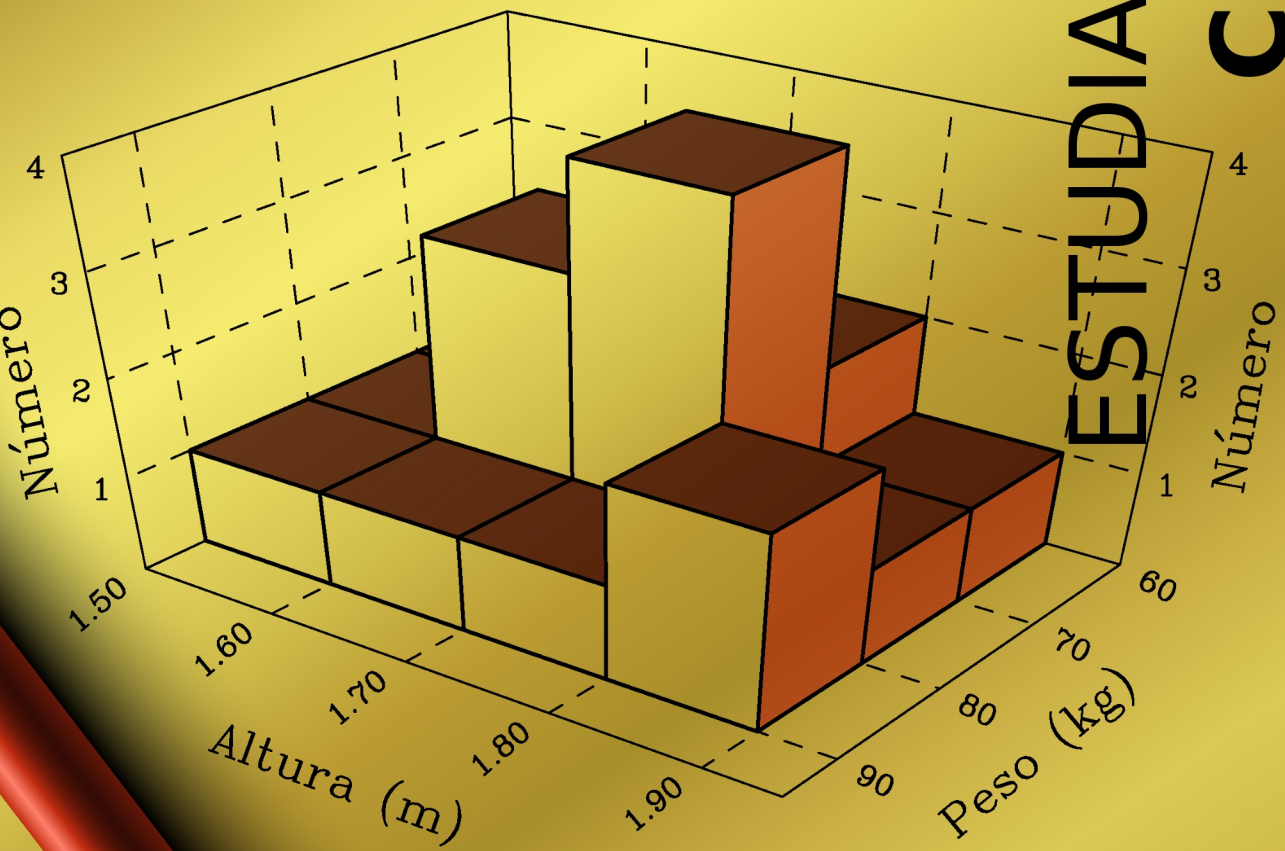
ESTADÍSTICA BÁSICA

PARA ESTUDIANTES DE **CIENCIAS**



Javier Gorgas García
Nicolás Cardiel López
Jaime Zamorano Calvo

ESTADÍSTICA BÁSICA
PARA ESTUDIANTES DE CIENCIAS

Javier Gorgas García
Nicolás Cardiel López
Jaime Zamorano Calvo

Departamento de Astrofísica y Ciencias de la Atmósfera
Facultad de Ciencias Físicas
Universidad Complutense de Madrid

ISBN 978-84-691-8981-8

Versión 17 de febrero de 2011

© Javier Gorgas, Nicolás Cardiel y Jaime Zamorano

"No confíes en lo que la estadística te dice hasta haber considerado con cuidado qué es lo que no dice."

William W. Watt

Índice general

Prefacio

Este libro recoge el material didáctico utilizado por los autores para la impartición de la asignatura *Estadística* en la Facultad de CC. Físicas de la Universidad Complutense de Madrid. Esta asignatura se introdujo en el Plan de Estudios del año 1995 y desde entonces ha demostrado aportar un conocimiento esencial para la formación de los estudiantes de la Licenciatura en Física. Estamos convencidos de que este tipo de conocimiento es básico para cualquier estudiante de ciencias.

Aunque la bibliografía en este campo es extensa, hemos considerado oportuno redactar un libro restringido a los contenidos específicos que se incluyen en un curso introductorio de Estadística. Pretendemos así delimitar, y en lo posible simplificar, el trabajo del estudiante, mostrándole de forma precisa los conceptos más fundamentales. Una vez consolidados estos conceptos, esperamos que los estudiantes de ciencias encuentren menos dificultades para aprender y profundizar en las técnicas estadísticas más avanzadas que son de uso común en el trabajo diario de un científico.

Queremos agradecer a los diferentes profesores que durante estos años han dedicado su esfuerzo a enseñar Estadística en la Facultad de CC. Físicas. El temario que finalmente se plasma en este libro ha evolucionado y se ha enriquecido de las conversaciones mantenidas con ellos: Natalia Calvo Fernández, Andrés Javier Cenarro Lagunas, Manuel Cornide Castro-Piñeiro, Jesús Fidel González Rouco, Ricardo García Herrera, Gregorio Maqueda Burgos, Mª Luisa Montoya Redondo, Mª Belén Rodríguez de Fonseca, Encarnación Serrano Mendoza y, de forma muy especial y con todo el afecto, nuestro agradecimiento a Elvira Zurita García. Una excelente profesora y mejor persona, para quien la calidad de la enseñanza fue siempre una prioridad constante. Siempre la recordaremos con cariño.

Los autores

Madrid, febrero de 2009

Capítulo 1

Introducción

"La Ciencia es más una forma de pensar que una rama del conocimiento."

Carl Sagan (1934-1996)

1.1. La Estadística como ciencia

La Estadística es la ciencia que se encarga de recoger, organizar e interpretar los datos. Es la ciencia de los datos. En la vida diaria somos bombardeados continuamente por datos estadísticos: encuestas electorales, economía, deportes, datos meteorológicos, calidad de los productos, audiencias de TV. Necesitamos una formación básica en Estadística para evaluar toda esta información. Pero la utilidad de la Estadística va mucho más allá de estos ejemplos.

La Estadística es fundamental para muchas ramas de la ciencia desde la medicina a la economía. Pero sobre todo, y en lo que a nosotros importa, es esencial para interpretar los datos que se obtienen de la investigación científica. Es necesario leer e interpretar datos, producirlos, extraer conclusiones, en resumen saber el significado de los datos. Es por lo tanto una **herramienta de trabajo profesional**.

Se recomienda leer la Introducción de *Estadística: modelos y métodos* de Daniel Peña, para conocer el desarrollo histórico de la Estadística. La Estadística (del latín, *Status* o ciencia del estado) se ocupaba sobre todo de la descripción de los datos fundamentalmente sociológicos: datos demográficos y económicos (censos de población, producciones agrícolas, riquezas, etc.), principalmente por razones fiscales. En el siglo XVII el cálculo de probabilidades se consolida como disciplina independiente aplicándose sobre todo a los juegos de azar. Posteriormente (s. XVIII) su uso se extiende a problemas físicos (principalmente de Astronomía) y actuariales (seguros marítimos). Posteriormente se hace imprescindible en la investigación científica y es ésta la que la hace avanzar. Finalmente, en el siglo XIX, nace la Estadística como ciencia que une ambas disciplinas.

El objetivo fundamental de la estadística es obtener conclusiones de la investigación empírica usando modelos matemáticos. A partir de los datos reales se construye un modelo que se confronta con estos datos por medio de la Estadística. Esta proporciona los métodos de evaluación de las discrepancias entre ambos. Por eso es necesaria para toda ciencia que requiere análisis de datos y diseño de experimentos.

1.2. Para qué sirve la Estadística

Ya hemos visto que la Estadística se encuentra ligada a nuestras actividades cotidianas. Sirve tanto para pronosticar el resultado de unas elecciones, como para determinar el número de ballenas que viven en nuestros océanos, para descubrir leyes fundamentales de la Física o para estudiar cómo ganar a la ruleta.

La Estadística resuelve multitud de problemas que se plantean en ciencia:

- *Análisis de muestras*. Se elige una muestra de una población para hacer inferencias respecto a esa población a partir de lo observado en la muestra (sondeos de opinión, control de calidad, etc).
- *Descripción de datos*. Procedimientos para resumir la información contenida en un conjunto (amplio) de datos.
- *Contraste de hipótesis*. Metodología estadística para diseñar experimentos que garanticen que las conclusiones que se extraigan sean válidas. Sirve para comparar las predicciones resultantes de las hipótesis con los datos observados (medicina eficaz, diferencias entre poblaciones, etc).
- *Medición de relaciones* entre variables estadísticas (contenido de gas hidrógeno neutro en galaxias y la tasa de formación de estrellas, etc)
- *Predicción*. Prever la evolución de una variable estudiando su historia y/o relación con otras variables.

1.3. El método científico

Citando a Martin Gardner: "La ciencia es una búsqueda de conocimientos fidedignos acerca del mundo: cómo se estructura y cómo funciona el universo (incluyendo los seres vivos)". La informacion que maneja la ciencia es amplia, al ser amplio su ámbito. Pero se suele reunir en tres apartados: los hechos, las leyes y las teorías. No es una partición estanca, aunque podemos entender aquí nos referimos con algún ejemplo. Los hechos se refiere a casos específicos y/o únicos. Por ejemplo la Tierra tiene una luna (satélite natural).

La primera ley de Kepler (ya que estamos con planetas) es un buen ejemplo de ley: los planetas describen orbitas elípticas en torno al Sol, que ocupa uno de los focos de la elipse. Como se ve, frente al hecho, concreto y único, la ley se refiere a muchos casos, como lo son los planetas que orbitan en torno al Sol. La generalización de la ley de Kepler permite aplicarla a cualquier par de cuerpos ligados por la gravedad.

Una teoría es una abstracción, con entidades inobservables, que explica hechos y leyes. Por ejemplo la teoría newtoniana de la gravitación. En ella se habla de fuerzas (o de campos gravitatorios) que no son entes observables, pero esta teoría explica hechos y leyes.

Sucede que el conocimiento científico no es completamente seguro en ninguna de las precedentes categorías. Podría existir otra luna en torno a la Tierra. O, como sabemos, la teoría newtoniana de la gravitación no es completa, porque no da cuenta de algunos fenómenos. De ahí vino su evolución a nuevas teorías de la gravitación. No hay así un conocimiento completamente seguro: los enunciados absolutamente ciertos sólo existen en el ámbito de las matemáticas o la lógica. Pero la ciencia usa una *correspondencia* con estas dos disciplinas. La matemática y la lógica aplicadas a las ciencias facilitan poder establecer hechos, leyes y teorías con coherencia interna y con un alto grado de certeza. Y la Estadística proporciona una herramienta para poder evaluar esta certeza, o proporcionar pautas para realizar inferencias a partir de lo que se conoce.

Lo que distingue a una teoría científica es que ésta, a diferencia de la que no lo es, puede ser refutada: puede existir un conjunto de circunstancias que si son observadas demuestran que la teoría está equivocada. A continuación se ofrece una visión simplificada del método científico.

Hacemos observaciones en la naturaleza y a través de un proceso creativo generamos una hipótesis de cómo funciona cierto aspecto de la naturaleza (modelos). Basándonos en esa hipótesis diseñamos un experimento que consiste en que un conjunto de observaciones deben tener lugar, bajo ciertas condiciones, si la hipótesis es cierta. En el caso de que estas observaciones no ocurran nos enfrentamos a varias posibilidades: nuestras hipótesis necesitan ser revisadas, el experimento se llevó a cabo de forma incorrecta, o nos hemos equivocado en el análisis de los resultados del experimento.

Hace algunos cientos de años se estableció un método para encontrar respuestas a los interrogantes que nos planteamos al contemplar la naturaleza. Este método, conocido como **método científico**, se basa en tres pilares fundamentales: *observación*, *razonamiento* y *experimentación*.
El método científico no es una simple receta, sino que es un proceso exigente que requiere, entre otros ingredientes, juicio crítico. De forma resumida, el método científico incorpora las siguientes facetas:

Observación: aplicación atenta de los sentidos a un objeto o a un fenómeno, para estudiarlos tal como se presentan en realidad.

Descripción: las mediciones deben ser fiables, es decir, deben poder repetirse. Las observaciones únicas e irrepetibles no permiten predecir futuros resultados. En este sentido la Cosmología se enfrenta, a priori, a un grave problema. El Universo es único y no podemos volver a repetirlo modificando las condiciones iniciales.

Predicción: las predicciones de cualquier fenómeno deben ser válidas tanto para observaciones pasadas, como presentes y futuras.

Control: capacidad de modificar las condiciones del experimento para estudiar el impacto de los diferentes parámetros participantes. Esto se opone a la aceptación pasiva de datos, que puede conducir a un importante *sesgo* (bias) empírico.

Falsabilidad o eliminación de alternativas plausibles: Este es un proceso gradual que requiere la repetición de los experimentos (preferiblemente por investigadores independientes, quienes deben ser capaces de replicar los resultados iniciales con la intención de corroborarlos). Todas las hipótesis y teorías deben estar sujetas a la posibilidad de ser refutadas. En este sentido, a medida que un área de conocimiento crece y las hipótesis o teorías sobre la que se sustenta van realizando predicciones comprobables, aumenta la confianza en dichas hipótesis o teorías (uno de los defensores fundamentales del criterio de falsabilidad es Karl Popper (1902–1994); ver, por ejemplo, *La lógica de la investigación científica* en Popper 1935).

Explicación causal: los siguientes requisitos son normalmente exigibles para admitir una explicación como científica:

- Identificación de las causas.
- Las causas identificadas deben correlacionarse con los observables.
- Las causas deben preceder temporalmente a los efectos medidos.

1.4. El proceso experimental

La experimentación está lejos de estar carente de dificultades. Algunas técnicas experimentales exigen un aprendizaje largo y, en muchas ocasiones, el volumen de datos a manejar puede ser tan grande que sea necesario un trabajo de análisis intenso. La paciencia y la perseverancia son grandes aliadas en este sentido.

Las razones para realizar un experimento son diversas y de alcance muy variable. Preguntas típicas son, por ejemplo: ¿Cómo de aplicable es una teoría particular? ¿Es posible mejorar una técnica de medida? ¿A qué temperatura debe fundir una nueva aleación? ¿Qué ocurre con las propiedades magnéticas de un material al someterlo a temperaturas de trabajo muy bajas? ¿Se ven alteradas las propiedades de un semiconductor debido al bombardeo por radiación nuclear?

De una forma esquemática, el proceso experimental suele desarrollarse siguiendo el siguiente esquema:

1. Definir la pregunta o problema a resolver. Cuanto más claro y definido sea el objetivo del experimento, mucho más fácil será realizar su planificación y ejecución.

2. Obtener información y recursos. Una vez definido el objetivo del experimento, es necesario elaborar un plan de trabajo para poder alcanzarlo. Hay que identificar qué equipos son necesarios, qué cantidades hay que medir, y de qué manera se va a realizar el experimento.

3. Formular hipótesis, acerca de los resultados de nuestro experimento. Hacerlo antes de su ejecución evita el *sesgo* personal de identificar los resultados que ya se conocen como objetivos iniciales (no debemos engañarnos a nosotros mismos).

4. Realizar el experimento y obtener las medidas. Esta tarea se subdivide en varios pasos:

   - Preparación: el equipo debe ser puesto a punto para su utilización. Si el experimento requiere la utilización de aparatos con los que no estamos familiarizados, es necesario leer atentamente los manuales de utilización, e incluso consultar a experimentadores con experiencia previa en su manejo. Todo ello evita perder tiempo y cometer errores *de bulto*, a la vez que preserva la integridad del equipo (¡y la nuestra!).
   - Experimentación preliminar: suele ser muy aconsejable realizar una pequeña experimentación de prueba antes de iniciar la toma definitiva de medidas. Esto facilita el uso correcto del equipo instrumental, permitiendo identificar los aspectos más difíciles o en los que resulta más fácil cometer errores.
   - Toma de datos: el trabajo cuidadoso y detallado son fundamentales en todo proceso experimental. Ejecutar dicha labor siguiendo un plan de trabajo bien definido resulta básico. No hay nada más frustrante que descubir, tras largas horas de medidas, que hemos olvidado anotar algún parámetro esencial o sus unidades. En este sentido resulta imprescindible tener presentes varias cuestiones
     - ¿Cuáles son las unidades asociadas a cada medida?
     - ¿Cuál es la incertidumbre asociada?
     - ¿Qué variabilidad presentan las medidas?
     - ¿Cómo puedo tener una idea del orden de magnitud de una medida antes de realizarla y saber así que los resultados que se van obteniendo son razonables?
     - ¿Qué información debe ser incluida en la tabla de datos?
   - Comprobación de la repitibilidad: siempre que sea posible, todo experimento debería repetirse varias veces para comprobar que los resultados obtenidos son repetibles y representativos. Y aunque, obviamente, la repetición de un experimento no proporciona exactamente los *mismos* números, discrepancias muy grandes deben alertarnos acerca de la existencia de efectos sistemáticos que pueden estar distorsionando el experimento.

5. Analizar los datos: una vez obtenidas las medidas es necesario su tratamiento estadístico para poder obtener magnitudes (e incertidumbres asociadas) representativas del objeto de nuestro estudio.

6. Interpretar los datos y extraer conclusiones que sirvan como punto de partida para nuevas hipótesis. El éxito de esta interpretación dependerá, básicamente, de la calidad de las medidas y de su análisis. **Las herramientas estadísticas que se describen en este libro nos permitirán tomar decisiones de manera objetiva.**

7. Publicar los resultados. Los resultados de cualquier proceso experimental deben ser comunicados de manera clara y concisa. Esto incluye desde un sencillo *informe de laboratorio*, como el que se exigirá en los diversos laboratorios en los que se trabajará durante la licenciatura de Físicas, hasta la publicación de un *artículo científico* en una revista reconocida.

No es extraño que, aunque la pregunta inicial a responder haya sido establecida de una forma clara, tras el desarrollo del experimento y el análisis de los resultados, se descubran fenómenos no previstos que obliguen a modificar y repetir el proceso descrito. De hecho, si el resultado de un experimento fuera completamente predecible, tendría poco sentido llevarlo a cabo. Por ello, de forma práctica el esquema anterior se ejecuta siguiendo un proceso iterativo entre los puntos 3 y 6. Una vez obtenido un conocimiento *significativo*, éste ha de ser explicado en una publicación, permitiendo a nuevos investigadores corroborar o refutar las conclusiones.

1.5. Bibliografía complementaria

La consulta de libros es necesaria para conocer diferentes enfoques y, desde luego, se hace imprescindible para ampliar la colección de ejemplos y ejercicios, ya que la Estadística es una disciplina eminentemente práctica. A continuación se enumeran algunos de los textos en castellano más frecuentes en las bibliotecas de las Facultades de Ciencias, con una pequeña descripción en relación a los contenidos cubiertos por este libro:

- *Curso y ejercicios de estadística*, Quesada, Isidoro & Lopez, Alhambra 1988.
  Cubre casi todos los temas. Buen formalismo matemático. Amplia colección de problemas.
- *Probabilidad y Estadística*, Walpole & Myers, McGraw-Hill 1992.
  Muy bien explicado. Con multitud de ejemplos. Es más amplio. De carácter práctico. Válido para todos los temas excepto el primero.
- *Probabilidad y Estadística*, Spiegel, McGraw-Hill 1991.
  Con muchos problemas. La teoría se encuentra muy resumida. Vale para todos los temas excepto el primero. Este tema se desarrola en otro libro de Spiegel: *Estadística (Teoría y Problemas)*.
- *Métodos Estadísticos*, Viedma, Ediciones del Castillo 1990.
  Muy sencillo. Cubre todos los temas, aunque algunos no de forma completa.

Tema I

ESTADÍSTICA DESCRIPTIVA

Capítulo 2

Fundamentos de Estadística Descriptiva

"Se cometen muchos menos errores usando datos inadecuados que cuando no se utilizan datos."

Charles Babbage (1792-1871)

La aplicación del tratamiento estadístico tiene dos fases fundamentales:

1. Organización y análisis inicial de los datos recogidos.
2. Extracción de conclusiones válidas y toma de decisiones razonables a partir de ellos.

Los objetivos de la Estadística Descriptiva son los que se abordan en la primera de estas fases. Es decir, su misión es ordenar, describir y sintetizar la información recogida. En este proceso será necesario establecer medidas cuantitativas que reduzcan a un número manejable de parámetros el conjunto (en general grande) de datos obtenidos.

La realización de gráficas (visualización de los datos en diagramas) también forma parte de la Estadística Descriptiva dado que proporciona una manera visual directa de organizar la información.

La finalidad de la Estadística Descriptiva no es, entonces, extraer conclusiones generales sobre el fenómeno que ha producido los datos bajo estudio, sino solamente su descripción (de ahí el nombre).

2.1. Variables estadísticas

El concepto de variable estadística es, sin duda, uno de los más importantes en Estadística. Pero antes de abordar su definición, es necesario introducir anteriormente diversos conceptos básicos.

2.1.1. Población y muestra

Se denomina **población** al conjunto completo de elementos, con alguna característica común, que es el objeto de nuestro estudio. Esta definición incluye, por ejemplo, a todos los sucesos en que podría concretarse un fenómeno o experimento cualesquiera. Una población puede ser *finita* o *infinita*.

Ejemplo I–1

Los habitantes de un país, los planetas del Sistema Solar, las estrellas en la Vía Láctea, son elementos de una población finita. Sin embargo, el número de posibles medidas que se puedan hacer de la velocidad de la luz, o de tiradas de un dado, forman poblaciones infinitas.

Cuando, aunque la población sea finita, su número de elementos es elevado, es necesario trabajar con solo una parte de dicha población. A un subconjunto de elementos de la población se le conoce como **muestra**.

Ejemplo I–2

Si se quiere estudiar las propiedades de las estrellas en nuestra Galaxia, no tendremos la oportunidad de observarlas todas; tendremos que conformarnos con una muestra representativa. Obviamente, elegir de forma *representativa* los elementos de una muestra es algo muy importante. De hecho existe un grave problema, conocido como efecto de selección, que puede condicionar el resultado de un estudio si uno no realiza una selección correcta de los elementos que forman parte de una muestra.

