} < d^*_{\hat{p}}/H_0\right) = P\left(d_{\hat{p}} < 2,75/H_0\right) = 0,9$ que llevaría a no rechazar la hipótesis nula H_0.

[Estúdiese cuál sería la conclusión en caso de considerar como alternativa $H_1 : p = 0,28$]

Obsérvese que, si bien podría llevarse a cabo un planteamiento más general del problema, para una mayor claridad expositiva esta ilustración considera dos únicos valores posibles en el espacio paramétrico $\Theta = \{p = 0,2\,,\, p = 0,15\}$.

J. Neyman y E.S. Pearson (1928, 1933) fueron los primeros en reconocer explícitamente la importancia de la hipótesis alternativa en el diseño de contrastes adecuados. Ambos autores establecieron bases teóricas sólidas para la consideración de la hipótesis alternativa, a partir de las cuales desarrollaron un nuevo enfoque en la teoría del contraste, con importantes rasgos diferenciales respecto a los contrastes de significación.

El contraste de hipótesis es una técnica inferencial, que por tanto lleva inherente un riesgo. Para cuantificar dicho riesgo, podemos examinar los posibles errores cometidos al adoptar una conclusión sobre determinada hipótesis, tal y como describe la siguiente

tabla:

		¿Es la hipótesis H_0 cierta?	
		SI	NO
¿Rechazamos H_0?	SI	ERROR I	———
	NO	———	ERROR II

La tabla anterior recoge tanto las distintas situaciones poblacionales (la hipótesis planteada puede ser cierta o falsa, pero nosotros siempre lo ignoraremos) como las diferentes conclusiones de nuestro contraste (según la evidencia muestral podríamos optar por rechazar la hipótesis o por no rechazarla). De las cuatro casillas posibles, puede observarse que en dos de ellas (marcadas con guiones) la decisión será acertada; sin embargo existen dos posibilidades de error: el denominado *error tipo I* o *error I* aparece cuando se rechaza una hipótesis que es cierta y el *error tipo II* o *error II* consiste en no rechazar una hipótesis cuando ésta es falsa.

La comparación de los contrastes de hipótesis con los procesos judiciales nos llevaría a definir el error I como "condenar a un inocente" mientras que el error II sería equivalente a "absolver a un culpable".

Dado que en la práctica no podemos aspirar a conocer si la hipótesis planteada era o no cierta tampoco podremos saber si hemos cometido alguno de los errores. En cambio, sí podemos estudiar sus correspondientes probabilidades, que aparecen recogidas a continuación:

		¿Es la hipótesis H_0 cierta?	
		SI	NO
¿Rechazamos H_0?	SI	$P(error\, I) = \alpha$	$1 - \beta$
	NO	1-α	$P(error\, II) = \beta$

Las probabilidades recogidas en esta tabla corresponden a las decisiones correcta e incorrecta bajo las distintas situaciones poblacionales. Así, si la hipótesis planteada resulta ser cierta, la probabilidad α aparece asociada al error I, ya que sería la probabilidad de que, siendo H_0 cierta, nuestra información muestral nos situase en la región crítica, llevándonos al rechazo de la hipótesis:

$$\alpha = P(error\, I) = P(\text{Rechazar}\, H_0/H_0) = P(T \in RC/H_0)$$

Por su parte, la probabilidad del error II puede ser expresada como:

$$\beta = P(error\, II) = P(\text{No Rechazar}\ H_0/H_1) = P(T \notin RC/H_1)$$

Para ilustrar esta situación supongamos que tenemos dos hipótesis simples: $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$, siendo $\theta_1 > \theta_0$, y que la región de rechazo viene delimitada por

Figura 8.9.: Probabilidad de errores I y II

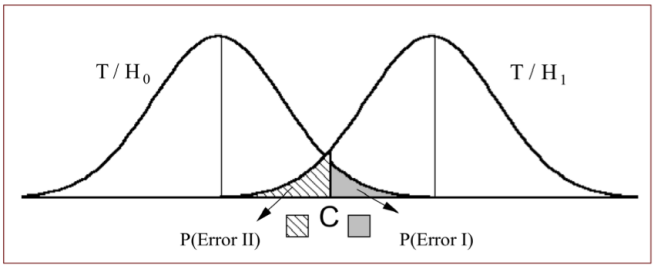
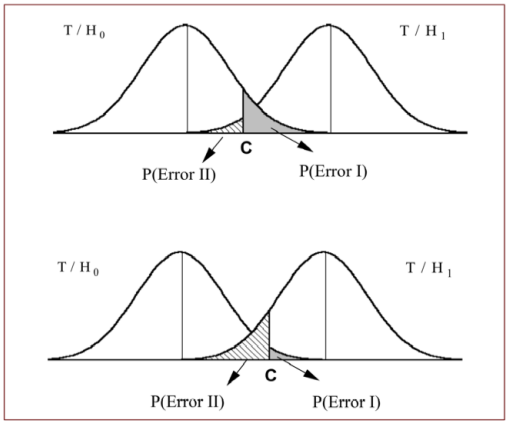


Figura 8.10.: Probabilidades de errores y Regiones Críticas



el valor crítico C. En esta situación ambos tipos de error pueden ser representados gráficamente tal y como recoge la figura 8.9

Si las curvas representadas recogen las distribuciones de probabilidad T/H_0 y T/H_1, el área sombreada representa la probabilidad de error tipo I, esto es, la probabilidad de que T se sitúe en la región crítica (a la derecha de C) siendo H_0 cierta; por su parte, el área rayada representa la probabilidad de error tipo II, es decir, de que T se sitúe en la región de aceptación siendo H_1 cierta.

En la figura 8.10 puede observarse que si disminuimos la probabilidad de error tipo I entonces estamos desplazando C hacia la derecha y por tanto aumentamos la probabilidad tipo II y recíprocamente.

Las gráficas anteriores ponen de manifiesto la relación de sustitución existente entre las probabilidades de los dos tipos de error considerados, que impide diseñar un procedimiento en el que podamos minimizar simultáneamente ambas probabilidades y por tanto una región crítica óptima en sentido global.

Una solución posible sería construir un óptimo condicionado; esto es, acotar una de las probabilidades de error y elegir, entre todas las regiones críticas que verifiquen la restricción anterior, aquélla que haga mínima la probabilidad del otro error.

Este tratamiento asimétrico de los errores exige establecer un criterio de prioridad entre ellos. Si fijamos una cota baja de probabilidad al error que sirve de restricción, al menos estaremos garantizando que la probabilidad de equivocarnos en ese sentido nunca será alta.

La probabilidad del otro error será la menor posible, pero a priori no podemos saber si será alta o baja; así pues, parece que debemos orientar nuestro mayor énfasis hacia el error de restricción que sea nuestro error principal, actuando el otro como un error secundario.

Aunque puede haber dudas conceptuales sobre la gravedad de ambos tipos de error, partiendo del supuesto de inocencia de la hipótesis nula parece deseable buscar contrastes que minimicen la probabilidad de condenar a un inocente, ya que este error I parece ser más grave que el de absolver a un culpable.

Otra ilustración que clarifica estos conceptos consiste en identificar la hipótesis nula con un alumno que domina cierta asignatura ("merece aprobar") siendo la información muestral el examen. En este caso el error I equivale a "suspender cuando merece aprobar" mientras el error II "aprobar cuando merece suspender" suele ser considerado menos grave.

No obstante, conviene aclarar que también existen casos en los que el error II resulta especialmente preocupante, por lo cual debemos prestar atención a su probabilidad β. Este tipo de situaciones se presentan en algunas pruebas de control estadístico de calidad y así parece claro que asumir como válidos un material de construcción o un medicamento que en realidad no cumplen los requisitos necesarios (error tipo II) seria más grave, por sus potenciales consecuencias, que rechazarlos cuando sí cumplen los requisitos de calidad (error tipo I).

Partiendo de la consideración del error tipo I como riesgo prioritario del contraste, la forma habitual de proceder para diseñar un buen test consiste en acotar la probabilidad de este error (α) minimizando a continuación la probabilidad de error tipo II (β). Este criterio de optimización fue desarrollado por Neyman y Pearson.

La probabilidad de error I se denomina *nivel de significación* o *tamaño del test*, y suele fijarse en valores del 5 % o el 1 %. Por otra parte, minimizar la probabilidad del error tipo II es equivalente a maximizar la de su complementario, que se denomina *potencia del test*, y es la probabilidad de rechazar una hipótesis falsa, esto es:

$$P = 1 - \beta = P(\text{Rechazar } H_0/H_1) = P(T \notin RC/H_1)$$

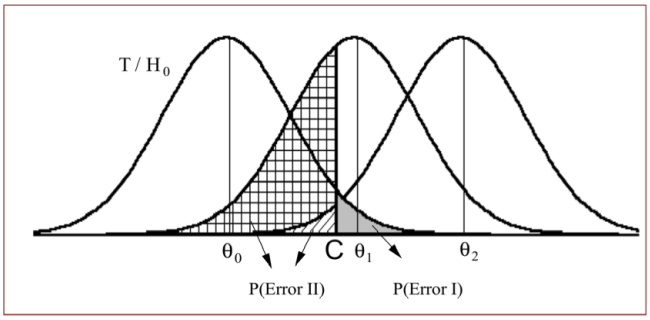
La potencia del test puede ser considerada como un indicador de eficiencia del contraste.

Aunque hemos denotado por β la probabilidad de error tipo II, esto es válido para las situaciones en las que la alternativa es simple. Sin embargo, cuando ésta es compuesta, la probabilidad depende del valor particular de la alternativa.

A modo de ilustración consideremos la hipótesis nula $H_0 : \theta = \theta_0$ frente a la alternativa compuesta $H_1 : \theta \in \{\theta_1, \theta_2\}$, que podemos representar en la figura 8.11 para un nivel de significación α.

Al igual que en figuras anteriores, hemos sombreado en gris la probabilidad de error tipo I y en rayado y con cuadrículas las probabilidades de error tipo II correspondientes a los dos valores de la

Figura 8.11.: Probabilidades de errores y Regiones Críticas



alternativa, que como vemos dependen de los valores adoptados por el parámetro bajo H_1. Por este motivo, la probabilidad de error tipo II será una función que a cada valor θ de la alternativa le asigna un valor$\beta(\theta)$.

Enlazando con el párrafo anterior, cuando la hipótesis alternativa es simple la potencia adoptará un valor constante $1-\beta$; sin embargo, para hipótesis H_1 compuestas la potencia será una función del parámetro que se denomina función de potencia $P(\theta) = 1 - \beta(\theta)$.

Si la hipótesis nula también fuera compuesta ($H_0 : \theta \in \Theta_0$), entonces para cada valor del parámetro en esta hipótesis obtendríamos una probabilidad de error tipo I, por lo cual definimos el nivel de significación o tamaño del test, α, como la mayor de esta probabilidades (o en su defecto si ésta no existe, como el supremo de las mismas):

$$\alpha = \sup_{\theta \in \Theta_0} \alpha(\theta)$$

Una vez fijado cierto nivel de significación α para los contrastes, estamos interesados en comparar los distintos tests que garanticen ese tamaño o nivel de significación, seleccionando los más adecuados.

Definición 8.2. Sean R y R' dos regiones críticas definidas para un contraste al mismo nivel de significación o tamaño α.

- Se dice que R es *más potente* que R' para un valor de la alternativa si la potencia de R supera a la de R' para ese valor.
- Cuando la potencia de R supera a la de R' para todos los valores de la alternativa, diremos que R es *uniformemente más potente* que R'.
- En el caso de que un test sea uniformemente más potente que cualquier test de su mismo tamaño α diremos que es *uniformemente de máxima potencia* (UMP) para el nivel α.

En los apartados que siguen estudiaremos las posibilidades de diseñar contrastes óptimos o al menos adecuados para distintos tipos de hipótesis.

Hipótesis nula simple frente a alternativa simple. Lema de Neyman-Pearson

Consideremos en primer lugar, el caso más sencillo con hipótesis nula y alternativa simples:

$$\boxed{\begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{array}}$$

Sea X una población con función de densidad $f(x, \theta)$. Dada una m.a.s. $(X_1, \ldots, X_n)$ sobre esta población, denotaremos por $L_0 = L(x_1, \ldots, x_n, \theta_0)$ y $L_1 = L(x_1, \ldots, x_n, \theta_1)$ las funciones de verosimilitud para los parámetros θ_0 y θ_1 respectivamente. En esta situación, la determinación de regiones críticas óptimas queda resuelta mediante el *Lema de Neyman-Pearson* cuyo enunciado es el siguiente:

Lema 8.1. *(Neyman-Pearson) Si R es una región crítica al nivel α y existe una constante k para la que se cumple:*

$$\frac{L(x_1, \ldots, x_n, \theta_1)}{L(x_1, \ldots, x_n, \theta_0)} \geq k \ , \ si\,(x_1, \ldots, x_n) \in R$$

$$\frac{L(x_1, \ldots, x_n, \theta_1)}{L(x_1, \ldots, x_n, \theta_0)} < k \ , \ si\,(x_1, \ldots, x_n) \notin R$$

entonces R es una región crítica óptima al nivel α para el contraste de las hipótesis $H_0 : \theta = \theta_0$ y $H_1 : \theta = \theta_1$.

El lema de Neyman-Pearson no proporciona un método directo para la construcción de test óptimos, pero a partir de los cocientes anteriores podemos buscar un valor de k de manera que mediante algunas transformaciones lleguemos a encontrar la región crítica óptima.

Supongamos que X es una población normal con media μ desconocida y varianza 1. Deseamos encontrar un test óptimo para contrastar: $H_0 : \mu = \mu_0$ frente a la alternativa $H_1 : \mu = \mu_1$. Supongamos $\mu_1 > \mu_0$.

La función de verosimilitud bajo estas hipótesis vendrá dada por:

$$L(x_1, \ldots, x_n, \mu_0) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2} \ ; \ L(x_1, \ldots, x_n, \mu_1) = \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2}$$

Partiendo del lema anterior, tenemos:

$$\frac{L(x_1, \ldots, x_n, \mu_1)}{L(x_1, \ldots, x_n, \mu_0)} = \frac{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2}}{e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2}} = e^{-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2+\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2} \geq k \ , \ si\ (x_1, \ldots, x_n) \in R$$

Tomando logaritmos, podemos expresar la relación anterior como:

$$\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_0)^2-\frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_1)^2\geq \ln k$$

desarrollando los cuadrados del primer miembro, resulta:

$$n\left(\mu_0^2-\mu_1^2\right)+2\left(\mu_1-\mu_0\right)\sum_{i=1}^{n}x_i\geq 2\ln k$$

con lo cual:

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n}\geq\frac{2\ln k-n\left(\mu_0^2-\mu_1^2\right)}{2n(\mu_1-\mu_0)}$$

o equivalentemente:

$$d_{\bar{x}}=\frac{\bar{x}-\mu_0}{\frac{1}{\sqrt{n}}}\geq\sqrt{n}\left[\frac{2\ln k+n\left(\mu_1^2-\mu_0^2\right)}{2n(\mu_1-\mu_0)}-\mu_0\right]=\sqrt{n}\left[\frac{2\ln k+n(\mu_1-\mu_0)^2}{2n(\mu_1-\mu_0)}\right]=C$$

En este caso, el lema de Neyman-Pearson nos define la forma de la región crítica óptima: rechazaremos la hipótesis nula cuando la discrepancia asociada al test sea mayor o igual que C; esto es, $R=\{d_{\bar{x}}^*/d_{\bar{x}}^*\geq C\}$.

Por otra parte, podemos calcular el valor de C a partir de la probabilidad de error tipo I. Fijado un nivel de significación α, se tiene: $\alpha=P\left(d_{\bar{x}}\geq C/\mu=\mu_0\right)$, la existencia de C garantiza la de k y por tanto la región crítica obtenida es óptima.

Observemos que la forma de la región crítica $(C,+\infty)$ coincide con las que habíamos construido en los contrastes de significación anteriores.

En el ejemplo anterior, para obtener la mejor región crítica no hemos tenido en cuenta el valor de la alternativa. Esto se debe al tipo de hipótesis planteada.

Este enunciado es equivalente a considerar, en las condiciones anteriores, la discrepancia de mayor potencia que viene dada para cada muestra posible por la expresión:

$$d^*=\frac{L(x_1,\ldots,x_n,\theta_1)}{L(x_1,\ldots,x_n,\theta_0)}$$

Cuando esta discrepancia adopte valores elevados, llegaremos a la conclusión de que el valor propuesto no es suficientemente verosímil y en consecuencia rechazaremos la hipótesis nula.

Determinación de tests uniformemente más potentes

El lema de Neyman-Pearson no puede ser aplicado cuando las hipótesis son compuestas, supuesto que resulta muy frecuente en la práctica.

En el caso de hipótesis compuestas no siempre existen test óptimos (por ejemplo, cuando la alternativa es bilateral compuesta, en general no es posible determinar un test UMP). Sin embargo, para el caso de hipótesis nula unilateral (del tipo $H_0:\theta\geq\theta_0$ o bien $H_0:\theta\leq\theta_0$) es posible obtener test óptimos; el método para resolver estos contrastes de forma óptima consiste en imponer restricciones al modelo probabilístico de la población o bien sobre el estadístico o discrepancia asociada al test (supuesto de suficiencia, en algunos casos).

Determinadas distribuciones de probabilidad dan lugar a cocientes de verosimilitud monótonos.

Definición 8.3. Dada una función de verosimilitud $L(x_1,\ldots,x_n,\theta_1)$ se dice que tiene un cociente de verosimilitudes monótono en una función muestral $T(X_1,\ldots,X_n)$, si para para todo par de valores del parámetro $\theta_1<\theta_2$, en los que la función de verosimilitud es distinta, se verifica que el cociente de verosimilitudes $\frac{L_2}{L_1}$ es una función monótona en T.

La monotonía puede ser estricta o no, creciente o decreciente según el comportamiento de la función muestral.

Los modelos binomial, Poisson, normal, exponencial, Gamma, etc. verifican la propiedad de que su razón de verosimilitudes es monótona (RVM).

Para los modelos que verifiquen esta propiedad es válido el siguiente resultado:

Corolario 8.1. *Sea X una población cuyo modelo de probabilidad verifica la propiedad de RVM (no decreciente), entonces la región crítica: $R = \{t/t = T(x_1, \ldots, x_n) \geq C\}$ es óptima (UMP) al tamaño α para contrastar las hipótesis: $H_0 : \theta \leq \theta_0$ frente a la alternativa $H_1 : \theta > \theta_0$.*

En el caso de que la monotonía fuera no creciente la región crítica vendría dada por: $R = \{t/t = T(x_1, \ldots, x_n) \leq C\}$. De forma complementaria se construyen las regiones para las hipótesis: $H_0 : \theta \geq \theta_0$ frente a la alternativa $H_1 : \theta < \theta_0$.

El valor de C que determina la región crítica puede obtenerse a partir de la probabilidad de error tipo I.

Test de la razón de verosimilitudes

Otro método importante para contrastar hipótesis estadísticas es el de la razón de verosimilitudes, cuyo punto central es la función de verosimilitud del parámetro. Este procedimiento consiste en calcular la función de verosimilitud cuando la hipótesis nula es cierta ($\theta \in \Theta_0$) y también para todo el espacio paramétrico ($\theta \in \Theta$), definiendo una medida de discrepancia como el cociente entre ambas verosimilitudes.
El contraste de dos hipótesis H_0 y H_1 puede ser planteado mediante el principio de la razón de verosimilitudes, consistente en comparar la mejor explicación de la información muestral dentro de la hipótesis H_0 con la mejor explicación posible.

Consideremos un contraste del tipo:

$$\boxed{\begin{array}{l} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{array}}$$

Definición 8.4. La *razón de verosimilitudes* viene dada por la expresión:

$$\lambda(x_1, \ldots, x_n) = \frac{\sup\limits_{\theta \in \Theta_0} L(x_1, \ldots x_n, \theta)}{\sup\limits_{\theta \in \Theta} L(x_1, \ldots x_n, \theta)}$$

cuyo numerador recoge la mejor explicación de la realización muestral bajo la hipótesis nula, mientras que el denominador es la mejor explicación de la muestra.

La función anterior λ es una v.a. por serlo la muestra (y por tanto las verosimilitudes) que puede tomar valores entre 0 y 1 (es un cociente de dos magnitudes no negativas donde el denominador es mayor o igual que el numerador).

En la medida que esta v.a. se aproxime a 0 querrá indicar que la muestra es suficientemente inconsistente con la hipótesis nula para rechazar la hipótesis. El razonamiento contrario se adoptaría si su valor se aproximase a 1.

Para muestras grandes la distribución de la variable $-2\log\lambda$ sigue aproximadamente un modelo χ^2 con $r-k$ g.l., donde r representa la dimensión de Θ y k la de Θ_0. Así pues, fijado un nivel de significación α, podemos elegir λ_0 de forma que $P(\lambda > \lambda_0) = \alpha$.

Este procedimiento nos permite abordar alguno de los tests anteriores de forma sistemática. Así, el test de la razón de verosimilitudes para contrastar $H_0 : \theta \in \Theta_0$ frente a $H_1 : \theta \in \Theta_1$ consiste en calcular λ, que en ciertos casos contendrá en su expresión las discrepancias tipificadas asociadas a los contrastes particulares que estemos realizando. En otros casos tendremos que aplicar la convergencia asintótica de λ para obtener la región crítica, consistente en rechazar H_0 si se cumple $\lambda(x_1, \ldots, x_n) < c$ para una constante c; $0 < c < 1$.

La constante c se determina, una vez fijado el tamaño del contraste α, mediante la expresión:

$$\sup_{\theta\in\Theta_0} P\left(\lambda(X_1, \ldots, X_n) < c\right) = \alpha$$

y si ese valor c no existe entonces se elige el mayor valor de c tal que

$$P\left(\lambda(X_1, \ldots, X_n) < c\right) = \alpha \ ; \ \forall\theta \in \Theta_0$$

Parte III.

Introducción a la Econometría

9. Modelos econométricos. El modelo lineal simple

La descripción de la realidad económica no es una tarea sencilla. Con el objetivo de representar esa realidad de forma simplificada pero adecuada, los modelos econométricos se han convertido en una herramienta habitual en el análisis económico.

Los modelos econométricos se basan en los modelos económicos, a los que incorporan un componente de incertidumbre o aleatoriedad que, como hemos visto en capítulos anteriores, es inherente al ámbito socioeconómico y que habitualmente denotaremos por u.

Consideremos a modo de ilustración uno de los modelos económicos más emblemáticos: la función de Consumo de Keynes, que describe el comportamiento de los consumidores en función de la renta.

En su Teoría General (1936), J. M. Keynes enuncia la ley psicológica fundamental según la cual el consumo es función creciente de la renta $C = f(R)$ y además, un incremento de renta provocará siempre un incremento de menor magnitud en el consumo: $0 < \frac{dC}{dR} < 1$ donde $\frac{dC}{dR}$ es la *Propensión Marginal al Consumo*.

Además, Keynes considera que una vez cubiertas las necesidades primarias se tenderá a acumular, hecho que provoca que la proporción de renta que se ahorra sea mayor a medida que la renta real aumenta: $\frac{dC}{dR} < \frac{C}{R}$, es decir, la propensión marginal al consumo es menor que la propensión media.

A partir del modelo económico anteriormente descrito podemos especificar el siguiente modelo econométrico:

$$C = \beta_1 + \beta_2 R + u \ , \ \ 0 < \beta_2 < 1 \ , \ \beta_1 > 0$$

Existe una amplia variedad de modelos econométricos, que llegan a alcanzar niveles de complejidad y sofisticación muy elevados. Lógicamente, en este tema nos limitaremos a presentar un tratamiento introductorio, estudiando únicamente modelos lineales uniecuacionales.

9.1. Los modelos econométricos

Los modelos econométricos recogen en una o varias ecuaciones las relaciones existentes entre las magnitudes económicas. Siguiendo un criterio de causalidad, las variables que intervienen en los modelos se clasifican en *endógenas* (aquéllas que son explicadas por el modelo) y *predeterminadas* que abarcan tanto las variables *exógenas* (determinadas externamente al fenómeno que se modeliza) como las *endógenas retardadas* (determinadas dentro del proceso pero en momentos anteriores al considerado).

Definición 9.1. Denominamos *modelo econométrico* a una descripción no determinista de la relación entre varias magnitudes económicas, mediante una expresión funcional

concreta y variables especificadas en términos estadísticos.
La parte funcional o sistemática de los modelos econométricos trata de recoger las relaciones entre los agentes económicos, bien sea mediante modelos de comportamiento (funciones de consumo o ahorro), relaciones tecnológicas (funciones de producción), relaciones institucionales (funciones de impuestos o gastos públicos) ... En cualquiera de los casos, estas relaciones se verán también afectadas por el factor humano y por tanto existirá un componente no determinista o aleatorio, que en ocasiones se denomina también *variable latente*.

Como consecuencia, los modelos econométricos conllevan una extensión o ampliación de los modelos económicos, ya que incorporan junto a la componente teórica o sistemática la presencia de incertidumbre o aleatoriedad.

Aunque no existe una división radical entre modelos económicos y econométricos, parece claro que estos últimos exigen una especificación funcional concreta que no siempre aparece en los modelos económicos, e incorporan además un componente aleatorio. Así pues, los modelos econométricos son modelos económicos que incluyen las especificaciones necesarias para su aplicación empírica.

Un modelo econométrico para la magnitud Y viene dado por la expresión genérica $Y = f(X) + u$, en la que se incluyen un componente sistemático $f(X)$, que recoge la relación causal entre las variables postulada por la teoría económica y una *perturbación aleatoria* u, sobre cuyo comportamiento solamente es posible establecer hipótesis.
La presencia de la perturbación aleatoria u viene justificada por diversas vías: en primer lugar, recoge la aleatoriedad del comportamiento humano, por otra parte resume la influencia conjunta de distintos factores que no pueden ser incluidos explícitamente en el modelo y además recoge los errores de medida en la observación de la variable Y.

En el modelo de Consumo Keynesiano introducido como ilustración $C = \beta_1 + \beta_2 R + u$ hemos incorporado explícitamente la perturbación aleatoria u, ya que la relación entre consumo y renta no es exacta (Keynes enunció una relación entre renta y consumo creciente por término medio). Además, el modelo aparece especificado en términos lineales y los enunciados keynesianos sobre el consumo autónomo y la propensión marginal al consumo se traducen en restricciones sobre los parámetros β_1 y β_2.

Habitualmente se distinguen dentro de la modelización econométrica tres fases diferenciadas. La primera de ellas es la *especificación* y consiste en traducir el modelo económico teórico, proponiendo una forma matemática que establezca cierta relación causal, haciendo explícita la perturbación aleatoria.
El punto de partida para la especificación de un modelo es la teoría económica, que facilita orientaciones sobre los tipos de relaciones existentes y la influencia que cada variable explicativa debe tener en la endógena. Sin embargo, raramente la teoría económica informa sobre la forma funcional del modelo.

Habitualmente las relaciones se formulan, al menos en una primera versión, en términos lineales o bien linealizables que proporcionan una descripción sencilla de la realidad. Ello no impide que puedan elaborarse modelos más complejos en cuanto a su formulación o, si ello resulta necesario, sistemas de varias ecuaciones que puedan describir más adecuadamente las interrelaciones entre magnitudes.

Sin embargo, en este capítulo nos limitaremos a analizar modelos econométricos lineales uniecuacionales.