Al número de elementos de la muestra se le llama **tamaño** de la muestra. Es fácil adelantar que para que los resultados de nuestro estudio estadístico sean fiables es necesario que la muestra tenga un tamaño mínimo. El caso particular de una muestra que incluye a todos los elementos de la población es conocido como **censo**.

2.1.2. Caracteres cuantitativos o cualitativos

El objeto de nuestra medida pueden ser caracteres de tipos muy diversos. De ahí que normalmente se clasifiquen en:

- caracteres **cuantitativos**: aquellos que toman valores numéricos. Por ejemplo la altura o la velocidad de un móvil.
- caracteres **cualitativos**: también llamados atributos, son aquellos que no podemos representar numéricamente y describen cualidades. Por ejemplo, un color o el estado civil.

Aunque existen algunas diferencias, el tratamiento para ambos casos es similar, pudiéndose asignar, en muchas ocasiones, valores numéricos a los diferentes caracteres cualitativos.

2.1.3. Variable estadística

Se entiende por **variable estadística** al símbolo que representa al dato o carácter objeto de nuestro estudio de los elementos de la muestra y que puede tomar un conjunto de valores. En el caso de que estemos tratando con caracteres cuantitativos, la variables estadísticas pueden clasificarse en: **discretas**, cuando solo pueden tomar una cantidad (finita o infinita) numerable de valores, y **continuas**, cuando pueden tomar teóricamente infinitos valores entre dos valores dados. Es la diferencia básica que existe entre *contar* y *medir*.

Ejemplo I–3

El número de electrones de un átomo es una variable discreta. La velocidad o la altura de un móvil son variables continuas.

Por otra parte, las variables se pueden asimismo clasificar en **unidimensionales**, cuando solo se mida un carácter o dato de los elementos de la muestra, o **bidimensionales**, tridimensionales, y en general **n–dimensionales**, cuando se estudien simultáneamente varios caracteres de cada elemento.

Ejemplo I–4

La temperatura o la presión atmosférica (por separado), son variables monodimensionales. La temperatura y la presión atmosférica (estudiadas conjuntamente), o la longitud y el peso de una barra conductora, son ejemplos de variables bidimensionales. La velocidad, carga eléctrica y masa de un ión es tridimensional.

2.2. Distribuciones de frecuencias

El primer paso para el estudio estadístico de una muestra es su ordenación y presentación en una tabla de frecuencias.

2.2.1. Tabla de frecuencias de una variable discreta

Supongamos que tenemos una muestra de tamaño N, donde la variable estadística x toma los valores distintos $x_1, x_2, \ldots, x_k$. En primer lugar hay que ordenar los diferentes valores que toma la variable estadística en orden (normalmente creciente). La diferencia entre el valor mayor y menor que toma la variable se conoce como **recorrido**, o rango.

En el caso de variables discretas, generalmente, un mismo valor de la variable aparecerá repetido más de una vez (es decir $k < N$). De forma que el siguiente paso es la construcción de una tabla en la que se indiquen los valores posibles de la variable y su frecuencia de aparición. Esta es la **tabla de frecuencias de una variable discreta**:

Valores de la variable estadística x_i	Frecuencias absolutas n_i	Frecuencias relativas f_i	Frecuencias absolutas acumuladas N_i	Frecuencias relativas acumuladas F_i
x_1	n_1	f_1	N_1	F_1
x_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_k	f_k	N_k	F_k

En la primera columna de esta tabla se escriben los distintos valores de la variable, x_i, ordenados de mayor a menor. Es posible hacer también una tabla de frecuencias de una variable cualitativa. En ese caso, en la primera columna se escribirán las diferentes cualidades o atributos que puede tomar la variable. En las siguientes columnas se escriben para cada valor de la variable:

- **Frecuencia absoluta** n_i: Definida como el número de veces que aparece repetido el valor en cuestión de la variable estadística en el conjunto de las observaciones realizadas. Si N es el número de observaciones (o tamaño de la muestra), las frecuencias absolutas cumplen las propiedades

$$0 \leq n_i \leq N \qquad ; \qquad \sum_{i=1}^{k} n_i = N.$$

  La frecuencia absoluta, aunque nos dice el número de veces que se repite un dato, no nos informa de la importancia de éste. Para ello se realiza la siguiente definición.

- **Frecuencia relativa** f_i: Cociente entre la frecuencia absoluta y el número de observaciones realizadas N. Es decir

$$f_i = \frac{n_i}{N}, \tag{2.1}$$

  cumpliéndose las propiedades

$$0 \leq f_i \leq 1 \qquad ; \qquad \sum_{i=1}^{k} f_i = \sum_{i=1}^{k} \frac{n_i}{N} = \frac{\sum_{i=1}^{k} n_i}{N} = 1.$$

Esta frecuencia relativa se puede expresar también en tantos por cientos del tamaño de la muestra, para lo cual basta con multiplicar por 100

$$(\,\%)_{x_i} = 100\mathrm{x}f_i.$$

Por ejemplo, si $f_i = 0.25$, esto quiere decir que la variable x_i se repite en el 25 % de la muestra.

- **Frecuencia absoluta acumulada** N_i: Suma de las frecuencias absolutas de los valores inferiores o igual a x_i, o número de medidas por debajo, o igual, que x_i. Evidentemente la frecuencia absoluta acumulada de un valor se puede calcular a partir de la correspondiente al anterior como

$$N_i = N_{i-1} + n_i \qquad \text{y} \qquad N_1 = n_1. \tag{2.2}$$

Además la frecuencia absoluta acumulada del último valor será

$$N_k = N.$$

- **Frecuencia relativa acumulada** F_i: Cociente entre la frecuencia absoluta acumulada y el número de observaciones. Coincide además con la suma de las frecuencias relativas de los valores inferiores o iguales a x_i

$$F_i = \frac{N_i}{N} = \frac{\sum_{j=1}^{i} n_j}{N} = \sum_{j=1}^{i} \frac{n_j}{N} = \sum_{j=1}^{i} f_i, \tag{2.3}$$

y la frecuencia relativa acumulada del último valor es 1

$$F_k = 1.$$

Se puede expresar asimismo como un porcentaje (multiplicando por 100) y su significado será el tanto por ciento de medidas con valores por debajo o igual que x_i.

Ejemplo I–5

Supongamos que el número de hijos de una muestra de 20 familias es el siguiente:

2	1	1	3	1	2	5	1	2	3
4	2	3	2	1	4	2	3	2	1

El tamaño de la muestra es $N = 20$, el número de valores posibles $k = 5$, y el recorrido es $5 - 1 = 4$.

x_i	n_i	f_i $n_i/20$	N_i $\sum_1^i n_j$	F_i $\sum_1^i f_j$
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

2.2.2. Agrupamiento en intervalos de clase

Cuando el número de valores distintos que toma la variable estadística es demasiado grande o la variable es continua no es útil elaborar una tabla de frecuencias como la vista anteriormente. En estos casos se realiza un **agrupamiento de los datos en intervalos** y se hace un recuento del número de observaciones que caen dentro de cada uno de ellos. Dichos intervalos se denominan **intervalos de clase**, y al valor de

la variable en el centro de cada intervalo se le llama **marca de clase**. De esta forma se sustituye cada medida por la marca de clase del intervalo a que corresponda. A la diferencia entre el extremo superior e inferior de cada intervalo se le llama **amplitud del intervalo**. Normalmente se trabajará con intervalos de amplitud constante. La tabla de frecuencias resultante es similar a la vista anteriormente. En el caso de una distribución en k intervalos ésta sería:

Intervalos de clase	Marcas de clase	Frecuencias absolutas	Frecuencias relativas	Frecuencias absolutas acumuladas	Frecuencias relativas acumuladas
$a_i - a_{i+1}$	c_i	n_i	$f_i = n_i/N$	N_i	$F_i = N_i/N$
$a_1 - a_2$	c_1	n_1	f_1	N_1	F_1
$a_2 - a_3$	c_2	n_2	f_2	N_2	F_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$a_k - a_{k+1}$	c_k	n_k	f_k	N_k	F_k

El realizar el estudio mediante el agrupamiento en intervalos de clase simplifica el trabajo, pero también supone una pérdida de información, ya que no se tiene en cuenta cómo se distribuyen los datos dentro de cada intervalo. Para que dicha pérdida sea mínima es necesario elegir con cuidado los intervalos. Aunque no existen ningunas reglas estrictas para la elección de los intervalos, los pasos a seguir son:

1. Determinar el recorrido, o rango, de los datos. Esto es, la diferencia entre el mayor y el menor de los valores que toma la variable.

2. Decidir el número k de intervalos de clase en que se van a agrupar los datos. Dicho número se debe situar normalmente entre 5 y 20, dependiendo del caso. En general el número será más grande cuanto más datos tenga la muestra. Una regla que a veces se sigue es elegir k como el entero más próximo a $\sqrt{N}$, donde N es el número total de medidas.

3. Dividir el recorrido entre el número de intervalos para determinar la amplitud (constante) de cada intervalo. Dicha amplitud no es necesario que sea exactamente el resultado de esa división, sino que normalmente se puede redondear hacia un número algo mayor.

4. Determinar los extremos de los intervalos de clase. Evidentemente el extremo superior de cada intervalo ha de coincidir con el extremo inferior del siguiente. Es importante que ninguna observación coincida con alguno de los extremos, para evitar así una ambiguedad en la clasificación de este dato. Una forma de conseguir esto es asignar a los extremos de los intervalos una cifra decimal más que las medidas de la muestra. Por ejemplo, si la variable estadística toma valores enteros: $10, 11, 12, \ldots$, los intervalos se podrían elegir: $9.5 - 11.5, 11.5 - 13.5, \ldots$.

5. Calcular las marcas de clase de cada intervalo como el valor medio entre los límites inferior y superior de cada intervalo de clase. Otra consideración a tomar en cuenta a la hora de elegir los intervalos es intentar que las marcas de clase coincidan con medidas de la muestra, disminuyéndose así la pérdida de información debida al agrupamiento.

Una vez determinados los intervalos se debe hacer un recuento cuidadoso del número de observaciones que caen dentro de cada intervalo, para construir así la tabla de frecuencias.

Ejemplo I–6

En la tabla siguiente se listan los datos medidos por James Short en 1763 sobre la paralaje del Sol en segundos de arco. La paralaje es el ángulo subtendido por la Tierra vista desde el Sol. Se midió observando tránsitos de Venus desde diferentes posiciones y permitió la primera medida de la distancia Tierra-Sol, que es la unidad básica de la escala de distancias en el Sistema Solar (la unidad astronómica).

Datos (en segundos de arco):

8.63	10.16	8.50	8.31	10.80	7.50	8.12
8.42	9.20	8.16	8.36	9.77	7.52	7.96
7.83	8.62	7.54	8.28	9.32	7.96	7.47

1. Recorrido: máximo−mínimo= $10.80 - 7.47 = 3.33$.
2. Número de intervalos: $k = \sqrt{21} = 4.53 \Rightarrow k = 5$. Como se redondea por exceso, la amplitud del intervalo multiplicada por el número de intervalos será mayor que el recorrido y no tendremos problemas en los extremos.
3. Amplitud del intervalo: $3.33/5 = 0.666 \Rightarrow 0.7$.
4. Extremos de los intervalos. Para evitar coincidencias se toma un decimal más. El primer extremo se toma algo menor que el valor mínimo, pero calculándolo de forma que el último extremo sea algo mayor que el valor máximo.
Si tomamos $a_1 = 7.405$ se verifica que es < 7.47 (mínimo), y el último extremo será $7.405 + 5 \times 0.7 = 10.905$ que resulta ser > 10.80 (máximo). Ahora ya podemos calcular los extremos para cada intervalo de clase y las marcas de clase correspondientes.
5. Recuento y construcción de la tabla.

a_i—a_{i+1}	c_i	n_i	f_i	N_i	F_i
7.405 — 8.105	7.755	7	0.333	7	0.333
8.105 — 8.805	8.455	9	0.429	16	0.762
8.805 — 9.505	9.155	2	0.095	18	0.857
9.505 — 10.205	9.855	2	0.095	20	0.952
10.205— 10.905	10.555	1	0.048	21	1.000
Suma		21	1.000		

2.3. Representaciones gráficas

Después de construir la tabla de frecuencias correspondiente es conveniente la representación gráfica de la distribución de los datos en un diagrama. Estas representaciones gráficas permiten una visualización rápida de la información recogida. Veamos los diferentes tipos de diagramas.

2.3.1. Representaciones gráficas para datos sin agrupar

El diagrama principal para representar datos de variables discretas sin agrupar es el **diagrama de barras**. En éste se representan en el eje de abscisas los distintos valores de la variable y sobre cada uno de ellos se levanta una barra de longitud igual a la frecuencia correspondiente. Pueden representarse tanto las frecuencias absolutas n_i como las relativas f_i. En la práctica se puede graduar simultáneamente el eje de ordenadas tanto en frecuencias absolutas como en relativas en tantos por ciento.

Un diagrama similar es el **polígono de frecuencias**. Este se obtiene uniendo con rectas los extremos superiores de las barras del diagrama anterior. De la misma forma, pueden representarse frecuencias absolutas,

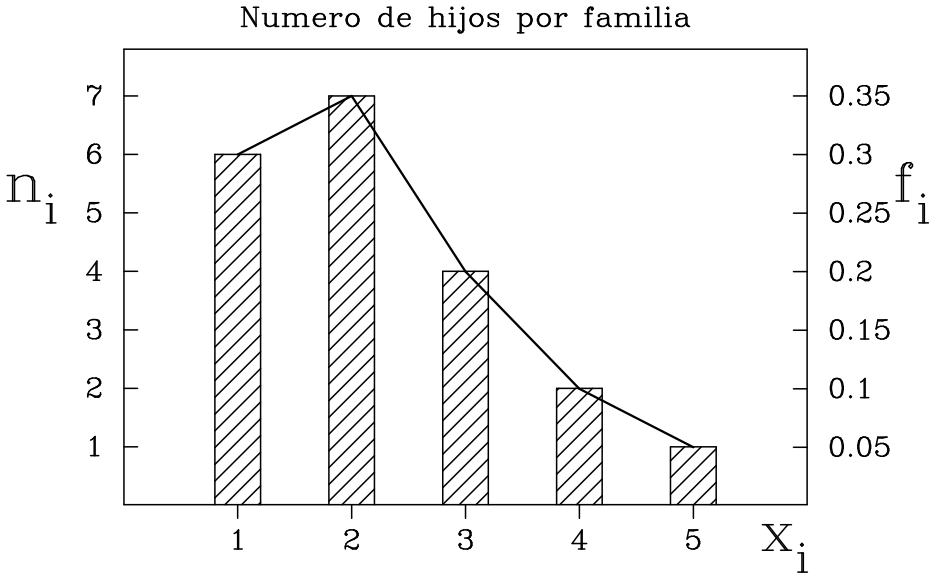


Figura 2.1: *Diagrama de barras y polígono de frecuencias. Se han usado los datos del ejemplo I–5.*

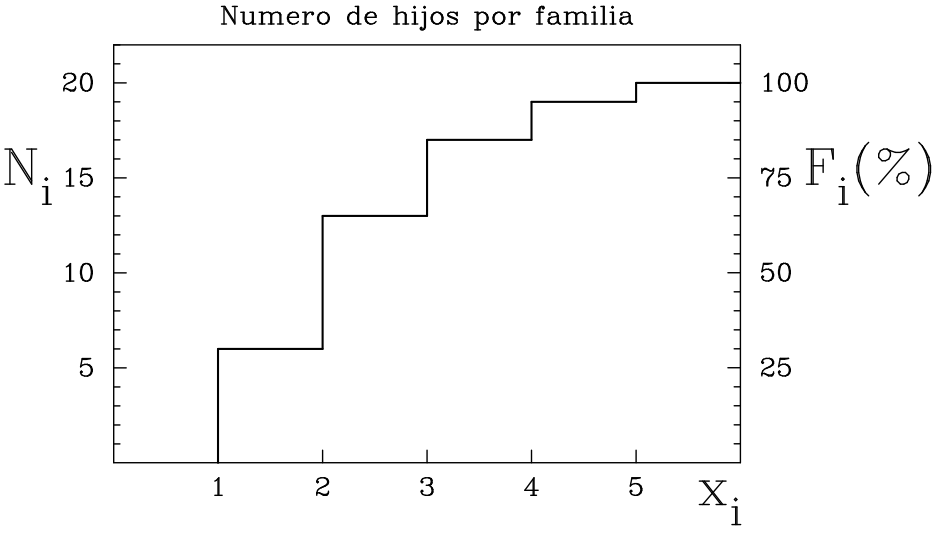


Figura 2.2: *Diagrama de frecuencias acumuladas. Se han usado los datos del ejemplo I–5.*

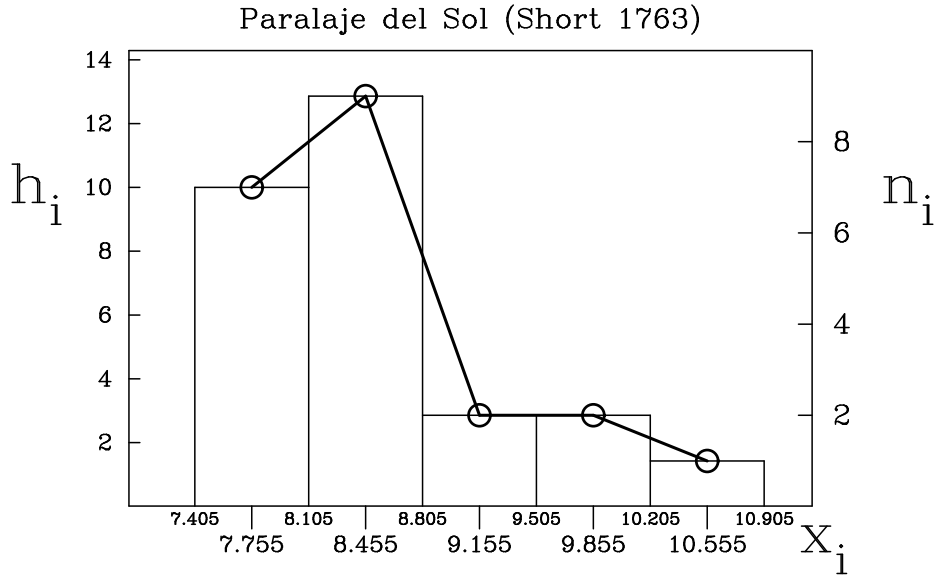


Figura 2.3: *Histograma y polígono de frecuencias de las medidas de la paralaje del Sol del ejemplo I–6. Las alturas de los rectángulos se obtienen como $h_i = n_i/\Delta$, siendo en este caso la amplitud del intervalo $\Delta = 0.7$. Nótese que el histograma tiene la misma forma si las alturas se hacen proporcionales a las frecuencias.*

relativas, o ambas a la vez. Ver Figura 2.1.

Para representar las frecuencias, tanto absolutas como relativas, acumuladas se usa el **diagrama de frecuencias acumuladas**. Este gráfico, en forma de escalera (ver Figura 2.2), se construye representando en abscisas los distintos valores de la variable y levantando sobre cada x_i una perpendicular cuya longitud será la frecuencia acumulada (N_i o F_i) de ese valor. Los puntos se unen con tramos horizontales y verticales como se muestra en la figura. Evidentemente la escalera resultante ha de ser siempre ascendente.

2.3.2. Representaciones gráficas para datos agrupados

La representación gráfica más usada para datos agrupados es el **histograma** de frecuencias absolutas o relativas (ver Figura 2.3). Un histograma es un conjunto de rectángulos adyacentes, cada uno de los cuales representa un intervalo de clase. Las base de cada rectángulo es proporcional a la amplitud del intervalo. Es decir, el centro de la base de cada rectángulo ha de corresponder a una marca de clase. La altura se suele determinar para que el área de cada rectángulo sea igual a la frecuencia de la marca de clase correspondiente. Por tanto, la altura de cada rectángulo se puede calcular como el cociente entre la frecuencia (absoluta o relativa) y la amplitud del intervalo. En el caso de que la amplitud de los intervalos sea constante, la representación es equivalente a usar como altura la frecuencia de cada marca de clase, siendo este método más sencillo para dibujar rápidamente un histograma.

Al igual que en las variables no agrupadas, otro tipo de representación es el **polígono de frecuencias**. Este se obtiene uniendo por líneas rectas los puntos medios de cada segmento superior de los rectángulos en el histograma. Ver Figura 2.4.

El **polígono de frecuencias acumuladas** sirve para representar las frecuencias acumuladas de datos agrupados por intervalos. En abscisas se representan los diferentes intervalos de clase. Sobre el extremo superior de cada intervalo se levanta una línea vertical de altura la frecuencia (absoluta o relativa) acumulada de ese intervalo. A continuación se unen por segmentos rectos los extremos de las líneas anteriores. El polígono parte de una altura cero para el extremo inferior del primer intervalo. Evidentemente, la altura que se alcanza al final del polígono es N, para frecuencias absolutas, o 1, para frecuencias relativas.

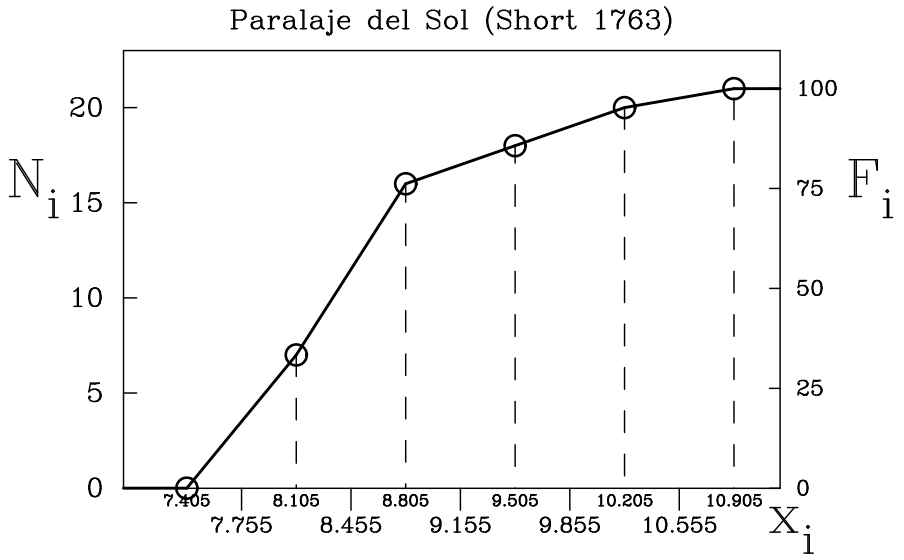


Figura 2.4: *Polígono de frecuencias acumuladas de las medidas de la paralaje del Sol del ejemplo I–6. Las zonas de mayor pendiente en este diagrama corresponden a las zonas más altas en el histograma (ver figura anterior).*

Mediante la interpolación en el polígono de frecuencias acumuladas (o leyendo sobre la escala de ordenadas) puede determinarse el número de observaciones mayores o menores que un valor dado, o incluso el número de datos comprendidos entre dos valores (restando las frecuencias acumuladas correspondientes), incluso aunque esos valores no sean marcas de clase.

2.3.3. Representaciones gráficas para variables cualitativas

Existe una gran variedad de representaciones para variables cualitativas, de las cuales vamos a describir únicamente las dos más usadas. El **diagrama de rectángulos** es similar al diagrama de barras y el histograma para las variables cuantitativas. Consiste en representar en el eje de abscisas los diferentes caracteres cualitativos y levantar sobre cada uno de ellos un rectángulo (de forma no solapada) cuya altura sea la frecuencia (absoluta o relativa) de dicho carácter.

Un diagrama muy usado es el **diagrama de sectores** (también llamado diagrama de tarta). En él se representa el valor de cada carácter cualitativo como un sector de un círculo completo, siendo el área de cada sector, o, lo que es lo mismo, el arco subtendido, proporcional a la frecuencia del carácter en cuestión. De forma práctica, cada arco se calcula como 360^0 multiplicado por la frecuencia relativa. Es además costumbre escribir dentro, o a un lado, de cada sector la frecuencia correspondiente. Este tipo de diagrama proporciona una idea visual muy clara de cuáles son los caracteres que más se repiten.

Ejemplo I–7

Las notas de una asignatura de Físicas (en la UCM) del curso académico 95/96 se distribuyeron de acuerdo a la siguiente tabla para los alumnos presentados en junio:

Nota	n_i	f_i	N_i	F_i	α_i
Suspenso (SS)	110	0.46	110	0.46	165.6
Aprobado (AP)	90	0.38	200	0.84	136.8
Notable (NT)	23	0.10	223	0.94	36.0
Sobresaliente (SB)	12	0.05	235	0.99	18.0
Matrícula de Honor (MH)	2	0.01	237	1.00	3.6

Los diagramas de rectángulos y de sectores correspondientes se muestran en la Figura 2.5.

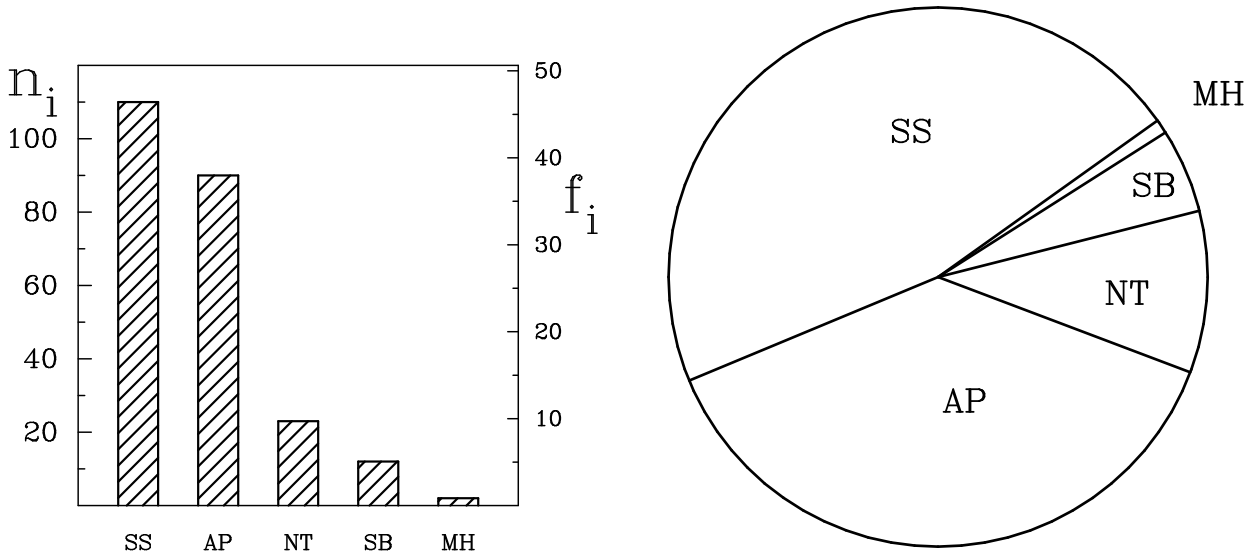


Figura 2.5: *Diagrama de rectángulos (izquierda) y de sectores (derecha) para las notas del ejemplo I–7. Las frecuencias relativas están dadas en tanto por ciento. Los ángulos de cada sector circular se determinan como* $\alpha_i = f_i \times 360$ *(grados).*

Capítulo 3

Medidas características de una distribución

"La percepción, sin comprobación ni fundamento, no es garantía suficiente de verdad."

Bertrand Russell (1872-1970)

Después de haber aprendido en el capítulo anterior a construir tablas de frecuencias y haber realizado alguna representación gráfica, el siguiente paso para llevar a cabo un estudio preliminar de los datos recogidos es el cálculo de diferentes magnitudes características de la distribución. Se definen entonces diversas medidas que serán capaces de resumir toda la información recogida a un pequeño número de valores. Estas *medidas resumen* van a permitir comparar nuestra muestra con otras y dar una idea rápida de cómo se distribuyen los datos. Es evidente que todas estas medidas solo pueden definirse para **variables cuantitativas**.

3.1. Medidas de centralización

Entre las medidas características de una distribución destacan las llamadas medidas de **centralización**, que nos indicarán el valor promedio de los datos, o en torno a qué valor se distribuyen estos.

3.1.1. Media aritmética

Supongamos que tenemos una muestra de tamaño N, donde la variable estadística x toma los valores $x_1, x_2, \ldots, x_N$. Se define la **media aritmética** $\overline{x}$, o simplemente **media**, de la muestra como

$$\overline{x} = \frac{\sum_{i=1}^{N} x_i}{N}. \tag{3.1}$$

Es decir, la media se calcula sencillamente sumando los distintos valores de x y dividiendo por el número de datos. En el caso de que los diferentes valores de x aparezcan *repetidos*, tomando entonces los valores $x_1, x_2, \ldots, x_k$, con frecuencias absolutas $n_1, n_2, \ldots, n_k$, la media se determina como

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_i}{N}, \tag{3.2}$$

pudiéndose expresar también en función de las frecuencias relativas mediante

$$\overline{x} = \sum_{i=1}^{k} x_i f_i. \tag{3.3}$$

Ejemplo I–5

(Continuación.) Calcularemos la media aritmética para los datos del ejemplo I–5.

x_i	n_i	f_i	$x_i \times n_i$	$x_i \times f_i$
1	6	0.30	6	0.30
2	7	0.35	14	0.70
3	4	0.20	12	0.60
4	2	0.10	8	0.40
5	1	0.05	5	0.25
Total	20	1.00	45	2.25

Aplicando la ecuación (3.2)

$$\overline{x} = \frac{\sum_1^5 x_i\, n_i}{N} = \frac{45}{20} = 2.25,$$

o también usando las frecuencias relativas mediante la ecuación (3.3)

$$\overline{x} = \sum_1^5 x_i\, f_i = 2.25.$$

En el caso de tener una muestra agrupada en k intervalos de clase la media se puede calcular, a partir de las marcas de clase c_i y el número n_i de datos en cada intervalo, utilizando una expresión similar a (3.2)

$$\overline{x} = \frac{\sum_{i=1}^k c_i n_i}{N}. \tag{3.4}$$

Sin embargo, hay que indicar que la expresión anterior es solamente aproximada. En el caso de que sea posible, es más exacto para el cálculo de la media, no realizar el agrupamiento en intervalos y usar la expresión (3.1).

Ejemplo I–6

(Continuación.) Calcularemos la media aritmética para el ejemplo I–6.

c_i	n_i	$c_i \times n_i$
7.755	7	54.285
8.455	9	76.095
9.155	2	18.310
9.855	2	19.710
10.555	1	10.555
Total	21	178.955

Aplicando la ecuación (3.4)

$$\overline{x} = \frac{\sum_1^5 c_i\, n_i}{N} = \frac{178.955}{21} = 8.522.$$

Si empleamos en su lugar la expresión correcta dada por la ecuación (3.1), se obtiene

$$\overline{x} = \frac{\sum_1^{21} x_i}{N} = \frac{178.43}{21} = 8.497.$$

Una propiedad importante de la media aritmética es que la suma de las desviaciones de un conjunto de datos respecto a su media es cero. Es decir, la media equilibra las desviaciones positivas y negativas respecto

a su valor

$$\sum_{i=1}^{N}(x_i - \overline{x}) = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \overline{x} = \sum_{i=1}^{N} x_i - N\overline{x} = 0. \tag{3.5}$$

La media representa entonces una especie de *centro de gravedad*, o centro geométrico, del conjunto de medidas. Una característica importante de la media como medida de tendencia central es que es muy poco *robusta*, es decir depende mucho de valores particulares de los datos. Si por ejemplo, en una muestra introducimos un nuevo dato con un valor mucho mayor que el resto, la media aumenta apreciablemente (dados los datos $1, 2, 1, 1, 100$, se tiene $\overline{x} = 21$). La media aritmética es por tanto muy dependiente de observaciones extremas.

Como el objetivo de la estadística descriptiva es describir de la forma más simple y clara la muestra obtenida, es importante siempre usar unas unidades que cumplan mejor dicho fin. Por este motivo, a veces es muy útil realizar un cambio de origen y unidades para simplificar los valores de la variable. Por ejemplo, supongamos que x es la altura en metros de una muestra de individuos. Tomará entonces valores típicos $x = 1.75, 1.80, 1.67, \ldots$. Si efectuamos aquí un cambio a una nueva variable y definida como $y = 100(x - 1.65)$, los nuevos valores serán $y = 10, 15, 2, \ldots$ y, por tanto, el análisis será más sencillo y se usarán menos dígitos. A este proceso de cambio de origen y unidades se le llama una **transformación lineal** y, en general, consistirá en pasar de una variable x a otra y definida como

$$y = a + bx. \tag{3.6}$$

Es fácil encontrar una relación entre la media aritmética de x e y, ya que

$$\overline{y} = \frac{\sum y_i}{N} = \frac{\sum(a + bx_i)}{N}. = \frac{aN + b\sum x_i}{N} = a + b\overline{x}$$

Es decir, una vez calculada la media aritmética de la nueva variable y, se puede encontrar la media de x haciendo

$$\overline{x} = \frac{\overline{y} - a}{b}.$$

Ejemplo I–8

Supongamos una serie de medidas experimentales con un péndulo simple para obtener el valor de la aceleración de la gravedad (en $\mathrm{m/s^2}$).

x_i	y_i
9.77	-3
9.78	-2
9.80	0
9.81	$+1$
9.83	$+3$
10.25	$+45$

Calculemos primero la media aritmética

$$\overline{x} = \frac{\sum_1^6 x_i}{N} = \frac{59.24}{6} = 9.873 \text{ m/s}^2.$$

Si hacemos un cambio de variable $y = a + b\ x = -980 + 100\ x$, y calculamos los valores de y_i (segunda columna de la tabla de la izquierda), el valor de la media sería

$$\overline{y} = \frac{\sum_1^6 y_i}{N} = \frac{44}{6} = 7.33,$$

$$\overline{x} = \frac{\overline{y} - a}{b} = \frac{7.33 + 980}{100} = 9.873 \text{ m/s}^2.$$

Nótese lo sensible que es la media de un valor extremo. Si no tuviésemos en cuenta el último valor, obtendríamos $\overline{x} = 9.798$.

3.1.2. Medias geométrica, armónica y cuadrática

Existen otras definiciones de media que pueden tener su utilidad en algunos casos. La primera de éstas es la **media geométrica** x_G. En el caso de una muestra con valores diferentes de la variable se define como la raíz enésima (N es el tamaño de la muestra) del producto de los valores de la variable

$$\overline{x_G} = \sqrt[N]{x_1 x_2 \dots x_N}. \tag{3.7}$$

Si los datos aparecen agrupados en k valores distintos la definición sería

$$\overline{x_G} = \sqrt[N]{x_1^{n_1} x_2^{n_2} \dots x_k^{n_k}}. \tag{3.8}$$

Esta media tiene la característica negativa de que si uno de los valores es nulo, la media sería asimismo cero, y por lo tanto sería poco representativa del valor central. Además si existen valores negativos es posible que no se pueda calcular. A la hora de calcularla es útil tener en cuenta que el logaritmo de la media geométrica es la media aritmética del logaritmo de los datos

$$\log \overline{x_G} = \frac{\sum_{i=1}^{k} n_i \log x_i}{N}.$$

La **media armónica** x_A se define como la inversa de la media aritmética de las inversas de los valores de la variable. Es decir, para variables no agrupadas y agrupadas, sería

$$\overline{x_A} = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}} \quad ; \quad \overline{x_A} = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}. \tag{3.9}$$

Es evidente que si una de las medidas es 0, la media armónica no tiene sentido.

Una tercera definición corresponde a la **media cuadrática** x_Q. Se define ésta como la raíz cuadrada de la media aritmética de los cuadrados de los valores

$$\overline{x_Q} = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \quad ; \quad \overline{x_Q} = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{N}}. \tag{3.10}$$

Esta media tiene su utilidad con frecuencia en la aplicación a fenómenos físicos.

Se puede demostrar que estas medias se relacionan con la media aritmética, en el caso de valores positivos de la variable, por

$$\overline{x_A} \leq \overline{x_G} \leq \overline{x} \leq \overline{x_Q}.$$

Ninguna de estas medias es muy robusta en general, aunque esto depende de cómo se distribuyan las variables. Por ejemplo, la media armónica es muy poco sensible a valores muy altos de x, mientras que a la media cuadrática apenas le afectan los valores muy bajos de la variable.

Ejemplo I–8

(Continuación.)

Media geométrica

$$\overline{x_G} = \sqrt[6]{x_1 x_2 \dots x_6} = \sqrt[6]{9.77 \times 9.78 \times \dots \times 10.25} = 9.872.$$

Media armónica

$$\overline{x_A} = \frac{6}{\sum_{i=1}^{6} \frac{1}{x_i}} = \frac{6}{\frac{1}{9.77} + \frac{1}{9.78} + \dots + \frac{1}{10.25}} = 9.871.$$

Media cuadrática

$$\overline{x_Q} = \sqrt{\frac{\sum_{i=1}^{6} x_i^2}{6}} = \sqrt{\frac{9.77^2 + 9.78^2 + \dots + 10.25^2}{6}} = 9.875.$$

Debe notarse que

$$\begin{array}{ccccccc} \overline{x_A} & \leq & \overline{x_G} & \leq & \overline{x} & \leq & \overline{x_Q} \\ 9.871 & \leq & 9.872 & \leq & 9.873 & \leq & 9.875 \end{array}$$

y que la media armónica es la menos afectada por el valor demasiado alto, mientras que la cuadrática es la más sensible a dicho número.

3.1.3. Mediana

Una medida de centralización importante es la **mediana** M_e. Se define ésta como una medida central tal que, con los datos ordenados de menor a mayor, el 50 % de los datos son inferiores a su valor y el 50 % de los datos tienen valores superiores. Es decir, la mediana divide en dos partes iguales la distribución de frecuencias o, gráficamente, divide el histograma en dos partes de áreas iguales. Vamos a distinguir diversos casos para su cálculo:

1. Supongamos en primer lugar que los diferentes valores de la variable no aparecen, en general, repetidos. En este caso, y suponiendo que tenemos los datos ordenados, la mediana será el valor central, si N is impar, o la media aritmética de los dos valores centrales, si N es par. Por ejemplo, si $x = 1, 4, 6, 7, 9$, la mediana sería 6. Por otro lado, si $x = 1, 4, 6, 7$ la mediana es $M_e = (4 + 6)/2 = 5$.

Ejemplo I–8

(Continuación.)

Para el ejemplo de las medidas de la gravedad, como el número de datos es par ($N = 6$), se situará entre los dos centrales (media aritmética)

$$9.77/9.78/9.80/\ *\ /9.81/9.83/10.25$$

$$M_e = \frac{9.80 + 9.81}{2} = 9.805$$

Nótese que no depende tanto del valor extremo. Es una medida más robusta. Compárese con el valor $\overline{x} = 9.873$ calculado anteriormente.

2. En el caso de que tengamos una variable discreta con valores repetidos sobre la cual hemos elaborado una tabla de frecuencias se calcula en primer lugar el número de observaciones N dividido entre 2. Podemos distinguir entonces dos casos. El primero de ellos es cuando dicho valor $N/2$ coincide con la frecuencia absoluta acumulada N_j de un valor x_j de la variable (o, lo que es lo mismo, cuando la frecuencia relativa acumulada $F_j = 0.5$). En este caso la mediana se ha de situar entre este valor de la variable y el siguiente ya que de esta forma dividirá la distribución de frecuencias en 2. Es decir, se calcula como la media aritmética de dicho valor de la variable y su superior

$$M_e = \frac{x_j + x_{j+1}}{2}$$

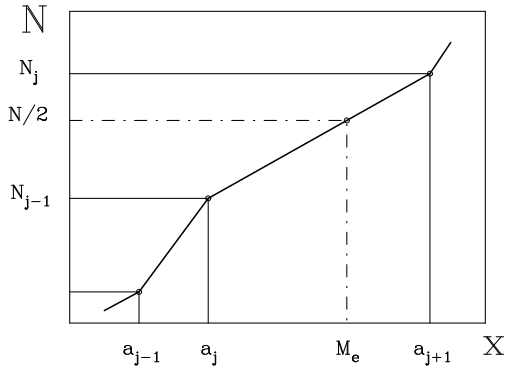


Figura 3.1: *Interpolación en el polígono de frecuencias para determinar la mediana en el caso de que $N/2$ no coincida con ninguna frecuencia acumulada N_j.*

Si $N/2$ no coincidiese con ningún valor de la columna de frecuencias acumuladas (como suele ocurrir) la mediana sería el primer valor de x_j con frecuencia absoluta acumulada N_j mayor que $N/2$, ya que el valor central de la distribución correspondería a una de las medidas englobadas en ese x_j.

Ejemplo I–5

(Continuación.)
Usando los datos del número de hijos del ejemplo I–5, tenemos

x_i	N_i
1	6
2	13
3	17
4	19
5	20

1–1–1–1–1–1–1–2–2–2–2–2–2–2–3–3–3–3–4–4–5

$$N/2 = 10$$

La mediana será el primer valor de x_i con frecuencia absoluta acumulada $N_i > 10$, es decir

$$M_e = x_2 = 2.$$

Modificando la tabla de datos para estar en el otro caso mencionado

x_i	N_i
1	6
2	10
3	15
4	17
5	20

1–1–1–1–1–1–2–2–2–2–3–3–3–3–3–4–4–5–5–5

$$N/2 = 10 = N_2,$$

entonces

$$M_e = \frac{x_2 + x_{2+1}}{2} = \frac{2+3}{2} = 2.5.$$

3. Supongamos ahora que tenemos una muestra de una variable continua cuyos valores están agrupados en intervalos de clase. En este caso pueden ocurrir dos situaciones. En primer lugar, si $N/2$ coincide con la frecuencia absoluta acumulada N_j de un intervalo (a_j, a_{j+1}) (con marca de clase c_j), la mediana será sencillamente el extremo superior a_{j+1} de ese intervalo. En el caso general de que ninguna frecuencia absoluta acumulada coincida con $N/2$ será necesario interpolar en el polígono de frecuencias acumuladas (Fig. 3.1). Supongamos que el valor $N/2$ se encuentra entre las frecuencias N_{j-1} y N_j, correspondientes a los intervalos (a_{j-1}, a_j) y (a_j, a_{j+1}) respectivamente, la mediana se situará en algún lugar del intervalo superior (a_j, a_{j+1}). Para calcular el valor exacto se interpola según se observa en la Figura 3.1

$$\frac{a_{j+1} - a_j}{N_j - N_{j-1}} = \frac{M_e - a_j}{N/2 - N_{j-1}}$$

$$\Rightarrow M_e = a_j + \frac{N/2 - N_{j-1}}{N_j - N_{j-1}}(a_{j+1} - a_j) = a_j + \frac{N/2 - N_{j-1}}{n_j}(a_{j+1} - a_j).$$

Ejemplo I–6

(Continuación.)
Volviendo de nuevo a las medidas agrupadas del ejemplo I–6, podemos calcular la mediana recordando el agrupamiento en intervalos que realizamos en su momento.

a_i—a_{i+1}	n_i	N_i
7.405—8.105	7	7
8.105—8.805	9	16
8.805—9.505	2	18
9.505—10.205	2	20
10.205—10.905	1	21

$$N/2 = 10.5 \neq N_i$$
$$(N_1 = 7) < (N/2 = 10.5) < (N_2 = 16)$$
La mediana se situará entonces en el intervalo 8.105—8.805,
$$8.105 < M_e < 8.805.$$

$$M_e = a_j + \frac{N/2 - N_{j-1}}{n_j}(a_{j+1} - a_j) = a_2 + \frac{10.5 - N_1}{n_2}(a_3 - a_2) =$$
$$= 8.105 + \frac{10.5 - 7}{9}(8.805 - 8.105) = 8.105 + 0.388 \times 0.7 = 8.38.$$

Compárese este resultado con $\overline{x} = 8.52$.

En comparación con la media aritmética la mediana, como medida de centralización, tiene propiedades muy distintas, presentando sus ventajas e inconvenientes. Por un lado, la mayor ventaja de la media es que se utiliza toda la información de la distribución de frecuencias (todos los valores particulares de la variable), en contraste con la mediana, que solo utiliza el orden en que se distribuyen los valores. Podría pues considerarse, desde este punto de vista, que la media aritmética es una medida más fiable del valor central de los datos. Sin embargo, como hemos visto anteriormente, la media es muy poco robusta, en el sentido de que es muy sensible a valores extremos de la variable y, por lo tanto, a posibles errores en las medidas. La mediana, por otro lado, es una medida robusta, siendo muy insensible a valores que se desvíen mucho. Por ejemplo, supongamos que la variable x toma los valores $x = 2, 4, 5, 7, 8$, la media y la mediana serían en este caso muy parecidas ($\overline{x} = 5.2$, $M_e = 5$). Pero si sustituimos el último valor 8 por 30, la nueva media se ve muy afectada ($\overline{x} = 9.6$), no siendo en absoluto una medida de la tendencia central, mientras que el valor de la mediana no cambia ($M_e = 5$). Podríamos poner como contraejemplo el caso de las longitudes de barras (en cm) inicialmente idénticas calentadas a temperaturas desconocidas en distintos recipientes: 1.80/1.82/1.85/1.90/2.00, cuya media y mediana son $\overline{x} = 1.874$ y $M_e = 1.85$. Si la temperatura de uno de esos recipientes varía, y la longitud mayor aumenta de 2.00 a 2.20 cm, la mediana no varía, pero la media pasa a $\overline{x} = 1.914$ y nos informa del cambio.