Una vez que el modelo econométrico ha sido especificado se dispone de una expresión genérica para las relaciones estudiadas. Sin embargo, en la práctica los modelos deben ser aproximados a partir de la información estadística relativa a las variables que intervienen en los mismos, etapa que se denomina estimación.

La *estimación* de un modelo econométrico consiste en la obtención de valores numéricos para sus parámetros a partir de la información estadística disponible. En esta etapa resulta imprescindible la información sobre todas las variables que aparecen en el modelo econométrico y la aplicación de un método de estimación adecuado.

Los datos tienen una importancia primordial ya que condicionan las inferencias que realicemos sobre los modelos econométricos. Sin embargo, contra lo que en un principio pudiera suponerse no es imprescindible que nuestra información muestral constituya una verdadera muestra aleatoria representativa de la población estudiada. El criterio esencial para la selección de los datos es que todas las observaciones procedan del mismo proceso económico, es decir, que sigan idénticos patrones de comportamiento.

Por lo que respecta al método de estimación, seguiremos los procedimientos estudiados en capítulos anteriores (mínimos cuadrados y máxima verosimilitud), que garantizan buenas propiedades para los estimadores de los parámetros.

La etapa de *validación* -que también se denomina *contraste o verificación*- tiene por objetivo comprobar el grado de coherencia que el modelo presenta con la realidad económica de partida. Por ello, es imprescindible en esta fase establecer los criterios para rechazar o aceptar un modelo.

La presencia de incoherencias o contradicciones se podría detectar tanto en los parámetros estimados (que podrían presentar signos distintos a los esperados o bien valores no adaptados a la teoría económica), como en los supuestos o hipótesis asumidos. En el caso de que aparezcan contradicciones entre nuestros resultados y las hipótesis iniciales es necesario investigar cuál es el fallo: podría tratarse de los datos utilizados en la estimación, la especificación de partida, los métodos de estimación empleados e incluso la teoría económica de la que partimos.

Aunque en principio las fases de especificación, estimación y contraste son secuenciales, es posible retroceder o avanzar según el resultado obtenido en cada etapa. De ahí que no haya normas generales sino que, como indica el esquema, serán en cada caso la propia dificultad del modelo y la información disponible los factores que determinen la secuencia y el ritmo de nuestro trabajo.

Una vez que el modelo ha superado la etapa de validación, estamos en condiciones de llevar a cabo su implementación práctica, que abarca tanto la realización de análisis estructurales como la elaboración de predicciones.

9.2. El modelo de regresión lineal simple

Los modelos lineales ocupan un lugar central en los análisis econométricos, tanto por su interés metodológico como por su aplicación practica. En el supuesto más sencillo, estos modelos describen el comportamiento de cierta magnitud que denotamos por Y (respuesta o variable dependiente) a partir de una única causa X (variable independiente). Para cada valor concreto de la causa, la expresión del modelo lineal sería $Y = \beta_1 + \beta_2 X$, cuyos parámetros β_1 y β_2 tienen a menudo interesantes interpretaciones económicas.

A modo de ejemplo, en el modelo de consumo keynesiano tendríamos β_1 = Consumo fijo o autónomo y β_2 = Propensión Marginal al Consumo (PMgC), parámetros ambos de gran interés conceptual.

Es importante señalar que la variable independiente X no tiene carácter aleatorio sino que sus valores son conocidos. De este modo, para valores prefijados de la variable X (X_i) estudiamos el comportamiento de Y, que sí es una variable aleatoria sobre la cual es posible definir la distribución de probabilidad condicionada Y/X_i .

En efecto, la incertidumbre presente en todo modelo econométrico hace que el valor Y/X_i sea aleatorio ya que para cada X_i se obtiene $Y = \beta_1 + \beta_2 X_i + u_i$. Así pues, los desarrollos del modelo lineal se realizan asumiendo determinadas condiciones sobre el comportamiento probabilístico de la perturbación aleatoria u, hipótesis que pueden ser trasladadas a la variable dependiente Y.

Los supuestos básicos del modelo lineal son los siguientes:

- La perturbación tiene esperanza nula: $E(u_i) = 0,\ \forall i = 1, ..., n$

Este primer supuesto resulta coherente con la propia naturaleza de la perturbación aleatoria, ya que si ésta no tiene ningún componente sistemático se espera que se compensen las perturbaciones positivas y negativas conduciendo a un valor esperado nulo.

Dicha hipótesis se traduce en la siguiente esperanza para la variable dependiente:

$$E(Y/X_i) = \beta_1 + \beta_2 X_i\,,\ \ \forall i = 1,...,\text{n}$$

que garantiza que los valores esperados de Y se encuentran en la denominada *línea de regresión poblacional* $E(Y/X_i) = \beta_1 + \beta_2 X_i$

- La perturbación tiene varianza constante: $Var(u_i) = E\left(u_i^2\right) = \sigma^2\ ,\ \forall i = 1, ..., n$

Este supuesto, conocido como *homoscedasticidad* puede también ser expresado sobre la variable dependiente:

$$Var(Y/X_i) = Var(u_i) = \sigma^2\ ,\ \forall i = 1, ..., n$$

- Las perturbaciones correspondientes a distintas observaciones no se encuentran correlacionadas: $Cov(u_i, u_j) = E(u_i u_j) = 0\ ,\ i \neq j$

Teniendo en cuenta la relación entre la perturbación y la variable dependiente, esta hipótesis de *ausencia de correlación* puede también ser expresada sobre Y: $Cov(Y/X_i, Y/X_j) = 0$, $i \neq j$

Junto a estos supuestos básicos, se asume a menudo la hipótesis adicional de *normalidad*:

- Las perturbaciones aleatorias -y, como consecuencia, también la variable dependiente- se distribuyen normalmente.

Este supuesto queda justificado teniendo en cuenta que las perturbaciones pueden ser generadas por un conjunto numeroso de factores independientes entre sí, cuya actuación conjunta -según los teoremas límites- conduce a un modelo normal.

En una visión sintética, los supuestos anteriormente recogidos permitirán afirmar que las perturbaciones u son variables aleatorias incorreladas $u_i \approx \mathcal{N}(0, \sigma)$ y, como consecuencia, para cada valor fijado de X se tiene $Y/X_i \approx \mathcal{N}(\beta_1 + \beta_2 X_i, \sigma)$.

Obsérvese que Y/X_i se obtiene a partir de u_i mediante un cambio de origen, ya que $Y/X_i = \beta_1 + \beta_2 X_i + u_i$ siendo $\beta_1 + \beta_2 X_i$ constante para cada observación.

Así pues, el modelo de Y será normal siempre que lo sea u, y basta aplicar las propiedades de esperanza y varianza ante cambios de origen para obtener:

$$E(Y/X_i) = \beta_1 + \beta_2 X_i + E(u_i) = \beta_1 + \beta_2 X_i$$

$$Var(Y/X_i) = Var(u_i) = \sigma^2$$

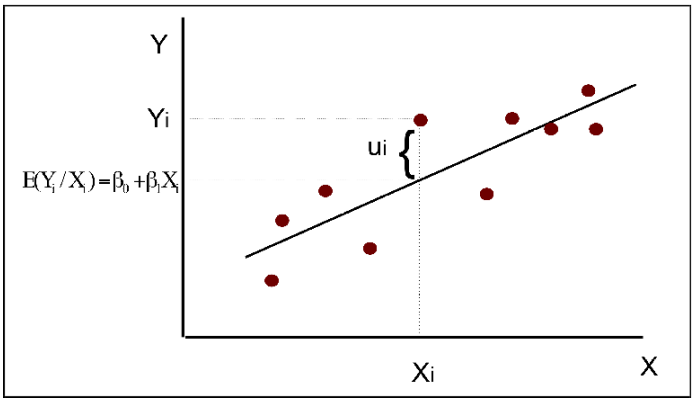
Como veremos en los epígrafes que siguen, los supuestos anteriormente enunciados resultan fundamentales en los procesos de estimación y contraste de los modelos econométricos.

	Supuestos básicos	
	Sobre u	Sobre Y
Esperanza	$E(u_i) = 0, \forall i = 1, \dots n$	$E(Y/X_i) = \beta_1 + \beta_2 X_i, \forall i$
Varianza	$Var(u_i) = \sigma^2, \forall i = 1, \dots n$	$Var(Y/X_i) = \sigma^2, \forall i$
Correlación	$Cov(u_i, u_j) = 0, \forall i \neq j = 1, \dots n$	$Cov(Y/X_i, Y/X_j) = 0, \forall i \neq j$
Modelo prob.	$u_i \approx \mathcal{N}(0, \sigma)$	$Y/X_i \approx \mathcal{N}(\beta_1 + \beta_2 X_i, \sigma)$

9.3. Estimación de los parámetros de regresión

En el modelo lineal simple, la línea de regresión poblacional viene dada por la recta $\beta_1 + \beta_2 X_i$ que recoge la componente sistemática del modelo y asigna a cada valor concreto de la variable explicativa (X_i) el correspondiente valor esperado de la variable dependiente: $E(Y/X_i) = \beta_1 + \beta_2 X_i$.

Figura 9.1.: Modelo de regresión lineal



Esta línea se corresponde con el lugar geométrico de las esperanzas condicionadas de la variable dependiente para cada una de las observaciones de la variable explicativa.

La diferencia entre los valores esperados recogidos por la línea de regresión poblacional y los verdaderos valores de Y que en la realidad aparecen asociados a la variable X es un error o perturbación que, como ya hemos comentado, tiene carácter aleatorio. Así, tal y como recoge la figura 9.1 es posible representar gráficamente los componentes sistemático $\beta_1 + \beta_2 X_i$ y aleatorio u_i asociados a cada X_i.

Dado que la línea de regresión poblacional es la traducción de un supuesto teórico, se trata de un valor poblacional desconocido que deberá ser estimado con base en la información muestral disponible. Así, a partir de un conjunto de observaciones de las variables estudiadas, se llegará a obtener la *línea de regresión muestral* $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$.

Esta línea de regresión estimada depende directamente de la información disponible, y por tanto adoptará valores diferentes para cada muestra. Como consecuencia, no disponemos de garantías referidas a cada recta concreta sino únicamente al procedimiento de estimación.

Los métodos más habituales para estimar las rectas de regresión son el mínimo cuadrático y el de máxima verosimilitud, que parten de filosofías distintas: en el primer caso, minimizar la suma de errores cuadráticos y en el segundo, maximizar la verosimilitud asociada a la muestra observada.

9.3.1. Estimación mínimo cuadrática

La línea de regresión muestral $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ proporciona un valor de la variable dependiente que en general no coincide con el verdadero, surgiendo así un error de estimación que denotamos por $\hat{u}_i$. De este modo, al igual que hemos visto para la población, es posible separar cada valor Y_i en sus componentes estimado y residual: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i$

La estimación por mínimos cuadrados se propone obtener rectas de regresión pró-

ximas a la información real, esto es, que minimicen los errores de estimación

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \left(\hat{\beta}_1 + \hat{\beta}_2 X_i\right)$$

Aunque una primera opción sería minimizar directamente la suma de los errores, debemos evitar que éstos se compensen, por lo cual acudimos a la agregación de errores cuadráticos. Esta decisión se debe a que la alternativa de agregar errores en valor absoluto presenta dificultades desde el punto de vista matemático (exigiría resolver un problema de programación lineal o bien un procedimiento de cálculo iterativo) y además no garantiza la existencia de una solución única.

Definición 9.2. Dado un modelo de regresión lineal simple $Y = \beta_1 + \beta_2 X + u$, los *estimadores mínimo cuadráticos* (EMC) de los parámetros de regresión vienen dados por las expresiones:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \ , \ \hat{\beta}_2 = \frac{S_{XY}}{S_X^2}$$

La deducción de estas expresiones se lleva a cabo minimizando la suma de los cuadrados de los residuos (SCR):

$$SCR = \sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

Aplicando la condición necesaria de extremo se igualan a cero las derivadas primeras de esta expresión respecto a los parámetros. Para llegar a las expresiones de los estimadores, basta desarrollar el sistema de ecuaciones normales mínimo cuadráticas

$$\frac{\partial SCR}{\partial \hat{\beta}_1} = 0 \Rightarrow -2\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) = 0 \Rightarrow \sum_{i=1}^{n} Y_i = \hat{\beta}_1 n + \hat{\beta}_2 \sum_{i=1}^{n} X_i \Rightarrow \bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$$

$$\frac{\partial SCR}{\partial \hat{\beta}_2} = 0 \Rightarrow -2\sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right) X_i = 0 \Rightarrow \sum_{i=1}^{n} Y_i X_i = \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2$$

obteniéndose de la primera ecuación $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ y, mediante sustitución de esta expresión en la segunda ecuación: $\hat{\beta}_2 = \frac{S_{XY}}{S_X^2}$. [Compruébese]

Los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ representan respectivamente el término independiente (ordenada en el origen) y la pendiente de la recta de regresión muestral. Así, $\hat{\beta}_1$ estima el valor esperado de Y para valores nulos de X mientras que $\hat{\beta}_2$ cuantifica la variación esperada en Y ante aumentos unitarios de X .

Obsérvese que los estimadores mínimo cuadráticos no podrían ser determinados en el caso de que la dispersión de X fuese nula (es decir, si la muestra sólo presenta un valor de X no tendría sentido buscar una explicación al comportamiento de Y en función de X).

Puede verse además que la pendiente estimada presenta igual signo que la covarianza entre las variables. Por lo que respecta al término independiente, en su estimación intervienen tanto la pendiente como los valores medios de las variables X e Y.

Proposición 9.1. *La estimación mínimo-cuadrática de un modelo lineal simple garantiza las siguientes propiedades:*

1. $\sum_{i=1}^{n} \hat{u}_i = 0$
2. $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$
3. $\sum_{i=1}^{n} X_i \hat{u}_i = 0$
4. $\sum_{i=1}^{n} Y_i \hat{u}_i = 0$

La demostración de estas propiedades es sencilla. Así, para comprobar la nulidad de la suma de los residuos de la regresión basta tener en cuenta la primera de las ecuaciones de mínimos cuadrados, ya que se tiene:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right) = 0$$

Partiendo de esta expresión se llega fácilmente a la segunda propiedad $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$
De modo similar, se comprueba la tercera propiedad, ya que:

$$\sum_{i=1}^{n} X_i \hat{u}_i = \sum_{i=1}^{n} X_i \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \right) = 0$$

donde hemos aplicado la segunda condición mínimo-cuadrática [Compruébese que se cumple $\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = 0$]

9.3.2. Estimación máximo verosímil

El método de máxima verosimilitud consiste en adoptar como estimadores aquellos valores de los parámetros que maximizan la probabilidad o verosimilitud de la muestra observada.

Definición 9.3. Dado un modelo de regresión lineal simple $Y = \beta_1 + \beta_2 X + u$, $u \approx \mathrm{N}(0,\sigma)$, los estimadores máximo verosímiles (EMV) de los parámetros β_1,β_2 y σ^2 vienen dados por las expresiones:

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} \quad , \quad \hat{\beta}_2 = \frac{S_{XY}}{S_X^2} \quad , \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n}$$

La obtención de los estimadores máximo verosímiles (abreviadamente EMV) se lleva a cabo a partir de la función de verosimilitud de la muestra que, asumiendo el supuesto de normalidad, viene dada por la expresión:

$$L(y_1, \cdots .y_n, \beta_1, \beta_2, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^n e^{-\frac{1}{2} \sum_{i=1}^{n} \frac{(y_i - \beta_1 - \beta_2 x_i)^2}{\sigma^2}}$$

Bajo el supuesto de normalidad para las perturbaciones se tiene $u \approx \mathcal{N}(0,\sigma)$ e $Y \approx \mathcal{N}(\beta_1+\beta_2 X,\sigma)$ con lo cual la función de verosimilitud depende de tres parámetros: β_1, β_2 y σ^2 y se obtiene mediante producto de funciones de densidad:

$$\begin{aligned} L(y_1,\cdots,y_n,\beta_1,\beta_2,\sigma^2) = & \prod_{i=1}^{n} f(y_i,,\beta_1,\beta_2,\sigma^2) = \prod_{i=1}^{n}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) e^{-\frac{1}{2}\frac{(y_i-\beta_1-\beta_2 x_i)^2}{\sigma^2}} = \\ = & \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\beta_1-\beta_2 x_i)^2}{\sigma^2}} \end{aligned}$$

Para obtener los EMV de los parámetros esta función se transforma mediante logaritmos, y posteriormente se aplican las correspondientes condiciones de extremo.

$$\ln L(y_1,\cdots,y_n,\beta_1,\beta_2,\sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\beta_1-\beta_2 x_i)^2}{\sigma^2}$$

$$\frac{\partial \ln L(y_1,\cdots,y_n,\beta_1,\beta_2,\sigma^2)}{\partial\beta_1} = 0 \Rightarrow -\frac{2}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 x_i) = 0$$

$$\frac{\partial \ln L(y_1,\cdots,y_n,\beta_1,\beta_2,\sigma^2)}{\partial\beta_2} = 0 \Rightarrow -\frac{2}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 x_i)\,x_i = 0$$

$$\frac{\partial \ln L(y_1,\cdots,y_n,\beta_1,\beta_2,\sigma^2)}{\partial\sigma^2} = 0 \Rightarrow -\frac{n}{2\sigma^2} + \frac{\sum_{i=1}^{n}(y_i-\beta_1-\beta_2 x_i)^2}{2\sigma^4} = 0$$

Las dos primeras ecuaciones coinciden con las obtenidas anteriormente por el procedimiento mínimo-cuadrático. Como consecuencia, los EMV para los parámetros β_1 y β_2 son coincidentes con los EMC anteriormente estudiados.

Por lo que se refiere al estimador máximo verosímil de la varianza poblacional σ^2 , de la tercera ecuación se obtiene -una vez conocidos $\hat{\beta}_1$ y $\hat{\beta}_2$ - la expresión

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{u}_i^2}{n}$$

que resulta sesgada, pero sin embargo es consistente.

9.3.3. Características y propiedades de los estimadores

Además de ser adecuado desde los puntos de vista metodológico y descriptivo, el método mínimo-cuadrático garantiza también un buen comportamiento inferencial para los estimadores.

Así, puede demostrarse fácilmente que los estimadores mínimo cuadráticos son insesgados, consistentes y óptimos, resultado este último que viene recogido en el denominado *Teorema de Gauss-Markov.*

Teorema 9.1. *Dentro de la familia de estimadores lineales e insesgados, los Estimadores Mínimo Cuadráticos son óptimos en el sentido de que presentan mínima varianza.*

Este resultado, que tiene gran trascendencia en la modelización econométrica, permite la designación de los estimadores mínimo cuadráticos como ELIO (Estimadores Lineales Insesgados Optimos) o, utilizando la terminología inglesa, BLUE (Best Linear Unbiased Estimators).

La demostración de la ausencia de sesgo y del teorema de Gauss-Markov se recogen en el epígrafe 10.1.2 para el modelo lineal múltiple.

En el teorema de Gauss-Markov se incluyen en realidad dos resultados distintos: el primero es el enfoque de mínimos cuadrados debido a Gauss (1821) mientras que posteriormente Markov (1900) planteó el enfoque de mínima varianza.

Debemos señalar además que la formulación del teorema de Gauss-Markov se basa únicamente en la dispersión de los estimadores pero no exige el supuesto de normalidad para las perturbaciones aleatorias. No obstante, bajo el supuesto adicional de normalidad, es posible garantizar que los EMC tienen varianza mínima para todas las clases de estimadores insesgados, sean éstos lineales o no. Este resultado, que se debe a Rao, es un postulado más fuerte que el de Gauss-Markov porque, asumiendo condiciones más restrictivas (normalidad), no se restringe a la clase de los estimadores lineales.

Las varianzas de los estimadores mínimo cuadráticos viene dadas por las siguientes expresiones:

$$Var\left(\hat{\beta}_2\right) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad , \quad Var\left(\hat{\beta}_1\right) = \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

donde la dispersión poblacional σ^2 es habitualmente desconocida y suele ser estimada por la varianza muestral S^2, dando lugar a las expresiones:

$$\widehat{Var}\left(\hat{\beta}_2\right) = \frac{S^2}{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} \quad , \quad \widehat{Var}^2\left(\hat{\beta}_1\right) = \frac{S^2 \sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}$$

Denominaremos *varianza muestral* S^2 al estimador insesgado de la varianza poblacional que se define como cociente entre la suma de errores cuadráticos y los grados de libertad $(n-2)$, es decir:

$$S^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}$$

Como ya hemos visto, los grados de libertad se obtienen como diferencia entre el total de observaciones muestrales y las restricciones lineales impuestas (en este caso, antes de calcular la varianza debemos conocer las estimaciones de los parámetros, que son dos restricciones).

En los análisis de regresión en principio podríamos considerar como medida adecuada la varianza del error o bien su desviación típica. Sin embargo, resulta más adecuado el error estándar de la regresión, dado por una expresión similar, en la que aparece como denominador (n-2), que es el número

de grados de libertad.

En las propiedades comentadas hasta ahora no hemos utilizado el supuesto de normalidad de las perturbaciones aleatorias. Sin embargo, esta hipótesis resulta necesaria para obtener los estimadores máximo verosímiles y también para garantizar modelos probabilísticos asociados a los estimadores mínimo cuadráticos y a la dispersión muestral.

Bajo la hipótesis de normalidad de la perturbación la variable dependiente sigue también una distribución normal y ya hemos visto que los estimadores máximo verosímiles de β_1 y β_2 coinciden con sus EMC. Además, estos EMV son de mínima varianza en la clase de los estimadores insesgados (sean éstos lineales o no), resultado conocido como *Teorema de Rao* que extiende las conclusiones de Gauss-Markov.

Por lo que se refiere a la distribución probabilística de los estimadores, bajo la hipótesis $u \approx \mathcal{N}(0, \sigma)$ es posible garantizar:

$$\hat{\beta}_1 \approx N(\beta_1, \sigma_{\hat{\beta}_1})$$

$$\hat{\beta}_2 \approx N(\beta_2, \sigma_{\hat{\beta}_2})$$

$$\frac{(n-2)S^2}{\sigma^2} \approx \chi^2_{n-2}$$

distribuciones en las que se basan las inferencias sobre el modelo lineal simple.

9.3.4. Construcción de las discrepancias tipificadas

A partir de los estimadores derivados anteriormente nos interesa llegar a obtener discrepancias tipificadas con modelo probabilístico conocido, que permitan realizar procesos inferenciales de estimación y contraste.

Adoptamos como punto de partida los errores aleatorios cometidos en la estimación de un parámetro $e_{\hat{\beta}} = \hat{\beta} - \beta$, cuyas características son:

$$E(e_{\hat{\beta}}) = 0 \ , \ \ Var(e_{\hat{\beta}}) = Var(\hat{\beta})$$

y podemos llevar a cabo una tipificación:

$$d_{\hat{\beta}} = \frac{e_{\hat{\beta}} - E(e_{\hat{\beta}})}{\sigma_{e_{\hat{\beta}}}} = \frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}}$$

obteniendo así una discrepancia tipificada con esperanza nula y dispersión unitaria.

Si además asumimos el supuesto de normalidad para la perturbación, los estimadores son normales y la discrepancia sigue también un modelo normal $d_{\hat{\beta}} \approx \mathcal{N}(0, 1)$.

No obstante, dado que en general se desconoce la varianza poblacional σ^2 debemos trabajar con sus correspondientes estimaciones S^2 . En este caso, se llega a la expresión de la discrepancia:

$$d_{\hat{\beta}} = \frac{\hat{\beta} - \beta}{S_{\hat{\beta}}}$$

distribuida según un modelo t de Student con $n-2$ grados de libertad.

Para comprobar que la expresión anterior sigue un modelo t de Student, basta tener presente el resultado anterior:

$$\frac{\hat{\beta} - \beta}{\sigma_{\hat{\beta}}} \approx \mathcal{N}(0,1)$$

y aplicar el teorema de Fisher que, gracias a la normalidad de la población, garantiza

$$\frac{(n-2)S^2}{\sigma^2} \approx \chi^2_{n-2}$$

Como consecuencia, es posible construir una nueva discrepancia normalizada en los siguientes términos:

$$d_{\hat{\beta}} = \frac{\left(\frac{\hat{\beta}-\beta}{\sigma_{\hat{\beta}}}\right)}{\sqrt{\frac{(n-2)S^2}{\sigma^2(n-2)}}} = \frac{\hat{\beta} - \beta}{S_{\hat{\beta}}}$$

Teniendo en cuenta que el numerador de esta expresión sigue una distribución $\mathcal{N}(0,1)$ y su denominador -que es independiente de la variable anterior- es el cociente de una chi-cuadrado entre sus grados de libertad, queda justificado que la expresión de la discrepancia se distribuye en este caso según un modelo t de Student con $n-2$ g.l.

Las expresiones deducidas son aplicables tanto al estimador $\hat{\beta}_2$ del coeficiente de regresión como a $\hat{\beta}_1$, estimador del término independiente β_1. De ahí que en el apartado que sigue abordemos la construcción de intervalos para un parámetro genérico β.

Si nuestro objetivo es la dispersión poblacional σ^2, la correspondiente discrepancia debe ser construida a partir del error relativo

$$e^R_{S^2} = \frac{S^2}{\sigma^2}$$

que, con sólo multiplicar por los g.l. da lugar a la expresión

$$d_{S^2} = \frac{(n-2)S^2}{\sigma^2} \approx \chi^2_{n-2}$$

9.3.5. Obtención de intervalos de confianza

Una vez conocidas las expresiones de las discrepancias es posible llevar a cabo estimaciones de los parámetros que complementen las estimaciones puntuales con sus correspondientes márgenes de error.

Siguiendo la metodología general para la construcción de intervalos de confianza de β comenzaremos por fijar el nivel de confianza deseado $1-\alpha$ determinando a

continuación constantes k tales que se cumpla:

$$P\left(\left|d_{\hat{\beta}}\right| \leq k\right) = 1 - \alpha$$

A partir de dicha igualdad se determinan las siguientes expresiones aleatorias en las que, con una probabilidad $1-\alpha$, se encuentra el parámetro desconocido:

$\left[\hat{\beta} - k\sigma_{\hat{\beta}}, \hat{\beta} + k\sigma_{\hat{\beta}}\right]$ con σ conocido y k calculado en las tablas $\mathcal{N}(0,1)$

$\left[\hat{\beta} - kS_{\hat{\beta}}, \hat{\beta} + kS_{\hat{\beta}}\right]$ con σ desconocido y k calculado en las tablas t_{n-2}

De modo similar, es posible construir intervalos de confianza para la varianza poblacional σ^2 con sólo tener presente que la discrepancia vendrá definida como:

$$d_{S^2} = \frac{(n-2)S^2}{\sigma^2} \approx \chi^2_{n-2}.$$

Siguiendo el procedimiento general de construcción de intervalos, para un nivel de confianza dado $1-\alpha$ se buscan en tablas dos valores k_1 y k_2 tales que:

$$P\left(\frac{(n-2)S^2}{\sigma^2} < k_1\right) = P\left(\frac{(n-2)S^2}{\sigma^2} > k_2\right) = \frac{\alpha}{2}$$

Se llega así al siguiente intervalo de confianza para σ^2:

$$\left[\frac{(n-1)S^2}{k_2}, \frac{(n-1)S^2}{k_1}\right]$$

9.4. Contrastes asociados a un modelo. Evaluación de la bondad

En numerosas ocasiones resulta interesante contrastar algún supuesto relativo a los parámetros de un modelo, problema que abordaremos aplicando la metodología general del contraste de hipótesis, y utilizando como punto de partida las expresiones deducidas para las discrepancias $d_{\hat{\beta}}$.

Cuando nuestro modelo se somete a un contraste estadístico resultan relevantes tanto los supuestos iniciales como las hipótesis que pretendemos contrastar, y a menudo están basadas en la teoría económica.

Entre los primeros es importante el supuesto de normalidad, que como hemos visto condiciona la distribución probabilística de los estimadores y cuyo contraste abordaremos en un apartado posterior. Nos ocuparemos aquí de los contrastes referidos a uno o varios parámetros del modelo, que admiten enunciados tanto unilaterales como bilaterales.

De entre estos contrastes nos interesan especialmente los denominados *tests básicos de significación*, que se diseñan para validar el modelo planteado.

El contraste básico aparece asociado a la pregunta ¿afecta verdaderamente X a Y?

y se plantea en los términos siguientes:

$$\boxed{\begin{array}{ll} H_0: & \beta_2 = 0 \\ H_1: & \beta_2 \neq 0 \end{array}}$$

de modo que, si la hipótesis nula es cierta, el modelo propuesto no tiene sentido al ser $E(Y/X_i) = \beta_1$ para cualquier posible valor de X.

Como consecuencia de este planteamiento, si la información muestral disponible conduce al rechazo de la hipótesis nula, concluiremos que β_2 es *significativamente distinto de 0*, con lo cual existe evidencia estadística de que X afecta a Y, y en consecuencia tiene sentido plantear un modelo $Y = \beta_1 + \beta_2 X + u$.

Si por el contrario se obtiene una información muestral que no resulta significativa para rechazar la hipótesis nula, entonces nuestro modelo no queda validado por la información estadística.

Obsérvese que el contraste básico de significación va referido tan sólo al coeficiente β_2. Resulta sencillo comprobar que la nulidad del parámetro β_1 no invalidaría el modelo sino que únicamente supondría que la recta pasa por el origen.

Al igual que en cualquier contraste de significación, es posible optar por el procedimiento clásico, determinando regiones críticas para el estimador $\hat{\beta}_2$. No obstante, resulta más habitual el método del nivel crítico, que parte de la observación de la discrepancia muestral $d^*_{\hat{\beta}}$ para la que se calcula el nivel crítico $p = P\left(\left|d_{\hat{\beta}}\right| > \left|d^*_{\hat{\beta}}\right| / H_0\right)$. Si esta probabilidad es suficientemente baja, el resultado podrá ser calificado de significativo para rechazar.

Aunque parezca contradictorio, un resultado significativo conlleva el rechazo de la hipótesis pero en cambio valida el modelo propuesto. En este sentido, los contrastes básicos de los modelos econométricos presentan rasgos diferenciales con respecto a los vistos hasta ahora, ya que plantean como hipótesis la nulidad del coeficiente β_2 que acompaña a la variable explicativa, con la esperanza de rechazar este supuesto.

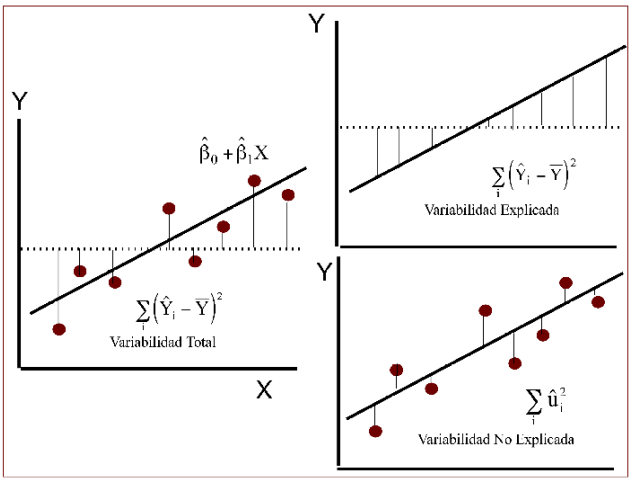
Habitualmente este contraste se lleva a cabo mediante el estadístico t de Student, ya que la dispersión poblacional es desconocida. Este es uno de los resultados básicos proporcionados por cualquier paquete econométrico y nos informa sobre si la muestra apoya la introducción de la variable X como explicativa.

Sin embargo, como más adelante justificaremos, cuando el modelo incluye varias variables explicativas debemos ser prudentes con la interpretación de los resultados de los contrastes t, ya que éstos se ven influidos por la existencia de correlación lineal entre las variables explicativas. Este fenómeno es lo que habitualmente se denomina problema de la *multicolinealidad*.

Un contraste de validez equivalente al anterior puede ser planteado a partir del análisis de la capacidad explicativa del modelo. En este caso, si para una misma observación consideramos su valor observado o verdadero (Y_i), su valor promedio en la muestra $\bar{Y}$ y su valor estimado por la línea de regresión $(\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i)$ se cumple:

$$\left(Y_i - \bar{Y}\right) = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

Figura 9.2.: Análisis de varianza



expresiones que, en términos cuadráticos y agregadas, conducen al procedimiento de *análisis de la varianza* (ANOVA), basado en la igualdad:

$$\sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

donde el primer término recoge la dispersión respecto a la media de la variable observada (denominado *variabilidad total*, VT, o *suma total de cuadrados*), el sumando

$$\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2 = \hat{\beta}_2^2 \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2$$

es la dispersión respecto a la media de la variable estimada, (*variabilidad explicada*, VE, o *suma de cuadrados explicados*) y por último

$$\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \hat{u}_i^2$$

es la *variabilidad no explicada*, VNE, o *suma de cuadrados de los residuos*.

La variabilidad explicada viene dada por la expresión $\sum_{i=1}^{n} \left(\hat{Y}_i - \bar{Y}\right)^2$ que recoge las desviaciones cuadráticas de los valores estimados respecto a su media [Compruébese que se cumple $\bar{Y} = \bar{\hat{Y}}$].

Teniendo en cuenta además la definición $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ se obtiene la expresión alternativa $\hat{\beta}_2^2 \sum_{i=1}^{n} (X_i - \bar{X})^2$ que a menudo resulta más cómoda.

La descomposición anterior es utilizada habitualmente para construir el *coeficiente de determinación* R^2 que se define como la proporción de la variación total de Y que

viene explicada por el modelo:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = 1 - \frac{\sum_{i=1}^{n}\hat{u}_i^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}$$

Este indicador adopta valores comprendidos entre 0 y 1, y su resultado aumenta con la bondad o capacidad explicativa del modelo.

Para los modelos lineales simples, el coeficiente de determinación coincide con el cuadrado del coeficiente de correlación lineal, ya que se cumple:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = \hat{\beta}_2^2\frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} = \left(\frac{S_{XY}}{S_X^2}\right)^2\frac{S_X^2}{S_Y^2} = \left(\frac{S_{XY}}{S_X S_Y}\right)^2 = r_{XY}^2$$

Además de su interpretación como proporción de la variación de Y que es explicada por el modelo es necesario tener presente que el coeficiente de determinación R^2 viene afectado por la naturaleza del modelo estudiado. Así, es habitual obtener valores elevados del coeficiente en modelos estimados a partir de datos de serie temporal, ya que la existencia de tendencia puede hacer que las variables investigadas evolucionen en paralelo. Por el contrario, cuando los modelos se estiman a partir de datos de corte transversal los coeficientes de determinación adoptan valores considerablemente más bajos.

Podemos ahora plantearnos cómo sería posible utilizar la expresión anterior del análisis de la varianza para contrastar la hipótesis nula $H_0 : \beta_2 = 0$. En principio, bajo dicho supuesto parece lógico esperar que el estimador $\hat{\beta}_2$, y como consecuencia también la variabilidad explicada, adopten valores bajos.

La expresión utilizada para llevar a cabo el contraste de validez del modelo a partir de información muestral se basa en la comparación por cociente entre las sumas de cuadrados explicados y residuales, es decir:

$$\frac{\hat{\beta}_2^2\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{\sum_{i=1}^{n}\hat{u}_i^2}$$

expresión que, debidamente tipificada (dividiendo el denominador entre sus grados de libertad $n-2$) conduce, bajo la hipótesis nula, a una discrepancia con distribución F de Snedecor:

$$\frac{\hat{\beta}_2^2\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}{S^2} \approx F_{n-2}^1$$

El contraste se llevará a cabo calculando el valor muestral de esta discrepancia y su correspondiente nivel crítico p. En el caso de que éste adopte un valor suficientemente

bajo rechazaremos la hipótesis de nulidad de β_2 y por tanto validaremos nuestro modelo.

Para la deducción de esta distribución debemos tener en cuenta que bajo H_0 el estimador presenta una distribución $\hat{\beta}_2 \approx \mathcal{N}\left(\beta_2, \sigma_{\hat{\beta}_2}\right)$ y como consecuencia:

$$\frac{\hat{\beta}_2}{\sigma_{\hat{\beta}_2}} = \frac{\hat{\beta}_2}{\frac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}}} \approx \mathcal{N}(0,1)$$

expresión que elevada al cuadrado seguirá un modelo chi-cuadrado:

$$\frac{\hat{\beta}_2^2 \sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2} \approx \chi_1^2$$

Utilizando además el resultado

$$\frac{(n-2)S^2}{\sigma^2} \approx \chi_{n-2}^2$$

y teniendo en cuenta la independencia entre ambas expresiones, es posible definir una F de Snedecor con grados de libertad 1 en el numerador y $n-2$ en el denominador:

$$\frac{\frac{\hat{\beta}_2^2 \sum_{i=1}^{n}(X_i-\bar{X})^2}{\sigma^2}}{\frac{(n-2)S^2}{(n-2)\sigma^2}} = \frac{\hat{\beta}_2^2 \sum_{i=1}^{n}(X_i-\bar{X})^2}{S^2} \approx F_{n-2}^1$$

Como hemos visto en el capítulo 6, cualquier modelo F de Snedecor con un solo grado de libertad en el denominador puede ser expresado como el cuadrado de una t de Student con grados de libertad los del denominador. Como consecuencia de esta propiedad, el estadístico F definido coincidirá con el cuadrado de la t de Student utilizado en los contrastes individuales, relación que garantiza la coherencia de resultados entre ambos tipos de contrastes.

Además de los tests básicos, puede resultar interesante llevar a cabo otros contrastes para los parámetros, que en general traducirán restricciones impuestas por la teoría económica o bien resultados obtenidos en investigaciones previas.

Así, por ejemplo, según la ley psicológica fundamental de la teoría keynesiana, la propensión marginal al Consumo debe ser no superior a la unidad, de modo que sobre el modelo $C = \beta_1 + \beta_2 R$ sería interesante contrastar $H_0 : \beta_2 \leq 1$ frente a $H_1 : \beta_2 > 1$.

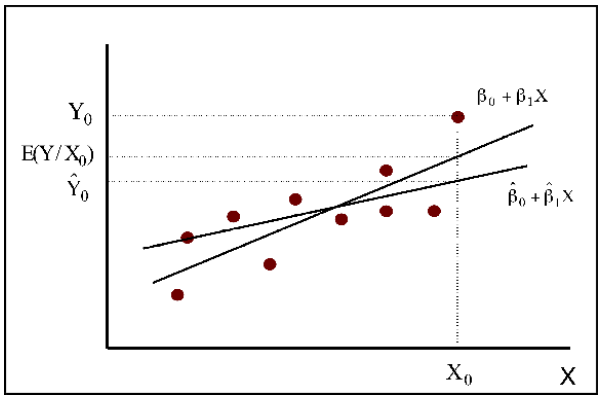
De modo similar, la teoría económica define funciones de demanda con pendiente negativa, de modo que sobre un modelo de demanda $C = \beta_1 + \beta_2 P$ deberíamos contrastar la hipótesis $H_0 : \beta_2 \leq 0$ frente a $H_1 : \beta_2 > 0$.

Por otra parte, si en estudios previos sobre la población que analizamos se ha estimado un gasto fijo en alimentación de 80 unidades monetarias, podríamos someter a contraste este valor del parámetro, planteando $H_0 : \beta_1 = 80$ frente a $H_1 : \beta_1 \neq 80$.

9.5. Predicción

Gran parte de los modelos econométricos tienen por objetivo la realización de predicciones. Una vez que el modelo ya ha sido validado, para realizar predicciones basta

Figura 9.3.: Predicción



con sustituir el valor de la variable explicativa X_0 en el modelo estimado:

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

Esta predicción es una variable aleatoria con las siguientes características:

$$E\left(\hat{Y}_0\right) = E\left(\hat{\beta}_1 + \hat{\beta}_2 X_0\right) = \beta_1 + \beta_2\, X_0 = E\left(Y/X_0\right)$$

$$Var\left(\hat{Y}_0\right) = Var\left(\hat{\beta}_1 + \hat{\beta}_2 X_0\right) = \sigma^2 \left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right)$$

Dado que la predicción $\hat{Y}_0$ es una aproximación del verdadero valor Y_0, debemos considerar el error de predicción $e_{\hat{Y}_0} = Y_0 - \hat{Y}_0$ que puede ser expresado como:

$$e_{\hat{Y}_0} = Y_0 - \hat{Y}_0 = (Y_0 - E(Y/X_0)) + \left(E(Y/X_0) - \hat{Y}_0\right)$$

Esta descomposición, que aparece recogida gráficamente en la figura siguiente, identifica dos componentes en el error de predicción: uno de ellos se debe a la propia dispersión poblacional ($u_0 = Y_0 - (Y/X_0)$), y viene representado gráficamente por la distancia del punto a la recta de regresión poblacional, mientras el otro componente es de carácter muestral ($E(Y/X_0) - \hat{Y}_0$), y se corresponde con la distancia entre la recta de regresión poblacional y la estimada en el punto considerado.

Según fijemos nuestra atención en el error total de predicción o sólo en el componente muestral, podemos construir diferentes bandas de confianza para la predicción, cuyos rasgos se recogen en la siguiente tabla:

Predicción	Error de predicción	Discrepancia tipificada	IC
Para Y_0	Total $e_{\hat{Y}_0} = Y_0 - \hat{Y}_0$	$d_{Y_0-\hat{Y}_0} = \frac{Y_0-\hat{Y}_0}{S_{Y_0-\hat{Y}_0}} \approx t_{n-2}$	$\hat{Y}_0 \pm kS_{Y_0-\hat{Y}_0}$
Para $E(Y/X_0)$	Muestral $\left(E(Y/X_i) - \hat{Y}_0\right)$	$d_{\hat{Y}_0} = \frac{E(Y/X_0)-\hat{Y}_0}{S_{\hat{Y}_0}} \approx t_{n-2}$	$\hat{Y}_0 \pm kS_{\hat{Y}_0}$

En la primera de las situaciones el objetivo de la predicción es el *valor verdadero* Y_0 (que aparece en la nube de puntos) con lo cual el error de predicción presenta las dos componentes muestral y poblacional. Como consecuencia se tiene:

$$E\left(e_{\hat{Y}_0}\right) = E(Y_0) - E\left(\hat{Y}_0\right) = 0$$

$$Var(e_{\hat{Y}_0}) = Var\left(Y_0 - \hat{Y}_0\right) = Var(Y_0) + Var\left(\hat{Y}_0\right) = \sigma^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

si bien, al ser en general σ^2 desconocida, esta última expresión debe ser estimada mediante:

$$S_{Y_0-\hat{Y}_0} = S^2\left(1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

Así pues, se obtiene el siguiente intervalo de confianza para Y_0:

$$\left[\hat{Y}_0 - kS_{Y_0-\hat{Y}_0}, \hat{Y}_0 + kS_{Y_0-\hat{Y}_0}\right]$$

donde k se obtiene en tablas de la distribución t de Student con $n-2$ g.l. para el nivel de confianza fijado.

Si en cambio deseamos construir bandas de confianza para el valor esperado $E(Y/X_0)$ estamos considerando tan sólo el error muestral, para el cual se tiene:

$$E\left(E(Y/X_0) - \hat{Y}_0\right) = 0$$

$$Var\left(E(Y/X_0) - \hat{Y}_0\right) = Var\left(\hat{Y}_0\right) = \sigma^2\left(\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum\limits_{i=1}^n (X_i - \bar{X})^2}\right)$$

Para la construcción de intervalos esta varianza se aproxima utilizando la información muestral y la expresión del IC para $E(Y/X_0)$ será entonces:

$$\left[\hat{Y}_0 - kS_{\hat{Y}_0}, \hat{Y}_0 + kS_{\hat{Y}_0}\right]$$

En la práctica resulta habitual obtener predicciones y bandas de confianza para el valor esperado. Así, en el ejemplo de la función de consumo keynesiana nuestro objetivo no sería predecir el consumo de una familia concreta, sino predecir el consumo esperado para las familias que perciben una renta determinada X_0.

10. El modelo lineal múltiple

Con frecuencia la especificación de un modelo resulta más realista si consideramos numerosas variables explicativas (de hecho, ya hemos comentado que raramente se podrá aglutinar en una variable explicativa única todas las posibles causas de cierto efecto). Por ello, la extensión natural del modelo de regresión lineal simple analizado en el epígrafe anterior será el modelo de regresión lineal múltiple.

Consideraremos una especificación lineal del tipo:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$$

en la que aparecen k parámetros $(\beta_1, \beta_2, \ldots, \beta_k)$ y $k-1$ variables explicativas, que designamos por $X_2, \ldots, X_k$.

Por lo que se refiere a u, como ya hemos visto anteriormente es una perturbación aleatoria originada por múltiples causas irrelevantes que, al actuar conjuntamente, tienen un efecto no despreciable. Por tanto, u es una v.a. no observable, y como consecuencia también el regresando Y es una variable aleatoria.

Una expresión más genérica sería $Y = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u$, que conduce a la anterior si asumimos la existencia de un término independiente, esto es, $X_1 = 1$.

Nuestro objetivo es aproximar los k parámetros que representan la relación existente entre las variables a partir de información muestral sobre las mismas. En concreto, asumimos que disponemos de una muestra de tamaño n, que nos conduce al sistema de ecuaciones:

$$\begin{array}{c} Y_1 = \beta_1 + \beta_2 X_{21} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 = \beta_1 + \beta_2 X_{22} + \cdots + \beta_k X_{k2} + u_2 \\ \vdots \quad \cdots \quad \ddots \quad \cdots \quad \vdots \\ Y_n = \beta_1 + \beta_2 X_{2n} + \cdots + \beta_k X_{kn} + u_n \end{array}$$

cuya expresión matricial es:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{10.1}$$

donde:

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} \; ; \; \mathbf{X} = \begin{pmatrix} 1 & X_{21} & \cdots & X_{k1} \\ 1 & X_{22} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{2n} & \cdots & X_{kn} \end{pmatrix} \; ; \; \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} \; ; \; \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}$$

Esta notación condensada resulta más operativa cuando se trabaja con un modelo

general, por lo cual será la que adoptemos a partir de ahora.

10.1. Estimación

El modelo genérico anteriormente introducido se denomina *modelo lineal múltiple*, y su estudio se lleva a cabo asumiendo ciertas hipótesis de trabajo que explicitaremos a continuación. No todas las hipótesis tienen el mismo carácter ni resultan igualmente restrictivas en la práctica. La más genérica de todas es la referida a la *forma funcional* del modelo, respecto a la que se asume la linealidad.

La expresión $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ conecta la variable dependiente (regresando) con las variables explicativas (regresores) y la perturbación mediante una relación lineal, y es equivalente al supuesto de linealidad en los parámetros.

Por lo que se refiere a la *perturbación aleatoria*, se asumen los siguientes supuestos, análogos a los vistos para el modelo simple:

- La perturbación $\mathbf{u}$ es una v.a. no observable de esperanza nula: $E(\mathbf{u}) = \mathbf{0}$

  Esta hipótesis, que no es contrastable, equivale a admitir que los efectos de las variables incluidas en el término de perturbación tienden a compensarse por término medio. Incluso en el caso de que los efectos no se compensasen exactamente y se obtuviesen valores esperados no nulos, éstos podrían ser acumulados en el término constante del modelo de regresión, con lo cual la hipótesis no plantea problemas.

- La matriz de varianzas-covarianzas de la perturbación viene dada por la expresión:

$$Cov(\mathbf{u}) = E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I_n}$$

  Esta hipótesis recoge dos supuestos ya vistos en el caso simple: la *homoscedasticidad* y la *ausencia de correlación* entre las perturbaciones. El primero de ellos exige que las perturbaciones aleatorias presenten varianza constante:

$$Var(u_i) = E(u_i^2) = \sigma^2. \ , \ \forall i = 1, ..., n$$

  mientras que la ausencia de correlación entre las perturbaciones exige:

$$Cov(u_i, u_j) = E(u_i u_j) = 0 \ , \ \forall i \neq j$$

  Así pues, la matriz de varianzas-covarianzas de las perturbaciones puede ser expresada como:

$$Cov(\mathbf{u}) = E(\mathbf{u}\mathbf{u}') = \begin{pmatrix} E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n) \\ E(u_2u_1) & E(u_2^2) & \cdots & E(u_2u_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(u_nu_1) & E(u_nu_2) & \cdots & E(u_n^2) \end{pmatrix}$$

$$= \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2\mathbf{I_n}$$

Las hipótesis de homoscedasticidad y de *ausencia de correlación* pueden ser contrastadas a partir de los residuos mínimo cuadráticos y, como veremos en un apartado posterior, su incumplimiento origina algunos problemas en el modelo de regresión.

La hipótesis de *homoscedasticidad* puede no cumplirse cuando se trabaja con datos de corte transversal, en cuyo caso las perturbaciones se denominan heteroscedásticas. Por su parte, el supuesto de ausencia de correlación entre las perturbaciones resulta especialmente restrictivo cuando trabajamos con datos en serie temporal, ya que a menudo la perturbación aleatoria asociada a un período t puede estar correlacionada con la correspondiente al período anterior $t-1$.

- En ocasiones se admite un supuesto adicional para el vector de perturbaciones aleatorias: la distribución *normal multivariante*. Esta hipótesis parece justificada si tenemos en cuenta la heterogeneidad de factores que contribuyen a generar el error aleatorio y que se pueden suponer independientes entre sí, haciendo posible la aplicación del TCL.

Podemos también realizar supuestos sobre los regresores, para los cuales se asume:

- La matriz de regresores $\mathbf{X}$ es fija, es decir, adopta los mismos valores para distintas muestras. Esta hipótesis de regresores no estocásticos, que es admisible para las ciencias experimentales, puede sin embargo resultar restrictiva en ciencias sociales, ya que los datos se obtienen habitualmente mediante encuestas y vienen afectados por numerosas fuentes de error.

  En el caso de que los regresores tuvieran carácter estocástico, el efecto sobre el modelo no sería grave siempre que los regresores no se encontrasen correlacionados con la perturbación aleatoria, supuesto que puede ser contrastado mediante el test de Hausman.

- La matriz de regresores tiene rango k, esto es, $\rho(\mathbf{X}) = k$. Dado que la matriz $\mathbf{X}$ tiene k columnas (tantas como parámetros) y n filas (observaciones muestrales), esta hipótesis resume dos supuestos: por una parte, la información estadística disponible sobre el conjunto de variables observables debe ser suficientemente amplia para llevar a cabo la solución del modelo. Así pues, el número de datos (n) debe ser superior al de parámetros del modelo (k). Por otra parte, las columnas de la matriz $\mathbf{X}$ deben ser linealmente independientes, es decir, no debe existir relación lineal exacta entre los regresores del modelo.

  En el caso de que existiera relación lineal entre algún subconjunto de regresores, el rango de $\mathbf{X}$ sería inferior a k y por tanto, como veremos más adelante, no sería posible determinar los estimadores del modelo. En la práctica no suelen presentarse relaciones lineales exactas entre las variables explicativas, pero en cambio sí resulta frecuente un cierto grado de relación lineal entre los regresores.

Por último, podemos especificar una *hipótesis referida a los parámetros*:

- β es un vector fijo.

Este supuesto, que puede ser contrastado, equivale a asumir la existencia de una estructura única válida para todo el período de observación y el horizonte de predicción del fenómeno estudiado, y resulta de gran utilidad.