En general, lo mejor es considerar media aritmética y mediana como medidas complementarias. Es más, la comparación de sus valores puede suministrar información muy útil sobre la distribución de los datos.

3.1.4. Moda

Se define la **moda** M_o de una muestra como aquel valor de la variable que tiene una frecuencia máxima. En otras palabras, es el valor que más se repite. Hay que indicar que puede suceder que la moda no sea única, es decir que aparezcan varios máximos en la distribución de frecuencias. En ese caso diremos que tenemos una distribución bimodal, trimodal, etc. Evidentemente, en el caso de una variable discreta que no toma valores repetidos, la moda no tiene sentido. Cuando sí existen valores repetidos su cálculo es directo ya que puede leerse directamente de la tabla de distribución de frecuencias.

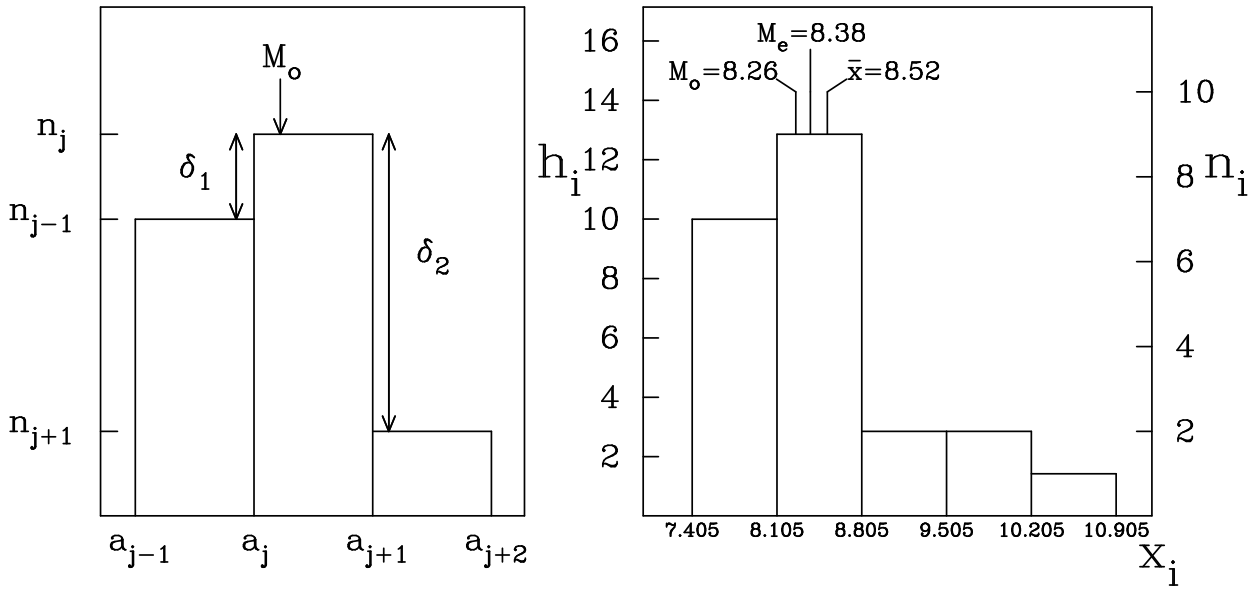


Figura 3.2: *Determinación de la moda utilizando las diferencias de frecuencias entre el intervalo modal y los adyacentes. Histograma con datos del ejemplo I–6 (también ejemplo I–2), y localización de la media, mediana y moda.*

Ejemplo I–5

(Continuación.)
Consideremos de nuevo el caso del número de hijos por familia.

x_i	n_i	f_i	N_i	F_i
1	6	0.30	6	0.30
2	7	0.35	13	0.65
3	4	0.20	17	0.85
4	2	0.10	19	0.95
5	1	0.05	20	1.00

El valor que más se repite es 2 hijos, que ocurre en siete familias de la muestra ($n_i = 7$). La moda es por tanto $M_o = 2$ y en este caso coincide con la mediana.

En el caso de variables continuas agrupadas en intervalos de clase existirá un intervalo en el que la frecuencia sea máxima, llamado intervalo modal. Es posible asociar la moda a un valor determinado de la variable dentro de dicho intervalo modal. Para ello supongamos que sea (a_j, a_{j+1}) el intervalo con frecuencia máxima n_j. Si n_{j-1} y n_{j+1} son las frecuencias de los intervalos anterior y posterior al modal, definimos $\delta_1 = n_j - n_{j-1}$ y $\delta_2 = n_j - n_{j+1}$ (ver el histograma de la Figura 3.2). En este caso, el valor exacto de la moda se puede calcular como

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}(a_{j+1} - a_j)$$

(ver demostración en el libro de Quesada). Es decir, la moda estará más próxima a a_j cuanto menor sea la diferencia de frecuencias con el intervalo anterior, y al revés. Si, por ejemplo, $n_{j-1} = n_j$ ($\delta_1 = 0$), la moda será efectivamente a_j. Por el contrario si $n_{j+1} = n_j$ ($\delta_2 = 0$) la moda será a_{j+1}, estando situada entre dos intervalos.

Ejemplo I–6

(Continuación.)

Para el caso de las medidas de la paralaje solar (ejemplo I–6), se estudia el intervalo con frecuencia máxima (intervalo modal) que en este caso es $(a_j, a_{j+1}) = (8.105\ ,\ 8.805)$,

a_i—a_{i+1}	c_i	n_i
7.405—8.105	7.755	7
8.105—8.805	8.455	9 ←
8.805—9.505	9.155	2
9.505—10.205	9.855	2
10.205—10.905	10.555	1

$$j = 2; \quad n_{j-1} = 7; \quad n_j = 9; \quad n_{j+1} = 2$$

$$\delta_1 = n_j - n_{j-1} = 9 - 7 = 2$$

$$\delta_2 = n_j - n_{j+1} = 9 - 2 = 7$$

$$M_o = a_j + \frac{\delta_1}{\delta_1 + \delta_2}(a_{j+1} - a_j) = 8.105 + \frac{2}{2+7}(8.805 - 8.105) = 8.26.$$

En el caso de que tuviésemos una distribución perfectamente simétrica, las tres medidas de centralización media aritmética, mediana y moda coincidirían en el mismo valor. Sin embargo, cuando la distribución de las medidas es claramente asimétrica las posiciones relativas entre las tres medidas suelen ser típicamente como se representa en el polígono de frecuencias de la Figura 3.2. Es decir, la mediana se suele situar entre la moda y la media.

3.1.5. Cuartiles, deciles y percentiles

Vamos a generalizar ahora el concepto de mediana. Vimos que ésta era el valor de la variable que dividía a la muestra (ordenada) en dos mitades iguales. Definimos ahora los **cuartiles** como los tres valores que divididen la muestra en cuatro partes iguales. Así el primer cuartil $Q_{1/4}$ será la medida tal que el 25 % de los datos sean inferiores a su valor y el 75 % de los datos sean superiores. El segundo cuartil $Q_{1/2}$ coincide con la mediana, mientras que el tercer cuartil $Q_{3/4}$ marcará el valor tal que las tres cuartas partes de las observaciones sean inferiores a él y una cuarta parte sea superior. La forma de calcular los cuartiles es igual a la ya vista para la mediana pero sustituyendo $N/2$ por $N/4$ y $3N/4$ para $Q_{1/4}$ y $Q_{3/4}$ respectivamente.

Ejemplo I–5

(Continuación.)

En el ejemplo del número de hijos de una muestra de 20 familias tenemos

x_i	N_i
1	6
2	13
3	17
4	19
5	20

1–1–1–1–1 — 1–2–2–2–2 — 2–2–2–3–3 — 3–3–4–4–5

$$N/4 = 20/4 = 5 \Rightarrow Q_{1/4} = 1$$

$$N/2 = 20/2 = 10 \Rightarrow Q_{1/2} = M_e = 2$$

$$3 \times N/4 = 15 \Rightarrow Q_{3/4} = 3$$

Ejemplo I–6

(Continuación.)

En el caso de las medidas agrupadas en intervalos de clase se trabaja igual que para determinar la mediana.

a_i—a_{i+1}	n_i	N_i
7.405—8.105	7	7
8.105—8.805	9	16
8.805—9.505	2	18
9.505—10.205	2	20
10.205—10.905	1	21

$$N/4 = 5.25 < 7 \qquad 3 \times N/4 = 15.75 < 16$$

$Q_{1/4}$ se sitúa en el primer intervalo 7.405—8.105.

$Q_{3/4}$ se sitúa en el segundo intervalo 8.105—8.805.

$$Q_{1/4} = a_j + \frac{N/4 - N_{j-1}}{n_j}(a_{j+1} - a_j) = 7.405 + \frac{5.25 - 0}{7}0.7 = 7.93.$$

$$Q_{3/4} = a_j + \frac{3 \times N/4 - N_{j-1}}{n_j}(a_{j+1} - a_j) = 8.105 + \frac{15.75 - 7}{9}0.7 = 8.79.$$

De la misma forma podemos definir los **deciles** como aquellos valores de la variable que dividen la muestra, ordenada, en 10 partes iguales. Estos valores, denotados por D_k, con $k = 1, 2, \ldots, 9$, tienen entonces un valor tal que el decil k–esimo deja por debajo de él al 10xk por ciento de los datos de la muestra. De la misma manera se definen los **percentiles**, también llamados centiles, como aquellos valores P_k (con $k = 1, 2, \ldots, 99$) que dividen la muestra en 100 partes iguales. Es decir el percentil P_k deja por debajo de él al k por ciento de la muestra ordenada.

La forma de calcular deciles y percentiles es igual a la de la mediana y los cuartiles, sustituyendo $N/2$ por la fracción del número total de datos correspondiente. Evidentemente algunos valores de cuartiles, deciles y centiles coinciden, cumpliéndose por ejemplo

$$P_{50} = D_5 = Q_{1/2} = M_e$$

3.2. Medidas de dispersión

Las medidas de centralización vistas anteriormente reducen la información recogida de la muestra a un solo valor. Sin embargo, dicho valor central, o medio, será más o menos representativo de los valores de la muestra dependiendo de la dispersión que las medidas individuales tengan respecto a dicho centro. Para analizar la representatividad de las medidas de centralización se definen las llamadas medidas de dispersión. Estas nos indicarán la variabilidad de los datos en torno a su valor promedio, es decir si se encuentran muy o poco esparcidos en torno a su centro. Se pueden definir entonces, diversas medidas de desviación o dispersión, siendo éstas fundamentales para la descripción estadística de la muestra.

3.2.1. Recorridos

Una evaluación rápida de la dispersión de los datos se puede realizar calculando el **recorrido** (también llamado rango), o diferencia entre el valor máximo y mínimo que toma la variable estadística. Con el fin de eliminar la excesiva influencia de los valores extremos en el recorrido, se define el **recorrido intercuartílico** como la diferencia entre el trecer y primer cuartil

$$R_I = Q_{3/4} - Q_{1/4}. \tag{3.11}$$

Está claro que este recorrido nos dará entonces el rango que ocupan el 50 % central de los datos. En ocasiones se utiliza el **recorrido semiintercuartílico**, o mitad del recorrido intercuartílico

$$R_{SI} = \frac{Q_{3/4} - Q_{1/4}}{2}.$$

3.2.2. Desviación media

Otra manera de estimar la dispersión de los valores de la muestra es comparar cada uno de estos con el valor de una medida de centralización. Una de las medidas de dispersión más usada es la **desviación media**, también llamada con más precisión desviación media respecto a la media aritmética. Se define ésta como la media aritmética de las diferencias absolutas entre los valores de la variable y la media aritmética de la muestra. Suponiendo que en una muestra de tamaño N los k distintos valores x_i de la variable tengan

frecuencias absolutas n_i, la expresión de la desviación media será

$$D_{\overline{x}} = \frac{\sum_{i=1}^{k} |x_i - \overline{x}| n_i}{N}. \tag{3.12}$$

Evidentemente, en el caso de que la variable no tome valores repetidos, ni esté agrupada en intervalos, la expresión anterior se simplifica a

$$D_{\overline{x}} = \frac{\sum_{i=1}^{N} |x_i - \overline{x}|}{N}. \tag{3.13}$$

Hay que destacar la importancia de tomar valores absolutos de las desviaciones. Si no se hiciese así unas desviaciones se anularían con otras, alcanzando finalmente la desviación media un valor de 0, debido a la propiedad de la media aritmética vista en (3.5).

En ocasiones se define una desviación media en términos de desviaciones absolutas en torno a una medida de centralización diferente de la media aritmética. Cuando se utiliza la mediana se obtiene la llamada **desviación media respecto a la mediana**, definida como

$$D_{M_e} = \frac{\sum_{i=1}^{k} |x_i - M_e| n_i}{N}. \tag{3.14}$$

Ejemplo I–5

(Continuación.)
Calculemos el recorrido semiintercuartílico y las desviación respecto a la media aritmética.

$$R_{SI} = \frac{Q_{3/4} - Q_{1/4}}{2} = \frac{3-1}{2} = 1$$
$$D_{\overline{x}} = \frac{\sum_{1}^{k} |x_i - \overline{x}| n_i}{N} = \frac{\sum_{1}^{5} |x_i - 2.25| n_i}{20} = 0.925$$

Ejemplo I–6

(Continuación.)
Calculemos el recorrido semiintercuartílico y las desviación respecto a la media aritmética.

$$R_{SI} = \frac{Q_{3/4} - Q_{1/4}}{2} = \frac{8.79 - 7.93}{2} = 0.43$$
$$D_{\overline{x}} = \frac{\sum_{1}^{k} |x_i - \overline{x}| n_i}{N} = \frac{\sum_{1}^{5} |x_i - 8.52| n_i}{21} = 0.57$$

3.2.3. Varianza y desviación típica

Sin lugar a dudas la medida más usada para estimar la dispersión de los datos es la desviación típica. Esta es especialmente aconsejable cuando se usa la media aritmética como medida de tendencia central. Al igual que la desviación media, está basada en un valor promedio de las desviaciones respecto a la media. En este caso, en vez de tomar valores absolutos de las desviaciones, para evitar así que se compensen desviaciones positivas y negativas, se usan los cuadrados de las desviaciones. Esto hace además que los datos con desviaciones grandes influyan mucho en el resultado final. Se define entonces la **varianza** de una muestra con datos repetidos como

$$s^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N-1}. \tag{3.15}$$

Evidentemente la varianza no tiene las mismas unidades que los datos de la muestra. Para conseguir las mismas unidades se define la **desviación típica** (algunas veces llamada desviación estándar) como la raíz cuadrada de la varianza

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_i}{N-1}}. \tag{3.16}$$

En el caso de que los datos no se repitan, estas definiciones se simplifican a

$$s^2 = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1} \qquad ; \qquad s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}. \tag{3.17}$$

En muchas ocasiones se definen varianza y desviación típica utilizando N en vez de $N-1$ en el denominador, representando entonces la varianza una verdadera media aritmética del cuadrado de las desviaciones. Está claro que ambas definiciones llevan a valores muy parecidos cuando N es grande. El motivo de haber optado aquí por la definición con $N-1$ es que ésta da una mejor estimación de la dispersión de los datos. Téngase en cuenta que como la suma de las desviaciones $x_i - \overline{x}$ es siempre 0 (ver (3.5)), la desviación del último dato puede calcularse una vez que se conozcan las $N-1$ anteriores. Es decir, sólo se tienen $N-1$ desviaciones independientes (se dice que el sistema tiene $N-1$ grados de libertad) y se promedia entonces dividiendo por $N-1$, ya que no tiene mucho sentido promediar N números no independientes. Notesé además que cuando solo se tiene un dato ($N=1$), en el caso de la definición con N en el denominador se obtendría una varianza 0, que no tiene mucho sentido, mientras que en la definición con $N-1$ la varianza estaría indeterminada. En cualquier caso, siempre se puede obtener una desviación típica a partir de la otra multiplicando (o dividiendo) por $\sqrt{(N-1)/N}$

$$\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N}} = \sqrt{\frac{N-1}{N}}\sqrt{\frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1}}.$$

La expresión (3.15) no es muy comoda para calcular la desviación típica de forma rápida. A efectos prácticos, dicha expresión se puede transformar en otra más fácil de aplicar

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N-1} = \frac{\sum x_i^2 n_i - 2\sum x_i \overline{x} n_i + \sum \overline{x}^2 n_i}{N-1} =$$
$$= \frac{\sum x_i^2 n_i - 2\overline{x}\sum x_i n_i + N\overline{x}^2}{N-1},$$

donde se ha usado que $\sum_{i=1}^{k} n_i = N$. Utilizando ahora la expresión (3.2) para la media

$$s^2 = \frac{\sum x_i^2 n_i - 2\frac{1}{N}\sum x_i n_i \sum x_i n_i + \frac{N}{N^2}(\sum x_i n_i)^2}{N-1} = \frac{\sum_{i=1}^{k} x_i^2 n_i - \frac{1}{N}(\sum_{i=1}^{k} x_i n_i)^2}{N-1}.$$

La expresión anterior es más facil de aplicar ya que bastará con calcular los sumatorios de los datos al cuadrado y de los datos, habiéndose calculado ya este último para la media.

Ejemplo I–5

(Continuación.)
En el caso de una variable discreta

x_i	n_i	$x_i \times n_i$	$x_i^2 \times n_i$
1	6	6	6
2	7	14	28
3	4	12	36
4	2	8	32
5	1	5	25
Total	20	45	127

$$s^2 = \frac{\sum_1^5 x_i^2\, n_i - \frac{1}{20}(\sum_1^5 x_i\, n_i)^2}{20-1}$$

$$s^2 = \frac{127 - \frac{1}{20}45^2}{19} = 1.355$$

$$s = \sqrt{1.355} = 1.16$$

Ejemplo I–6

(Continuación.)
En el caso de datos agrupados en intervalos de clase

c_i	n_i	$c_i \times n_i$	$c_i^2 \times n_i$
7.755	7	54.285	420.980
8.455	9	76.095	643.383
9.155	2	18.310	167.628
9.855	2	19.710	194.242
10.555	1	10.555	111.408
Total	21	178.955	1537.641

$$s^2 = \frac{\sum_1^5 c_i^2\, n_i - \frac{1}{20}(\sum_1^5 c_i\, n_i)^2}{21-1}$$

$$s^2 = \frac{1537.641 - \frac{1}{21}178.955^2}{20} = 0.632$$

$$s = \sqrt{0.632} = 0.795$$

(sin agrupar en intervalos se obtiene $s = 0.900$)

En cuanto a las propiedades de la desviación típica, es fácil ver que ésta será siempre positiva y sólo tendrá un valor nulo cuando todas las observaciones coincidan con el valor de la media. Además, si se define la desviación cuadrática respecto a un promedio a como

$$D^2 = \frac{\sum_{i=1}^k (x_i - a)^2 n_i}{N-1}.$$

Se puede demostrar que dicha desviación cuadrática será mínima cuando $a = \overline{x}$. Es decir, la varianza (y, por tanto, la desviación típica) es la mínima desviación cuadrática. Para demostrarlo derivamos la expresión anterior respecto a a, e igualamos la derivada a 0 (condición necesaria para que D^2 sea mínimo)

$$\frac{\partial D^2}{\partial a} = 0 = \frac{-2\sum (x_i - a) n_i}{N-1}$$

$$\Rightarrow \quad \sum (x_i - a) n_i = 0 \quad \Rightarrow \quad \sum x_i n_i - a \sum n_i = 0$$

$$\Rightarrow \quad \sum x_i n_i - aN = 0 \quad \Rightarrow \quad a = \frac{\sum x_i n_i}{N} = \overline{x},$$

como queríamos demostrar. Esta propiedad le da además más sentido a la definición de la desviación típica.

Hay que indicar que la desviación típica no es una medida robusta de la dispersión. El hecho de que se calcule evaluando los cuadrados de las desviaciones hace que sea muy sensible a observaciones extremas, bastante más que la desviación media (dado que aparece un cuadrado). En definitiva, la desviación típica no es una buena medida de dispersión cuando se tiene algún dato muy alejado de la media. El rango intercuartílico nos daría en ese caso una idea más aproximada de cuál es la dispersión de los datos. El que la desviación típica sea la medida de dispersión más común se debe a su íntima conexión con la distribución normal, como se verá en sucesivos capítulos.

En la discusión sobre la media aritmética se vió cómo su cálculo se podía simplificar a veces si se realizaba una transformación lineal de la variable x a una nueva variable y, definida en (3.6). En este caso, existe una relación muy sencilla entre las desviaciones típicas (s_x y s_y) de ambas variables, ya que

$$s_y = \sqrt{\frac{\sum (y_i - \overline{y})^2}{N-1}} = \sqrt{\frac{\sum (a + bx_i - a - b\overline{x})^2}{N-1}} = \sqrt{\frac{b^2 \sum (x_i - \overline{x})^2}{N-1}} = b s_x.$$

De esta forma, una vez calculada la desviación típica de y, se puede evaluar la de x haciendo

$$s_x = \frac{s_y}{b}.$$

Se demuestra así además que, aunque la desviación típica depende de la unidades elegidas (a través de b), es independiente de un cambio de origen (dado por a).

Ejemplo I–8

(Continuación.)
En el ejemplo de las medidas con el péndulo simple, ya vimos que para el cálculo de la media aritmética efectuábamos un cambio de variable $y = a + b\,x = -980 + 100\,x$.

x_i	y_i
9.77	−3
9.78	−2
9.80	0
9.81	+1
9.83	+3
10.25	+45

$$s_x^2 = \frac{\sum_1^6 (x_i - \overline{x})^2}{N-1} \quad ; \quad s_y^2 = \frac{\sum_1^6 (y_i - \overline{y})^2}{N-1}$$

$$s_y^2 = \frac{\sum_1^6 (y_i - 7.33)^2}{5} = 345.07$$

$$\Rightarrow s_y = \sqrt{345.07} = 18.58$$

$$s_x = \frac{s_y}{b} = \frac{18.58}{100} = 0.186 \text{ m/s}^2.$$

Nótese que es mucho mayor que la desviación media $D_{\overline{x}} = 0.125$. La desviación típica es poco robusta y fuertemente dependiente de los valores extremos.

3.2.4. Coeficientes de variación

Un problema que plantean las medidas de dispersión vistas es que vienen expresadas en las unidades en que se ha medido la variable. Es decir, son medidas absolutas y con el único dato de su valor no es posible decir si tenemos una dispersión importante o no. Para solucionar esto, se definen unas medidas de dispersión relativas, independientes de la unidades usadas. Estas dispersiones relativas van a permitir además comparar la dispersión entre diferentes muestras (con unidades diferentes). Entre estas medidas hay que destacar el **coeficiente de variación de Pearson**, definido como el cociente entre la desviación típica y la media aritmética

$$CV = \frac{s}{|\overline{x}|}. \tag{3.18}$$

Nótese que este coeficiente no se puede calcular cuando $\overline{x} = 0$. Normalmente CV se expresa en porcentaje, multiplicando su valor por 100. Evidentemente, cuanto mayor sea CV, mayor dispersión tendrán los datos.