En síntesis, el *modelo básico de regresión lineal* puede ser descrito mediante las siguientes expresiones:

$$\boxed{\begin{array}{c}\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \\ E(\mathbf{u}) = \mathbf{0} \\ E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I}_n \\ \rho(\mathbf{X}) = k < n\end{array}}$$

10.1.1. Estimadores mínimo cuadráticos y máximo verosímiles

Definición 10.1. Dado un modelo lineal $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, $\mathbf{u} \simeq \mathbf{N}\left(\mathbf{0}, \sigma^2\mathbf{I_n}\right)$, la estimación mínimo cuadrática (MC) del vector de parámetros coincide con su estimación máximo verosímil (MV) y ambas vienen dadas por la expresión:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

La obtención de los estimadores mínimo cuadráticos ordinarios (MCO) en el modelo básico es análoga a la ya vista para los modelos simples, esto es, parte de la minimización de la expresión

$$\sum_{i=1}^{n} \hat{u}_i^2$$

que en notación matricial viene dada por:

$$\begin{aligned}\hat{\mathbf{u}}'\hat{\mathbf{u}} = \left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y} - X\hat{\boldsymbol{\beta}}\right) &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'X'\mathbf{y} - \mathbf{y}'X\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}} \\ &= \mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'X'\mathbf{y} + \hat{\boldsymbol{\beta}}'X'X\hat{\boldsymbol{\beta}}\end{aligned}$$

expresión en la que hemos tenido en cuenta que $\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = \mathbf{y}'\mathbf{X}\hat{\boldsymbol{\beta}}$. Exigiendo a esta expresión la condición necesaria de mínimo, se obtiene el vector de estimadores mínimo cuadráticos como:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Para la determinación de este vector se iguala a cero la primera derivada:

$$\frac{\partial \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\partial \hat{\boldsymbol{\beta}}} = \frac{\partial\left(\mathbf{y}'\mathbf{y} - 2\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} + \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}\right)}{\partial \hat{\boldsymbol{\beta}}} = 0 \;\Rightarrow\; -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = 0 \;\Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

donde hemos tenido en cuenta la expresión de las derivadas de formas lineales y cuadráticas

respecto a un vector:

$$\frac{\partial \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}}{\partial \hat{\boldsymbol{\beta}}} = \mathbf{X}'\mathbf{y} \quad , \quad \frac{\partial \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}}{\partial \hat{\boldsymbol{\beta}}} = 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

La obtención de los estimadores máximo verosímiles (MV) se lleva a cabo partiendo de la función de verosimilitud que, bajo el supuesto de normalidad ($\mathbf{u} \approx \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, $\mathbf{y} \approx \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$), puede expresarse matricialmente como sigue:

$$L(\mathbf{y}, \sigma^2, \boldsymbol{\beta}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} e^{-\frac{1}{2\sigma^2}[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}$$

y conduce a las expresiones:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \quad , \quad \hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n}$$

En primer lugar, expresamos la función de verosimilitud como:

$$L(\mathbf{y}, \sigma^2, \boldsymbol{\beta}) = (2\pi)^{-\frac{n}{2}} \left(\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}[(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})]}$$

Para llegar a estos estimadores efectuamos una transformación logarítmica:

$$\ln L(y, \beta, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

aplicando a continuación las condiciones de extremo:

$$\frac{\partial \ln L}{\partial \boldsymbol{\beta}} = -\frac{1}{2\sigma^2}(-2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = 0 \;\Rightarrow\; \mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y} \Rightarrow \hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y}$$

Sustituyendo ahora el EMC de $\boldsymbol{\beta}$ se tiene para la varianza:

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})}{2\sigma^4} = 0 \Rightarrow \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{2\sigma^4} = \frac{n}{2\sigma^2} \Rightarrow \hat{\sigma}^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n}$$

Proposición 10.1. *Dado un modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ los estimadores $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ cumplen un serie de propiedades (similares a las vistas en el modelo lineal simple):*

- *La suma de los residuos es nula $\sum_{i=1}^{n} \hat{u}_i = 0$. Como consecuencia puede afirmarse que la media de las observaciones coincide con la media de las estimaciones: $\bar{Y} = \bar{\hat{Y}}$*
- *El hiperplano de regresión pasa por el punto $\left(\bar{X}_2, \cdots, \bar{X}_k\right)$ denominado "*centro de gravedad*". Estas dos primeras propiedades exigen que la especificación de la regresión contenga un término independiente.*
- *Los momentos de segundo orden entre los regresores y los residuos son nulos: $\mathbf{X}'\hat{\mathbf{u}} = 0$.*
- *Los momentos de segundo orden entre $\hat{\mathbf{y}}$ y los residuos son nulos, esto es: $\hat{\mathbf{y}}'\hat{\mathbf{u}} = 0$.*

10.1.2. Propiedades y características de los estimadores

Los estimadores mínimo cuadráticos y máximo verosímiles de los parámetros del modelo de regresión lineal resultan adecuados ya que son *insesgados, consistentes y*

de mínima varianza dentro de la clase de los *estimadores lineales insesgados.*

Los estimadores MCO pueden ser expresados como combinación lineal de y: $\hat{\boldsymbol{\beta}} = \mathbf{Ay}$, siendo $\mathbf{A} = (\mathbf{X'X})\mathbf{X'}$ una matriz fija (no aleatoria) de dimensión $k \times n$. Sustituyendo $\mathbf{y}$ por su expresión se tiene:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'u}$$

comprobándose fácilmente que el valor esperado de $\hat{\boldsymbol{\beta}}$ coincide con $\boldsymbol{\beta}$, y por tanto los estimadores MCO son insesgados:

$$E\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta} + (\mathbf{X'X})^{-1}\mathbf{X'}E(\mathbf{u}) = \boldsymbol{\beta}$$

La matriz de dispersión o de varianzas-covarianzas de los estimadores viene dada por: $Cov\left(\hat{\beta}\right) = \sigma^2(X'X)^{-1}$ cuya expresión matricial es:

$$Cov\left(\hat{\beta}\right) = \begin{pmatrix} \sigma^2_{\hat{\beta}_1} & \sigma_{\hat{\beta}_1\hat{\beta}_2} & \cdots & \sigma_{\hat{\beta}_1\hat{\beta}_k} \\ \sigma_{\hat{\beta}_2\hat{\beta}_1} & \sigma^2_{\hat{\beta}_2} & \cdots & \sigma_{\hat{\beta}_2\hat{\beta}_k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{\hat{\beta}_k\hat{\beta}_1} & \sigma_{\hat{\beta}_k\hat{\beta}_2} & \cdots & \sigma^2_{\hat{\beta}_k} \end{pmatrix}$$

La deducción de esta matriz se lleva a cabo partiendo de la definición de covarianza:

$$\begin{aligned} Cov\left(\hat{\boldsymbol{\beta}}\right) &= E\left[\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\right)'\right] = E\left[(\mathbf{X'X})\mathbf{X'u}\left((\mathbf{X'X})^{-1}\mathbf{X'u}\right)'\right] \\ &= (\mathbf{X'X})^{-1}\mathbf{X'}E[\mathbf{uu'}]\mathbf{X}(\mathbf{X'X}) = \sigma^2(\mathbf{X'X}) \end{aligned}$$

La propiedad de consistencia garantiza que el estimador converge al parámetro cuando la muestra aumenta indefinidamente su tamaño. Por tanto este requisito es equivalente a la anulación asintótica de la matriz de dispersión:

$$\lim_{n\to\infty} Cov\left(\hat{\boldsymbol{\beta}}\right) = 0$$

Teorema 10.1. *(*Gauss-Markov*)* $\hat{\boldsymbol{\beta}}$ *es un estimador lineal insesgado óptimo (ELIO), es decir, dentro de la clase de estimadores lineales e insesgados,* $\hat{\boldsymbol{\beta}}$ *presenta mínima varianza.*

Bajo la hipótesis de normalidad, este resultado puede ser extendido mediante la cota de Frechet-Cramer-Rao a toda la clase de estimadores insesgados.

Bajo las condiciones de regularidad es posible aplicar a cualquier estimador $\tilde{\boldsymbol{\beta}}$ del parámetro $\boldsymbol{\beta}$ la acotación de Frechet-Cramer-Rao. Según dicha acotación, se cumple para todo $\tilde{\boldsymbol{\beta}}$ insesgado:

$$Var\left(\tilde{\boldsymbol{\beta}}\right) \geq \frac{1}{-E\left[\frac{\partial^2 \ln L(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right]}$$

Para esta cota inferior se obtiene:

$$\begin{aligned} \frac{\partial \ln L(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -\frac{1}{2\sigma^2}\left(2\mathbf{X'y} + 2\mathbf{X'X}\boldsymbol{\beta}\right) \Rightarrow \frac{\partial^2 \ln L(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \frac{1}{\sigma^2}(\mathbf{X'X}) \\ &\Rightarrow \frac{1}{-E\left[\frac{\partial^2 \ln L(\mathbf{y},\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right]} = \sigma^2(\mathbf{X'X}) \end{aligned}$$

observándose que dicha expresión coincide con la matriz de dispersión de $\hat{\boldsymbol{\beta}}$ y por tanto los EMC resultan ser óptimos entre todos los estimadores insesgados.

Por lo que se refiere a la distribución probabilística, el supuesto de normalidad para la perturbación aleatoria garantiza la normalidad para los estimadores: $\hat{\boldsymbol{\beta}} \approx \mathcal{N}\left(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\right)$ o, equivalentemente, para cada estimador individual $\hat{\beta}_j \approx \mathcal{N}(\beta_j, \sigma_{\hat{\beta}_j})$.

En realidad, los estimadores MCO serán aproximadamente normales incluso si no se cumple la hipótesis de normalidad para la perturbación aleatoria, siempre que el tamaño de muestra n sea suficientemente elevado para aplicar el teorema central del límite.

Como hemos visto, la expresión de la matriz de varianzas-covarianzas de los estimadores depende de la varianza poblacional σ^2, parámetro que en general resulta desconocido y deberá ser por tanto estimado. Dicha estimación de la varianza se lleva a cabo a partir del vector de residuos mínimo cuadráticos, que tiene carácter aleatorio y aproxima el vector de perturbaciones $\mathbf{u}$.

Concretamente, el estimador insesgado de σ^2 viene dado por:

$$S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$$

Desarrollando la expresión del vector de residuos se obtiene:

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{y} = \mathbf{M}\mathbf{y}$$

donde M es una matriz definida como $[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']$ que cumple las propiedades de ser idempotente (el producto de M por sí misma da como resultado la misma matriz, esto es, $\mathbf{MM} = \mathbf{M}$, semidefinida positiva ($\mathbf{a}'\mathbf{Ma} \geq 0$, $\forall \mathbf{a}$) y simétrica ($\mathbf{M}' = \mathbf{M}$). Una expresión alternativa a la anterior para el vector de residuos es:

$$\hat{\mathbf{u}} = \mathbf{M}\mathbf{y} = \left[\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right]\mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{u} = \mathbf{M}\mathbf{u}$$

a partir de la cual se obtienen fácilmente las características del vector:

$$E(\hat{\mathbf{u}}) = E(\mathbf{M}\mathbf{u}) = \mathbf{M}E(\mathbf{u}) = \mathbf{0}$$

$$Cov(\hat{\mathbf{u}}) = E\left(\hat{\mathbf{u}}\hat{\mathbf{u}}'\right) = E(\mathbf{M}\mathbf{u}\mathbf{u}'\mathbf{M}) = \mathbf{M}E(\mathbf{u}\mathbf{u}')\mathbf{M} = \mathbf{M}\sigma^2\mathbf{I}\mathbf{M} = \sigma^2\mathbf{M}$$

expresiones que permiten construir un estimador insesgado para el parámetro σ^2. Partiendo de la expresión $\hat{\mathbf{u}}'\hat{\mathbf{u}} = (\mathbf{M}\mathbf{u})'\mathbf{M}\mathbf{u} = \mathbf{u}'\mathbf{M}\mathbf{u}$ se obtiene el valor esperado $E(\hat{\mathbf{u}}'\hat{\mathbf{u}}) = \sigma^2(n-k)$. Como consecuencia, el estimador insesgado de la varianza σ^2 será: $S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$.

La deducción del valor esperado $E(\hat{\mathbf{u}}'\hat{\mathbf{u}})$ se lleva a cabo teniendo en cuenta dos propiedades de la traza: la traza de un escalar es el mismo escalar y se cumple $tr(\mathbf{AB}) = tr(\mathbf{BA})$:

$$\begin{aligned} E[\hat{\mathbf{u}}'\hat{\mathbf{u}}] &= E[\mathbf{u}'\mathbf{M}\mathbf{u}] = tr\left[E(\mathbf{u}'\mathbf{M}\mathbf{u})\right] = E[tr(\mathbf{u}'\mathbf{M}\mathbf{u})] = E[tr(\mathbf{M}\mathbf{u}\mathbf{u}')] \\ &= tr\mathbf{M}E[\mathbf{u}\mathbf{u}'] = tr\mathbf{M}\sigma^2 I = \sigma^2 tr\mathbf{M} = \sigma^2(n-k) \end{aligned}$$

El estimador S^2 permite también obtener estimaciones de la matriz de varianzas-covarianzas mediante la expresión: $S^2(\mathbf{X}'\mathbf{X})^{-1}$. Por lo que se refiere al modelo probabilístico asociado a la varianza muestral, asumiendo la normalidad de las perturbaciones $\mathbf{u} \approx \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ se cumple:

$$\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma^2} = \frac{(n-k)S^2}{\sigma^2} \approx \chi^2_{n-k}$$

10.2. Contrastes y análisis de la bondad del modelo

En el modelo lineal básico la presencia de múltiples regresores amplía considerablemente las posibilidades inferenciales.

Al igual que en el modelo simple, los contrastes más interesantes son los de significación que estudian la validez del modelo propuesto, bien sea a nivel global o para ciertos parámetros.

Un elemento clave en estos contrastes es el vector aleatorio de perturbaciones $\mathbf{u}$. A partir de dicho vector y de la matriz idempotente $\mathbf{M}$ es posible obtener una forma cuadrática $\frac{\mathbf{u}'\mathbf{M}\mathbf{u}}{\sigma^2}$, expresión que, como ya hemos visto, sigue un modelo chi-cuadrado con $(n-k)$ grados de libertad.

Partiendo de los errores estimados se obtiene por tanto la expresión aleatoria:

$$\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\sigma^2} \approx \chi^2_{n-k}$$

que resulta de notable utilidad en los procesos inferenciales sobre los parámetros que analizaremos a continuación.

10.2.1. Contrastes individuales

El planteamiento de los contrastes individuales de significación es análogo al visto para la regresión simple y se basa en una discrepancia con distribución t de Student:

$$d_{\hat{\beta}} = \frac{\hat{\beta} - \beta}{S_{\hat{\beta}}} \approx t_{n-k}$$

El supuesto de normalidad de las perturbaciones aleatorias u garantiza que los estimadores mínimo cuadráticos se distribuyen normalmente:

$$\mathbf{u} \approx N(\mathbf{0}, \sigma^2\mathbf{I}) \Rightarrow \hat{\boldsymbol{\beta}} \approx \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

donde $\boldsymbol{\beta}$ recoge el vector de esperanzas y $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ es la matriz de varianzas y covarianzas de $\hat{\boldsymbol{\beta}}$.

Como en la práctica el parámetro σ^2 resulta desconocido debemos trabajar con su estimación muestral S^2, pasando a considerar para cada parámetro β_j estadísticos del tipo:

$$d_{\hat{\beta}_j} = \frac{\hat{\beta}_j - \beta_j}{S_{\hat{\beta}_j}} \approx t_{n-k}$$

A partir de las discrepancias entre el estimador $\hat{\beta}_j$ y el parámetro β_j es posible llevar a cabo contrastes individuales del tipo:

$$\boxed{H_0 : \beta_j = 0 \ , \ \ H_1 : \beta_j \neq 0} \tag{10.2}$$

que se resuelven calculando el valor muestral $\dfrac{\hat{\beta}}{S_{\hat{\beta}}} = d^*_{\hat{\beta}}$ y obteniendo posteriormente el correspondiente nivel crítico:

$$p = P\left(|t_{n-k}| > \left|d^*_{\hat{\beta}}\right| / H_0\right)$$

Si esta probabilidad adopta valores pequeños conduce al rechazo de la hipótesis nula, respaldando por tanto la introducción de la variable X_j como explicativa en nuestro modelo. En otras situaciones puede resultar interesante contrastar valores concretos del parámetro, para lo cual se sigue un procedimiento análogo al anterior sustituyendo el valor hipotético de β en la expresión de la discrepancia.

Debemos tener presente que el hecho de que aparezcan valores elevados de $\hat{\beta}_j$ no significa que la variable X_j tenga gran influencia sobre Y (de hecho, con sólo efectuar un cambio de escala se modificaría la estimación del parámetro β_j). Lo importante será por tanto el producto $\hat{\beta}_j X_j$ o bien la variable normalizada, que da lugar a la discrepancia.

10.2.2. Contrastes globales de significación

El contraste de significación global del modelo puede ser planteado en los siguientes términos:

$$\boxed{\begin{array}{l} H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \ , \text{ para algún } j = 2, \cdots, k \end{array}}$$

donde la hipótesis nula equivale a afirmar que ninguno de los regresores tiene capacidad explicativa sobre Y, mientras el complementario se recoge en la hipótesis alternativa.

En principio este contraste podría plantearse globalmente sobre el vector de coeficientes $\boldsymbol{\beta}$:

$$\begin{array}{ll} H_0 : \boldsymbol{\beta} = \mathbf{0} & H_0 : \beta_1 = \beta_2 = ... = \beta_k = 0 \\ H_0 : \boldsymbol{\beta} \neq \mathbf{0} & H_1 : \beta_j \neq 0 \ , \text{ para algún } j = 1, ..., k \end{array}$$

sin embargo, en general se excluye el término independiente al que no es posible asignar capacidad explicativa sino únicamente *impactos fijos*.

Por lo que se refiere a la relación de este test con los contrastes individuales anteriormente vistos, se observa que el cumplimiento de la hipótesis múltiple $H_0 : \beta_2 = \beta_3 = ... = \beta_k = 0$ equivale al cumplimiento simultáneo de todas las hipótesis $\beta_2 = 0, \beta_3 = 0, ..., \beta_k = 0$ mientras que la aceptación de todas las hipótesis simples no garantiza el cumplimiento de la conjunta al mismo nivel de significación.

En sentido contrario, el rechazo de cualquiera de las hipótesis simples se traduce en el rechazo de la conjunta. Así pues, el test global de significación sólo permite afirmar que el modelo "*tiene sentido*" pero no que dicho modelo sea "*totalmente correcto*".

Al igual que hemos visto en el capítulo anterior para el modelo simple, los contrastes

globales de significación se basan en el análisis de la varianza, tal y como describe la tabla 10.1:

Tabla 10.1.: Análisis de varianza

VARIACIÓN	EXPRESIÓN	G.L.	RATIO
EXPLICADA	$\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - n\bar{Y}^2$	k-1	$\frac{\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}-n\bar{Y}^2}{k-1}$
NO EXPLICADA	$\hat{\mathbf{u}}'\hat{\mathbf{u}}$	n-k	$\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$
TOTAL	$\mathbf{y}'\mathbf{y} - n\bar{Y}^2$	n-1	$\frac{\mathbf{y}'\mathbf{y}-n\bar{Y}^2}{n-1}$

A partir de esta descomposición de variaciones es posible definir una discrepancia dada por el ratio:

$$\frac{\frac{\left(\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}-n\bar{Y}^2\right)}{k-1}}{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}} \approx F_{n-k}^{k-1}$$

que bajo la hipótesis nula sigue un modelo F de Snedecor con $k-1$ g.l. en el numerador y $n-k$ g.l. en el denominador. Por lo que respecta a la interpretación de esta expresión, es fácil comprobar que cuantifica la relación entre la parte de variación explicada y no explicada del modelo, ajustadas ambas por sus grados de libertad. A medida que los valores de la discrepancia F aumentan se reduce el nivel crítico asociado a las observaciones muestrales y en consecuencia aumentan los argumentos para rechazar la hipótesis conjunta planteada.

10.2.3. Bondad del modelo. Coeficientes de determinación

La bondad de los modelos econométricos puede ser analizada mediante el *coeficiente de determinación*, que para un modelo lineal genérico viene dado por la expresión:

$$R^2 = 1 - \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}$$

Es decir, en el análisis de varianza anterior, R^2 se obtiene como 1 menos la proporción de variación no explicada, o lo que es lo mismo la proporción de variación explicada.

Este coeficiente de determinación aparece conectado con la expresión del ratio F asociado al contraste de significación global del modelo, ya que se cumple:

$$F_{n-k}^{k-1} = \frac{\frac{R^2}{k-1}}{\frac{1-R^2}{n-k}}$$

[Compruébese].

Por consiguiente, los modelos con gran capacidad explicativa llevarán asociado un coeficiente de determinación cercano a la unidad y en consecuencia valores elevados de F, con lo cual se rechaza la hipótesis de nulidad de los parámetros.

El coeficiente de determinación es una función no decreciente del número de variables explicativas del modelo. Como consecuencia, la fiabilidad aparente de un modelo aumenta a medida que introducimos nuevas variables y de hecho, en el caso extremo n =k el coeficiente de determinación alcanza un valor unitario siempre que no exista relación lineal entre los regresores.

En efecto, partiendo de la expresión del vector de residuos:

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

si asumimos n=k sin que exista relación lineal entre los regresores se obtiene el rango de $\mathbf{X}$, $\rho(\mathbf{X}) = n$ con lo cual la matriz $\mathbf{X}$ es invertible y por tanto:

$$\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{y} - \mathbf{X}\left[\mathbf{X}^{-1}(\mathbf{X}')^{-1}\right]\mathbf{X}'\mathbf{y} = \mathbf{y} - \mathbf{y} = \mathbf{0} \Rightarrow R^2 = 1$$

Para evitar que el coeficiente de determinación se eleve artificialmente, resulta conveniente introducir el *coeficiente de determinación corregido o ajustado*, que penaliza la inclusión de nuevas variables explicativas en el modelo. El coeficiente de determinación ajustado se define como:

$$\tilde{R}^2 = 1 - \frac{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}}{\frac{\mathbf{y}'\mathbf{y}-n\bar{Y}^2}{n-1}} = 1 - \frac{n-1}{n-k}\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}$$

expresión que resulta de ajustar por sus grados de libertad las variaciones total y residual del modelo, y que puede también ser formulada como $\tilde{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$ [Compruébese].

A partir de la expresión anterior se comprueba fácilmente la desigualdad $\tilde{R}^2 < R^2$ para todo $k > 1$:

$$\begin{aligned} \tilde{R}^2 < R^2 \quad &\Rightarrow 1 - (1 - R^2)\tfrac{n-1}{n-k} < R^2 \Rightarrow (1 - R^2) < (1 - R^2)\tfrac{n-1}{n-k} \\ &\Rightarrow n - k < n - 1 \Rightarrow k > 1 \end{aligned}$$

A medida que aumenta el número de variables explicativas de un modelo, su coeficiente ajustado se aleja del inicial, pudiendo incluso llegar a adoptar valores negativos.

Conviene tener presente que al comparar dos modelos mediante sus coeficientes de determinación ajustados resulta imprescindible que la variable dependiente sea la misma y que los modelos tengan carácter causal.

Otros indicadores utilizados para comparar la bondad de modelos alternativos son los basados en criterios de información. Estas medidas resumen los errores de estimación asociados a cada modelo y penalizan además la inclusión de parámetros.

- ***Logaritmo de verosimilitud:*** Dada una muestra de tamaño n el logaritmo de la función de verosimilitud viene dado por la expresión

$$\ln L = -\frac{n}{2}\left[1 + \ln\left(2\pi\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n}\right)\right]$$

  Dado que el resultado de esta medida guarda una relación inversa con la suma

de los residuos cuadráticos, la comparación de varios modelos alternativos nos llevará a elegir aquél que presenta un mayor resultado de $\ln L$.

- ***Criterios de información:*** La evaluación de bondad de los modelos puede realizarse también a partir de los criterios de información, cuyas expresiones penalizan tanto los errores de estimación (cuantificados a través de $\hat{\mathbf{u}}'\hat{\mathbf{u}}$) como la inclusión de parámetros (k). Por lo tanto, al comparar entre varios modelos alternativos optaremos por aquél que presente valores más reducidos de las medidas de información. Las expresiones más habituales para estas medidas son las propuestas por H. Akaike (1974), E.S. Schwarz (1997) y E.J. Hannan y G.G. Quinn (1979):

- ***Criterio de Akaike:***

$$AIC = n \ln \left(\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n} \right) + 2k + n[1 + \ln(2\pi)]$$

- ***Criterio de Schwarz:***

$$SIC = -2 \ln L + k \ln(n)$$

- ***Criterio de Hannan-Quinn:***

$$HQC = -2 \ln L + 2k \ln(\ln(n))$$

Aunque el coeficiente de determinación ajustado se utiliza con gran generalidad, en ciertas ocasiones resultan necesarios otros ajustes en los coeficientes de bondad de los modelos. Este será el caso de los modelos temporales, donde la existencia de una tendencia común en las variables analizadas puede dar lugar a valores elevados del coeficiente de determinación, incluso del ajustado. En este caso puede resultarnos más útiles los anteriores criterios de información o incluso otras adaptaciones del coeficiente de determinación.

Como consecuencia de la presencia de varios regresores, en los modelos de regresión múltiple podemos distinguir distintos tipos de coeficientes de determinación. Entre ellos el de carácter más global es el *coeficiente de determinación múltiple* (R^2 o $R^2_{Y,X_2,...,X_k}$) que recoge la parte de variación de Y explicada conjuntamente por todas las variables incluidas en el modelo $X_2, ..., X_k$. Tal y como hemos visto anteriormente, este coeficiente viene dado por la expresión:

$$R^2 = \frac{\hat{\mathbf{y}}'\hat{\mathbf{y}} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2} = \frac{\hat{\boldsymbol{\beta}}\mathbf{X}'\hat{\mathbf{y}} - n\bar{Y}^2}{\mathbf{y}'\mathbf{y} - n\bar{Y}^2}$$

El *coeficiente de determinación parcial* $R^2_{Y,X_k/X_2,\cdots,X_{k-1}}$ tiene en cuenta la relación entre Y y X_k, una vez descontado el efecto de las otras variables explicativas del

modelo. Se trata de un coeficiente acotado entre 0 y 1 cuya expresión es la siguiente:

$$R^2_{Y,X_k/X_2,\ldots,X_{k-1}} = \frac{R^2_{Y,X_2,\ldots,X_k} - R^2_{Y,X_2,\ldots,X_{k-1}}}{1 - R^2_{Y,X_2,\ldots,X_{k-1}}}$$

Como consecuencia de su definición, el coeficiente de determinación parcial permite conocer qué proporción de la variación residual de un modelo conseguimos explicar con la introducción de una variable adicional. Así pues, es un instrumento útil para construir un modelo econométrico en etapas, valorando la ganancia explicativa de cada nueva variable.

Por último, es posible definir los *coeficientes de determinación simples* que sólo tienen en cuenta una de las variables explicativas ignorando por completo la existencia de las restantes. Estos coeficientes van asociados a los modelos lineales simples y su carácter es bidireccional por coincidir con los cuadrados de los coeficientes de correlación lineal.

10.2.4. Contrastes relativos a subconjuntos de parámetros

En ocasiones nos interesa contrastar hipótesis relativas a ciertos subconjuntos de parámetros del modelo, bien sea especificando valores concretos para algunos de ellos o bien relaciones entre parámetros que a menudo son postuladas por la propia teoría económica.

A modo de ejemplo, si especificamos el siguiente modelo econométrico para la inversión: $I_t = \beta_1 + \beta_2 PIB_t + \beta_3 Int_t + \beta_4 IPC_t + u_t$ podríamos proponer sobre el mismo hipótesis como las siguientes:
a) $H_0 : \beta_2 = 1$, la propensión marginal a invertir es unitaria
b) $H_1 : \beta_3 + \beta_4 = 0$, la inversión tiene en cuenta el tipo de interés real (es decir, la inversión no variará si un aumento en el tipo de interés nominal viene acompañado de un aumento en los precios, ceteris paribus las restantes variables).

En este tipo de contrastes la hipótesis puede ser expresada en forma genérica como $H_0 : \mathbf{R}\boldsymbol{\beta} = \boldsymbol{\beta}^*$, donde $\mathbf{R}$ es una matriz de r filas (tantas como restricciones impuestas en la hipótesis) y k columnas (tantas como parámetros para que sea multiplicable por $\boldsymbol{\beta}$).

Según cuál sea el tipo de contraste planteado sobre los coeficientes cambiará la expresión de la matriz **R.** Así, cuando deseamos contrastar valores concretos para los parámetros del modelo, la matriz **R** contiene únicamente valores 0 y 1, mientras que si el contraste es de restricciones lineales, los componentes de **R** son los coeficientes que recogen las relaciones entre los distintos parámetros del modelo.

En los ejemplos anteriores, la formulación matricial vendría dada en los siguientes términos:
a) $H_0 : (0\,1\,0\,0) \begin{pmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{pmatrix} = 1$

b)$H_0 : (0\,0\,1\,1)\begin{pmatrix}\beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = 0$

Si estudiamos la distribución del vector $\mathbf{R}\hat{\boldsymbol{\beta}}$ se obtiene, bajo la hipótesis de normalidad de las perturbaciones, un modelo normal r-dimensional con:

$$E(\mathbf{R}\hat{\boldsymbol{\beta}}) = \mathbf{R}\boldsymbol{\beta}$$
$$Cov(\mathbf{R}\hat{\boldsymbol{\beta}}) = \mathbf{R}Cov(\hat{\boldsymbol{\beta}})\mathbf{R}' = \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

y de modo similar al contraste de significación global del modelo se construye una discrepancia normalizada con distribución F de r g.l. en el numerador y (n-k) g.l. en el denominador cuya expresión, bajo la hipótesis nula, es la siguiente:

$$\left(\frac{(\mathbf{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)'(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)}{\hat{\mathbf{u}}'\hat{\mathbf{u}}}\right)\frac{n-k}{r} \approx F^r_{n-k}$$

La discrepancia asociada a este contraste, también, puede ser expresada como:

$$\left(\frac{\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}}\right)\left(\frac{n-k}{r}\right) \simeq F^r_{n-k}$$

donde $\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}}$ es la suma de residuos cuadráticos asociados al modelo sometido a las restricciones de la hipótesis propuesta y $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ recoge la suma de cuadrados de los residuos para el modelo estimado sin restricciones.

Este contraste de restricciones puede ser también resuelto mediante contrastes chi-cuadrado con r grados de libertad, basados en la maximización de la función de verosimilitud bajo la restricción recogida en la hipótesis. Más concretamente, las expresiones serían en este caso:

$$LM = \frac{\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\frac{\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}}}{n}}$$

$$LR = n\ln\left(\frac{\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}}}{\hat{\mathbf{u}}'\hat{\mathbf{u}}}\right)$$

$$W = \frac{\hat{\mathbf{u}}'_{\mathbf{R}}\hat{\mathbf{u}}_{\mathbf{R}} - \hat{\mathbf{u}}'\hat{\mathbf{u}}}{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n}}$$

entre las que se verifica la desigualdad $W > LR > LM$ y para la expresión W (asociada al test de Wald) se garantiza para tamaños elevados de muestra la proporcionalidad con el estadístico F:

$$W \to rF^r_{n-k}$$

10.2.5. Predicción

Una vez que disponemos de uno o varios modelos que superaron la etapa de validación y que por lo tanto parecen ser adecuados debemos pasar a la etapa de predicción de las variables dependientes; esta etapa resulta especialmente útil en los modelos

temporales.

Si bien existen muchos métodos distintos para realizar predicciones, nosotros nos referiremos en todo momento a predicciones científicas; es decir, aquéllas basadas en modelos y que tienen una metodología transparente, de modo que cualquier persona en las condiciones iniciales puede replicar la predicción y debería obtener el mismo valor.

Existen diversas formas de clasificar las predicciones, según el uso que se haga de la información disponible (condicionada o no condicionada), según el período al que vayan referidas (ex-post y ex-ante) o bien según los valores que se encuentran registrados en cada etapa (estática y dinámica).

Predicción condicionada y no condicionada

Cuando realizamos una predicción científica, disponemos de un modelo y unos datos iniciales o inputs; al aplicar el modelo estimado a los datos disponibles se generan unos valores de Y que serán nuestras estimaciones o predicciones.

Partiendo de un modelo, generalmente realizaremos una *predicción condicionada*, entendida como aquélla que incorpora la información disponible en el momento actual. Así, si disponemos de información sobre el vector de datos $\mathbf{x}_0 = (1, X_{20}, \cdots, X_{k0})$, entonces la predicción condicionada sería $E(\hat{Y}/\mathbf{x}_0)$.

Cuando se ignora el valor informativo de los inputs, la predicción se dice *no condicionada*, y en esta situación actuaríamos como si no existiesen esos datos, asumiendo hipótesis ingenuas sobre el comportamiento de las variables explicativas.

Predicción ex-post y ex-ante

La *predicción ex-post* es aquélla que va referida a valores de Y para los cuales disponemos de datos registrados. La principal ventaja de esta técnica es que, al disponer de la información real de la variable en el horizonte de predicción, permite evaluar la capacidad predictiva de un modelo.

En cambio, la *predicción ex-ante* se realiza de cara a futuro, y por tanto va referida a períodos para los cuales no hay datos registrados. Si bien esto se corresponde con lo que en el lenguaje común se entiende por predicción, y tiene como finalidad reducir nuestra incertidumbre futura, debemos tener en cuenta que en este caso no es posible evaluar la calidad de los resultados obtenidos. De ahí que en la práctica resulte recomendable combinar ambos tipos de predicciones, contrastando la capacidad predictiva de los modelos como paso previo a la obtención de predicciones ex-ante.

Predicción estática y dinámica

Denominamos *predicción estática* a aquélla en la que los inputs son siempre datos registrados, mientras que la *predicción dinámica* utiliza las predicciones como inputs del modelo.

Como consecuencia, la *predicción dinámica* entraña un mayor riesgo que la estática, puesto que combina dos fuentes de error: la referida a los inputs (que no son datos registrados, sino predicciones, con lo cual tienen un margen de error), y la inherente a toda predicción.

Las consideraciones efectuadas para la predicción con modelos simples son aplicables en gran medida al modelo lineal general. Así, dado un modelo $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ podemos estar interesados en efectuar predicciones para la variable dependiente una vez cono-

cidos los valores adoptados por las variables explicativas, recogidos en el vector $\mathbf{x_0'}$ que viene dado por la expresión $\mathbf{x_0'} = (1, X_{20}, \cdots, X_{k0})$.
En primera instancia puede obtenerse la predicción puntual: $\hat{Y}_0 = \mathbf{x_0'}\hat{\boldsymbol{\beta}}$ que proporciona un valor individual de la variable dependiente afectado por un error de predicción $e_{\hat{Y}_0} = Y_0 - \hat{Y}_0$.

Como ya hemos visto en el caso simple, esta predicción $\hat{Y}_0$ es una variable aleatoria para la que se cumple:
$\hat{Y}_0 = \mathbf{x_0'}\hat{\boldsymbol{\beta}} = \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{X'y} = \mathbf{Zy}$ es decir, presenta carácter lineal.
$E\left(\hat{Y}_0\right) = E\left(\mathbf{x_0'}\hat{\boldsymbol{\beta}}\right) = \mathbf{x_0'}\boldsymbol{\beta} = E(Y/\mathbf{x_0})$ es un predictor insesgado del valor esperado.

$$\begin{aligned} Var\left(\hat{Y}_0\right) = & \; E\left[\left(\hat{\mathbf{Y}}_\mathbf{0} - \mathbf{E(Y/x_0)}\right)'\left(\hat{\mathbf{Y}}_\mathbf{0} - \mathbf{E(Y/x_0)}\right)\right] = E\left[\left(\mathbf{x_0'}\hat{\boldsymbol{\beta}} - \mathbf{x_0'}\boldsymbol{\beta}\right)'\left(\mathbf{x_0'}\hat{\boldsymbol{\beta}} - \mathbf{x_0'}\boldsymbol{\beta}\right)\right] = \\ = & \; \mathbf{x_0'}\sigma^2(\mathbf{X'X})^{-1}\mathbf{x_0} \end{aligned}$$

Puede además demostrarse que esta varianza es mínima en la clase de predictores lineales insesgados, por lo cual se trata de un predictor óptimo.

Utilizando los resultados anteriores pueden construirse intervalos de confianza para el verdadero valor Y_0 y para el valor esperado $E(Y/\mathbf{x_0})$ que se basan en las discrepancias recogidas en el tabla 10.2:

Tabla 10.2.: Intervalos de confianza para la predicción

Predicción	**Discrepancia tipificada**	**Intervalo de confianza**
Para Y_0 Error Total $e_{\hat{Y}_0} = Y_0 - \hat{Y}_0$	$d_{Y_0-\hat{Y}_0} = \frac{Y_0-\hat{Y}_0}{S_{Y_0-\hat{Y}_0}} \approx t_{n-k}$	$\left[\hat{Y}_0 \pm k\sqrt{S^2\left(1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)}\right]$
Para $E(Y/\mathbf{x_0})$ Error Muestral $e_{\hat{Y}_0} = \left(E(Y/x_0) - \hat{Y}_0\right)$	$d_{\hat{Y}_0} = \frac{E(Y/x_0)-\hat{Y}_0}{S_{\hat{Y}_0}} \approx t_{n-k}$	$\left[\hat{Y}_0 \pm k\sqrt{S^2\left(\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)}\right]$

En el primer caso se elaboran bandas de predicción para un valor individual Y_0, obteniéndose las características:
$E\left(e_{\hat{Y}_0}\right) = E(Y_0) - E(\hat{Y}_0) = 0$
$Var\left(e_{\hat{Y}_0}\right) = Var\left(Y_0 - \hat{Y}_0\right) = Var(Y_0) + Var(\hat{Y}_0) = \sigma^2\left[1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right]$
teniendo en cuenta que esta varianza σ^2 puede ser estimada como $S^2 = \frac{\hat{u}'\hat{u}}{n-k}$
Asumiendo la normalidad para las perturbaciones se tiene:

$$e_{\hat{Y}_0} = Y_0 - \hat{Y}_0 \approx N\left(\mathbf{0}, \sigma^2\left(1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)\right) \Rightarrow \frac{Y_0 - \hat{Y}_0}{\sqrt{S^2\left(1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)}} \approx t_{n-k}$$

obteniéndose el siguiente intervalo de confianza para Y_0:

$$\left[\hat{Y}_0 - k\sqrt{S^2\left(1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)}\;,\; \hat{Y}_0 + k\sqrt{S^2\left(1 + \mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0}\right)}\right]$$

donde k se calcula en tablas de la distribución t de Student con $n-k$ g.l. para el nivel de confianza fijado.

Siguiendo un procedimiento análogo se obtienen las bandas de confianza para el valor esperado $E(Y/\mathbf{x_0})$, para las cuales se obtiene la expresión:

$$\left[\hat{Y}_0 - k\sqrt{S^2\,(\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0})}\;,\; \hat{Y}_0 + k\sqrt{S^2\,(\mathbf{x_0'}(\mathbf{X'X})^{-1}\mathbf{x_0})}\right]$$

[Efectuar la deducción correspondiente]

Como hemos visto anteriormente, para evaluar un modelo econométrico podemos consultar medidas de bondad asociadas al mismo (coeficientes de determinación, medidas de información, error estándar, ...). Sin embargo, en ocasiones puede ser conveniente valorar la capacidad predictiva a *posteriori*, comparando las predicciones proporcionadas por el modelo con los valores adoptados por la variable dependiente Y una vez que éstos sean conocidos. Este planteamiento es adecuado para las observaciones temporales cuando realizamos predicciones ex-post. Así, si designamos por T el horizonte de predicción, por $\hat{Y}_t$ la predicción de la variable para el período t y por Y_t el valor efectivamente observado, es posible definir varias medidas del tipo siguiente:

- Error estándar de las predicciones o raíz del error cuadrático medio:

$$RECM = \sqrt{\frac{\sum_{t=1}^{T}\left(\hat{Y}_t - Y_t\right)^2}{T}}$$

- Error absoluto medio:

$$EAM = \frac{\sum_{t=1}^{T}\left|\hat{Y}_t - Y_t\right|}{T}$$

- Error absoluto porcentual medio:

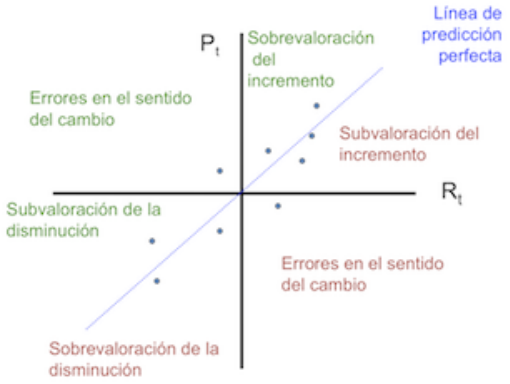
$$EAPM = \sum_{t=1}^{T}\frac{\left|\hat{Y}_t - Y_t\right|}{TY_t} \times 100$$

Coeficiente de desigualdad de Theil:

$$U = \frac{\sqrt{\frac{\sum_{t=1}^{T}(P_t - R_t)^2}{T}}}{\sqrt{\frac{\sum_{t=1}^{T}R_t^2}{T}}}$$

donde P_t es el porcentaje de cambio previsto para la variable, $P_t = \frac{\hat{Y}_t - Y_{t-1}}{Y_{t-1}}$, y R_t el porcentaje de cambio real, $R_t = \frac{Y_t - Y_{t-1}}{Y_{t-1}}$.
Es fácil comprobar que el índice de Theil adopta valor nulo cuando las predicciones son perfectas ($P_t = R_t$) y valor unitario en el caso de que optemos por un *modelo simplista o naive*, que asignaría como predicción el valor presente (*status quo*). Como consecuencia, valores del índice superiores a la unidad indican que el modelo no es útil para fines predictivos.

Figura 10.1.: Diagrama predicción-realidad



Para ampliar las conclusiones del coeficiente de Theil es posible representar en un sistema de coordenadas los incrementos previstos y verdaderos. En el caso de que las observaciones estuvieran situadas sobre la diagonal, el modelo sería adecuado para predecir mientras que en otras situaciones permitiría detectar el signo de las desviaciones en la predicción, tal y como indica la figura 10.1.

El principal atractivo del coeficiente de Theil es que permite conocer las causas de la inexactitud de las predicciones, gracias a la igualdad:

$$U^2 = U_S^2 + U_V^2 + U_C^2$$

donde:

$U_S^2 = \frac{(\bar{P}-\bar{R})^2}{\sum_{t=1}^{T}(P_t-R_t)^2}$, recoge el componente de sesgo

$U_V^2 = \frac{(S_P-S_R)^2}{\sum_{t=1}^{T}(P_t-R_t)^2}$, es el componente de varianza

$U_C^2 = \frac{2(1-r_{PR})S_P S_R}{\sum_{t=1}^{T}(P_t-R_t)^2}$, es el componente de covarianza

Este último sumando suele ser considerado especialmente preocupante, ya que refleja discrepancias entre predicciones y realizaciones que no es posible corregir al no presentar origen sistemático.

10.3. Modelos con variables cualitativas

En los apartados anteriores hemos asumido el carácter cuantitativo de las variables que intervienen en los modelos econométricos. Sin embargo, somos conscientes de que las relaciones socioeconómicas dependen a menudo de factores cualitativos, y de ahí el interés de introducirlos en nuestros modelos.

10.3.1. Variables explicativas cualitativas.

Las magnitudes de interés en el ámbito económico (consumo, inversión, ..) vienen frecuentemente influidas por factores causales de carácter cualitativo, que deberán ser introducidos entre las variables explicativas de un modelo.

Este sería el caso de características como el sexo, la ideología, la cualificación profesional, el sector económico... no cuantificables de modo inmediato pero que pueden desempeñar un papel relevante en la descripción de una realidad económica.
Además, incluso en algunas ocasiones en las que los modelos incluyen variables numéricas, lo que nos interesa no es tanto su valor concreto como la clase a la que pertenecen: por ejemplo, frecuentemente la población aparece clasificada en grupos de edad, las empresas se agrupan en pequeñas, medianas y grandes, ...
Una vez efectuadas estas agrupaciones ignoramos el valor numérico exacto para interesarnos únicamente por la categoría a la que las variables pertenecen. Por tanto, las variables que inicialmente eran cuantitativas adquieren así un carácter claramente cualitativo.

La situación más sencilla consiste en considerar una característica con dos modalidades, a la que asociamos una variable dicotómica que denotamos por D. Así, para la variable Y, podríamos plantear un modelo $Y = \beta_1 + \beta_2 X + \beta_3 D + u$, donde el parámetro β_3 recoge el efecto que tiene sobre la variable dependiente la pertenencia a una u otra modalidad.

Supongamos a modo de ejemplo que deseamos explicar el salario percibido por un colectivo de trabajadores para lo cual hemos especificado un modelo lineal $Y = \beta_1 + \beta_2 X + u$ donde Y es el salario mensual, en euros y X la experiencia laboral, en años.
Si ahora deseamos introducir como explicativa el sexo de cada trabajador podemos definir la variable cualitativa:

$$D = \begin{cases} 0, & \text{si la observación corresponde a una mujer} \\ 1, & \text{si la observación corresponde a un hombre} \end{cases}$$

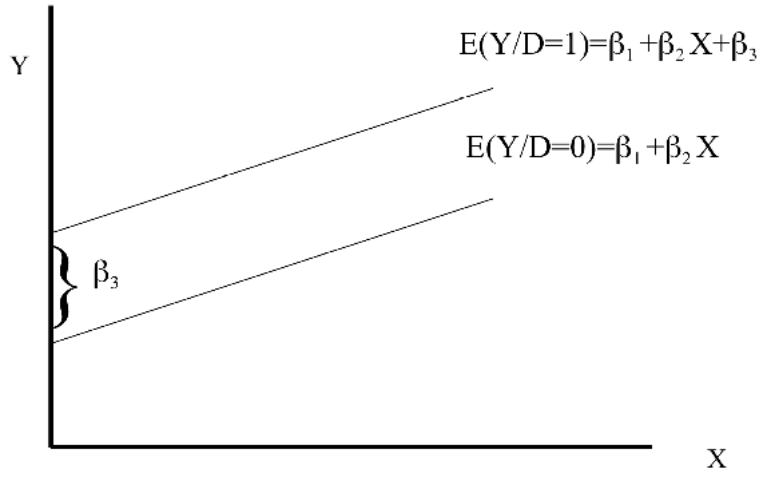
con lo cual el modelo se completaría como: $Y = \beta_1 + \beta_2 X + \beta_3 D + u$, y β_3 representa el aumento de salario que se produce como consecuencia de que el trabajador tenga sexo masculino.

La inclusión de una variable cualitativa (también llamada *ficticia o dummy*) no altera los métodos de estimación y contraste ni las medidas de bondad de un modelo econométrico, sino que equivale a una extensión del mismo a las dos categorías o modalidades consideradas, ya que se obtienen ahora dos líneas poblacionales diferenciadas:

$$\begin{cases} E(Y/D=0) = \beta_1 + \beta_2 X \\ E(Y/D=1) = \beta_1 + \beta_2 X + \beta_3 \end{cases}$$

con lo cual el coeficiente de la variable D, β_3, se interpreta como el efecto sobre el valor esperado del cambio de categoría: $\beta_3 = E(Y/D=1) - E(Y/D=0)$:

Figura 10.2.: Modelo con variable dummy



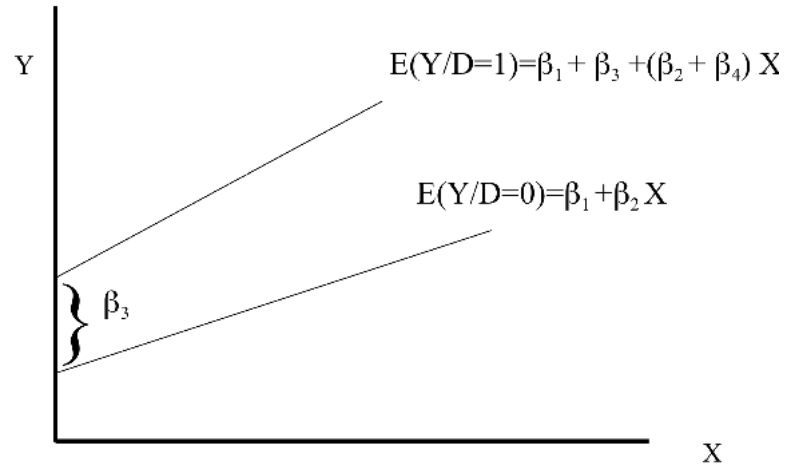
La interpretación del parámetro β_3 dependerá de la definición dada a D. Así, en el ejemplo considerado este coeficiente recoge el aumento salarial esperado para los hombres respecto al de las mujeres. Dado que habitualmente la discriminación laboral favorece a los hombres, se espera que este parámetro presente signo positivo pero es fácil observar que la interpretación sería la opuesta si cambiásemos la definición de la variable D (D=1 para las mujeres y D=0 para los hombres).

El modelo anterior podría ser completado si consideramos que la pertenencia a una modalidad condiciona no sólo la ordenada en el origen sino también el impacto que X tiene sobre Y (esto es, la pendiente de la recta). En tales situaciones cabe considerar un nuevo término, que recoge la interacción entre la variable ficticia D y X, variable explicativa de carácter cuantitativo: $Y = \beta_1 + \beta_2 X + \beta_3 D + \beta_4 (DX) + u$, con lo cual se obtiene:

$$\begin{cases} E(Y/D=0) = \beta_1 + \beta_2 X \\ E(Y/D=1) = \beta_1 + \beta_2 X + \beta_3 + \beta_4 X = (\beta_1 + \beta_3) + (\beta_2 + \beta_4) X \end{cases}$$

y las rectas asociadas a las dos categorías difieren en cuanto a su ordenada en el origen (en β_2) y a su pendiente (en β_4).

Figura 10.3.: Modelo con variable dummy y término de interacción



En nuestro ejemplo podríamos introducir un término de interacción entre la variable D y la experiencia laboral X con lo cual se obtiene

$$Y = \beta_1 + \beta_2 X + \beta_3 D + \beta_4 (DX) + u$$

recogiendo el parámetro β_4 la diferencia en el efecto marginal que cada año de experiencia tiene sobre el salario para hombres y mujeres.

El planteamiento anterior puede ser extendido al caso de una variable cualitativa con más de dos categorías. Esta situación se presenta cuando consideramos múltiples modalidades: variables del tipo "categorías socioeconómicas", "ramas de actividad", "tamaño de las empresas", "estaciones del año", ... La práctica para incluir una variable cualitativa en el modelo consiste entonces en adoptar una de las modalidades como referencia, definiendo variables dummy (0, 1) para todas las categorías restantes.

En general, si deseamos considerar el efecto sobre Y de una característica cualitativa con m modalidades o categorías, el modelo planteado sería:

$$Y = \beta_1 + \beta_2 X + \beta_3 D_1 + \cdots + \beta_{m+1} D_{m-1} + u$$

donde han sido introducidas m-1 variables cualitativas (una menos que las modalidades consideradas).

La especificación de un modelo de m-1 variables ficticias (en lugar de m, como en principio podría parecer lógico) evita la denominada *trampa de las variables ficticias.* Puede comprobarse que la inclusión de m variables ficticias llevaría llevaría a una matriz X de rango no pleno (por ser sus columnas linealmente dependientes) impidiendo así la estimación del modelo propuesto.

Imaginemos que proponemos ahora para el salario una explicación en función de la antigüedad laboral y del sector de actividad económica, característica para la que consideramos las modalidades

agricultura, industria, construcción y servicios. Si definiéramos cuatro variables ficticias se tendría:

$$D_A = \begin{cases} 1, & \text{si el trabajador pertenece al sector } \textit{agricultura} \\ 0, & \text{en otro caso} \end{cases}$$

$$D_I = \begin{cases} 1, & \text{si el trabajador pertenece al sector } \textit{industria} \\ 0, & \text{en otro caso} \end{cases}$$

$$D_C = \begin{cases} 1, & \text{si el trabajador pertenece al sector } \textit{construcción} \\ 0, & \text{en otro caso} \end{cases}$$

$$D_S = \begin{cases} 1, & \text{si el trabajador pertenece al sector } \textit{servicios} \\ 0, & \text{en otro caso} \end{cases}$$

El modelo planteado sería entonces $Y = \beta_1 + \beta_2 X + \beta_3 D_A + \beta_4 D_I + \beta_5 D_C + \beta_6 D_S + u$ y la matriz $\mathbf{X}$ de datos vendría dada por la expresión:

$$\mathbf{X} = \begin{pmatrix} 1 & X_1 & D_{A1} & D_{I1} & D_{C1} & D_{S1} \\ 1 & X_2 & D_{A2} & D_{I2} & D_{C2} & D_{S2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & D_{An} & D_{In} & D_{Cn} & D_{Sn} \end{pmatrix}$$

observándose fácilmente que se cumple para cada observación: $D_{Ai}+D_{Ii}+D_{Ci}+D_{Si} = 1$, $\forall i = 1, ..., n$ (ya que la actividad de cada individuo se incluye en uno y sólo un sector de actividad) y como consecuencia la suma de las cuatro últimas columnas es unitaria, coincidiendo con la primera columna.

Como consecuencia de esta situación de relación lineal entre las columnas o *multicolinealidad* la matriz $\mathbf{X}$ no es de rango pleno incumpliéndose uno de los supuestos del modelo lineal básico (se tiene $\rho(\mathbf{X}) < k$) y no es posible llevar a cabo la estimación por mínimos cuadrados al ser $\mathbf{X}'\mathbf{X}$ no invertible ($|\mathbf{X}'\mathbf{X}| = 0$).

Para evitar este tipo de situaciones se adopta una de las modalidades como referencia introduciendo variables cualitativas para las restantes. En este caso, si adoptásemos como referencia la agricultura consideraríamos el modelo $Y = \beta_1 + \beta_2 X + \beta_3 D_I + \beta_4 D_C + \beta_5 D_S + u$ en el que ya no aparece problema de relación lineal entre las variables. De este modo, los términos independientes de la recta adoptarían los valores:

CATEGORÍA	TÉRMINO INDEPENDIENTE
AGRICULTURA	β_1
INDUSTRIA	$\beta_1 + \beta_3$
CONSTRUCCIÓN	$\beta_1 + \beta_4$
SERVICIOS	$\beta_1 + \beta_5$

que permiten interpretar cada coeficiente de las variables cualitativas como el efecto que origina sobre la variable dependiente Y (salario) la pertenencia al sector económico correspondiente.

Los modelos de variables ficticias con m modalidades pueden ser completados si se desea incorporar las posibles interacciones de cada una de estas categorías cualitativas con las variables cuantitativas X_i. La significación de cada uno de estos términos puede ser contrastada mediante las correspondientes discrepancias, que seguirán una distribución t de Student.

Como hemos visto, según cuál sea el modo en el que las variables ficticias afectan a la variable dependiente aparecerán distintas especificaciones alternativas para un modelo.

Dado que las expresiones e interpretación de los contrastes de significación coinciden con los estudiados en apartados anteriores, resulta aconsejable plantear inicialmente modelos con especificaciones completas, seleccionando el modelo definitivo según los resultados de los contrastes a partir de nuestra información muestral.

La introducción de variables cualitativas puede resultar de gran ayuda para modelizar factores de dos o más categorías: factores estacionales, cambios estructurales, cambios metodológicos, valores atípicos,...

10.3.2. Variables cualitativas dependientes. Introducción a los modelos *logit y probit*

Hasta ahora las variables cualitativas han sido consideradas como explicativas pero es evidente que pueden también constituir el objeto de un estudio. Por ejemplo, es indudable el interés de explicar la situación laboral (si un individuo está o no en paro), el resultado de determinada política (si se alcanza o no cierto nivel de crecimiento), la decisión de un consumidor (comprar o no un artículo) etc.

Dentro de los modelos de variable dependiente cualitativa existen varias alternativas que no vamos a estudiar en detalle:

- Si queremos explicar una variable con dos modalidades (como las de los ejemplos anteriores) los modelos resultantes son de tipo *binomial*.
- Si la variable puede adoptar más de dos modalidades, tendríamos un modelo *multinomial*.
- Cuando la característica presenta varias modalidades que siguen un orden natural, se trata de modelos *ordenados*.
- En el caso de que la característica que explicamos corresponda a una decisión que condiciona las siguientes se trataría de modelos *secuenciales*.

En principio podríamos considerar que los modelos de variable cualitativa dependiente forman parte de la teoría general de la regresión, con la única salvedad de que la variable a explicar no es continua.

Consideremos a modo de ejemplo una variable Y que recoge si determinado individuo se encuentra o no en paro. Esta variable que tratamos de explicar es dicotómica adoptando valor 1 si el individuo está en paro y 0 si no lo está (o viceversa si se definen de modo simétrico).

Una vez seleccionadas las variables (cuantitativas o cualitativas) que se consideran adecuadas como explicativas, el modelo adopta la expresión:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u \quad \text{o de forma compacta: } Y = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Se trata de un modelo para una variable que adopta dos únicos valores, 0 y 1, con probabilidades desconocidas condicionadas a los valores de $\mathbf{X}$. Aplicando la esperanza

matemática, se obtiene:

$$E(Y/X) = p = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \mathbf{X}\boldsymbol{\beta}$$

Aunque este modelo resulta sencillo de interpretar (cada coeficiente recogerá el efecto sobre la probabilidad de cada variable considerada), se aprecian en él dos inconvenientes:

- El error, al igual que la variable Y, es dicotómico:

$$\begin{cases} Y = 0, & \Rightarrow \mathbf{u} = -\mathbf{X}\boldsymbol{\beta} \\ Y = 1, & \Rightarrow \mathbf{u} = 1 - \mathbf{X}\boldsymbol{\beta} \end{cases}$$

  Así pues, el modelo no se adapta a las hipótesis planteadas en regresión lineal (supuestos de homoscedasticidad y de normalidad para las perturbaciones), y por lo tanto la eficiencia de los estimadores no viene garantizada por el método de los mínimos cuadrados.