Ejemplo I–*

(Continuación.)
Calculemos el coeficiente de variación de los ejemplos anteriores.

Ejemplo I–5: $CV = s/|\overline{x}| = 1.16/2.25 = 0.516$ 52 %.

Ejemplo I–6: $CV = s/|\overline{x}| = 0.795/8.52 = 0.093$ 9 %.

Ejemplo I–8: $CV = s/|\overline{x}| = 0.186/9.873 = 0.019$ 2 %.

Asimismo se pueden definir otras medidas de dispersión relativas, como el **coeficiente de variación media**. Éste es similar al coeficiente de variación de Pearson, pero empleando una desviación media en vez de la media aritmética. Se tienen entonces dos coeficientes de variación media dependiendo de que se calcule respecto a la desviación media respecto a la media aritmética o respecto a la mediana

$$CVM_{\overline{x}} = \frac{D_{\overline{x}}}{|\overline{x}|} \quad ; \quad CVM_{M_e} = \frac{D_{M_e}}{|M_e|}. \tag{3.19}$$

3.3. Momentos

Algunas de las definiciones vistas hasta ahora, como la de la media aritmética y la varianza, son en realidad casos particulares de una definición más general. Si tenemos una muestra de la variable estadística

x, la cual toma los valores $x_1, x_2, \ldots, x_k$ con frecuencias absolutas $n_1, n_2, \ldots, n_k$, se define el **momento de orden r respecto al parámetro c** como

$$M_r(c) = \frac{\sum_{i=1}^{k}(x_i - c)^r n_i}{N}. \tag{3.20}$$

3.3.1. Momentos respecto al origen

Un caso particular especialmente interesante de la definición de momento es cuando $c = 0$. De esta forma se define el **momento de orden r respecto al origen** como

$$a_r = \frac{\sum_{i=1}^{k} x_i^r n_i}{N}. \tag{3.21}$$

Los momentos respecto al origen suministran entonces medidas de tendencia central. Es fácil ver que los primeros momentos respecto al origen son

$$a_0 = \frac{\sum_{i=1}^{k} n_i}{N} = 1 \quad ; \quad a_1 = \frac{\sum_{i=1}^{k} x_i n_i}{N} = \overline{x} \quad ; \quad a_2 = \frac{\sum_{i=1}^{k} x_i^2 n_i}{N} = \overline{x_Q}^2$$

Es decir, la media aritmética es el momento de primer orden respecto al origen.

3.3.2. Momentos respecto a la media

De la misma manera, se pueden obtener medidas de dispersión sustituyendo c por la media aritmética en la definición de momento. Se tiene así los **momentos de orden r respecto a la media**

$$m_r = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^r n_i}{N}, \tag{3.22}$$

donde los primeros momentos son entonces

$$m_0 = \frac{\sum_{i=1}^{k} n_i}{N} = 1 \quad , \quad m_1 = \frac{\sum_{i=1}^{k}(x_i - \overline{x}) n_i}{N} = 0,$$

$$m_2 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 n_i}{N} = \frac{N-1}{N} s^2.$$

El momento de orden 1 se anula por la propiedad de la media aritmética expresada en (3.5). Puede observarse que el momento de orden 2 respecto a la media es, aproximadamente, la varianza.

3.4. Asimetría y curtosis

La descripción estadística de una muestra de datos no concluye con el cálculo de su tendencia central y su dispersión. Para dar una descripción completa es necesario estudiar también el grado de simetría de los datos respecto a su medida central y la concentración de los datos alrededor de dicho valor.

3.4.1. Coeficientes de asimetría

Se dice que una distribución de medidas es **simétrica** cuando valores de la variable equidistantes, a uno y otro lado, del valor central tienen la misma frecuencia. Es decir, en este caso tendremos simetría en el histograma (o en el diagrama de barras) alrededor de una vertical trazada por el punto central. En el

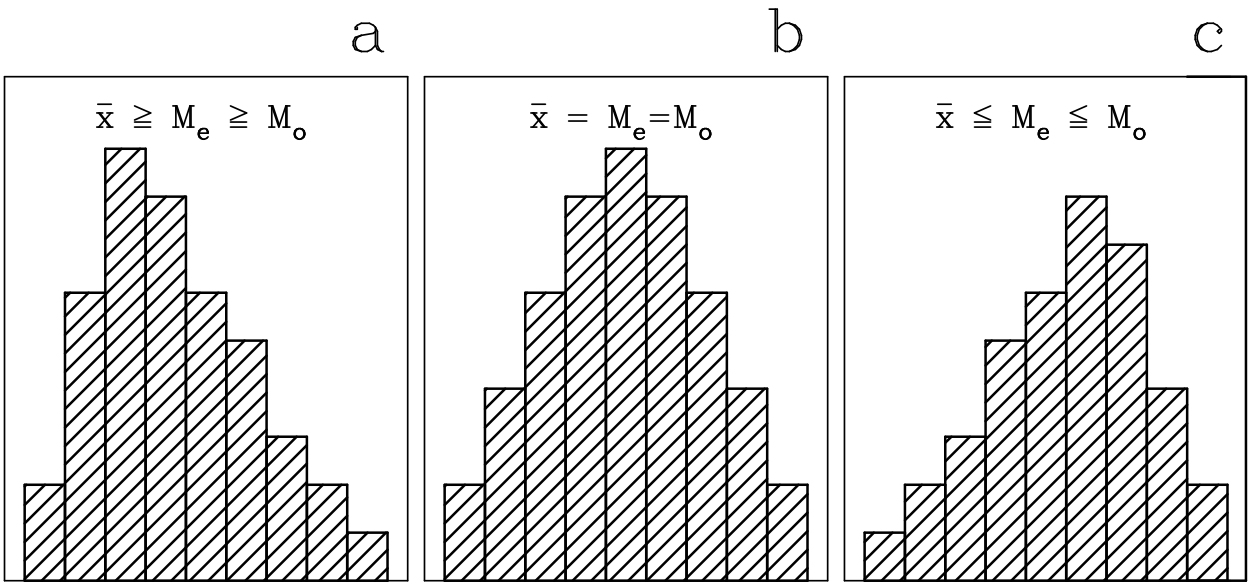


Figura 3.3: *Distribución con asimetría hacia la derecha, positiva, (panel a), simétrica (panel b) y con asimetría hacia la izquierda, negativa (panel c).*

caso de una distribución perfectamente simétrica los valores de media aritmética, mediana y moda coinciden ($\overline{x} = M_e = M_o$).

En el caso de no tener simetría, diremos que tenemos asimetría a la derecha (o positiva) o a la izquierda (o negativa) dependiendo de que el histograma muestre una cola de medidas hacia valores altos o bajos de la variable respectivamente. También se puede decir que la distribución está sesgada a la derecha (sesgo positivo) o a la izquierda (sesgo negativo). En el caso de una distribución asimétrica, la media, mediana y moda no coinciden, siendo $\overline{x} \geq M_e \geq M_o$ para una asimetría positiva y $\overline{x} \leq M_e \leq M_o$ para una asimetría negativa (ver Figura 3.3).

Con el fin de cuantificar el grado de asimetría de una distribución se pueden definir los coeficientes de asimetría. Aunque no son los únicos, existen dos coeficientes principales:

- **Coeficiente de asimetría de Fisher.** Se define como el cociente entre el momento de orden 3 respecto a la media y el cubo de la desviación típica

$$g_1 = \frac{m_3}{s^3} \qquad \text{donde} \qquad m_3 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^3 n_i}{N}. \tag{3.23}$$

  En el caso una distribución simétrica, las desviaciones respecto a la media se anularán (puesto que en m_3 el exponente es impar se sumarán números positivos y negativos) y el coeficiente de asimetría será nulo ($g_1 = 0$). En caso contrario, g_1 tendrá valores positivos para una asimetría positiva (a la derecha) y negativos cuando la asimetría sea en el otro sentido. Hay que indicar que la división por el cubo de la desviación típica se hace para que el coeficiente sea adimensional y, por lo tanto, comparable entre diferentes muestras.

- **Coeficiente de asimetría de Pearson.** Este coeficiente, también adimensional, se define como

$$A_P = \frac{\overline{x} - M_o}{s}. \tag{3.24}$$

  Su interpretación es similar a la del coeficiente de Fisher, siendo nulo para una distribución simétrica (en ese caso media y moda coinciden) y tanto más positivo, o negativo, cuando más sesgada esté la

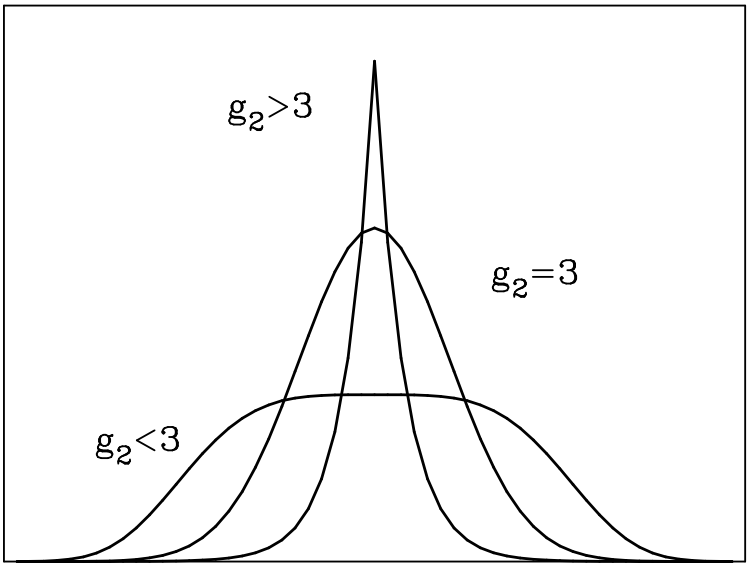


Figura 3.4: *Distribuciones con diferente grado de apuntamiento: leptocúrtica ($g_2 > 3$), mesocúrtica ($g_2 = 3$) y platicúrtica ($g_2 < 3$).*

distribución hacia la derecha, o hacia la izquierda.

Ejemplo I–*

(Continuación.)

Calculemos los coeficientes de asimetría en los ejemplos anteriores.

Ejemplo	$\overline{x}$	s	M_o	m_3	$g_1 = m_3/s^3$	$A_p = (\overline{x} - M_o)/s$
I–5	2.25	1.16	2	1.06	0.68 (positiva)	0.22
I–6	8.52	0.80	8.26	0.50	0.98 (positiva)	0.325

3.4.2. Coeficiente de curtosis

Además de la simetría, otra característica importante de la forma en que se distribuyen los datos de la muestra es cómo es el agrupamiento en torno al valor central. Como se observa en la Figura 3.4, los datos se pueden distribuir de forma que tengamos un gran apuntamiento (o pico en el histograma) alrededor del valor central, en cuyo caso diremos que tenemos una distribución **leptocúrtica**, o en el extremo contrario, el histograma puede ser muy aplanado, lo que corresponde a una distribución **platicúrtica**. En el caso intermedio, diremos que la distribución es **mesocúrtica** y el agrupamiento corresponderá al de una distribución llamada **normal**, o en forma de campana de Gauss.

Esta característica del agrupamiento de los datos se denomina **curtosis** y para cuantificarla se define el **coeficiente de curtosis** como el cociente entre el momento de cuarto orden respecto a la media y la cuarta potencia de la desviación típica

$$g_2 = \frac{m_4}{s^4} \qquad \text{donde} \qquad m_4 = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^4 n_i}{N}. \tag{3.25}$$

Este coeficiente adimensional alcanza valores mayores cuanto más puntiaguda es la distribución, teniendo un valor de 3 para la distribución mesocúrtica (o normal), mayor que 3 para la leptocúrtica y menor para la platicúrtica (ver Figura 3.4).

Capítulo 4

Variables estadísticas bidimensionales

> "Solíamos pensar que si sabíamos lo que significaba *uno*, sabríamos lo que es *dos*, porque uno y uno son dos. Ahora descubrimos que primero debemos aprender mucho más sobre lo que significa *y*."
>
> Sir Arthur Eddington (1882-1944)

Diremos que tenemos una muestra estadística bidimensional cuando sobre cada elemento de la muestra se realiza la observación simultánea de dos caracteres. Por ejemplo, una muestra bidimensional sería una serie de datos sobre altura y presión atmosférica, o la edad y el peso de un grupo de individuos. Tendremos en este caso una **variable estadística bidimensional**, representada por la pareja de símbolos (x, y) y que en general, para una muestra de N elementos, podrá tomar los valores $(x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)$. Evidentemente, los caracteres representados por las variables x e y no tienen porqué ser del mismo tipo, pudiendo ser cada uno de ellos de tipo cuantitativo o cualitativo. Además en el caso de ser ambas variables cuantitativas (caso en el que nos concentraremos en nuestro análisis) cada una de ellas podrá ser continua o discreta. En este capítulo se describirá en primer lugar cómo se puede estudiar la distribución de frecuencias de una variable bidimensional. En el Tema V se abordará el estudio de cómo se pueden analizar las posibles relaciones entre los dos caracteres de una variable bidimensional. Hay que indicar que el estudio de las variables bidimensionales es un caso particular del de las variables n-dimensionales, el cual se puede abordar con facilidad generalizando el primero.

4.1. Distribuciones de frecuencias de una variable bidimensional

De la misma manera que el análisis de la distribución de frecuencias de una variable unidimensional constituye un primer paso para la descripción estadística de la muestra, el estudio de la distribución de frecuencias de una variable bidimensional es de gran utilidad. Evidentemente este estudio solo tendrá sentido cuando tratemos con una variable discreta en la que haya repetición de valores o una variable continua agrupada en intervalos.

4.1.1. Tabla de frecuencias de doble entrada

Al igual que en el caso unidimensional, el primer paso para el estudio de la distribución de frecuencias es la construcción de una tabla de frecuencias. Supongamos que tenemos N pares de medidas de una variable bidimensional (x, y). Diremos que dos pares de medidas serán iguales (o estarán repetidos) cuando coincidan ambas componentes. Supongamos que x puede tomar los k valores distintos $x_1, x_2, \ldots, x_k$, y que y puede

tomar los l valores diferentes $y_1, y_2, \ldots, y_l$, donde k no tiene porqué ser igual a l. Para construir la tabla de frecuencias habrá que contabilizar el número de veces que cada par distinto de la variable bidimensional aparece repetido, ordenándose dichos valores en la llamada **tabla de frecuencias de doble entrada**, donde en ordenadas se escriben los diferentes valores de x y en abscisas los valores de y:

$x \setminus y$	y_1	y_2	y_3	$\cdots$	y_j	$\cdots$	y_l	Suma
x_1	n_{11}	n_{12}	n_{13}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{x_1}
x_2	n_{21}	n_{22}	n_{23}	$\cdots$	n_{2j}	$\cdots$	n_{2l}	n_{x_2}
x_3	n_{31}	n_{32}	n_{33}	$\cdots$	n_{3j}	$\cdots$	n_{3l}	n_{x_3}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	n_{i2}	n_{i3}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{x_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	n_{k1}	n_{k2}	n_{k3}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{x_k}
Suma	n_{y_1}	n_{y_2}	n_{y_3}	$\cdots$	n_{y_j}	$\cdots$	n_{y_l}	N

En esta tabla n_{ij} es la frecuencia absoluta, o número de veces que se repite el par (x_i, y_j). De la misma forma se podría construir una tabla de frecuencias relativas escribiendo los valores f_{ij}, definidos como

$$f_{ij} = \frac{n_{ij}}{N}.$$

Al igual que ocurría en las variables unidimensionales se cumplen las propiedades

$$\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} = N,$$

$$\sum_{i=1}^{k}\sum_{j=1}^{l} f_{ij} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{n_{ij}}{N} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij}}{N} = 1.$$

La tabla anterior se puede construir de la misma manera en el caso de que uno o los dos caracteres x e y correspondan a datos agrupados en intervalos.

Ejemplo I–9

Se tienen los siguientes datos para las alturas x_i (en m) y pesos y_j (en kg):

(1.64,64)	(1.76,77)	(1.79,82)	(1.65,62)	(1.68,71)
(1.65,72)	(1.86,85)	(1.82,68)	(1.73,72)	(1.75,75)
(1.59,81)	(1.87,88)	(1.73,72)	(1.57,71)	(1.63,74)
(1.71,69)	(1.68,81)	(1.73,67)	(1.53,65)	(1.82,73)

Generamos la tabla de frecuencias de doble entrada agrupando los datos.

$x_i \setminus y_j$	60–70	70–80	80–90	n_{x_i}
1.50–1.60	1	1	1	3
1.60–1.70	2	3	1	6
1.70–1.80	2	4	1	7
1.80–1.90	1	1	2	4
n_{y_j}	6	9	5	20

4.1.2. Distribuciones marginales

A veces es interesante analizar cuántas veces se repite un cierto valor de x sin tener en cuenta para nada a los posibles valores de y, o viceversa. Para estudiar cada una de las componentes de la variable bidimensional aisladamente de la otra se definen las **frecuencias marginales** n_{x_i} y n_{y_j} como

$$n_{x_i} = \sum_{j=1}^{l} n_{ij} \qquad ; \qquad n_{y_j} = \sum_{i=1}^{k} n_{ij}. \tag{4.1}$$

De esta forma, n_{x_i} representa el número de veces que x toma el valor x_i, independientemente de los posibles valores de y, y lo mismo para n_{y_j}. A la distribución formada por los diferentes valores de x y sus frecuencias marginales se le llama **distribución marginal** de x. Normalmente las frecuencias marginales de x e y se escriben respectivamente en la última columna y fila de la tabla de frecuencias de doble entrada. Su cálculo es muy sencillo ya que basta con sumar los correspondientes valores de cada fila y columna.

De la misma manera se pueden definir las frecuencias relativas marginales como

$$f_{x_i} = \frac{n_{x_i}}{N} \qquad ; \qquad f_{y_j} = \frac{n_{y_j}}{N}.$$

Algunas propiedades evidentes son

$$\sum_{i=1}^{k} n_{x_i} = N \qquad ; \qquad \sum_{j=1}^{l} n_{y_j} = N.$$

$$\sum_{i=1}^{k} f_{x_i} = 1 \qquad ; \qquad \sum_{j=1}^{l} f_{y_j} = 1.$$

Para caracterizar estas distribuciones marginales se pueden definir sus medias y varianzas como

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i n_{x_i}}{N} \qquad ; \qquad \overline{y} = \frac{\sum_{j=1}^{l} y_j n_{y_j}}{N}.$$

$$s_x^2 = \frac{\sum_{i=1}^{k} (x_i - \overline{x})^2 n_{x_i}}{N-1} \qquad ; \qquad s_y^2 = \frac{\sum_{j=1}^{l} (y_j - \overline{y})^2 n_{y_j}}{N-1}.$$

y las desviaciones típicas serían las correspondientes raíces cuadradas de las varianzas.

Hay que indicar que al evaluar las frecuencias marginales se está perdiendo información, ya que se obvian las distribuciones en la otra parte de la variable. Es más, el análisis de ambas distribuciones marginales no proporciona tanta información como la tabla de frecuencias completa.

Ejemplo I–9

(Continuación.) Calculemos las distribuciones marginales del ejemplo anterior. Determinamos las medias y varianzas usando las marcas de clase.

x_i	c_i	n_{x_i}
1.50–1.60	1.55	3
1.60–1.70	1.65	6
1.70–1.80	1.75	7
1.80–1.90	1.85	4
Suma		20

$$\overline{x} = \frac{\sum_{i=1}^{k} c_i \, n_{x_i}}{N} = 1.71 \text{ m}$$

$$\overline{y} = \frac{\sum_{j=1}^{l} c_j \, n_{y_j}}{N} = 74.5 \text{ kg}$$

y_j	c_j	n_{y_j}
60–70	65	6
70–80	75	9
80–90	85	5
Suma		20

$$s_x = \sqrt{\frac{\sum_{i=1}^{k} (c_i - \overline{x})^2 \, n_{x_i}}{N-1}} = 0.10 \text{ m} \qquad ; \qquad s_y = \sqrt{\frac{\sum_{j=1}^{l} (c_j - \overline{y})^2 \, n_{y_j}}{N-1}} = 7.6 \text{ kg}$$

4.1.3. Distribuciones condicionadas

En muchos casos es importante conocer la distribución de la variable x para todos aquellos pares de datos en los que la variable y toma un cierto valor y_j. Es decir, al contrario que en las distribuciones marginales en que no importaba el valor que tomase la otra variable, ahora se fija dicho valor. A este conjunto de valores que puede tomar la variable x para un cierto valor y_j de y se le llama **distribución de x condicionada a $y = y_j$** y las correspondientes frecuencias absolutas se representan por $n(x_i|y = y_j)$, cuyo significado es, entonces, el número de veces que aparece repetido el valor x_i entre aquellos pares de datos que tienen $y = y_j$. De la misma forma se puede definir la **distribución de y condicionada a $x = x_i$**. Los valores de estas frecuencias absolutas condicionadas pueden extraerse directamente de la tabla de doble entrada ya que es claro que

$$n(x_i|y = y_j) = n_{ij} \qquad ; \qquad n(y_j|x = x_i) = n_{ij}.$$

Es decir, la tabla de frecuencias para la distribución de x condicionada a $y = y_j$ sería:

x	$n(x\|y = y_j)$	$f(x\|y = y_j)$
x_1	n_{1j}	f_{1j}
x_2	n_{2j}	f_{2j}
$\vdots$	$\vdots$	$\vdots$
x_i	n_{ij}	f_{ij}
$\vdots$	$\vdots$	$\vdots$
x_k	n_{kj}	f_{kj}
	n_{y_j}	1

Para calcular las frecuencias relativas de x condicionadas a $y = y_j$ habrá que dividir por el número de datos que tienen $y = y_j$, es decir por la frecuencia marginal de y_j (n_{y_j})

$$f(x_i|y = y_j) = \frac{n(x_i|y = y_j)}{n_{y_j}} = \frac{n_{ij}}{n_{y_j}} \qquad ; \qquad f(y_j|x = x_i) = \frac{n(y_j|x = x_i)}{n_{x_i}} = \frac{n_{ij}}{n_{x_i}}.$$

Como es fácil de comprobar, se cumple que

$$\sum_{i=1}^{k} n(x_i|y = y_j) = n_{y_j} \qquad ; \qquad \sum_{j=1}^{l} n(y_j|x = x_i) = n_{x_i},$$

$$\sum_{i=1}^{k} f(x_i|y = y_j) = 1 \qquad ; \qquad \sum_{j=1}^{l} f(y_j|x = x_i) = 1.$$

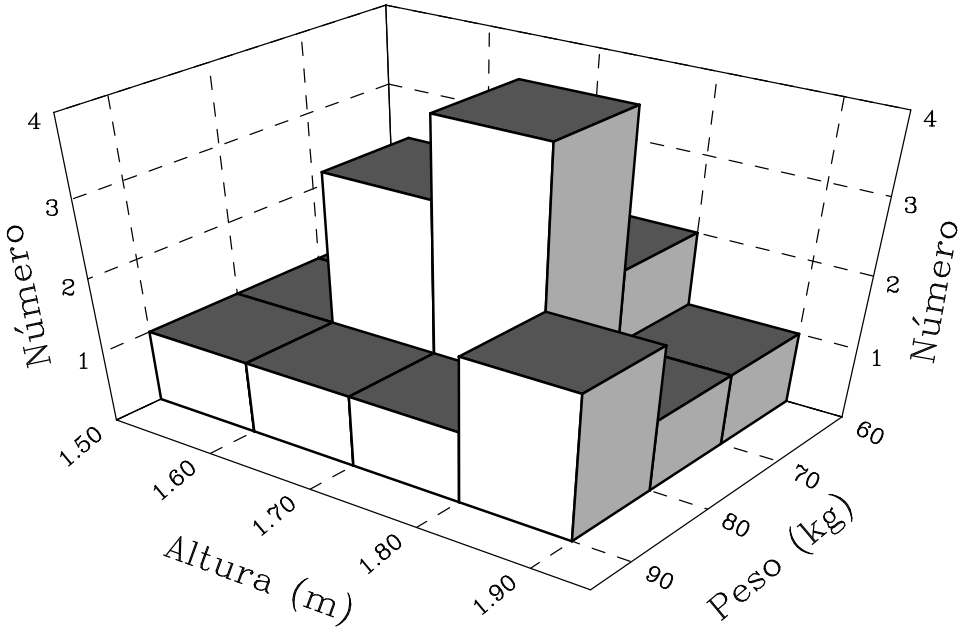


Figura 4.1: *Diagrama tridimensional para la muestra de pesos y alturas del ejemplo I–9.*

Ejemplo I–9

(Continuación.)