- Además, los valores esperados de Y se interpretan como probabilidades, por lo cual las estimaciones deberían estar comprendidas entre 0 y 1, requisito que no viene garantizado por mínimos cuadrados. De ahí que se incluyan modificaciones que dan lugar a los modelos denominados *logit, probit y modelo lineal restringido*. Otra posibilidad es pasar del modelo lineal a la *función lineal discriminante*, que permite clasificar cualquier nueva observación en una de las dos categorías o grupos considerados.

Supongamos que deseamos modelizar una variable dicotómica Y asociada a sucesos del tipo "estar en paro", "comprar un producto", "votar un candidato",... Vistos los inconvenientes anteriormente señalados la propuesta habitual consiste en especificar un modelo para cierta variable latente que denotamos por Z:

$$Z_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

donde $\mathbf{x}'$ es el vector fila de datos de variables explicativas y $\boldsymbol{\beta}$ el vector columna de k coeficientes a estimar. Por su parte la variable Z, que no es observable, aparece conectada con la dicotómica y representa una especie de "*propensión al suceso considerado*" (en nuestros ejemplos sería la propensión a estar en paro, la inclinación a comprar un artículo, la intención de votar a cierto candidato...).

En definitiva, Z incluye siempre una idea de potencialidad de modo que será un indicador que en principio se encuentra entre menos infinito y más infinito siendo posible establecer un umbral o *índice de propensión* Z^* tal que se cumpla:

$$Y = \begin{cases} 1, & si\ Z > Z^* \\ 0, & si\ Z \leq Z^* \end{cases}$$

A diferencia de Z, la variable Y sí resulta observable puesto que en la práctica nosotros

no conoceremos las propensiones de los individuos al suceso considerado, tan sólo sabremos si una persona está en paro, si ha consumido un producto, si ha votado a un candidato, ...

Este valor límite se fija habitualmente en $Z^* = 0$ y para calcular las probabilidades se tiene por tanto:

$$p_i = P(Y = 1) = P(Z > Z^*) = 1 - F(-\mathbf{x}'\boldsymbol{\beta})$$
$$1 - p_i = P(Y = 0) = P(Z \leq Z^*) = F(-\mathbf{x}'\boldsymbol{\beta})$$

Ambas expresiones muestran probabilidades en función de los coeficientes $\boldsymbol{\beta}$ y las variables explicativas $\mathbf{X}$. Teniendo en cuenta que nuestra información muestral serían realizaciones de una variable dicotómica, obtendríamos la función de verosimilitud muestral:

$$L = \prod_{Y_i=1}^{n} p_i(1 - p_i)$$

Esta expresión depende de p_i, valores que a su vez se obtendrán según la distribución probabilística especificada para los errores u (F). Así surgen los modelos logit, probit y de probabilidad lineal.

El modelo *logit* surge cuando la distribución considerada es de tipo logístico, el *probit* cuando es de tipo normal y el de *probabilidad lineal* cuando se trata de un modelo uniforme.

Comenzando por el modelo logit, debemos tener en cuenta que la función logística viene dada por una probabilidad acumulada $F(x) = \frac{1}{1+e^{-x}}$. Por tanto, si asumimos que los errores u se distribuyen según un modelo logístico se tiene:

$$p_i = P(Y_i = 1) = 1 - F(-x_i'\beta) = 1 - \frac{1}{1 + e^{x_i'\beta}} = \frac{e^{x_i'\beta}}{1 + e^{x_i'\beta}}$$

Esta expresión puede ser linealizada mediante logaritmos:

$$p_i\left(1 + e^{x_i'\beta}\right) = e^{x_i'\beta} \;\Rightarrow\; \left(p_i + p_i e^{x_i'\beta}\right) = e^{x_i'\beta} \;\Rightarrow\; p_i = e^{x_i'\beta}(1 - p_i)$$

con lo cual se tiene:

$$e^{x_i'\beta} = \frac{p_i}{1 - p_i}$$

Luego:

$$\log\left(\frac{p_i}{1 - p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta} = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

Una vez efectuada la transformación anterior, el problema se reduce a estimar los coeficientes $\boldsymbol{\beta}$ con la información muestral. Teniendo en cuenta que el modelo es no lineal el método habitualmente empleado consiste en maximizar la función de verosimilitud L que es cóncava por lo cual tiene un único óptimo; el cálculo se realiza utilizando algoritmos numéricos como Newton-Rapson o *scoring*.

Cuando los errores u siguen una distribución normal se tiene el modelo probit, cuyas

características y método de estimación resultan similares a los del logit. Teniendo en cuenta que el modelo logístico y el normal son cercanos excepto en las colas, será habitual obtener resultados similares salvo para muestras grandes.

Interpretación de parámetros Los parámetros estimados permiten interpretar los efectos ocasionados por cambios en las variables explicativas sobre las probabilidades. Asi, si expresamos la variable latente como:

$$Z_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i = \beta_1 + \sum_{j=2}^{k} \beta_j X_{ji} + u_i$$

entonces dichos efectos vienen dados por las expresiones:

$$\frac{\partial p_i}{\partial X_{ij}} = \begin{cases} \beta_j & \text{modelo lineal} \\ \beta_j p_i(1-p_i) & \text{modelo logit} \\ \beta_j f(Z_i) & \text{modelo probit} \end{cases}$$

donde f recoge la función de densidad normal estándar. Puede observarse fácilmente que en el modelo lineal las derivadas son constantes. En cambio, en el modelo logit serían constantes los efectos de los cambios en el ratio $\log\left(\frac{p_i}{1-p_i}\right)$ ya que:

$$\frac{\partial\left(\log\frac{p_i}{1-p_i}\right)}{\partial X_{ij}} = \beta_i$$

Los coeficientes de los modelos logit y probit no admiten una interpretación inmediata como en el caso del modelo lineal básico. La interpretación debe hacerse en términos relativos entre pares de variables de forma que los cocientes $\frac{\hat{\beta}_i}{\hat{\beta}_j}$ indican la importancia relativa que los efectos de las variables X_i y X_j tienen sobre la probabilidad de escoger la alternativa $Y_i = 1$. Por este motivo perdemos la interpretación habitual de las variables económicas.

Significación y evaluación de la bondad Una vez especificado uno o varios modelos de variable cualitativa, la significación de sus coeficientes puede ser contrastada calculando el cociente entre el valor del coeficiente y su desviación estándar estimada. El resultado puede ser comparado con el valor crítico de una distribución t de student con los grados de libertad correspondientes (número de observaciones menos coeficientes estimados), aunque en este caso la validez es únicamente asintótica.

Por lo que se refiere a la evaluación del modelo, existen varios planteamientos para comparar las alternativas de modelización de características cualitativas:

- Calcular la suma de cuadrados de las desviaciones para las probabilidades previstas

- Comparar los porcentajes correctamente predichos en distintos modelos
- Observar las derivadas de las probabilidades con respecto a alguna variable independiente

Conviene tener presente que el comportamiento del coeficiente de determinación para variables dicotómicas no es adecuado, por lo cual distintos autores han propuesto modificaciones de esta medida. En el caso del modelo de regresión lineal todas estas expresiones serían equivalentes pero no sucede lo mismo en los modelos de variables cualitativas.

Denotando por Y los valores observados (que serán necesariamente 1 o 0) y por $\hat{Y}$ los valores estimados por el modelo, que representan probabilidades, las medidas más extendidas son las siguientes:

Tabla 10.3.: Medidas de bondad de modelos logit

MEDIDA	**DEFINICIÓN**
Medida de Effron (1978)	$R^2 = 1 - \frac{n}{n_1 n_2} \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$
R^2	Cuadrado del coeficiente de correlación entre Y e $\hat{Y}$
Medida de Amemiya	$R^2 = 1 - \frac{\frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\hat{Y}_i(1-\hat{Y}_i)}}{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}$
Medida basada en verosimilitudes	$R^2 = -2 \ln \frac{L_{NR}}{L_R}$ L_{NR}: Máx. de $\mathbf{L}$ respecto a todos los parámetros L_R: Máximo de $\mathbf{L}$ con $\beta_i = 0$, $\forall i$ $0 \leq R^2 \leq 1 - L_R^{\frac{2}{n}}$
Medida de Cragg y Uhler (1970)	$R^2 = \frac{L_{NR}^{\frac{2}{n}} - L_R^{\frac{2}{n}}}{\left(1 - L_R^{\frac{2}{n}}\right) L_{NR}^{\frac{2}{n}}}$ L_{NR}: Máx. de $\mathbf{L}$ respecto a todos los parámetros L_R: Máximo de $\mathbf{L}$ con $\beta_i = 0$, $\forall i$ $0 \leq R^2 \leq 1$
Medida de Mc Fadden (1974)[1]	$R^2 = 1 - \frac{\log L_{NR}}{\log L_R}$ L_{NR}: Máx. de $\mathbf{L}$ respecto a todos los parámetros L_R: Máximo de $\mathbf{L}$ con $\beta_i = 0$, $\forall i$ $0 \leq R^2 \leq 1$
Proporción de aciertos	$R^2 = \frac{\text{núm. predicciones correctas}}{\text{núm. observaciones}}$

10.4. Alteración de supuestos del modelo lineal

Una vez que hemos desarrollado el modelo de regresión lineal múltiple bajo las hipótesis habituales de trabajo (es decir, el *modelo lineal básico*), vamos a examinar

las posibles alteraciones de los supuestos, analizando sus causas, las consecuencias sobre el modelo, los métodos para su detección y las posibles soluciones.

10.4.1. Errores de especificación

La especificación es la primera etapa de la modelización econométrica y condiciona por tanto en gran medida el éxito de una investigación. Esta etapa incluye la definición de los elementos que intervienen en el modelo y las relaciones existentes entre los mismos, que traducen las hipótesis de comportamiento sobre la población. Así pues, en principio el *error de especificación* es un término amplio que hace referencia a la discrepancia entre el modelo propuesto y la realidad.

En general, al hablar de errores de especificación hacemos referencia a equivocaciones cometidas en la elección de la forma funcional del modelo o de las variables explicativas.

10.4.1.1. Forma funcional del modelo

Las consecuencias de una *confusión en la forma funcional* del modelo son difíciles de evaluar. Para intentar evitarlas resulta conveniente tener presente la teoría económica referida al fenómeno analizado y examinar la nube de puntos que representa nuestra información.

La admisión de la relación lineal entre las variables no resulta muy restrictiva en la práctica ya que la experiencia econométrica ha demostrado que mediante relaciones lineales entre variables se consiguen a menudo aproximaciones válidas de la realidad.

Además, ciertas relaciones no lineales de gran interés en economía (funciones de producción Cobb-Douglas, curvas de indiferencia, modelos exponenciales...) pueden transformarse fácilmente en lineales mediante cambios de variable.

Así, para un modelo $Y_i = \beta_1 X_{2i}^{\beta_2} X_{3i}^{\beta_3} e^{u_i}$ podríamos efectuar una transformación logarítmica del tipo $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_{2i} + \beta_3 \ln X_{3i} + u_i$ llegando al modelo logarítmico (o doble logarítmico).

Si por el contrario la especificación de partida fuese del tipo $Y_i = e^{\beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i}$, entonces llegaríamos mediante transformación logarítmica al modelo log-lineal (o semi-log), dado por la expresión: $\ln Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$

La transformación sería aun más sencilla para funciones de tipo hiperbólico $Y_i = \beta_1 + \frac{\beta_2}{X_i} + u_i$, en las que bastaría con plantear el cambio $Z_i = \frac{1}{X_i}$ para llegar a un modelo lineal.

Las transformaciones anteriores permiten aplicar los desarrollos de MCO a los modelos linealizados, consiguiendo así una estimación de los parámetros por el procedimiento habitual[2]. Es importante tener presente que estas transformaciones son auxiliares para llevar a cabo la estimación, pero el modelo objetivo es el inicialmente propuesto en cada caso. Por ello, los errores deben ser cuantificados sobre dicho modelo (potencial, exponencial,...) y no sobre el linealizado (estos últimos son los que habitualmente proporcionan los paquetes econométricos al efectuar una regresión lineal sobre logaritmos).

[2] En el caso de que los modelos no fuesen linealizados sería necesario plantear una estimación por mínimos cuadrados no lineales (nonlinear least squares o NLS) que exigen procedimientos iterativos.

10.4.1.2. Omisión de variables explicativas relevantes e inclusión de variables irrelevantes

Por lo que se refiere a las variables incluidas en la especificación, pueden producirse errores tanto por exceso como por omisión, siendo estos últimos los más preocupantes.

Supongamos que el modelo correcto para el vector $\mathbf{y}$ viene dado por la expresión $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, distinguiendo para las variables explicativas las matrices $\mathbf{X}_1$ y $\mathbf{X}_2$:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

Si nuestra especificación propuesta es $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{v}$, existe un *error de omisión de variables* con lo cual los estimadores mínimo cuadráticos son sesgados e inconsistentes.

Para la especificación propuesta los estimadores MCO son del tipo $\hat{\boldsymbol{\beta}}_1 = (\mathbf{X_1'X_1})^{-1}\mathbf{X_1'y}$ que, sustituyendo $\mathbf{y}$ por su valor verdadero, conducen a la expresión:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= (\mathbf{X_1'X_1})^{-1}\mathbf{X_1'y} = (\mathbf{X_1'X_1})^{-1}\mathbf{X_1'}(\mathbf{X_1}\boldsymbol{\beta_1} + \mathbf{X_2}\boldsymbol{\beta_2} + \mathbf{u}) = \\ &= \boldsymbol{\beta_1} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'X_2}\boldsymbol{\beta_2} + (\boldsymbol{X_1'X_1})^{-1}\boldsymbol{X_1'u}\end{aligned}$$

Como consecuencia de esta especificación equivocada, los estimadores MCO pasan a ser sesgados ya que se obtiene un valor esperado:

$$E\left(\hat{\boldsymbol{\beta}}_1\right) = \boldsymbol{\beta}_1 + \mathbf{P}\boldsymbol{\beta}_2$$

donde $\mathbf{P} = (\mathbf{X_1'X_1})^{-1}\mathbf{X_1'X_2}$ es la expresión del estimador mínimo cuadrático correspondiente a la regresión de $\mathbf{X}_2$ sobre $\mathbf{X}_1$.

A partir de esta expresión se observa que los estimadores resultarían insesgados si no existiera correlación entre las variables excluidas y las incluidas en el modelo (esto es si $\mathbf{P} = \mathbf{0}$). Sin embargo aún en ese caso se presentarían problemas, ya que la omisión de $\mathbf{X}_2$ daría lugar a residuos superiores a los correspondientes al modelo verdadero y en consecuencia conduciría a contrastes de significación excesivamente exigentes.

En efecto, si planteásemos como modelo verdadero:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{10.3}$$

la omisión de la variable X_k nos llevaría a plantear la relación:

$$Y_i = \beta_1^* + \beta_2^* X_{2i} + \cdots + \beta_{k-1}^* X_{k-1,i} + u_i \tag{10.4}$$

Si asumimos que la variable excluida aparece relacionada con las restantes por la expresión: $X_{ki} = \alpha_1 + \alpha_2 X_{2i} + \cdots + \alpha_{k-1}X_{k-1,i} + v_i$, entonces es posible comprobar sobre el modelo propuesto la relación:

$$\beta_j^* = \beta_j + \beta_k\alpha_j \ , \ \forall j = 2, \cdots, k-1 \quad , y \quad u_i^* = u_i + \beta_{ki} \tag{10.5}$$

y, si $\mathbf{v}$ es una v.a. no correlacionada con $\mathbf{u}$ y cuyo comportamiento se adapta a los supuestos habituales, se tiene:

$$\boxed{\begin{array}{c} E(u^*) = E(u + \beta v) \\ E(u^{*2}) = \sigma_u^2 + \beta_k^2\sigma_v^2 \\ E(u_r u_s) = 0 \quad , \forall r \neq s \end{array}}$$

por lo cual el modelo propuesto 10.4 puede estimarse por MCO bajo las hipótesis habituales, y cada parámetro β_j^* diferirá del correspondiente en el modelo verdadero 10.3, β_j según la relación 10.5,

siempre que la variable esté relacionada con la excluida. Como consecuencia, el coeficiente de X_j ya no representa el efecto de un incremento de esta variable sobre Y, sino *el efecto de un incremento de esta variable sobre Y bajo la condición de que X_k se incremente en un valor igual a α_j y de que ello provoque un incremento adicional en la variable Y de magnitud $\beta_k\alpha_j$.*

Por lo que respecta a la bondad del modelo, como ya hemos comentado los residuos aumentan respecto a los del modelo verdadero 10.3 y con ellos las varianzas del error y de los estimadores, dependiendo la cuantía de este aumento del grado de relación lineal que guarde la variable excluída con las restantes. Lógicamente, si esta relación es muy alta entonces la exclusión de la variable apenas afectaría a la bondad del modelo.

Puede comprobarse que el sesgo asociado a los estimadores MCO no desaparece al aumentar el tamaño muestral n por lo cual los estimadores también son inconsistentes.

Por lo que se refiere a la estimación de la varianza, los residuos asociados al modelo propuesto 10.4 serían superiores a los del modelo verdadero 10.3 y como consecuencia se sobreestimarán la varianza residual y las varianzas estimadas para los estimadores de los parámetros. En estas situaciones los contrastes resultan más exigentes para rechazar la nulidad de los parámetros y por tanto, el proceso de contraste de significación queda invalidado.

Si por el contrario se plantease el problema en los términos opuestos, la *inclusión de variables irrelevantes* en el modelo propuesto da lugar a estimadores mínimo cuadráticos insesgados y consistentes, pero que sin embargo presentan varianzas estimadas superiores a las que les corresponderían. Así pues, en este caso los contrastes de significación serán también excesivamente exigentes, y además podrían aparecer problemas asociados a la pérdida de grados de libertad y la presencia de multicolinealidad entre las variables.

En efecto, si sobre el modelo anteriormente planteado como verdadero 10.3, proponemos ahora la inclusión de una variable X_{k+1} que resulta irrelevante para explicar Y, entonces el coeficiente de dicha variable sería nulo y el modelo a estimar sería:

$$Y_i = \beta_1^* + \beta_2^* X_{2i} + \cdots + \beta_{k-1}^* X_{k+1,i} + u_i$$

con $u_i^* = u_i - \beta_{k+1} X_{k+1,i} = u_i$. Por tanto, en este caso no existirán diferencias en las perturbaciones de ambos modelos.

Las consecuencias sobre la estimación del modelo dependerán del grado de relación lineal de la variable irrelevante incluida con las excluidas. En general la varianza estimada se verá poco afectada mientras que la matriz inversa de $X\,'X$ tendrá en general diagonales mayores con el consiguiente aumento de las varianzas asociadas a los estimadores y a los errores de predicción.

El problema de especificación inadecuada puede ser contrastado partiendo de un modelo ampliado, en el que se incluyen como explicativas las variables que dudamos incluir en el modelo, contrastando sobre dicho modelo la nulidad de los coeficientes asociados a dichas variables (que sería un caso particular de restricción referida a un subconjunto de parámetros). Así pues, bajo la hipótesis nula serían nulos los coeficientes de las variables sobre las que dudamos y por tanto, únicamente si rechazamos la hipótesis deberíamos incluir dichas variables en el modelo.

Este contraste, es un caso particular del test de restricciones sobre coeficientes y por tanto puede ser resuelto mediante la expresión

$$\left(\frac{\widehat{u}'_R\hat{u}_R - \hat{u}'\hat{u}}{\hat{u}'\hat{u}}\right)\frac{n-k}{r} \approx F^r_{n-k}$$

o bien mediante una razón de verosimilitudes (LR)

$$LR = -2\left(\ln L_R - \ln L\right) = n\ln\left(\frac{\hat{\mathbf{u}}'_R\hat{\mathbf{u}}_R}{\hat{\mathbf{u}}'\hat{\mathbf{u}}}\right) \rightarrow \chi^2_r$$

donde $\hat{\mathbf{u}}'_R\hat{\mathbf{u}}_R$ son los residuos cuadráticos del modelo restringido (es decir, sin las variables explicativas sobre las que dudamos).

Las expresiones anteriores pueden aplicarse en dos modalidades: partiendo de un modelo restringido nos plantearíamos si es conveniente añadir nuevas variables explicativas o bien partiendo de un modelo ampliado contrastaríamos si es aconsejable eliminar del mismo algunas variables explicativas por ser irrelevantes (como hemos señalado, en ambos casos la hipótesis nula equivale a afirmar que las variables sobre las que dudamos no son relevantes para explicar Y, y lógicamente la conclusión a la que llegaremos será la misma con independencia del planteamiento de partida).

10.4.1.3. Test de especificación RESET de Ramsey

El test RESET (*Regression Error Specification Test*) detecta errores de especificación en sentido amplio (es decir, forma funcional incorrecta, variables relevantes omitidas, variables irrelevantes incluidas...). Para plantear este contraste se parte de un modelo inicial

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_k X_k \tag{10.6}$$

Si sospechamos que este modelo no está bien especificado pero desconocemos qué nuevas variables se podrían incorporar, consideraremos como proxies los exponentes de la variable estimada: $\hat{Y}^2$ o $\hat{Y}^3$ y propondremos un modelo ampliado:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \delta_1\hat{Y}^2 + \delta_2\hat{Y}^3 + \boldsymbol{v} \tag{10.7}$$

sobre el que planteamos como hipótesis nula que todos los coeficientes de la parte no lineal del modelo son nulos, esto es, que el modelo inicial está correctamente especificado.

$$\begin{array}{|ll|}\hline H_0: & \delta_1 = \delta_2 = 0 \\ H_1: & \delta_1 \neq 0\ y/o\ \delta_2 \neq 0 \\ \hline\end{array}$$

Para realizar este contraste se utiliza la prueba F comparando los residuos del modelo base (restringido) y del modelo ampliado (se trataría por tanto de un caso particular del test de restricciones lineales)

$$\left(\frac{\widehat{u}'_R\hat{u}_R - \hat{u}'\hat{u}}{\hat{u}'\hat{u}}\right)\frac{n-k}{r} \approx F^r_{n-k}$$

10.4.2. Alteración de las hipótesis sobre la perturbación

Hemos enunciado diferentes supuestos sobre la perturbación aleatoria $\mathbf{u}$: esperanza nula, varianza constante, ausencia de autocorrelación y distribución normal. Dado que la perturbación $\mathbf{u}$ es un vector aleatorio de carácter no observable, la única posibilidad de analizar el cumplimiento de los supuestos será adoptar como referencia el vector de residuos $\hat{\mathbf{u}}$ asociados al modelo.

10.4.2.1. Perturbaciones de media no nula

El supuesto de esperanza nula para los residuos resulta fácilmente admisible, ya que parece lógica la compensación de desviaciones por exceso con otras por defecto. Sin embargo, podría ocurrir que, debido por ejemplo a especificaciones incorrectas del modelo, las perturbaciones presentasen una componente sistemática con lo cual su valor esperado ya no sería nulo.

Las consecuencias de este hecho son distintas según que $E(\mathbf{u})$ sea constante o variable: en el primer caso, el efecto se produce solamente sobre el término independiente por lo que no suele resultar grave pero en cambio si la esperanza de las perturbaciones es variable, los estimadores pasarán a ser sesgados e inconsistentes.

$\mathbf{E(u)}$ constante

Si $E(\mathbf{u})$ es constante, esta componente afecta tan sólo a las conclusiones sobre el término independiente ya que se obtendría:

$$Y = \beta_1 + \beta_2 X_2 + \cdots + \beta_k X_k + u \;\Rightarrow\; E(Y) = (\beta_1 + E(u)) + \beta_2 X_2 + \cdots + \beta_k X_k$$

$E(\mathbf{u})$variable

En cambio, cuando las perturbaciones presentan esperanza variable la situación resulta más grave, ya que en este supuesto se comprueba que los estimadores mínimo cuadráticos resultan sesgados e inconsistentes. De hecho, las perturbaciones de media variable son una consecuencia de la omisión de variables relevantes analizada anteriormente.

Como hemos visto, en este caso se propondrían expresiones del tipo $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{v}$, en lugar de la especificación correcta $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$. Tomando esperanzas se obtiene:

$$E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + E(\mathbf{v})$$
$$E(\mathbf{y}) = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

y como consecuencia $E(\mathbf{v}) = \mathbf{X}_2\boldsymbol{\beta}_2 \neq 0$.

La hipótesis de esperanza nula de la perturbación no puede ser contrastada mediante la información muestral, ya que el procedimiento mínimo cuadrático garantiza un valor nulo para la suma (y en consecuencia también para la media) de los errores de estimación.

10.4.2.2. Matriz de varianzas-covarianzas no escalar

En el modelo lineal básico se asume el supuesto de una matriz de covarianzas escalar: $Cov(\mathbf{u}) = E(\mathbf{u}\mathbf{u}') = \sigma^2\mathbf{I_n}$, expresión que resume las hipótesis de *homoscedasticidad y*

no autocorrelación. Como consecuencia de la primera, la diagonal de la matriz está formada por varianzas constantes $E(u_i^2) = \sigma^2$,$\forall i = 1, \ldots, n$, mientras que el segundo supuesto garantiza la nulidad de los restantes elementos de la matriz: $E(u_i u_j) = 0$,$\forall i \neq j$.

Resulta interesante estudiar las consecuencias del incumplimiento del supuesto $E(\mathbf{uu'}) = \sigma^2\mathbf{I_n}$, para lo cual asumiremos que se cumple $E(\mathbf{uu'}) = V$ o bien $E(\mathbf{uu'}) = \sigma^2\mathbf{\Omega}$ donde $\mathbf{\Omega}$ es una matriz definida positiva no escalar. En este caso los elementos de la diagonal principal de la matriz de varianzas-covarianzas no son coincidentes y sus restantes elementos pueden ser no nulos (en estas condiciones las perturbaciones se denominan *no esféricas*).

Si en esta situación se lleva a cabo la estimación por MCO se obtiene: $\hat{\boldsymbol{\beta}}^{MCO} = (\mathbf{X'X})^{-1}\mathbf{X'y}$, expresión que resulta insesgada y consistente pero no óptima para estimar $\boldsymbol{\beta}$.

- Los estimadores $\hat{\boldsymbol{\beta}}^{MCO}$ son insesgados: $E\left(\hat{\boldsymbol{\beta}}^{MCO}\right) = \beta$
- Los estimadores son consistentes
- Los estimadores no son óptimos ya que su matriz de varianzas-covarianzas es ahora:

$$E\left[\left(\hat{\boldsymbol{\beta}}^{MCO}\text{-}\beta\right)\left(\hat{\boldsymbol{\beta}}^{MCO}\text{-}\beta\right)'\right] = E\left[(\mathbf{X'X})^{-1}\mathbf{X'uu'X}(\mathbf{X'X})^{-1}\right] = \sigma^2\left[(\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1}\right]$$

expresión que no coincide con $\sigma^2(\mathbf{X'X})^{-1}$ que, como ya hemos visto es el valor mínimo (obsérvese que la coincidencia se produciría si $\mathbf{\Omega} = \mathbf{I}$, en cuyo caso la matriz sería escalar).