Distribuciones condicionadas en el ejemplo anterior. Calculamos la distribución de x condicionada a y_j =(70–80) kg.

x	$n(x\|y = 70\text{–}80)$	$f(x\|y = 70\text{–}80)$
1.50–1.60	1	0.11 (1/9)
1.60–1.70	3	0.33 (3/9)
1.70–1.80	4	0.44 (4/9)
1.80–1.90	1	0.11 (1/9)
Suma	$9 = n_{y_j}$	1

La distribución de y condicionada a x_i =(1.70–1.80) será:

y	$n(y\|x = 1.70\text{–}1.80)$	$f(y\|x = 1.70\text{–}1.80)$
60–70	2	0.29 (2/7)
70–80	4	0.57 (4/7)
80–90	1	0.14 (1/7)
Suma	$7 = n_{x_i}$	1

4.1.4. Representaciones gráficas

Al igual que para las variables unidimensionales, existen diversas formas de representar gráficamente los datos de una muestra bidimensional de forma que se pueda obtener una idea rápida de cómo se distribuyen los valores.

En el caso de variables discretas con repeticiones de valores y de datos agrupados en intervalos, los diagramas más usuales son los **diagramas de barras** e **histogramas tridimensionales**. Para ello se dibuja en perspectiva un plano XY donde se marcan los valores de la variable y se levanta, en el caso del diagrama de barras (para variables discretas), sobre cada par una barra de altura proporcional a la frecuencia (ver Figura 4.1).

El histograma, para variables agrupadas en intervalos, se construye sustituyendo las barras por parale-

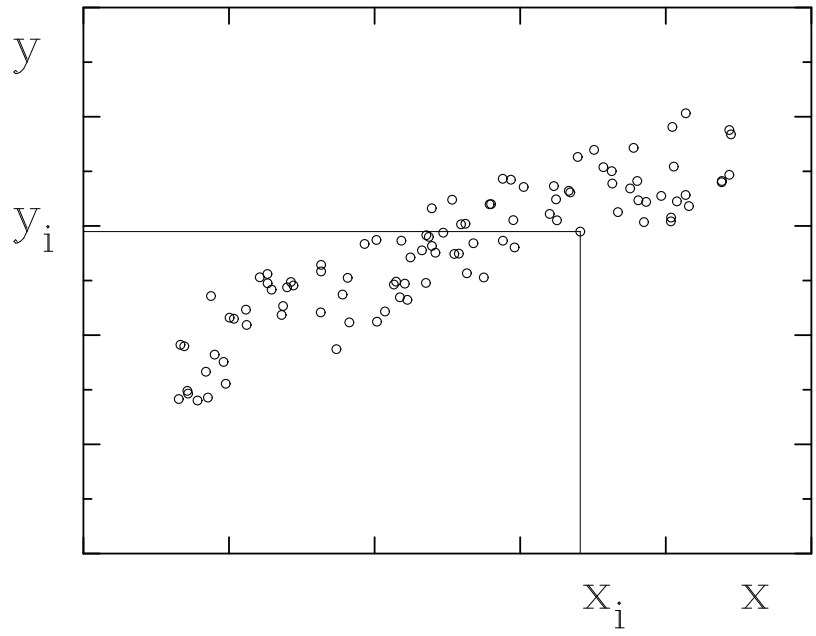


Figura 4.2: *Ejemplo de diagrama de dispersión.*

lepípedos solapados. En general se hace que los volúmenes de los paralelepípedos sean proporcionales a las frecuencias de cada intervalo o, para intervalos de amplitud constante y de forma más sencilla, con alturas proporcionales a las frecuencias.

Cuando no existen apenas valores repetidos y no se hace agrupamiento por intervalos, la representación se hace sobre un **diagrama de dispersión** (ver Figura 4.2). Este diagrama bidimensional se construye dibujando para cada par (x, y) un punto sobre un plano cartesiano. Como se verá posteriormente, este diagrama permite examinar de forma rápida si puede haber alguna relación entre las dos partes de la variable bidimensional.

Tema II

DISTRIBUCIONES DE PROBABILIDAD

Capítulo 5

Leyes de probabilidad

"La vida es una escuela sobre probabilidad."

Walter Bagehot (1826-1877)

El objetivo fundamental de la Estadística es inferir las propiedades de una población a partir de la observación de una muestra, o subconjunto, de ésta. La construcción y estudio de los modelos estadísticos están entonces íntimamente ligados al cálculo de probabilidades, a cuyas bases están dedicados este tema y los tres siguientes.

5.1. Sucesos aleatorios

La teoría de la probabilidad surge para poder estudiar los, llamados, **experimentos aleatorios**. Se dice que un experimento es aleatorio si puede dar lugar a varios resultados sin que se pueda predecir con certeza el resultado concreto. Es decir, al repetir el experimento bajo condiciones similares se obtendrán resultados que, en general, serán diferentes. Un ejemplo de un experimento aleatorio puede ser la tirada de un dado, ya que no se puede predecir el número que aparecerá en su cara superior.

Al conjunto de todos los resultados posibles de un experimento aleatorio se le llama **espacio muestral**, que representaremos por el símbolo S. Por ejemplo, en el lanzamiento del dado, el espacio muestral sería el conjunto $S = \{1, 2, 3, 4, 5, 6\}$. No siempre es posible describir el espacio muestral enumerando sus diferentes elementos. A veces se define por medio de una condición, o regla, que han de cumplir sus elementos (ej. puntos que se sitúan en una circunferencia). Dependiendo del número de resultados posibles del experimento aleatorio, el espacio muestral podrá ser: finito (ej. resultados de la tirada de un dado), infinito numerable (cuando a cada elemento del espacio se le puede hacer corresponder un número entero sin límite, ej. vida en años de un componente electrónico), e infinito no numerable (ej. números reales en el intervalo $0 - 1$).

Se define un **suceso** como un subconjunto A del espacio muestral, es decir es un subconjunto de resultados posibles. Los sucesos más simples son los **sucesos elementales**, que consisten en un único punto del espacio muestral. De forma más exacta se puede definir los sucesos elementales de un experimento aleatorio como aquellos sucesos que verifican: **a)** siempre ocurre alguno de ellos, y **b)** son mutuamente excluyentes. Por ejemplo, obtener un 4 es un suceso elemental del experimento de lanzar un dado. Por otra parte, diremos que un suceso es **compuesto** cuando, al contrario que con los sucesos elementales, puede ser descompuesto en sucesos más simples. Es decir, serían los sucesos construídos a partir de la unión de sucesos elementales. Por ejemplo, en el experimento de lanzar el dado, al suceso compuesto A de obtener un número par le corresponde el siguiente conjunto de puntos del espacio muestral $A = \{2, 4, 6\}$.

Existen dos sucesos particulares especialmente interesantes. El primero es el **suceso imposible** $\emptyset$, definido como el subconjunto vacío del espacio muestral. Es decir, será el suceso que no ocurrirá nunca. Por otra parte, el propio espacio muestral también puede considerarse como un suceso. Será el **suceso seguro** S, que ocurrirá siempre. Cuando un suceso no coincide ni con el suceso imposible ni con el seguro, diremos que el suceso es **probable**.

Puesto que los sucesos aleatorios se definen como conjuntos, podemos definir entre ellos las mismas operaciones que se realizan sobre los conjuntos abstractos. Se definen así:

- La **unión** de dos sucesos A y B como el suceso, representado por $A \cup B$, que ocurrirá siempre que ocurra el suceso A o el suceso B.
- La **intersección** de dos sucesos A y B como el suceso, representado por $A \cap B$, que ocurrirá siempre que ocurran simultáneamente los sucesos A y B.
- Dado un suceso A, llamaremos suceso **complementario** de A al suceso A' que ocurrirá siempre que no ocurra A. Evidentemente, se cumplen las propiedades

$$A \cup A' = S \quad ; \quad A \cap A' = \emptyset \quad ; \quad S' = \emptyset \quad ; \quad \emptyset' = S.$$

- Diremos que dos sucesos A y B son **incompatibles**, o mutuamente excluyentes, si nunca pueden ocurrir a la vez. Es decir cuando

$$A \cap B = \emptyset.$$

- Dados dos sucesos A y B, diremos que A está contenido en B, y lo representaremos por $A \subset B$, cuando se cumpla que siempre que ocurre A ocurre a la vez B. Es evidente que para cualquier suceso A se cumple

$$\emptyset \subset A \subset S.$$

Además, la unión e intersección de sucesos cumplirán las conocidas propiedades conmutativa, asociativa y distributiva[1]. Podemos afirmar además que la clase formada por los sucesos de un experimento aleatorio tiene estructura de álgebra de Boole.

Para facilitar el estudio de los sucesos se pueden utilizar los conocidos **diagramas de Venn** (Figura 5.1), donde el espacio muestral se representa por un rectángulo, y cada suceso como un recinto incluído en él.

[1] En álgebra abstracta, un álgebra booleana es una estructura algebraica (una colección de elementos y operaciones que obedecen unos axiomas definidos) que engloban las propiedades esenciales de las operaciones lógicas y de conjuntos. Específicamente, se encarga de las operaciones de conjuntos denominadas intersección, unión y complemento; y las operaciones lógicas AND, OR y NOT.
— Propiedad conmutativa: $A \cup B = B \cup A$; $A \cap B = B \cap A$
— Propiedad asociativa: $A \cup (B \cup C) = (A \cup B) \cup C$; $A \cap (B \cap C) = (A \cap B) \cap C$
— Propiedad distributiva: $A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$; $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
— Ley de Morgan #1: $(A \cup B)' = A' \cap B'$: lo opuesto a que al menos uno de los eventos ocurra es que no ocurra ninguno de ellos.
— Ley de Morgan #2: $(A \cap B)' = A' \cup B'$: ambos eventos no ocurren simultáneamente si al menos uno de ellos no ocurre.

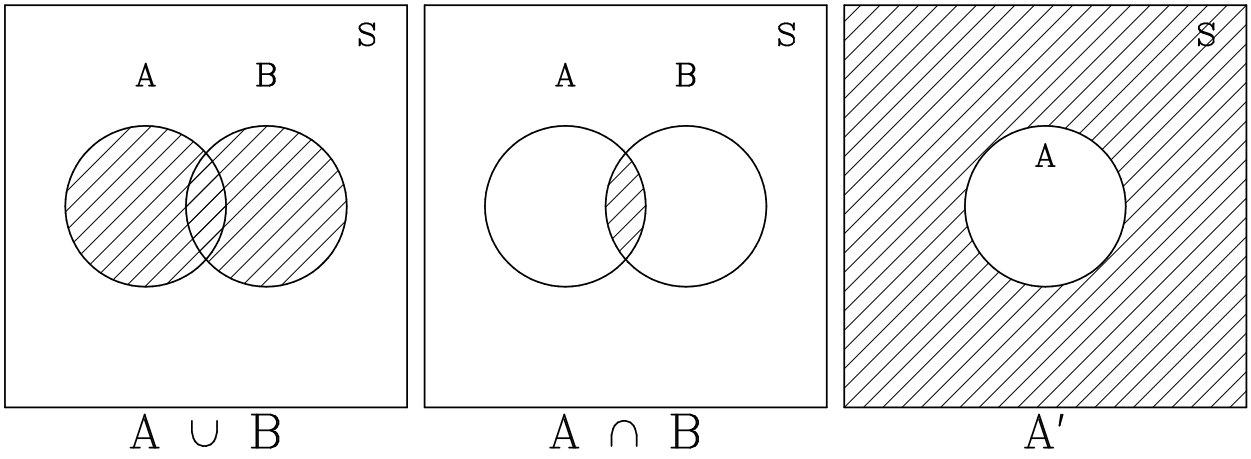


Figura 5.1: *Diagramas de Venn: este tipo de diagramas son ilustraciones utilizadas en el campo de las matemáticas conocido como Teoría de Conjuntos. Se emplean para mostrar las relaciones matemáticas o lógicas entre diferentes conjuntos de cosas.*

5.2. Definición y propiedades de la probabilidad

5.2.1. Concepto clásico de probabilidad

El concepto de probabilidad surge para medir la certeza o incertidumbre de un suceso de un experimento aleatorio. Históricamente, la teoría de la probabilidad se desarrolló en primer lugar para encontrar estrategias óptimas para los juegos de azar, aunque, rápidamente, su utilidad desbordó este campo. Evidentemente, la forma más directa de saber la posibilidad de que ocurra un suceso en un experimento aleatorio es repetir dicho experimento muchas veces. De esta forma, supongamos que se repita n veces el experimento y llamemos n_A, o frecuencia absoluta de A, al número de veces en que ocurre el suceso A. Se puede definir entonces la **probabilidad** $P(A)$ **del suceso** A como

$$P(A) \equiv \lim_{n\to\infty} \frac{n_A}{n} = \lim_{n\to\infty} \frac{\text{frecuencia absoluta del suceso } A}{\text{número de veces que se repite el experimento}}, \tag{5.1}$$

es decir, $P(A)$ es el límite cuando n tiende a infinito de la frecuencia relativa del suceso A. Puede observarse que si el suceso ocurre siempre $n_A = n$ y $P(A) = 1$, y, al contrario, si el suceso no ocurre nunca, su probabilidad $P(A) = 0$. De esta forma, la probabilidad de un suceso estará comprendida entre 0 y 1 ($0 \leq P(A) \leq 1$), y el suceso será tanto más probable cuanto más se acerque a 1 su probabilidad.

Ejemplo II–1

El lanzamiento de la moneda al aire es clásico. La probabilidad de obtener cara o cruz es $P(A) = 1/2$. En 1900 el estadístico Pearson realizó el experimento con un número total de lanzamientos de 24000 (tardó unas 40 horas). Obtuvo un resultado de 12012 caras (y 11988 cruces). Esto significa $P(A) = 12012/24000 = 0.5005$ que es un valor muy próximo a la probabilidad teórica.

La definición anterior implica, evidentemente, que hay que repetir un gran número de veces el experimento para calcular la probabilidad de un suceso. Afortunadamente, el cálculo de la probabilidad se puede simplificar mucho en el caso en que todos los sucesos elementales sean equiprobables (es decir, sus frecuencias sean iguales cuando el experimento se repite un gran número de veces). En este caso, la probabilidad de un suceso se puede establecer a partir de la definición, introducida por Laplace, según la cual $P(A)$ es el cociente entre el número a de casos favorables al suceso A (o número de sucesos elementales en que se da A) y el número N

de casos posibles (o número de sucesos elementales del espacio muestral)

$$P(A) = \frac{a}{N} = \frac{\text{casos favorables}}{\text{casos posibles}}. \tag{5.2}$$

En particular, en este caso de sucesos equiprobables, la probabilidad de un suceso elemental será: $P(A) = \frac{1}{N}$.

Ejemplo II–2

El lanzamiento de un dado no trucado supone que los sucesos son equiprobables. Así la probabilidad de obtener un 4 al lanzar un dado será 1/6. Como ejemplo de un suceso compuesto, la probabilidad de obtener un número par en dicho lanzamiento será $P(A) = 3/6 = 1/2$, ya que hay tres casos favorables $\{2, 4, 6\}$ de seis posibles $\{1, 2, 3, 4, 5, 6\}$.

A veces sucesos que parecen equiprobables no lo son. Por ejemplo si se estudia una ruleta en particular durante el tiempo suficiente, se comprueba que no todos los números son equiprobables. Esto es debido a pequeñas imperfecciones en la propia ruleta. Por esta causa los casinos no permiten la entrada a los jugadores que anotan sistemáticamente los resultados de sus ruletas ya que éstos jugarían con ventaja si conocieran bien su comportamiento.

5.2.2. Definición axiomática de la probabilidad

Las definiciones anteriores presentan serias dificultades: o bien se necesita repetir el experimento un número muy grande de veces, o se ha de estar seguro que todos los sucesos elementales son equiprobables (lo cual no siempre es obvio). Por estos motivos se utiliza la siguiente definición, más correcta, de probabilidad:

Dado un experimento aleatorio con un espacio muestral S y representando por A a un suceso, o subconjunto, cualquiera del espacio muestral, se define la **probabilidad** $P(A)$ como una función real que hace corresponder a cada A un número real de forma que se cumplen los tres axiomas siguientes:

1. Para cada suceso A

$$P(A) \geq 0, \tag{5.3}$$

es decir, la probabilidad de cualquier suceso es mayor o igual que cero.

2. Para el suceso seguro S

$$P(S) = 1. \tag{5.4}$$

3. Dados dos sucesos A y B incompatibles ($A \cap B = \emptyset$)

$$P(A \cup B) = P(A) + P(B). \tag{5.5}$$

Es decir, la probabilidad del suceso unión de dos incompatibles es la suma de las probabilidades de ambos sucesos. Esto se puede generalizar a cualquier número de sucesos incompatibles

$$P(A_1 \cup A_2 \cup \ldots \cup A_n \cup \ldots) = P(A_1) + P(A_2) + \ldots + P(A_n) + \ldots$$

Estos axiomas constituyen la base sobre la que se puede construir toda la teoría del cálculo de probabilidades. Nótese que las propiedades anteriores son coherentes con la definición de la probabilidad basada en las frecuencias relativas de un gran número de experimentos.

5.2.3. Propiedades de la probabilidad

A partir de los axiomas anteriores se pueden deducir algunas propiedades importantes de la probabilidad. Estas propiedades van a ser útiles para calcular la probabilidad de sucesos a partir de las probabilidades conocidas de otros sucesos más sencillos, simplificando así el cálculo. Hay que indicar además que estas propiedades son consistentes con las propiedades de las frecuencias relativas.

- Si A' es el suceso complementario de A, entonces

$$P(A') = 1 - P(A). \tag{5.6}$$

Efectivamente, puesto que $A \cup A' = S$ y teniendo en cuenta que A y su complementario son incompatibles ($A \cap A' = \emptyset$)

$$P(A \cup A') = P(S) \qquad \Rightarrow \qquad P(A) + P(A') = 1$$

Ejemplo II–3

En el caso del lanzamiento de un dado,
A: obtener un 6 $\quad P(A) = 1/6$
A': que no salga un 6 $\quad P(A') = 1 - P(A) = 1 - (1/6) = 5/6$.
Lo que ya sabíamos ya que éste es el cociente entre casos favorables (5) y posibles (6).

- La probabilidad del suceso imposible es cero

$$P(\emptyset) = 0. \tag{5.7}$$

Se demuestra a partir de la propiedad anterior y teniendo en cuenta que el suceso imposible es el complementario del suceso seguro ($\emptyset' = S$)

$$P(\emptyset) = 1 - P(S) = 1 - 1 = 0.$$

- A partir del primer axioma y la propiedad anterior, se puede ver que para cualquier suceso A

$$0 \leq P(A) \leq 1. \tag{5.8}$$

- Si un suceso A está contenido en otro B, se cumple (por definición de un suceso contenido en otro)

$$A \subset B \qquad \Rightarrow \qquad P(A) \leq P(B) \tag{5.9}$$

- Si A y B son dos sucesos cualesquiera, siempre se cumple

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{5.10}$$

En el caso particular de que los sucesos fuesen incompatibles ($A \cap B = \emptyset$) esta propiedad se reduciría al tercer axioma de la probabilidad.

Ejemplo II–4

Calcular la probabilidad de obtener o un número par o un número mayor que 3 en el lanzamiento de un dado.
A : obtener un número par $\quad$ P(A) = 3/6 = 1/2 $\quad$ {2,4,6}
B : obtener un número mayor que 3 $\quad$ P(B) = 3/6 = 1/2 $\quad$ {4,5,6}

$$P(A \cap B) = 2/6 \qquad ; \qquad (\{4,6\}\ \textit{es el espacio muestral})$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = \frac{1}{2} + \frac{1}{2} - \frac{2}{6} = \frac{4}{6} = \frac{2}{3}$$

que era lo esperado ya que el espacio muestral es en este caso $\{2, 4, 5, 6\}$, es decir, $4/6 = 2/3$.