Para llevar a cabo la estimación de la matriz anterior, es necesario sustituir σ^2 por su estimador insesgado, que viene ahora dado por la expresión: $S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{tr(\mathbf{M\Omega})}$

Este estimador supone un cambio respecto a la expresión utilizada en el modelo básico $S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}$. La justificación es que, sustituyendo los residuos por su expresión $\hat{\mathbf{u}} = \mathbf{Mu}$, se obtiene ahora:

$$E\left(\hat{\mathbf{u}}'\hat{\mathbf{u}}\right) = tr\mathrm{M}E\left(\mathbf{u}'\mathbf{u}\right) = \sigma^2 tr\mathbf{M\Omega}$$

Así pues, si sobre un modelo con perturbaciones no esféricas aplicamos los desarrollos de MCO ignorando esta violación de supuestos, introduciríamos un doble sesgo al estimar la matriz de varianzas-covarianzas, ya que asumiríamos que ésta viene dada por la expresión $S^2_{\hat{\beta}^{MCO}} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{n-k}(\mathbf{X'X})^{-1}$ cuando la expresión correcta es: $S^2_{\hat{\beta}^{MCO}} = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{tr\mathbf{M\Omega}}\left[(\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1}\right]$. Se aprecia fácilmente que estaríamos cuantificando los dos componentes de esta expresión incorrectamente, sin que sea posible a priori avanzar el signo del sesgo cometido.

Con el objetivo de evitar los problemas señalados y conseguir estimadores que mejoren los MCO se propone una transformación del modelo hasta obtener una matriz de

varianzas-covarianzas que sea escalar. El método seguido para ello es buscar una matriz $\mathbf{P}$ cuadrada no singular de orden n, tal que $\mathbf{P}'\mathbf{P} = \mathbf{\Omega}^{-1}$, donde $\mathbf{\Omega}^{-1}$ es simétrica y definida positiva por ser la inversa de $\mathbf{\Omega}$.

Si premultiplicamos el modelo de regresión por esta matriz $\mathbf{P}$ se obtiene:

$$\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pu} \tag{10.8}$$

modelo transformado para el que se cumplen los supuestos básicos sobre las perturbaciones (esperanza nula y matriz de varianzas-covarianzas escalar).

En efecto, se tiene:

$$\begin{aligned} E(\mathbf{Pu}) &= \mathbf{0} \\ Var(\mathbf{Pu}) &= E\left[\mathbf{Pu}(\mathbf{Pu})'\right] = E(\mathbf{Puu'P'}) = \sigma^2\mathbf{P\Omega P'} \\ &= \sigma^2\mathbf{P}(\mathbf{P'P})^{-1}\mathbf{P'} = \sigma^2\mathbf{PP}^{-1}(\mathbf{P'})^{-1}\mathbf{P'} = \sigma^2\mathbf{I} \end{aligned}$$

La aplicación a este modelo transformado 10.8 de los estimadores mínimos cuadrados proporciona los *estimadores de mínimos cuadrados generalizados* (MCG) denominados también *estimadores de Aitken* en honor del autor que en 1935 planteó por primera vez esta estimación.

La expresión de los estimadores MCG se obtiene aplicando la expresión mínimo cuadrática al modelo 10.8, es decir:

$$\hat{\boldsymbol{\beta}}^{MCG} = \left[(\mathbf{PX})'(\mathbf{PX})\right]^{-1}(\mathbf{PX})'\mathbf{Py} = \left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}y}$$

y teniendo en cuenta que el modelo transformado cumple todas las hipótesis relativas a las perturbaciones $\mathbf{u}$, podemos garantizar que estos estimadores serán lineales insesgados y óptimos (ELIO).

Si expresamos el modelo transformado 10.8 como $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*$ sería posible aplicar las expresiones deducidas en el modelo básico para la estimación mínimo-cuadrática, equivalentes a las anteriores:

	Modelo Transformado	Modelo Inicial
	$\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*$	$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
Estimadores	$\hat{\boldsymbol{\beta}}^{MCG} = \left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}\mathbf{X}^{*'}\mathbf{y}$	$\hat{\boldsymbol{\beta}}^{MCG} = \left(\mathbf{X'\Omega^{-1}X}\right)^{-1}\mathbf{X'\Omega^{-1}y}$
Matriz Var-Cov	$Cov\left(\hat{\boldsymbol{\beta}}^{MCG}\right) = \sigma^2\left(\mathbf{X}^{*'}\mathbf{X}^*\right)^{-1}$	$Cov\left(\hat{\boldsymbol{\beta}}^{MCG}\right) = \sigma^2\left(\mathbf{X'\Omega^{-1}X}\right)^{-1}$

La matriz de varianzas-covarianzas viene dada en este caso por la expresión:

$$Var\left(\hat{\boldsymbol{\beta}}^{MCG}\right) = E\left[\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{MCG}} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{MCG}} - \boldsymbol{\beta}\right)'\right] = \sigma^2\left(\mathbf{X'\Omega X}\right)^{-1}$$

siendo $S^2 = \frac{\hat{\mathbf{u}}'\mathbf{\Omega}^{-1}\hat{\mathbf{u}}}{n-k}$ el estimador insesgado de σ^2, con los residuos obtenidos mediante el modelo de MCG: $\hat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}^{MCG}$

Se comprueba fácilmente que los estimadores $\hat{\boldsymbol{\beta}}^{\boldsymbol{MCG}}$ son insesgados: $E\left(\hat{\boldsymbol{\beta}}^{\boldsymbol{MCG}}\right) = \boldsymbol{\beta}$ y consistentes.

Por su parte, la matriz de varianzas covarianzas se obtiene como:

$$\begin{aligned}
Cov\left(\hat{\boldsymbol{\beta}}^{MCG}\right) &= E\left[\left(\hat{\boldsymbol{\beta}}^{MCG} - \boldsymbol{\beta}\right)\left(\hat{\boldsymbol{\beta}}^{MCG} - \boldsymbol{\beta}\right)'\right] \\
&= E\left[\left(\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} - \boldsymbol{\beta}\right)\left(\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y} - \boldsymbol{\beta}\right)'\right] \\
&= E\left[\left(\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}(\mathbf{X}\beta + \mathbf{u}) - \boldsymbol{\beta}\right)\left(\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}(\mathbf{X}\beta + \mathbf{u}) - \boldsymbol{\beta}\right)'\right] \\
&= E\left[\begin{array}{l}\left(\left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}\left(X'\boldsymbol{\Omega}^{-1}X\right)\beta + \left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}X'\boldsymbol{\Omega}^{-1}u - \boldsymbol{\beta}\right)\times \\ \times\left(\left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}\left(X'\boldsymbol{\Omega}^{-1}X\right)\beta + \left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}X'\boldsymbol{\Omega}^{-1}u - \boldsymbol{\beta}\right)'\end{array}\right] \\
&= E\left[\left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}X'\boldsymbol{\Omega}^{-1}u\left(\left(X'\boldsymbol{\Omega}^{-1}X\right)^{-1}X'\boldsymbol{\Omega}^{-1}u\right)'\right] \\
&= E\left[\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)'\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{u}\mathbf{u}'\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\right] \\
&= \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)'\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{E}(\mathbf{u}\mathbf{u}')\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \\
&= \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)'\mathbf{X}'\boldsymbol{\Omega}^{-1}\sigma^{2}\boldsymbol{\Omega}\boldsymbol{\Omega}^{-1}\mathbf{X}\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \\
&= \sigma^2\left[\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\left(\boldsymbol{\Omega}^{-1}\boldsymbol{\Omega}\right)\boldsymbol{\Omega}^{-1}\boldsymbol{X}\left(\mathbf{X}\cdot\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}\right] = \sigma^2\left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1}
\end{aligned}$$

expresión que resulta mínima en la clase de estimadores lineales insesgados para el modelo transformado.

De forma más sencilla, se puede llegar a esta expresión partiendo del modelo transformado $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta} + \mathbf{u}^*$

$$Cov\left(\hat{\beta}^{MCG}\right) = \sigma^2\left(X^{*'}X^*\right)^{-1} = \sigma^2\left(X'P'PX\right)^{-1} = \sigma^2\left(X'\Omega^{-1}X\right)^{-1}$$

Del mismo modo, para la estimación de la varianza se tiene:

$$\begin{aligned}
S^2 &= \frac{\hat{\mathbf{u}}'_{MCG}\hat{\mathbf{u}}_{MCG}}{n-k} = \frac{1}{n-k}\left[\left(\mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}\right)'\left(\mathbf{y}^* - \mathbf{X}^*\hat{\boldsymbol{\beta}}\right)\right] \\
&= \frac{1}{n-k}\left[\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'P'P\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)\right] = \frac{1}{n-k}\left[\hat{\mathbf{u}}'\boldsymbol{\Omega}^{-1}\hat{\mathbf{u}}\right]
\end{aligned}$$

La presencia de la matriz $\boldsymbol{\Omega}$ en las expresiones de los estimadores de MCG resulta problemática, ya que difícilmente se concibe que, desconociendo los parámetros, conozcamos los valores de las varianzas-covarianzas de las perturbaciones aleatorias.

Bajo la hipótesis de normalidad para las perturbaciones, es posible comprobar que, en el caso de una matriz de varianzas-covarianzas no escalar, los estimadores máximo verosímiles de $\boldsymbol{\beta}$ coinciden con las expresiones de MCG anteriormente deducidas.

A la vista de los análisis anteriores, podemos concluir que, si las perturbaciones aleatorias presentan matriz de varianzas-covarianzas no escalar, la utilización de MCO conducirá a estimadores no óptimos y a varianzas muestrales incorrectas, con lo cual los

contrastes de significación sobre el modelo carecerán de validez. Estos inconvenientes son solucionados por los estimadores de MCG tal y como resume en la tabla siguiente:

MCO	MCG
$E\left(\hat{\boldsymbol{\beta}}^{MCO}\right) = \boldsymbol{\beta}$	$E\left(\hat{\boldsymbol{\beta}}^{MCG}\right) = \boldsymbol{\beta}$
$Cov\left(\hat{\boldsymbol{\beta}}^{MCO}\right) = \sigma^2\left[(\mathbf{X'X})^{-1}\mathbf{X'\Omega X}(\mathbf{X'X})^{-1}\right]$	$Cov\left(\hat{\boldsymbol{\beta}}^{MCG}\right) = \sigma^2\left[\mathbf{X'\Omega^{-1}X}\right]^{-1}$
$S^2 = \frac{\hat{\mathbf{u}}'\hat{\mathbf{u}}}{tr(\mathbf{M\Omega})}$	$S^2 = \frac{\hat{\mathbf{u}}'\mathbf{\Omega}^{-1}\hat{\mathbf{u}}}{n-k}$

10.4.2.3. Heteroscedasticidad. Detección y soluciones

Según el supuesto de *homoscedasticidad*, todas las perturbaciones u_i presentan idéntica varianza. Sin embargo, en la práctica -especialmente en los modelos de corte transversal- sucede con frecuencia que estas varianzas se ven afectadas por los valores de la variable, cumpliéndose entonces que $\exists i \neq j \quad \sigma_i \neq \sigma_j$, fenómeno denominado *heteroscedasticidad*.

Causas Las *causas de la heteroscedasticidad* son de diversa índole: una de ellas puede ser la omisión de alguna variable relevante en la especificación del modelo, que introduce un efecto que se acumula en el residuo. Como consecuencia, si la variable presenta tendencia, ésta origina mayores residuos al aumentar su valor con el consiguiente incumplimiento del supuesto de homoscedasticidad.

En otros casos puede producirse un cambio estructural, que da lugar a una alteración en la dimensión de las perturbaciones y en su varianza antes y después de determinado acontecimiento.

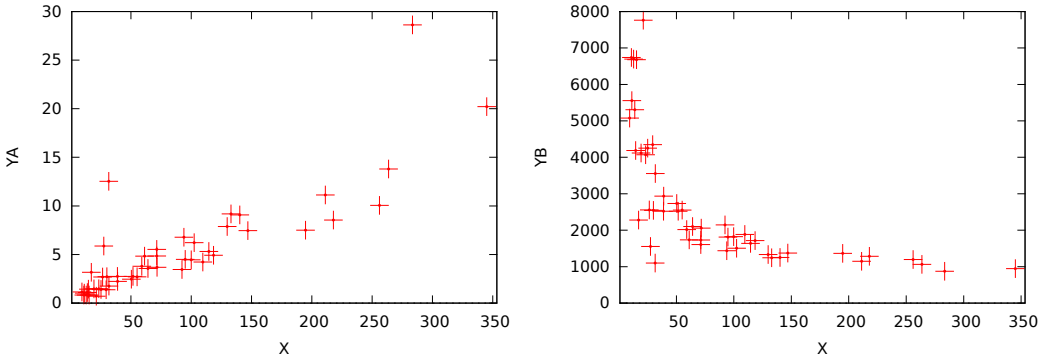
Bajo el supuesto de homoscedasticidad se cumplía $E(\mathbf{uu'}) = \sigma^2\mathbf{I_n}$, pero si se viola esta hipótesis la diagonal de la matriz estará formada por elementos σ_i^2 no coincidentes. Dado que se trata de un caso particular del supuesto de matriz $E(\mathbf{uu'})$ no escalar, *las consecuencias de la heteroscedasticidad* serán la obtención de estimadores que -aunque insesgados y consistentes- no son óptimos, la presencia de un sesgo en el cálculo de las varianzas muestrales y la aplicación incorrecta de los contrastes de significación.

Con la presencia de heteroscedasticidad, los estimadores MCO seguirán siendo insesgados y consistentes, pero dado que la matriz de Covarianzas es no escalar, sería válido lo desarrollado en el epígrafe anterior.

Por lo que se refiere a la *detección de la heteroscedasticidad*, una primera aproximación consiste en observar la nube de puntos, para saber si la dispersión varía con las observaciones de alguna variable explicativa. Además, existen diversos contrastes para identificar el problema a partir de procedimientos tanto paramétricos como no paramétricos.

Entre los segundos se encuentran el *contraste de picos*, basado en la representación gráfica de los residuos del modelo y el *contraste de rangos*. Por su parte, los contrastes paramétricos incluyen el test de Goldfeld y Quandt, el de White o el de Breusch-Pagan.

Figura 10.4.: Gráficos de heterocedasticidad



Análisis gráfico Una de las herramientas más claras para observar el cumplimiento de la hipótesis de homoscedasticidad es la representación gráfica de los residuos del modelo.

Bajo la hipótesis nula de homoscedasticidad esperamos un comportamiento de los residuos en el que la variabilidad no aumente ni disminuya con los valores de X, o no detectemos la existencia de varias subpoblaciones en la muestra.

Contraste de Goldfeld y Quandt El contraste de Goldfeld y Quandt (1965) plantea como hipótesis nula el supuesto de homoscedasticidad:

$$\begin{array}{ll} H_0 : \sigma_i^2 = \sigma_j^2 & \forall i = 1, \cdots, n \\ H_1 : \sigma_i^2 = g(X_i) & \text{siendo } g \text{ una función monótona} \end{array}$$

Para llevar a cabo este test resulta necesario ordenar los datos de la muestra según los valores de la variable explicativa ligada a la heteroscedasticidad (es decir, la que presumiblemente está relacionada con las varianzas de los perturbaciones).

A continuación, se eliminan p valores centrales, al objeto de tener dos submuestras claramente diferenciadas: la primera integrada por los $\frac{n-p}{2}$ primeros datos y la segunda por otros tantos correspondientes a las observaciones finales. Sobre cada una de estas submuestras se llevan a cabo estimaciones del modelo, obteniendo las correspondientes distribuciones de residuos, que denotamos respectivamente por $\hat{\mathbf{u}}_1$ y $\hat{\mathbf{u}}_2$.

El hecho de realizar dos regresiones diferenciadas permite separar en la medida de lo posible los valores extremos de la variable a la que se asocia la heteroscedasticidad y además garantiza la independencia necesaria entre las formas cuadráticas para definir a partir de ellas una F de Snedecor.

La comparación de estas distribuciones de errores se lleva a cabo por cociente, obteniendo una distribución F de Snedecor con $\frac{n-p}{2} - k$ grados de libertad tanto en el numerador como en el denominador:

$$\frac{\hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2}{\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1} \approx F_{\frac{n-p}{2}-k}^{\frac{n-p}{2}-k}$$

Si el valor del estadístico es elevado (y por tanto el nivel crítico es reducido) el resultado indicará que la dispersión de la segunda submuestra es muy superior a la de la primera, por lo cual conducirá al rechazo de la hipótesis nula de homoscedasticidad.

Contraste de White El test de White (1980) establece como el anterior la hipótesis nula de homocedasticidad si bien en este caso la alternativa es la heteroscedasticidad en un sentido más amplio

$$\begin{array}{ll} H_0 : \sigma_i^2 = \sigma_j^2 & \forall i = 1, \cdots, n \\ H_1 : \sigma_i^2 \neq \sigma_j^2 & \text{para algún i}\neq\text{j} \end{array}$$

La resolución del contraste se basa en la regresión de los cuadrados de los residuos del modelo estimado sobre las variables explicativas del modelo, sus cuadrados y todos los productos cruzados, es decir:

$$\hat{u}^2 = \alpha_0 + \sum_{i,j=1}^{k} \alpha_{ij} X_i X_j + \eta$$

Bajo la hipótesis nula de homoscedasticidad se cumple $nR^2 \approx \chi_m^2$, donde n es el tamaño muestral, R^2 el coeficiente de determinación en la regresión auxiliar y m el número de regresores de este modelo auxiliar sobre los residuos cuadráticos (es decir, número de parámetos menos 1, ya que se excluye el término independiente),

En caso de que el número de grados de libertad fuera reducido se podría proponer un modelo similar al anterior pero excluyendo los productos cruzados de variables.

Si la hipótesis nula de homoscedasticidad es cierta se esperan valores bajos del coeficiente de determinación del modelo auxiliar, y en consecuencia también del estadístico chi-cuadrado. Así pues, a medida que esta expresión aumenta su valor nos proporciona argumentos para el rechazo del supuesto de homoscedasticidad.

El test de White se dice que es el más general por cuanto no exige supuestos previos al comportamiento de los residuos (no exige normalidad) ni tampoco hay que pronunciarse con antelación sobre las variables X que pueden estar causando esta heteroscedasticidad.

Una vez identificado el problema y su detección resulta interesante apuntar soluciones para la heteroscedasticidad. En este sentido, sería aplicable la posibilidad ya estudiada de sustituir el método de mínimos cuadrados por una generalización que proporcione matrices de varianzas-covarianzas escalares. En concreto, resulta conveniente investigar la relación entre las varianzas y las variables explicativas para plantear un

modelo transformado o ponderado al que se aplican mínimos cuadrados.

A modo de ilustración, si asumimos el supuesto teórico $\sigma_i^2 = \sigma^2 X_i$ se buscaría una matriz de ponderaciones $\mathbf{P}$ tal que el modelo transformado $\mathbf{Py} = \mathbf{PX}\boldsymbol{\beta} + \mathbf{Pu}$ sea homoscedástico, esto es, $E[(\mathbf{Pu})'(\mathbf{Pu})] = \sigma^2\mathbf{I}$.

En consecuencia, podríamos buscar las ponderaciones P_i necesarias para cada valor:

$$E\left(P_i u_i\right)^2 = \sigma^2 \;\Rightarrow\; P_i^2 E\left(u_i\right)^2 = \sigma^2 \;\Rightarrow\; P_i^2 \sigma^2 X_i = \sigma^2 \;\Rightarrow\; P_i = \frac{1}{\sqrt{X_i}}$$

[¿Cuál sería el razonamiento si $\sigma_i^2 = \sigma^2 X_i^2$? ¿Y si $\sigma_i^2 = \frac{\sigma^2}{X_i^2}$?]

En la práctica la determinación de ponderaciones no resulta sencilla al ser desconocidas las varianzas poblacionales. Por ello suele seguirse el procedimiento propuesto por Glejser (1969), que consiste en plantear diferentes regresiones de los residuos (en términos absolutos o cuadráticos) respecto a la variable explicativa asociada a la heteroscedasticidad, seleccionando entre todas ellas la más significativa.

10.4.2.4. Autocorrelación. Contraste de Durbin-Watson

La alteración de la hipótesis de no autocorrelación entre las perturbaciones puede producirse por diferentes causas, como la presencia de "inercia" en los acontecimientos económicos y sociales (que extiende las consecuencias de cierta acción a varios períodos de tiempo), la especificación errónea del modelo, o ciertos cambios estructurales que pueden producir errores sistemáticos y autocorrelados.

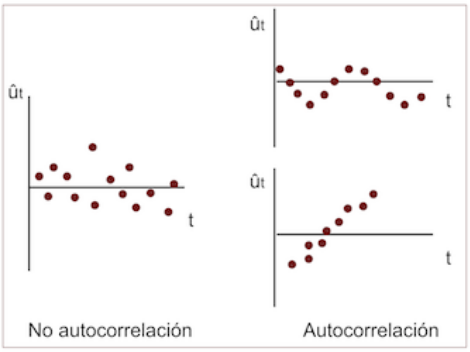
La presencia de autocorrelación se estudia en general en datos de serie temporal. Sin embargo, también para estudios de corte transversal la proximidad entre observaciones puede introducir correlaciones entre las correspondientes perturbaciones.

Por lo que respecta a los *efectos introducidos por la autocorrelación*, las consecuencias son las propias de una matriz no escalar, que como ya hemos comentado son la ineficiencia de los estimadores y la presencia de un sesgo en la varianza muestral, que desaconseja realizar contrastes de significación.

Para *detectar la presencia de autocorrelación*, resulta muy aconsejable como primera aproximación un examen gráfico de los residuos, cuyo patrón de comportamiento temporal puede delatar la presencia de correlaciones lineales tal y como recoge la figura 10.5

Aunque en principio la presencia de autocorrelación vendría descrita de forma genérica como $E(u_i u_j) \neq 0$, resulta conveniente especificar ciertos esquemas concretos de correlación entre los residuos. Así, el *contraste de autocorrelación de Durbin y Watson* de utilización generalizada, considera las perturbaciones relacionadas según el esquema: $u_t = \rho u_{t-1} + \epsilon_t$, donde u_t recoge la perturbación asociada al instante t, y se cumple:

Figura 10.5.: Autocorrelación



$$|\rho| < 1 \ , \ \epsilon_t \approx \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$$

Esta especificación se conoce como *modelo autorregresivo de orden 1* o $AR(1)$ debido a que cada perturbación se relaciona consigo misma desfasada un período. Por tanto, se adopta un esquema de comportamiento de los errores en el que aparece una componente sistemática -la incidencia del instante anterior- y otra aleatoria (ϵ), para la que se asumen los supuestos de normalidad, esperanza nula, homoscedasticidad e incorrelación.

A las variables ϵ que cumplen estos requisitos se las denomina *ruidos blancos*, siendo su utilización frecuente en la modelización estocástica de series temporales. Sustituyendo en la expresión de u_t se obtiene:

$$u_t = \rho u_{t-1} + \epsilon_t = \rho(\rho u_{t-2} + \epsilon_{t-1}) + \epsilon_t = \epsilon_t + \rho\epsilon_{t-1} + \rho^2\epsilon_{-2} + \cdots = \sum_{i=0}^{\infty} \rho^i \epsilon_{t-i}$$

con las características siguientes:

$$E\left(u_t\right) = E\left(\sum_{i=0}^{\infty} \rho^i \epsilon_{t-i}\right) = 0$$

$$\begin{aligned} Var(u_t) = \ & E\left(\sum_{i=0}^{\infty} \rho^i \epsilon_{t-i}\right) = E\left(\sum_{i=0}^{\infty} \rho^{2i} \epsilon_{t-i}^2 + \sum_{i \neq j} \underbrace{\rho^i \rho^j \epsilon_{t-i}\epsilon_{t-j}}_{E(\epsilon_{t-i}\epsilon_{t-j})=0}\right) \\ = \ & \sum_{i=0}^{\infty} \rho^{2i} E\left(\epsilon_{t-i}^2\right) = \frac{\sigma_\epsilon^2}{1-\rho^2} \end{aligned}$$

Esta varianza

$$\sigma_u^2 = \frac{\sigma_\epsilon^2}{1-\rho^2}$$

depende de la varianza de ϵ (que se asume constante) y de la autocorrelación ρ.

El contraste de ausencia de autocorrelación se expresa:

$$\boxed{\begin{array}{c} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{array}}$$

y dado que las perturbaciones u no son observables el contraste debe llevarse a cabo con sus errores de estimación $\hat{u}$. Partiendo de estos residuos, Durbin y Watson (1950) definieron la expresión:

$$d_{DW} = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

que, para tamaños elevados de muestra podrá escribirse aproximadamente como: $d \approx (1-\hat{\rho})$ siendo $\hat{\rho}$ el coeficiente de autocorrelación estimado.

Los valores de este coeficiente estimado permiten extraer distintas conclusiones sobre la autocorrelación:

- Si $\hat{\rho} \backsimeq 1$ entonces $d \approx 0$ y estaremos en situación de autocorrelación positiva, esto es, la perturbación en cada instante está muy influida por la del instante anterior.
- Si $\hat{\rho} \backsimeq -1$ entonces se obtiene d próximo al valor 4, asociado a la autocorrelación negativa (en este caso a un valor alto de la perturbación le seguirá uno bajo y viceversa).
- Si $\hat{\rho} \backsimeq 0$ el estadístico d toma valores cercanos a 2, indicativos de la ausencia de correlación serial.

Es posible comprobar que $0 \leq d \leq 4$ y que el caso particular $d = 2$ indica ausencia de autocorrelación. Además, el estadístico d es función del tamaño muestral, de la matriz $\mathbf{X}$ de datos y del número de regresores $k' = k - 1$.

Para tratar de solucionar este inconveniente, Durbin y Watson demostraron que la distribución de d está comprendida entre otras dos distribuciones auxiliares que denominan d_L y d_U y que no dependen de $\mathbf{X}$ sino sólo del tamaño muestral (n) y del número de parámetros (k) o de variables $(k' = k - 1)$. La comparación del estadístico d con estas distribuciones auxiliares conduce a las siguientes conclusiones para el contraste de la hipótesis de ausencia de autocorrelación:

Tabla 10.4.: Contraste de Durbin-Watson. Valores significativos al 5 %

$k =$	2		3		4		5		6		10	
n	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
10	0,879	1,320	0,697	1,641	0,525	2,016	0,376	2,414	0,243	2,822		
11	0,927	1,324	0,758	1,604	0,595	1,928	0,444	2,283	0,316	2,645		
12	0,971	1,331	0,812	1,579	0,658	1,864	0,512	2,177	0,379	2,506		
13	1,010	1,340	0,861	1,562	0,715	1,816	0,574	2,094	0,445	2,390		
14	1,045	1,350	0,905	1,551	0,767	1,779	0,632	2,030	0,505	2,296	0,127	3,360
15	1,077	1,361	0,946	1,543	0,814	1,750	0,685	1,977	0,562	2,220	0,175	3,216
20	1,201	1,411	1,100	1,537	0,998	1,676	0,894	1,828	0,792	1,991	0,416	2,704
25	1,288	1,454	1,206	1,550	1,123	1,654	1,038	1,767	0,953	1,886	0,621	2,419
30	1,352	1,489	1,284	1,567	1,214	1,650	1,143	1,739	1,071	1,833	0,782	2,251
35	1,402	1,519	1,343	1,584	1,283	1,653	1,222	1,726	1,160	1,803	0,908	2,144
40	1,442	1,544	1,391	1,600	1,338	1,659	1,285	1,721	1,230	1,786	1,008	2,072
45	1,475	1,566	1,430	1,615	1,383	1,666	1,336	1,720	1,287	1,776	1,089	2,022
50	1,503	1,585	1,462	1,628	1,421	1,674	1,378	1,721	1,335	1,771	1,156	1,986
100	1,654	1,694	1,634	1,715	1,613	1,736	1,592	1,758	1,571	1,780	1,484	1,874

$d < d_L$	Rechazo de la hipótesis nula (Zona de autocorrelación positiva)
$d_U < d < 4 - d_U$	No rechazo de la hipótesis nula (Zona de no autocorrelación)
$d > 4 - d_L$	Rechazo de la hipótesis nula (Zona de autocorrelación negativa)

Las distribuciones de d_L y d_U aparecen tabuladas para distintos tamaños muestrales y número de variables explicativas. En dichas tablas puede observarse que existen zonas de duda en las que no es posible llegar a conclusiones definitivas (serían de aceptación para una distribución auxiliar y de rechazo para otra).