Para demostrar esta propiedad hacemos uso del diagrama de Venn (Figura 5.2), en el cual es fácil de comprobar que se verifica

$$A = (A \cap S) = (A \cap (B \cup B') = (A \cap B) \cup (A \cap B').$$

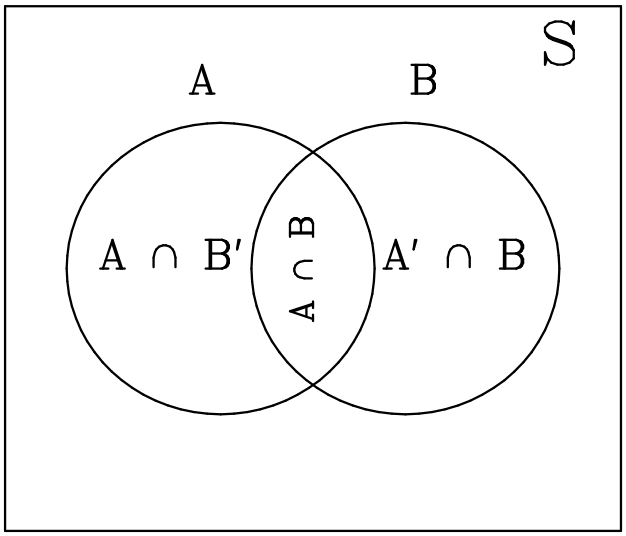


Figura 5.2: *Diagrama de Venn representando la probabilidad de un suceso unión de dos sucesos no incompatibles.*

De la misma forma

$$B = (A \cap B) \cup (A' \cap B).$$

Por tanto

$$A \cup B = (A \cap B) \cup (A \cap B') \cup (A' \cap B).$$

Puesto que en cada una de las expresiones anteriores, los sucesos del término de la derecha son incompatibles entre sí, usando el tercer axioma podemos escribir

$$P(A) = P(A \cap B) + P(A \cap B') \quad \Rightarrow \quad P(A \cap B') = P(A) - P(A \cap B)$$

$$P(B) = P(A \cap B) + P(A' \cap B) \quad \Rightarrow \quad P(A' \cap B) = P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A \cap B) + P(A \cap B') + P(A' \cap B)$$

Sustituyendo las dos primeras expresiones en la tercera

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

$$= P(A) + P(B) - P(A \cap B),$$

como queríamos demostrar.

La propiedad anterior se puede generalizar a la unión de más de dos sucesos. En el caso de tres sucesos cualesquiera tendríamos

$$P(A \cup B \cup C) =$$

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

5.3. Probabilidad condicionada

5.3.1. Definición de probabilidad condicionada

En muchos casos interesa conocer la probabilidad de un suceso A en el caso de que se haya cumplido otro suceso B. A esta probabilidad de que se cumpla A bajo la condición de que se cumpla B se le llama **probabilidad de A condicionada a B**, y se denota por $P(A|B)$. La definición matemática de la probabilidad condicionada es

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{5.11}$$

Como es lógico, esta definición sólo tiene sentido si $P(B) > 0$. El significado de la definición anterior se ve claro utilizando un diagrama de Venn (Figura 5.2; es una versión geométrica de casos favorables entre casos posibles). Al calcular la probabilidad condicionada hemos sustituido el espacio muestral S por el suceso B, de forma que, haciendo corresponder probabilidades a áreas en el espacio muestral, $P(A|B)$ será la fracción del nuevo espacio muestral B en que ocurre A.

Vamos a comprobar que la probabilidad condicionada cumple los tres axiomas de la definición general de probabilidad.

1. Es evidente que se satisface el primer axioma puesto que el cociente de dos números no negativos es un número no negativo
$$P(A|B) \geq 0.$$

2. La probabilidad condicionada del suceso seguro es también la unidad
$$P(S|B) = \frac{P(S \cap B)}{P(B)} = \frac{P(B)}{P(B)} = 1.$$

3. Dados dos sucesos A_1 y A_2 incompatibles ($A_1 \cap A_2 = \emptyset$)
$$P(A_1 \cup A_2|B) = \frac{P((A_1 \cup A_2) \cap B)}{P(B)} = \frac{P((A_1 \cap B) \cup (A_2 \cap B))}{P(B)}.$$
Los dos sucesos del numerador son incompatibles ya que
$$(A_1 \cap B) \cap (A_2 \cap B) = (A_1 \cap A_2) \cap B = \emptyset \cap B = \emptyset,$$
de forma que, aplicando el tercer axioma para la probabilidad
$$P(A_1 \cup A_2|B) = \frac{P(A_1 \cap B) + P(A_2 \cap B)}{P(B)} = \frac{P(A_1 \cap B)}{P(B)} + \frac{P(A_2 \cap B)}{P(B)}$$
$$\Rightarrow \qquad P(A_1 \cup A_2|B) = P(A_1|B) + P(A_2|B),$$
como queríamos demostrar.

Ejemplo II–5

En el caso del lanzamiento de un dado,

A: obtener un par	$\{2,4,6\}$	$P(A)=1/2$
B: idem un número mayor que 3	$\{4,5,6\}$	$P(B)=1/2$

$$P(A\cap B)=2/6 \quad (\textit{ejemplo anterior}) \quad ; \quad P(A|B)=\frac{P(A\cap B)}{P(B)}=\frac{2/6}{1/2}=\frac{4}{6}=\frac{2}{3}$$

Que coincide con el cociente entre casos favorables 2 ($\{4,6\}$) y casos posibles 3 ($\{4,5,6\}$).

5.3.2. Sucesos dependientes e independientes

La definición (5.11) de la probabilidad condicionada permite calcular la probabilidad de la intersección de dos sucesos (todavía no sabíamos cómo), es decir, la probabilidad de que se den ambos sucesos A y B a la vez

$$P(A\cap B)=P(A|B)P(B) \tag{5.12}$$

o

$$P(A\cap B)=P(B|A)P(A). \tag{5.13}$$

De esta forma, la probabilidad de que tanto A como B ocurran es igual a la probabilidad de que A ocurra dado que B haya ocurrido multiplicado por la probabilidad de que B ocurra. Esto se puede generalizar a la intersección de más sucesos. En el caso particular de 3 sucesos

$$P(A\cap B\cap C)=P(A|B\cap C)P(B|C)P(C).$$

Un caso importante es cuando se cumple

$$P(A|B)=P(A) \tag{5.14}$$

En este caso, la probabilidad de que A ocurra no está afectada por la ocurrencia o no ocurrencia de B y se dice que los dos sucesos son **independientes**. Aplicando (5.12) es fácil ver que en este caso se cumple

$$P(A\cap B)=P(A)P(B). \tag{5.15}$$

Es decir, la probabilidad de la intersección de dos sucesos independientes (en otras palabras, la probabilidad de que se den ambos sucesos) es el producto de sus probabilidades. Esta última relación se toma usualmente como condición necesaria y suficiente para la existencia de independencia. El concepto de independencia se puede generalizar a una familia de n sucesos. Se dice que son mutuamente independientes cuando cualquier pareja de sucesos es independiente y la probabilidad de la intersección de cualquier número de sucesos independientes es el producto de sus probabilidades. En el caso de tres sucesos independientes

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

Cuando no se cumple la relación (5.14) hay que utilizar la expresión general (5.12) para calcular la probabilidad de la intersección. En este caso se dice que los sucesos son **dependientes**, es decir, la probabilidad de que ocurra uno de ellos depende de que haya ocurrido o no el otro.

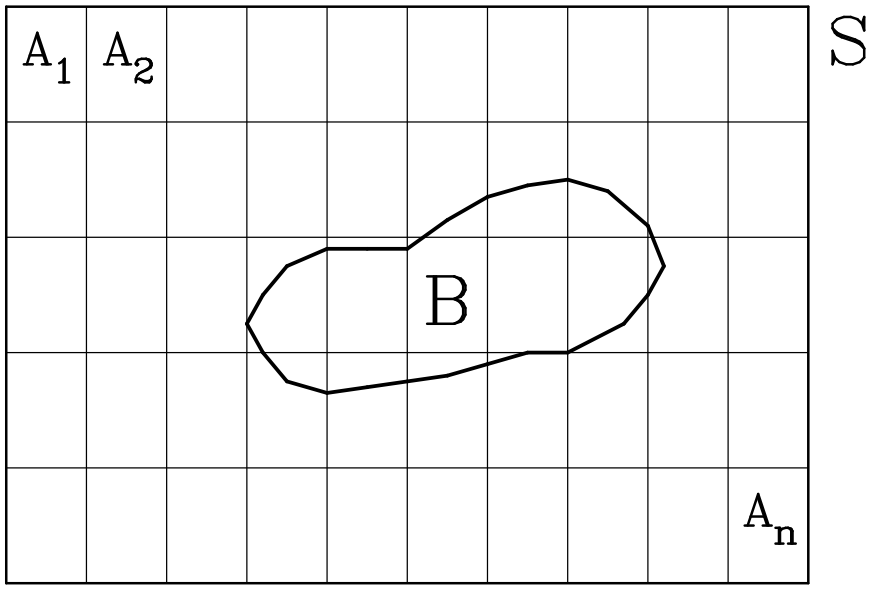


Figura 5.3: *Diagrama de Venn representando el Teorema de la Probabilidad Total.*

Ejemplo II–6

Tenemos en una urna 4 bolas blancas y 6 bolas negras. Si extraemos 2 bolas sucesivamente, calcular la probabilidad de que las 2 sean blancas. Consideremos dos casos:

a) Se reemplaza la 1^a despues de sacarla.

Entonces los dos sucesos son independientes: la naturaleza de la 2^a bola no está condicionada por la naturaleza de la 1^a.

A: bola blanca en la primera extracción

B: idem en la segunda

$$P(A \cap B) = P(A)\ P(B) = \frac{4}{10} \times \frac{4}{10} = \frac{16}{100} = 0.16$$

b) No se reemplaza la 1^a despues de sacarla.

Entonces los dos sucesos ya no son independientes y el color de la 2^a bola sí está condicionada por el color de la 1^a.

$$P(A \cap B) = P(A)\ P(B|A) = \frac{4}{10} \times \frac{3}{9} = \frac{12}{90} = 0.13$$

Es importante no confundir sucesos incompatibles ($A \cap B = \emptyset$) con sucesos independientes (la probabilidad de que ocurra el suceso A no está afectada por la ocurrencia o no del suceso B).

5.3.3. Teorema de la probabilidad total

Sea un conjunto de sucesos $A_i, i = 1, \ldots, n$ tales la unión de todos ellos es el suceso seguro y además son incompatibles entre sí. Es decir

$$\bigcup_{i=1}^{n} A_i = S \qquad ; \qquad A_i \cap A_j = \emptyset \quad \text{para} \quad i \neq j.$$

Este conjunto de sucesos recibe el nombre de **conjunto completo de sucesos** y se dice que constituye una partición del espacio muestral. Supongamos además que, para todo i, $P(A_i) > 0$. Entonces, el **teorema de**

la probabilidad total establece que la probabilidad de cualquier suceso B se puede calcular como

$$P(B) = \sum_{i=1}^{n} P(A_i)P(B|A_i), \tag{5.16}$$

es decir, la probabilidad de que ocurra B es la suma de las probabilidades de los sucesos A_i por las probabilidades de B condicionadas a cada A_i.

Para demostrar el teorema aplicamos las condiciones del conjunto completo de sucesos y expresamos el suceso B como

$$B = B \cap S = B \cap (\bigcup_{i=1}^{n} A_i) = \bigcup_{i=1}^{n} (B \cap A_i).$$

Al ser los sucesos A_i incompatibles también lo son los diferentes $(B \cap A_i)$, de forma que la probabilidad de B, utilizando (5.12), se puede expresar

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i) = \sum_{i=1}^{n} P(A_i)P(B|A_i),$$

como queríamos demostrar.

Ejemplo II–7

Supongamos que en unas elecciones las probabilidades de que ganen tres partidos A_1, A_2 y A_3 son 0.5, 0.3 y 0.2 respectivamente. Si ganara A_1, la probabilidad de que suban los impuestos es 0.8, mientras que en los casos en que salgan elegidos A_2 y A_3 son 0.2 y 0.5 respectivamente. ¿Cual es la probabilidad de que suban los impuestos?.

$$P(A_1) = 0.5 \qquad P(A_2) = 0.3 \qquad P(A_3) = 0.2$$

sea B subida de impuestos,

$$P(B|A_1) = 0.8 \qquad P(B|A_2) = 0.2 \qquad P(B|A_3) = 0.5$$

Por el teorema de la probabilidad total,

$$P(B) = P(A_1)\ P(B|A_1) + P(A_2)\ P(B|A_2) + P(A_3)\ P(B|A_3) =$$

$$P(B) = 0.5 \times 0.8 + 0.3 \times 0.2 + 0.2 \times 0.5 = 0.56$$

5.3.4. Teorema de Bayes

Supongamos que tenemos un conjunto completo de sucesos $A_i, i = 1, \ldots, n$ y un suceso B cualquiera del espacio muestral. A veces es necesario conocer la probabilidad de uno de los sucesos A_j condicionada a que haya ocurrido B. Esto se puede hacer por el **teorema de Bayes**, que establece

$$P(A_j|B) = \frac{P(A_j)P(B|A_j)}{\sum_{i=1}^{n} P(A_i)P(B|A_i)}. \tag{5.17}$$

El teorema es útil cuando, conociéndose que se cumple un cierto suceso B, queremos conocer la probabilidad de que la causa que lo haya producido sea el suceso A_j.

La demostración del teorema es sencilla, partiendo de la definición (5.11) y, aplicando la relación (5.12), podemos expresar

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{P(B)}.$$

Sustituyendo ahora $P(B)$ por su expresión según el teorema de la probabilidad total (5.16) llegamos a la expresión que queremos demostrar.

Ejemplo II–7

(Continuación.)
Continuando el ejemplo 5–7, si se sabe que han subido los impuestos ¿cual es la probabilidad de que haya ganado el partido A_1?

$$P(A_1|B) = \frac{P(A_1)\ P(B|A_1)}{\sum P(A_i)\ P(B|A_i)} = \frac{0.5 \times 0.8}{0.56} = 0.71$$

El sumatorio del denominador es simplemente la probabilidad de que se de el suceso B: $P(B) = 0.5 \times 0.8 + 0.3 \times 0.2 + 0.2 \times 0.5 = 0.56$.

Ejemplo II–8

Se dispone de dos urnas que contienen un 70 % de bolas blancas y 30 % de negras la primera y 30 % de blancas y 70 % de negras la segunda. Seleccionamos una de las urnas al azar y se extraen 10 bolas con reemplazamiento resultando B={bnbbbbnbbb} siendo b: bola blanca y n: bola negra. Determinar la probabilidad de que esta muestra proceda de la urna primera.
Como la urna se selecciona al azar

$$P(U_1) = P(U_2) = 1/2.$$

Como la extracción con reemplazamiento de 10 bolas son sucesos independientes

$$\begin{aligned} P(b|U_1) = 0.7 \quad ; \qquad & P(n|U_1) = 0.3 \\ P(b|U_2) = 0.3 \quad ; \qquad & P(n|U_2) = 0.7 \end{aligned}$$

luego

$$\begin{aligned} P(B|U_1) = P(bnbbbbnbbb|U_1) = P(b|U_1) \times P(n|U_1) \times \ldots P(b|U_1) = 0.7^8 \times 0.3^2 \\ P(B|U_2) = P(bnbbbbnbbb|U_2) = P(b|U_2) \times P(n|U_2) \times \ldots P(b|U_2) = 0.3^8 \times 0.7^2 \end{aligned}$$

Entonces la probabilidad que nos piden puede determinarse con la ayuda del teorema de Bayes

$$P(U_1|B) = \frac{P(B|U_1)P(U_1)}{P(B|U_1)P(U_1) + P(B|U_2)P(U_2)} =$$

$$= \frac{0.7^8 \times 0.3^2 \times 0.5}{0.7^8 \times 0.3^2 \times 0.5 + 0.3^8 \times 0.7^2 \times 0.5}$$

$$\Rightarrow P(U_1|B) = \frac{0.7^6}{0.7^6 + 0.3^6} = 0.994 \rightarrow 99.4\,\%,$$

resultado lógico, puesto que es la urna con mayor proporción de bolas blancas.

Ejemplo II–9

El problema de las tres puertas. (Daniel Peña, *Estadística Modelos y Métodos*, p. 111).
Un concursante debe elegir entre tres puertas, detrás de una de las cuales se encuentra el premio. Hecha la elección y antes de abrir la puerta, el presentador le muestra que en una de las dos puertas no escogidas no está el premio y le da la posibilidad de reconsiderar su decisión. ¿Qué debe hacer el concursante?
Definamos los dos sucesos siguientes:
A_i = el concursante elige inicialmente la puerta i; i=1,2,3
R_i = el premio realmente está en la puerta i; i=1,2,3
El espacio muestral está formado por 9 sucesos $(A_i \cap R_j)$, cada uno de ellos con probabilidad 1/9. Si, por ejemplo, se da A_1, la probabilidad de ganar es:

$$P(R_1|A_1) = \frac{P(R_1 \cap A_1)}{P(A_1)} = \frac{1/9}{1/3} = \frac{3}{9} = \frac{1}{3}$$

Supongamos que el concursante ha elegido la puerta A_1. Sea:
B_j = el presentador abre la puerta j y muestra que no contiene el premio (con $j = 2$ ó 3).
Según lo enunciado el espacio muestral está formado por los cuatro sucesos $\{B_2 \cap R_1,\ B_2 \cap R_3,\ B_3 \cap R_1,\ B_3 \cap R_2\}$. Podemos representar gráficamente las probabilidades de los sucesos elementales $\{B_j \cap R_i\}$ cuando se ha elegido la puerta 1 (ocurre A_1) de la siguiente manera:

Ejemplo II–9

(Continuación.)
(Ha ocurrido A_1)

	R_1	R_2	R_3
B_1	—	—	—
B_2	$P(B_2 \cap R_1) = 1/6$	—	$P(B_2 \cap R_3) = 1/3$
B_3	$P(B_3 \cap R_1) = 1/6$	$P(B_3 \cap R_2) = 1/3$	—

Veamos cómo se han calculado las probabilidades indicadas. Inicialmente el coche se ubica al azar en cualquiera de las tres puertas, es decir,

$$P(R_1) = P(R_2) = P(R_3) = 1/3$$

Cuando el premio está en la puerta elegida, R_1, tan probable es que el presentador muestre la puerta 2 como la 3, luego

$$P(B_2|R_1) = P(B_3|R_1) = 1/2,$$

y por lo tanto,

$$P(B_2 \cap R_1) = P(B_2|R_1)P(R_1) = \frac{1}{2} \times \frac{1}{3} = \frac{1}{6}$$

y lo mismo para $P(B_3 \cap R_1)$.
Cuando el concursante elige A_1 y el premio está en la puerta 2 (R_2) el presentador debe necesariamente mostrar la puerta 3 (B_3),

$$P(B_3|R_2) = 1 \quad ; \quad P(B_3 \cap R_2) = P(B_3|R_2)P(R_2) = 1 \times \frac{1}{3} = \frac{1}{3}$$

Análogamente, cuando el concursante elige A_1 y el premio está en la puerta 3 (R_3) el presentador debe necesariamente mostrar la puerta 2 (B_2),

$$P(B_2|R_3) = 1 \quad ; \quad P(B_2 \cap R_3) = P(B_2|R_3)P(R_3) = 1 \times \frac{1}{3} = \frac{1}{3}$$

Entonces la probabilidad de ganar que tienen los concursantes que no cambian su elección es 1/3 (la que tenían). Se comprueba viendo que tras elegir la puerta 1 (A_1) y abriendo el presentador la j (j=2,3),

$$P(R_1|B_j) = \frac{P(R_1)P(B_j|R_1)}{\sum P(R_i)P(B_j|R_i)} = \frac{\frac{1}{3} \times \frac{1}{2}}{\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1} = \frac{1}{3}$$

La probabilidad de ganar que tienen los concursantes que si cambian su elección es igual a la probabilidad de que el premio esté en la puerta que no muestra el presentador. Suponiendo que muestra la 3 (B_3),

$$P(R_2|B_3) = \frac{P(R_2)P(B_3|R_2)}{\sum P(R_i)P(B_3|R_i)} = \frac{\frac{1}{3} \times 1}{\frac{1}{3} \times \frac{1}{2} + \frac{1}{3} \times 1} = \frac{2}{3}$$

Este resultado es análogo si muestra la puerta 2, obteniéndose en ese caso $P(R_3|B_2) = 2/3$.

La razón por la que resulta rentable o conveniente cambiar de puerta es que el suceso B_j (presentador abre la puerta j) no es independiente de los sucesos R_i (el premio está en la puerta i), es decir *el suceso B_j da información sobre los R_i*. En efecto, $P(B_2) = P(B_3) = 1/2$ y $P(R_1) = P(R_2) = P(R_3) = 1/3$ pero en general $P(B_j \cap R_i) \neq 1/6$. Cuando se da A_1 los sucesos R_1 y B_j ($j = 2, 3$) sí son independientes ya que $P(R_1 \cap B_2) = P(R_1 \cap B_3) = 1/6$ (el presentador puede abrir las puertas 2 ó 3 indistintamente es, pues el premio está en la 1). Pero los sucesos R_i ($i = 2, 3$) y B_j ($j = 2, 3$) son dependientes (el presentador sólo puede mostrar la puerta 2/3 si el premio está en la 3/2). Esta dependencia conduce a que convenga reconsiderar la decisión y cambiar de puerta siempre. Si se juega muchas veces a la larga se gana 2/3 de las veces si se cambia de puerta y sólo 1/3 si se permanece en la primera elección.

5.4. Análisis combinatorio

Un caso especialmente interesante en los problemas de probabilidad es cuando todos los sucesos elementales son igualmente probables. Ya hemos visto que, en este caso, la probabilidad de un suceso elemental es $1/n$, donde n es el número de puntos del espacio muestral, o número de sucesos elementales en que se puede descomponer. Efectivamente, como el suceso seguro S se puede descomponer en los diferentes sucesos elementales A_i y todos estos tienen la misma probabilidad k

$$1 = P(S) = P(\bigcup_{i=1}^{n} A_i) = \sum_{i=1}^{n} P(A_i) = \sum_{i=1}^{n} k = kn$$

$$\Rightarrow \qquad P(A_i) = k = \frac{1}{n}$$

Una vez conocidas las probabilidades de los sucesos elementales de esta forma, las probabilidades de los sucesos compuestos se pueden calcular utilizando las propiedades de la probabilidad. El problema se reduce entonces a calcular n, o número de puntos del espacio muestral.

Una primera herramienta muy útil es el **regla de la multiplicación**, la cual establece que si una operación puede realizarse de n_1 formas y, por cada una de éstas, una segunda operación puede llevarse a cabo de n_2 formas, entonces las dos operaciones pueden realizarse juntas en $n_1 n_2$ formas (número de puntos del espacio muestral).

Para calcular n en el caso general se ha desarrollado el **análisis combinatorio**, el cual constituye una herramienta indispensable para estudiar los experimentos aleatorios. A continuación se ven sus principales conceptos y expresiones.

5.4.1. Variaciones

Dado un conjunto de m elementos, se llaman **variaciones de m elementos tomados de n en n** (con $n \leq m$) a todos los subconjuntos de n elementos que se pueden formar del conjunto original, con la condición de que dos subconjuntos se consideran distintos cuando difieren en algún elemento o en el orden de colocación de ellos. El número de variaciones se representa por $V_{m,n}$ y se calcula por

$$V_{m,n} = m(m-1)(m-2)\ldots(m-n+1). \tag{5.18}$$

Usando la definición de factorial: $n! = 1 \times 2 \times \ldots \times n$, se puede escribir la expresión anterior como

$$V_{m,n} = \frac{m!}{(m-n)!}, \tag{5.19}$$

(donde conviene recordar que el factorial del número cero es, por definición, igual a la unidad, $0! \equiv 1$.)