Dichas zonas no concluyentes son: $d_L \leq d \leq d_U$ y $4 - d_U \leq d \leq 4 - d_L$.

A modo de ilustración, recogemos algunos de los valores de estas variables auxiliares en la tabla 10.4

Por lo que se refiere a las *soluciones a la autocorrelación*, consisten en especificar un modelo autorregresivo -en el caso más sencillo un $AR(1)$- con lo cual el modelo podría ser expresado:

$$Y_t = \beta_1 + \beta_2 X_t + u_t = \beta_1 + \beta_2 X_t + (\rho u_{t-1} + \epsilon_t)$$

Dado que para el período anterior se tendría: $Y_{t-1} = \beta_1 + \beta_2 X_{t-1} + u_{t-1}$, es posible multiplicar esta segunda igualdad por ρ, obteniendo mediante diferencia el modelo transformado:

$$Y_t - Y_{t-1} = (1 - \rho)\beta_1 + (X_t - \rho X_{t-1})\beta_2 + \epsilon_t$$

es decir

$$Y_t^* = \beta_1^* + \beta_2 X_t^* + \epsilon_t$$

que si ϵ_t se adapta a los supuestos necesarios ya no presentará problemas de autocorrelación.

En la práctica, para el tratamiento de la autocorrelación se suele aplicar el procedimiento de Cochrane-Orcutt (1949), que abarca las etapas siguientes:

- Estimación por MCO y cálculo de los residuos, para estimar el valor $\hat{\rho}$
- Transformación del modelo $Y_t^* = \beta_1^* + \beta_2 X_t^* + \epsilon_t$ que se estima nuevamente por MCO
- Repetición de este procedimiento hasta que la diferencia entre dos estimaciones consecutivas de ρ sea muy pequeña (menos de 0,005)

No obstante, existe polémica respecto a la adecuación de este procedimiento porque no necesariamente conduce a un óptimo global.

10.4.2.5. No normalidad

La hipótesis de normalidad de la perturbación aleatoria es la base de todo el proceso inferencial sobre el modelo lineal básico. Por tanto, el incumplimiento de este supuesto podría afectar seriamente a los contrastes de significación desarrollados.

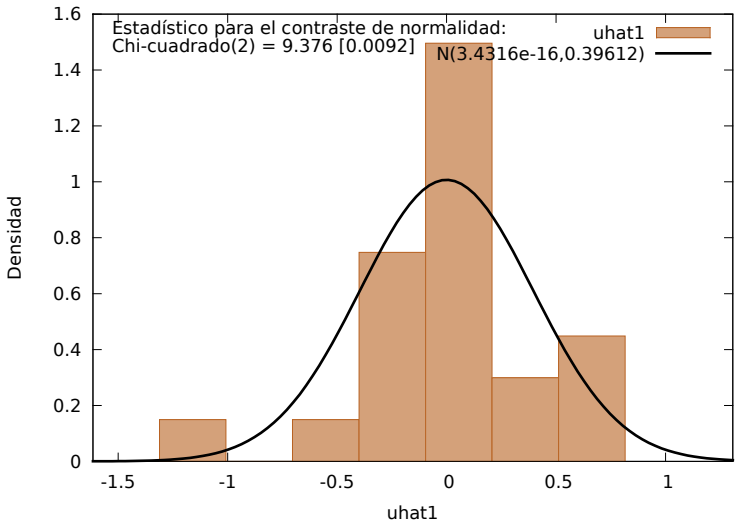
En apartados anteriores hemos comprobado que los estimadores mínimo cuadráticos son ELIO, es decir, lineales insesgados y óptimos. Estas propiedades se cumplen con independencia de la distribución probabilística de u y por tanto no se ven afectadas por el incumplimiento de la hipótesis de normalidad.

Sin embargo, el supuesto de normalidad añade afirmaciones más fuertes que el teorema de Gauss-Markov: si el vector de perturbaciones u es normal el vector de estimaciones máximo verosímiles (EMV) coincide con el vector de estimadores mínimo cuadráticos y además no existe ningún otro vector insesgado (lineal o no lineal) cuya varianza sea menor (teorema de Rao que extiende los resultados de Gauss-Markov).

Además, el incumplimiento de la hipótesis de normalidad impide conocer el modelo probabilístico seguido por el vector mínimo cuadrático y el de residuos. Como consecuencia, los estadísticos empleados en los procesos inferenciales, que seguían modelos chi cuadrado, t de Student o F de Snedecor, todos ellos derivados del normal, tampoco se adaptarán ahora a dichas distribuciones, con lo cual los contrastes habituales de hipótesis dejan de ser válidos (únicamente con tamaños muestrales elevados y bajo ciertas condiciones podrían obtenerse distribuciones asintóticas).

Parece por tanto aconsejable contrastar si el supuesto de normalidad de u es o no admisible y, dado que las verdaderas perturbaciones son desconocidas, el contraste de normalidad se lleva a cabo sobre los residuos del modelo $\hat{u}$ siguiendo los procedimientos

Figura 10.6.: Contraste de normalidad (bondad de ajuste)



habitualmente empleados (contrastes de Jarque-Bera, Kolmogorov-Smirnov o Chi-cuadrado).

La hipótesis nula es la normalidad de las perturbaciones y será contrastada a partir de los residuos. Por tanto nos estamos basando en la información muestral para estimar los parámetros, hecho que debe ser tenido en cuenta en el cálculo de los niveles críticos.

Así, si optásemos por el procedimiento de Kolmogorov-Smirnov para contrastar la normalidad de u, resultaría necesario llevar a cabo la estimación de los parámetros esperanza y varianza a partir de la muestra y las correspondientes probabilidades deberían ser consultadas en las tablas modificadas de Lilliefors.

10.4.3. Alteración de las hipótesis estructurales

Además de las hipótesis referidas a las perturbaciones es necesario examinar si el modelo se adapta a los supuestos de tipo estructural con los que habitualmente trabajamos.

10.4.3.1. Regresores estocásticos

En el desarrollo del modelo básico de regresión se asume que la matriz $\mathbf{X}$ de regresores es fija, es decir, adopta los mismos valores para distintas muestras. Esta hipótesis de regresores no estocásticos, que es admisible para las ciencias experimentales, puede sin embargo resultar restrictiva en ciencias sociales, ya que los datos se obtienen habitualmente por observación.

Teniendo en cuenta que en las investigaciones económicas dispondremos en general de muestras con información histórica sobre todas las magnitudes investigadas (regresando y regresores) resulta aconsejable estudiar qué efectos tendría sobre nuestros

resultados la consideración de *regresores estocásticos*.

Entre las razones que justifican la consideración de **X** como estocástica se encuentran la especificación de modelos que consideran como explicativas las variables endógenas retardadas. Así, si $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Y_{i-1} + u_i$ la variable Y_{i-1} es aleatoria por depender de la perturbación u_{i-1}.

Del mismo modo, cualquier modelo de ecuaciones simultáneas en el que aparezca como explicativa alguna variable endógena deberá considerarse por definición de regresores estocásticos. Otra posible razón del carácter estocástico de X es la presencia de errores en las variables del modelo, como consecuencia de una medición inadecuada de las mismas.

El carácter estocástico de la matriz **X** no afectará a nuestros resultados siempre que se cumplan dos condiciones:

- Las variables explicativas tienen distribución independiente de los parámetros de la regresión.
- Las variables explicativas tienen distribución independiente de la perturbación aleatoria.

Esta segunda condición no suele cumplirse en la práctica, hecho que afecta a los estimadores que pasan a ser sesgados y llevan asociadas matrices de varianzas-covarianzas inferiores a las reales.

Examinando las situaciones anteriormente planteadas se observa que en los modelos que incluyen como explicativas variables endógenas retardadas $Y_i = \beta_1 + \beta_2 X_i + \beta_3 Y_{i-1} + u_i$ tan sólo puede garantizarse la independencia entre valores contemporáneos de las variables explicativas y la perturbación aleatoria (X_i y u_i) en el caso de que no exista un proceso autorregresivo en las perturbaciones.

Por su parte, los modelos de ecuaciones simultáneas y los que contienen errores de medida en las variables incumplen sistemáticamente la condición de independencia entre las variables explicativas y la perturbación aleatoria.

10.4.3.2. Matrices X de rango no pleno

Hasta ahora hemos asumido que la matriz de regresores tiene rango k, esto es, $\rho(\mathbf{X}) = k$. Dado que la matriz **X** tiene k columnas (tantas como parámetros) y n filas (observaciones muestrales), esta hipótesis resume dos supuestos: por una parte, la información estadística disponible sobre el conjunto de variables observables debe ser suficientemente amplia para llevar a cabo la solución del modelo ($n > k$) y por otra, las columnas de la matriz **X** deben ser linealmente independientes, es decir, no debe existir relación lineal exacta entre los regresores del modelo.

El primer requisito va relacionado con el *tamaño muestral* que debe superar al número de parámetros k. A efectos operativos suele exigirse que el número de grados de libertad del modelo ($n-$k) sea suficientemente elevado para garantizar un proceso de estimación adecuado.

Debemos tener en cuenta que las expresiones utilizadas en los procesos inferenciales contienen explícitamente el número de los grados de libertad $n - k$. Por tanto, aunque un tamaño de muestra pequeño no viola ninguna de las hipótesis básicas del modelo, sí tiene consecuencias negativas al conducir a estimaciones que, aunque insesgadas y eficientes, presentan varianzas comparativamente más altas que las obtenidas con tamaños muestrales superiores.

Para evitar este tipo de problemas, es recomendable eliminar de un modelo las variables menos significativas, con lo cual se dispone de más grados de libertad. El principio de "parquedad" o "parsimonia" consiste en buscar el modelo que, con el mínimo número de variables explicativas, consiga un grado de eficacia explicativa comparable con otros más complejos.

Por otra parte, en el caso de que existiera relación lineal entre algún subconjunto de regresores, el rango de la matriz $\mathbf{X}$ sería inferior a k y por tanto no sería posible determinar los estimadores del modelo. Aparecería así una multicolinealidad perfecta, situación en la que se tiene:

$$\rho(\mathbf{X}) < k \Rightarrow \rho(\mathbf{X}'\mathbf{X}) < k \Rightarrow |\mathbf{X}'\mathbf{X}| = 0$$

con lo cual no resulta posible la determinación de los EMC.

10.4.3.3. Multicolinealidad

La presencia de relaciones lineales exactas entre los regresores no resulta frecuente en la práctica, por lo que la *multicolinealidad* perfecta se estudia tan sólo como un supuesto teórico extremo.

Como hemos visto, se presentaría un caso de multicolinealidad perfecta cuando en un modelo con variables explicativas de carácter cualitativo introdujéramos tantos regresores como modalidades tenga la característica investigada. Esta situación, denominada *trampa de la multicolinealidad de las variables ficticias* se soluciona reduciendo en uno el número de variables cualitativas introducidas en el modelo.

Otras situaciones en las que podría presentarse la multicolinealidad perfecta serían la presencia en el modelo de una variable explicativa con valor constante (perfectamente correlacionada por tanto con el término independiente) o de varias variables conectadas mediante una identidad.

En las investigaciones econométricas son frecuentes los modelos en los que aparece cierto grado de correlación (o *multicolinealidad aproximada*) entre las variables explicativas. Las razones de este hecho son la presencia de tendencias comunes a varios regresores o incluso la conexión teórica entre ellos y su principal consecuencia es el aumento en la matriz de varianzas-covarianzas de los estimadores.

Es importante destacar que las propiedades de los EMC no se ven afectadas por la presencia de una cierta multicolinealidad (siguen siendo insesgados, óptimos y consistentes) pero en cambio la matriz de varianzas-covarianzas, que depende de las relaciones existentes entre las variables explicativas, aumenta su valor. Como consecuencia, las expresiones de la t de Student aumentan su denominador, con lo cual resulta más difícil rechazar la hipótesis de no significación de los parámetros individuales. Además, la elevada varianza de los estimadores hace que éstos sean muy volátiles, por lo cual

podríamos cometer errores de interpretación estructural del modelo.

Las estimaciones obtenidas para modelos con un grado importante de multicolinealidad son muy poco estables, ya que al añadir nueva información muestral el modelo estimado podría cambiar radicalmente.

En cambio, esta multicolinealidad no afectará a las predicciones siempre que admitamos que las pautas de correlación se mantienen constantes en el período de predicción.

Las consecuencias comentadas proporcionan la base para la *detección de la multicolinealidad* en un modelo. En efecto, al tratarse de un problema muestral, la multicolinealidad no puede ser contrastada, pero sin embargo la observación conjunta de los resultados asociados a los contrastes individuales (t de Student) y global (F de Snedecor) permite comprobar si existen incoherencias entre ambos (podría ocurrir que el resultado del contraste global fuese significativo y en cambio ninguno de los individuales lo fuese) o bien si las estimaciones son muy volátiles respecto a la información muestral.

También resulta aconsejable llevar a cabo regresiones auxiliares para averiguar si alguna de las variables explicativas depende de las restantes.

De hecho, es posible comprobar que la varianza de los estimadores aumenta con la correlación entre las variables explicativas:

$$\sigma^2_{\hat{\beta}_j} = \frac{1}{n-k}\frac{\sigma^2_Y}{\sigma^2_X}\frac{1-R^2}{1-R^2_j}$$

donde R^2_j es el coeficiente de determinación de la regresión de X_j sobre las restantes variables independientes.

Un indicador habitual de la multicolinealidad son los Factores de Inflación de la Varianza (FIV) propuestos por Marquardt (1970) que vienen dados por la expresión:

$$FIV\left(\hat{\beta}_j\right) = \frac{1}{1-R^2_j}$$

donde R^2_j es el coeficiente de correlación múltiple entre la variable j y las restantes variables explicativas El FIV muestra en qué medida aumenta la varianza del estimador como consecuencia de no ortogonalidad de los regresores, y habitualmente se considera que existe un problema grave de multicolinealidad cuando el factor de inflación de varianza de algún coeficiente es mayor de 10.

Por lo que respecta a las posibles *soluciones al problema de la multicolinealidad*, podríamos plantear un aumento en la información muestral (o incluso la extramuestral), un cambio en el modelo especificado o en el método de estimación, ... En cualquier caso, conviene tener presente que se trata de un problema habitual con el que -siempre que no presente niveles excesivos- debemos convivir.

10.4.3.4. Cambio estructural

A partir de la expresión genérica de nuestro modelo de regresión $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$, el supuesto de que $\boldsymbol{\beta}$ es un vector fijo permite plantear la estimación de un modelo único con la información muestral disponible.

Sin embargo, a veces podemos tener sospechas sobre la existencia de alguna ruptura o cambio de estructura, que impediría admitir el supuesto de constancia de los parámetros. Este tipo de situaciones pueden deberse a un cambio en el sistema económico que se representa, o bien a una especificación errónea del modelo (omisión de variables o forma funcional inadecuada).

Para estudiar si este problema afecta a un modelo concreto, es posible llevar a cabo el *contraste de cambio estructural* de Chow (1960). La hipótesis nula de este contraste es la existencia de una estructura única válida para todo el período de observación y este supuesto se contrasta dividiendo la muestra en dos submuestras en las que el número de datos supere al de parámetros.

Así se plantean las tres regresiones siguientes:

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} & \text{con } \mathbf{u} \approx \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}_n\right) \\ \mathbf{y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u}_1 & \text{con } \mathbf{u}_1 \approx \mathcal{N}\left(\mathbf{0}, \sigma_1^2\mathbf{I}_{n_1}\right) \\ \mathbf{y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}_2 & \text{con } \mathbf{u}_2 \approx \mathcal{N}\left(\mathbf{0}, \sigma_2^2\mathbf{I}_{n_2}\right) \end{cases}$$

con lo cual la hipótesis nula de ausencia de cambio estructural equivale a afirmar que las dos muestras proceden de la misma población y puede ser expresada como $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}\,;\, \sigma_1 = \sigma_2 = \sigma$.

El método de Chow se basa en la discrepancia entre los residuos correspondientes a la regresión de la muestra total y la suma de residuos cuadráticos asociados a las regresiones de las dos submuestras, que convenientemente ajustadas por sus grados de libertad dan lugar a la expresión:

$$\frac{\dfrac{\hat{\mathbf{u}}'\hat{\mathbf{u}} - (\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2)}{k}}{\dfrac{\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1 + \hat{\mathbf{u}}_2'\hat{\mathbf{u}}_2}{n_1 + n_2 - 2k}} \approx F^k_{n_1+n_2-2k}$$

Esta discrepancia tipificada sigue un modelo F de Snedecor y su resultado se interpreta como el porcentaje de incremento en la suma de cuadrados de los residuos asociados a un modelo único, respecto a la que se obtendría con dos muestras.

[Justificar los grados de libertad de esta expresión]

Si el valor de la F es elevado está indicando un empeoramiento significativo del modelo como resultado de no dividir el período muestral, que lleva a rechazar la hipótesis nula.

Cuando el punto sobre el que se quiere contrastar el cambio estructural no permite disponer de dos muestras con tamaño suficientemente elevado ($n_2 < k$) el estadístico pasa a ser

$$\frac{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}} - \hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1}{n_2}}{\frac{\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1}{n_1 - k}} \approx F_{n_1-k}^{n_2}$$

con interpretación similar al anterior.

Los inconvenientes del contraste de Chow son que necesita conocer el punto de corte, y que pierde potencia a medida que dicho punto se acerca al extremo de la muestra global. Además, este contraste es sensible a la presencia de heteroscedasticidad por lo cual ésta debe ser corregida antes de contrastar los cambios estructurales.

En ocasiones se plantea el *contraste de Chow para la predicción*. En este caso, la hipótesis nula es la existencia de una estructura única válida para todo el período de observación y el horizonte de predicción del fenómeno estudiado.

El contraste se lleva a cabo estimando el modelo con las n_1 primeras observaciones y utilizándolo para predecir los n_2 últimos valores. Bajo la hipótesis nula se asume que las predicciones provienen del mismo modelo que los valores que dieron lugar a la estimación, y por tanto el estadístico de Chow viene dado por la expresión:

$$\frac{\frac{\hat{\mathbf{u}}'\hat{\mathbf{u}} - \hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1}{n_2}}{\frac{\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1}{n_1 - k}} \approx F_{n_1-k}^{n_2}$$

donde $\hat{\mathbf{u}}'\hat{\mathbf{u}}$ recoge la suma los residuos cuadráticos cometidos si la regresión se lleva a cabo para todos los valores muestrales mientras $\hat{\mathbf{u}}_1'\hat{\mathbf{u}}_1$ son los residuos cuadráticos cuando la regresión se extiende sólo a los n_1 primeros datos.

Bibliografía

[1] J. Aranda and J. Gómez. *Fundamentos de Estadística para Economía y Administración de Empresas*. Diego Martín, 2002.

[2] G. Arnáiz. *Introducción a la estadística teórica*. Lex Nova, 1986.

[3] F. Azorín and J.L. Sanchez Crespo. *Método y aplicaciones del muestreo*. Alianza Universidad, 1986.

[4] J. Baró Llinás. *Cálculo de probabilidades: aplicaciones económico-empresariales*. Parramón, 1985.

[5] M. Barrow. *Statistics for economics, accounting, and business studies*. Pearson Education, 2006.

[6] W.E. Becker and D.L. Harnett. *Business and economics statistics with computer applications*. Addison-Wesley, 1987.

[7] R. Behar and P. Grima. *55 Respuestas a dudas típicas de Estadística*. Díaz de Santos, 2004.

[8] M.L. Berenson and D.M. Levine. *Estadistica para administracion y economia: conceptos y aplicaciones*. Mc.Graw-Hill, 1991.

[9] D.A. Berry and B.W. Lindgren. *Statistics: Theory and Methodos*. Duxbury Press, 1996.

[10] H.D. Brunk. *Introducción a estadística matemática*. Trillas, 1987.

[11] G.C. Canavos. *Probabilidad y estadística: aplicaciones y métodos*. Mc.Graw-Hill, 2003.

[12] R. Cao, M.A. Presedo, and M.F. Fernández. *Introducción a la estadística y sus aplicaciones*. Pirámide, 2006.

[13] J.M. Casas. *Inferencia estadística para economía y administración de empresas*. Centro de Estudios Ramón Areces, 1996.

[14] J.M. Casas and J. Santos. *Introducción a la estadística para economía y administración de empresas*. Centro de Estudios Ramón Areces, 1995.

[15] G. Casella and R.L. Berger. *Statistical inference*. Textbook Reviews, 2006.

[16] Ya-Lun Chou. *Statistical analysis for business and economics.* Elsevier Science Publishing, 1989.

[17] H. Cramer. *Métodos matemáticos de estadística.* Aguilar, 1970.

[18] M. Cross and M.K. Starr. *Statistics for Business and Economics.* McGraw-Hill, 1983.

[19] N.M. Downie and R.W. Heath. *Métodos estadísticos aplicados.* Harla, 1986.

[20] R. Escuder. *Manual de teoría de la probabilidad con nociones de muestreo e inferencia estadística.* Tirant lo Blanch, 1992.

[21] H. Fernández, M.M. Guijarro, and J.L. Rojo. *Cálculo de probabilidades y estadística.* Ariel Economía, 1994.

[22] D. Freedman, R. Pisani, R. Purves, and A. Adhikari. *Estadística.* Antoni Bosh, 1993.

[23] J.E. Freund and F.J. Williams. *Elementos Modernos de Estadística Empresarial.* Prentice-Hall, 1989.

[24] A. García Barbancho. *Estadística teórica básica.* Ariel, 1992.

[25] J.D. Gibbons. *Nonparametric methods for quantitative analysis.* American Sciences Press, 1985.

[26] W.H. Greene. *Análisis econométrico.* Prentice-Hall, 1997.

[27] I. Herranz and L. Prieto. *¿Qué significa "estadísticamente significativo"?: la falacia del criterio del 5 % en la investigación científica.* Díaz de Santos, 2005.

[28] P.G. Hoel and R.J. Jessen. *Estadística básica para negocios y economía.* CECSA, 1986.

[29] P. Kauffman. *Statistique: information, estimation, tests.* Dunod, 1994.

[30] M. Kendall and A. Stuart. *The advanced theory of statistics (3 Vol.).* Charles Griffin, 1977.

[31] E.L. Lehmann. *Testing Statistical Hypotheses.* John Wiley and Sons, 1986.

[32] R.I. Levin. *Estadística para administradores.* Prentice Hall, 1988.

[33] F. Llorente and otros. *Inferencia estadística aplicada a la empresa.* Centro de Estudios Ramón Areces, 2001.

[34] M. López Cachero. *Fundamentos y Métodos de Estadística.* Pirámide, 1996.

[35] G.S. Maddala. *Econometría.* McGraw-Hill, 1985.

[36] J. Martín Pliego and L. Ruiz-Maya. *Estadística I: Probabilidad.* Paraninfo, 2004.

[37] R.D. Masson and D.A. Lind. *Estadística para Administración y la Economía.* Alfaomega, 1992.

[38] T. Mayer. *Truth versus precision in economics.* Edward Elgar Publishing Limited, 1983.

[39] W. Mendenhall and J.E. Reinmuth. *Estadística para administración y economía.* Wadsworth Internacional Iberoamericana, 1978.

[40] P. Meyer. *Probabilidad y aplicaciones estadísticas.* Fondo Educativo Interamericano, 1986.

[41] R.L. Mills. *Estadística para economía y administración.* McGraw-Hill, 1980.

[42] T.W. Mirer. *Economic statistics and econometrics.* Prentice-Hall, 1995.

[43] A.M. Mood and F.A. Graybill. *Introducción a la teoría de la estadística.* Aguilar, 1978.

[44] S. Murgui and R. Escuder. *Estadística aplicada. Inferencia estadística.* Tirant lo Blanch, 1994.

[45] P. Newbold and otros. *Estadística para administración y economía.* Prentice-Hall, 2008.

[46] R.L. Ott and W. Mendenhall. *Understanding statistics.* Duxbury Press, 1994.

[47] E. Parzen. *Teoría moderna de probabilidades y sus aplicaciones.* Limusa, 1987.

[48] J.A. Paulos. *El hombre anumerico.* Tisqiets, 1990.

[49] R. Pérez. *Nociones Básicas de Estadística.* Disponible desde Internet en: sites.google.com/a/uniovi.es/libros/nociones-basicas-estadistica, 2010.

[50] R. Pérez and A.J. López. *Análisis de datos económicos II. Métodos inferenciales.* Pirámide, Madrid, 1997.

[51] W.S. Peters. *Counting for Something.* Springer-Verlag, 1987.

[52] L. Prieto and I. Herranz. *Qué significa estadísticamente significativo?* Díaz de Santos, 2005.

[53] A. Pulido and J. Pérez. *Modelos Econométricos.* Pirámide, Madrid, 2001.

[54] R. Ramanathan. *Introductory Econometrics with Applications.* Harcourt College Publisher, 2002.

[55] D.G. Rees. *Foundations of Statistics.* Chapman and Hall, 1987.

[56] V.K. Rohatgi. *Statistical Inference.* Dover, 2003.

[57] L. Ruiz-Maya and F.J. Martín Pliego. *Estadística II: Inferencia.* Paraninfo, 2001.

[58] S. Siegel. *Estadística no paramétrica. Aplicada a las ciencias de la conducta.* Trillas, 1991.

[59] M.G. Sobolo and M.K. Starr. *Statistics for business and economics.* McGraw-Hill, 1983.

[60] A. Spooner and C. Lewis. *An Introduction to Statistics for Managers.* Prentice Hall, 1995.

[61] J. Tanur and otros. *La Estadística. Una guía de lo desconocido.* Alianza Editorial, 1992.

[62] A.F. Troconiz. *Probabilidades. Estadística. Muestreo.* Tebar Flores, 1987.

[63] E. Uriel and otros. *Econometría. El modelo lineal.* AC, 1990.

[64] R.E. Walpole and R.H. Myers. *Probabilidad y estadística.* McGraw-Hill, 1992.

[65] R.H. Wonnacott and T.H. Wonnacott. *Estadística básica práctica.* Limusa, 1991.

[66] R.H. Wonnacott and T.H. Wonnacott. *Fundamentos de estadística para Administración y Economía.* Limusa, 1993.

[67] J.M. Wooldridge. *Introducción a la econometría. Un enfoque moderno.* Paraninfo, 2008.

[68] M.V. Esteban y otros. *Econometría Básica Aplicada con Gretl.* Sarriko On, Universidad del País Vasco, 2008.

[69] T. Yamane. *Estadística.* Harla, 1979.

Index

View publication stats