Por otra parte, se llaman **variaciones con repetición de m elementos tomados de n en n** a las variaciones vistas anteriormente con la condición adicional de que un elemento puede aparecer repetido en el mismo subconjunto cualquier número de veces. Como en las variaciones normales, los subconjuntos son distintos si tienen diferentes elementos o diferente orden de colocación de estos. Su número se representa por V_m^n y es

$$V_m^n = m^n. \tag{5.20}$$

Ejemplo II–10

Dados los elementos a, b, c calculamos:
Variaciones de 3 elementos tomados de 2 en 2:

$$V_{3,2} \rightarrow \begin{matrix} ab & ac \\ ba & bc \\ ca & cb \end{matrix} \qquad V_{3,2} = \frac{m!}{(m-n)!} = \frac{3!}{1!} = 6$$

Variaciones con repetición de 3 elementos tomados de 2 en 2:

$$V_3^2 \rightarrow \begin{matrix} aa & ab & ac \\ ba & bb & bc \\ ca & cb & cc \end{matrix} \qquad V_3^2 = m^n = 3^2 = 9$$

5.4.2. Permutaciones

Las **permutaciones de n elementos** son el caso particular de las variaciones de m elementos tomados de n en n en que m es igual a n. Es decir, representan las diferentes formas de ordenar n elementos. Su número se representa por P_n y se calcula por

$$P_n = V_{n,n} = n(n-1)(n-2)\ldots 1 = n! \tag{5.21}$$

Para que esto sea consistente con la definición (5.19) de las variaciones, se toma por convenio que $0! = 1$.

Por otro lado, dado un conjunto de m elementos, se denominan **permutaciones con repetición** a los distintos subconjuntos de tamaño n que se pueden formar con los m elementos y en los que en cada subconjunto cada elemento aparece repetido $n_1, n_2, \ldots, n_m$ veces, con

$$n_1 + n_2 + \ldots + n_m = n$$

Por ejemplo, dado el conjunto *aabbbc* son permutaciones con repetición de él las siguientes: *abbcab*, *bcabab*, etc. El número de permutaciones con repetición se representa por $P_n^{n_1,n_2,\ldots,n_m}$ y se evalúa por

$$P_n^{n_1,n_2,\ldots,n_m} = \frac{n!}{n_1!\, n_2!\, \ldots n_m!} \tag{5.22}$$

Ejemplo II–10

(Continuación.)
Dados los elementos a, b, c calculamos:
Permutaciones de 3 elementos:

$$P_3 \rightarrow \begin{matrix} abc & acb \\ bac & bca \\ cab & cba \end{matrix} \qquad P_3 = 3! = 6$$

Permutaciones de 3 elementos con repetición:

$$P_5^{2,2,1} \rightarrow \begin{matrix} aabbc & aabcb \\ aacbb & acabb \\ cabab & etc \end{matrix} \qquad P_5^{2,2,1} = \frac{n!}{n_1!n_2!\ldots n:m!} = \frac{5!}{2!2!1!} = 30$$

5.4.3. Combinaciones

Dado un conjunto de m elementos, se llaman **combinaciones de m elementos tomados de n en n** a todos los subconjuntos de n elementos que se pueden formar del conjunto original, con la condición de que dos subconjuntos se consideran distintos cuando difieren en algún elemento. Es decir, a diferencia de las variaciones, no se considera el orden de colocación de los elementos. El número de combinaciones se representa por $C_{m,n}$ y se calcula por

$$C_{m,n} = \frac{V_{m,n}}{P_n} = \frac{m(m-1)(m-2)\ldots(m-n+1)}{1 \times 2 \times \ldots \times n}. \tag{5.23}$$

Esta expresión también se puede escribir como

$$C_{m,n} = \frac{m!}{(m-n)!\,n!} = \left(\begin{array}{c} m \\ n \end{array} \right), \tag{5.24}$$

donde el último término es el, llamado, número combinatorio.

Por otra parte, se conocen como **combinaciones con repetición de m elementos tomados de n en n** a todos los subconjuntos de tamaño n que se pueden formar con los m elementos, en los que pueden aparecer elementos repetidos, y con la condición de que dos subconjuntos se consideran distintos si tienen elementos diferentes, sin importar el orden. Se representan por C_m^n y su número se puede calcular utilizando

$$C_m^n = C_{m+n-1,n} = \left(\begin{array}{c} m+n-1 \\ n \end{array} \right) = \frac{(m+n-1)!}{(m-1)!\,n!} \tag{5.25}$$

Ejemplo II–10

(Continuación.)
Dados los elementos a, b, c calculamos:
Combinaciones de 3 elementos de 2 en 2:

$C_{3,2} \rightarrow$ ab, ac, bc

$$C_{3,2} = \frac{m!}{(m-n)!n!} = \frac{3!}{1!2!} = 3$$

Combinaciones de 3 elementos con repetición:

$C_3^2 \rightarrow$ aa bb, ab bc, ac cc

$$C_3^2 = \frac{(m+n-1)!}{(m-1)!n!} = \frac{4!}{2!2!} = 6$$

Capítulo 6

Variables aleatorias

"Claro que lo entiendo. Hasta un niño de cinco años podría entenderlo. ¡Que me traigan un niño de cinco años!"

Groucho Marx (1890-1977)

Con el fin de estudiar estadísticamente un cierto experimento aleatorio es imprescindible realizar una descripción numérica de los resultados de dicho experimento. Para ello se define una variable, llamada aleatoria, asignando a cada resultado del experimento aleatorio un cierto valor numérico. En este capítulo veremos cómo para describir el experimento aleatorio será necesario especificar qué valores puede tomar la variable aleatoria en cuestión junto con las probabilidades de cada uno de ellos. Las dos primeras secciones estarán dedicadas a las, llamadas, variables aleatorias unidimensionales, mientras que posteriormente se estudiarán brevemente las variables aleatorias bidimensionales.

6.1. Descripción de las variables aleatorias

6.1.1. Concepto de variable aleatoria

Dado un experimento aleatorio, definimos una **variable aleatoria** como una función definida sobre el espacio muestral que asigna un número real a cada uno de los puntos, o resultados posibles, de dicho espacio muestral. Por ejemplo en el lanzamiento de monedas podemos asignar 0 si sale cara y 1 si el resultado es cruz. De esta forma, la variable aleatoria toma valores (aleatorios) determinados por el resultado del experimento. Generalmente, la variable aleatoria se denota por una letra mayúscula (ej. X), reservándose las letras minúsculas (ej. x) para los distintos valores que puede tomar. Por ejemplo, en el experimento del lanzamiento de dos dados, se puede definir la variable aleatoria que asigna a cada resultado del experimento un número dado por la suma de los dos dados. En este caso, entonces, la variable aleatoria puede tomar los valores $X = \{2, 3, \ldots, 11, 12\}$.

Una variable aleatoria que toma un número finito o infinito, pero numerable, de valores, se denomina **variable aleatoria discreta**. Un ejemplo es la suma de las puntuaciones de los dados del experimento visto anteriormente. Por el contrario, cuando la variable puede tomar un número infinito no numerable de valores (o todos los valores posibles de un intervalo) se la denomina **variable aleatoria continua**. Un ejemplo sería la duración de un suceso, o el peso de una persona. En la mayoría de los casos, las variables aleatorias continuas representan datos *medidos*, mientras que las variables aleatorias discretas suelen representar datos que se *cuentan* (ej. número de veces que ha ocurrido un cierto suceso).

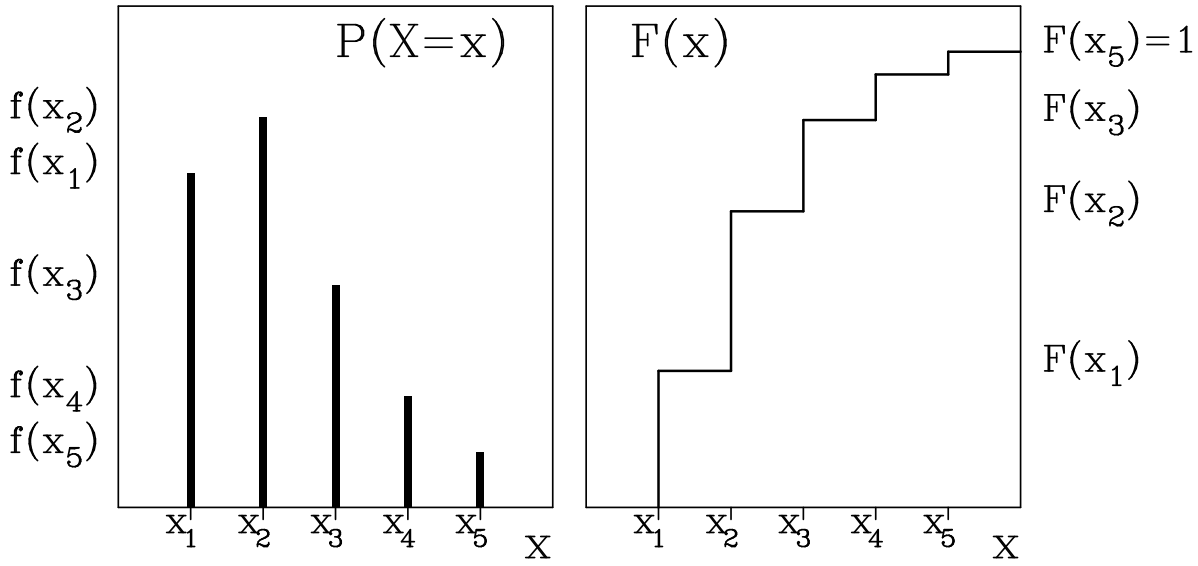


Figura 6.1: *Función de probabilidad, $f(x)$, y función de distribución, $F(x)$, para una variable aleatoria discreta $X = \{x_1, x_2, x_3, x_4, x_5\}$.*

6.1.2. Variable aleatoria discreta

Sea una variable aleatoria discreta X y supongamos que puede tomar los valores $x_1, x_2, x_3, \ldots$. Como ya se ha indicado, para describir completamente la variable aleatoria hay que indicar las probabilidades de que tome cada uno de sus valores posibles. De esta forma a cada valor de la variable aleatoria se le asigna como probabilidad la probabilidad de que ocurra el subconjunto del espacio muestral asociado con ese valor particular. Para esto se define una función $f(x)$ que indica la probabilidad de cada valor x de la variable aleatoria. Esta es la **función de probabilidad**, también llamada **distribución de probabilidad**, de la variable aleatoria discreta X

$$f(x) \equiv P(X = x). \tag{6.1}$$

En particular, para un valor x_i de la variable aleatoria: $f(x_i) = P(X = x_i)$. Además, por las propiedades de la probabilidad, la función de probabilidad cumple, para todo x_i

$$f(x_i) \geq 0 \qquad ; \qquad \sum_i f(x_i) = 1. \tag{6.2}$$

En muchas ocasiones, la distribución discreta de probabilidad se presenta en forma de tabla

x	x_1	x_2	$\cdots$	x_i	$\cdots$
$P(X = x)$	$f(x_1)$	$f(x_2)$	$\cdots$	$f(x_i)$	$\cdots$

Asimismo, gráficamente se suele representar usando un diagrama de barras donde en abscisas se sitúan los diferentes valores de X y en ordenadas las probabilidades correspondientes (Figura 6.1).

Otra forma de caracterizar la distribución de una variable aleatoria es mediante la **función de distribución** $F(x)$, o función de probabilidad acumulativa, definida para cada x como la probabilidad de que la variable aleatoria X tome un valor menor o igual que x. Es decir

$$F(x) = P(X \leq x), \tag{6.3}$$

donde x no se restringe a los valores que puede tomar la variable aleatoria y es cualquier número real ($-\infty \leq x \leq \infty$). Es fácil ver que, por su definición, $F(x)$ es una función no decreciente y toma los valores

extremos

$$F(-\infty) = 0 \qquad ; \qquad F(\infty) = 1.$$

La función de distribución se puede evaluar a partir de la función de probabilidad, y al contrario, ya que

$$F(x) = \sum_{x_i \leq x} f(x_i) = F(x_{i-1}) + f(x_i) \qquad ; \qquad f(x_i) = F(x_i) - F(x_{i-1}).$$

Si suponemos que la variable aleatoria puede tomar los valores $X = \{x_1, x_2, \ldots, x_n\}$, ordenados de menor a mayor, entonces la función de distribución para cada punto estará dada por

$$F(x) = \begin{cases} 0 & x < x_1 \\ f(x_1) & x_1 \leq x < x_2 \\ f(x_1) + f(x_2) & x_2 \leq x < x_3 \\ \vdots & \vdots \\ \sum_{i=1}^{n} f(x_i) = 1 & x_n \leq x \end{cases}$$

De modo que la representación gráfica de la función de distribución discreta tiene forma de escalera, con saltos en los valores aislados que toma la variable y con continuidad por la derecha (es decir, en cada salto el valor que toma $F(x)$ es el del escalón superior, ver Figura 6.1).

Conocida además la función de distribución puede calcularse la probabilidad de que la variable aleatoria esté comprendida entre dos valores x_i y x_j

$$P(x_i < X \leq x_j) = \sum_{k=i+1}^{j} f(x_k) = F(x_j) - F(x_i)$$

o de que la variable sea mayor que un determinado valor x_i

$$P(X > x_i) = 1 - F(x_i).$$

Ejemplo II–11

Suma de los puntos obtenidos al lanzar dos dados.

Espacio muestral o conjunto de sucesos posibles que se pueden obtener al lanzar dos dados comunes. Cada pareja de datos indica el valor facial de cada dado. En la tabla siguiente se han agrupado para obtener el número de combinaciones que dan lugar a un valor de la suma.

Resultados posibles ordenados	x_i	$f(x_i)$	$F(x_i)$	$x_i f(x_i)$	$x_i^2 f(x_i)$
(1,1)	2	1/36	1/36	2/36	4/36
(2,1) (1,2)	3	2/36	3/36	6/36	18/36
(3,1) (2,2) (1,3)	4	3/36	6/36	12/36	48/36
(4,1) (3,2) (2,3) (1,4)	5	4/36	10/36	20/36	100/36
(5,1) (4,2) (3,3) (2,4) (1,5)	6	5/36	15/36	30/36	180/36
(6,1) (5,2) (4,3) (3,4) (2,5) (1,6)	7	6/36	21/36	42/36	294/36
(6,2) (5,3) (4,4) (3,5) (2,6)	8	5/36	26/36	40/36	320/36
(6,3) (5,4) (4,5) (3,6)	9	4/36	30/36	36/36	324/36
(6,4) (5,5) (4,6)	10	3/36	33/36	30/30	300/36
(6,5) (5,6)	11	2/36	35/36	22/36	242/36
(6,6)	12	1/36	1	12/36	144/36
				252/36	1974/36

Ejemplo II–11

Si deseamos determinar la probabilidad de que este valor se encuentre en el rango $4 < x \leq 7$,

$$P(4 < x \leq 7) = F(7) - F(4) = \frac{21}{36} - \frac{6}{36} = \frac{15}{36} = (\frac{4}{36} + \frac{5}{36} + \frac{6}{36})$$

Analogamente para $x > 10$,

$$P(x > 10) = 1 - F(10) = 1 - \frac{33}{36} = \frac{3}{36} = (\frac{2}{36} + \frac{1}{36})$$

Como ejercicio adicional se puede demostrar que es más difícil obtener 9 tirando 3 dados que obtener 10. Galileo (1564-1642) demostró que hay 216 combinaciones posibles equiprobables: 25 conducen a 9 y 27 a 10. La diferencia es muy pequeña: $2/216 \sim 0.01$.

6.1.3. Variable aleatoria continua

Veamos ahora el caso de las variables aleatorias continuas, es decir, aquellas que pueden tomar cualquier valor en un intervalo (a, b), o incluso $(-\infty, \infty)$. En este caso, la probabilidad de que la variable X tome un valor determinado dentro de ese intervalo es cero, ya que existen infinitos valores posibles en cualquier intervalo, por pequeño que sea, alrededor del valor en cuestión. Por ejemplo, la probabilidad de que la altura de una persona sea exactamente 1.75 cm, con infinitos ceros en las cifras decimales, es cero. Por tanto no se puede definir una función de probabilidad igual que se hacía para las variables discretas, dando la probabilidad de cada valor de la variable. Lo que se si puede especificar es la probabilidad de que la variable esté en un cierto intervalo. Para ello se define una función $f(x)$ llamada **función de densidad**, o **distribución de probabilidad**, de la variable aleatoria continua X de forma que, para todo x, cumpla

$$f(x) \geq 0 \qquad ; \qquad \int_{-\infty}^{\infty} f(x)\ dx = 1. \tag{6.4}$$

De forma que la probabilidad de que X se encuentre entre dos valores x_1 y x_2 se puede calcular como

$$P(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)\ dx. \tag{6.5}$$

Las tres expresiones anteriores constituyen la definición de la función de densidad. Puede demostrarse que esta definición cumple los axiomas de la probabilidad. Puesto que la probabilidad de que X tome un determinado valor x_0 es nula ($\int_{x_0}^{x_0} f(x)\ dx = 0$), en la expresión anterior es indiferente escribir el signo $<$ ó $\leq$.

Puede observarse que, por la definición (6.4), la representación gráfica de la función de densidad (Figura 6.2) será la de una curva, normalmente continua, que toma siempre valores positivos o nulos, y con área, comprendida entre la curva y el eje x, unidad. De igual forma, por la expresión (6.5), la probabilidad de que la variable tome un valor entre x_1 y x_2 será el área bajo la función de densidad entre las abscisas x_1 y x_2. Esta asociación de probabilidad a área es sumamente útil para el estudio de la distribuciones continuas de probabilidad.

Al igual que para el caso discreto, se puede definir la **función de distribución** $F(x)$ en cada punto x de una variable aleatoria continua como la probabilidad de que la variable X tome un valor inferior a x

$$F(x) = P(X < x). \tag{6.6}$$

Por la definición de función de densidad, ésta se relaciona con la función de distribución por

$$F(x) = \int_{-\infty}^{x} f(t)\ dt. \tag{6.7}$$

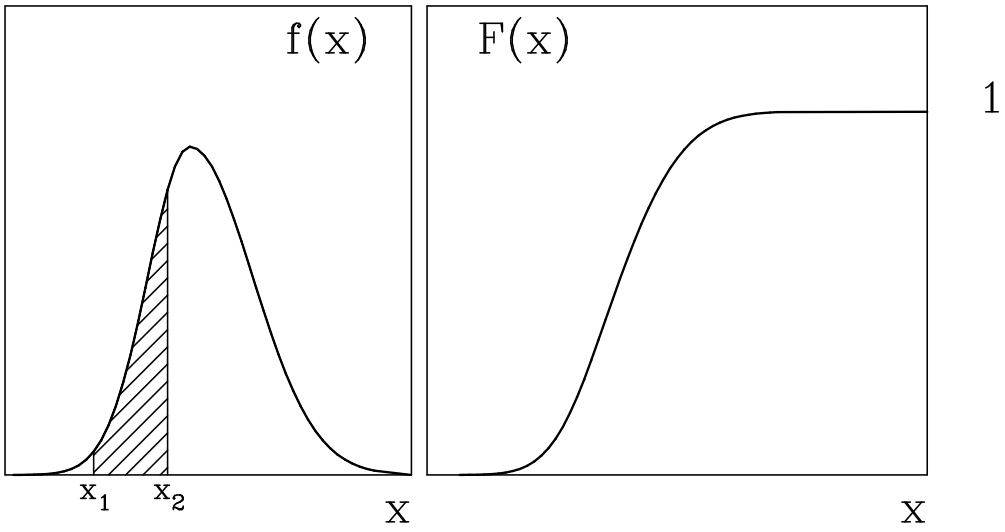


Figura 6.2: *Función de densidad, $f(x)$, y función de distribución, $F(x)$, para una variable aleatoria continua.*

También al igual que en el caso discreto, la probabilidad de que X esté en un cierto intervalo (x_1, x_2) se podrá expresar como

$$P(x_1 < X < x_2) = F(x_2) - F(x_1) = \int_{x_1}^{x_2} f(x)\ dx.$$

Si hacemos ese intervalo cada vez más pequeño, tendremos

$$F(x + \Delta x) - F(x) = P(x < X < x + \Delta x) \simeq f(x)\Delta x$$

$$\Rightarrow \qquad f(x) = \frac{dF(x)}{dx}.$$

Es decir, la derivada de la función de distribución es la función de densidad.

En general, la función de distribución será una función continua no decreciente que además cumple

$$F(-\infty) = \int_{-\infty}^{-\infty} f(x)\ dx = 0 \quad ; \quad F(\infty) = \int_{-\infty}^{\infty} f(x)\ dx = 1.$$

y, por tanto, su representación gráfica será como la mostrada en la Figura 6.2.

Evidentemente, la variable estadística puede que sólo tome valores en un intervalo (a, b). En este caso las integrales infinitas vistas anteriormente se reducen a integrales finitas y se cumple

$$\int_a^b f(x)\ dx = 1 \qquad \text{y} \qquad F(x) = \begin{cases} 0 & x < a \\ \int_a^x f(t)\ dt & a < x < b \\ 1 & x > b \end{cases}$$

6.2. Medidas características de una variable aleatoria

De la misma forma en que se definían medidas características de las distribuciones de frecuencias, se pueden definir también medidas características para la distribución de una variable aleatoria, dividiéndose éstas en medidas de centralización y medidas de dispersión. Por convenio, estas medidas teóricas se representan por letras griegas para así diferenciarlas de las medidas de las distribuciones de frecuencias, calculadas a partir de una muestra de datos, que se denotaban por letras latinas.

6.2.1. Media o esperanza matemática

La principal medida de centralización de la distribución de una variable aleatoria es la **media**, también conocida como **esperanza matemática**. Sea una variable aleatoria discreta X que toma los valores $x_1, x_2, \ldots$ y sea $f(x)$ su función de probabilidad. Por definición, la media o esperanza matemática μ (también representada por $E(X)$) de X viene dada por la expresión

$$\mu = E(X) = \sum_i x_i f(x_i). \tag{6.8}$$

Es decir, la media se obtiene multiplicando cada valor de X por su probabilidad y sumando estos productos para todos los posibles valores de X (el sumatorio se puede extender desde 1 hasta n ó ∞). Evidentemente, el significado de la media es que da un valor típico o promedio de la variable aleatoria. Nótese que esta definición es consistente con la de la media aritmética para una distribución de frecuencias ($\overline{x} = \sum_{i=1}^{k} x_i n_i/N$), ya que si hacemos tender el número de medidas a infinito y recordamos la definición de probabilidad dada en (5.1)

$$\lim_{N\to\infty} \overline{x} = \lim_{N\to\infty} \sum_{i=1}^{k} \frac{x_i n_i}{N} = \sum_{i=1}^{k} x_i \left(\lim_{N\to\infty} \frac{n_i}{N} \right) = \sum_{i=1}^{k} x_i P(X = x_i) = \sum_{i=1}^{k} x_i f(x_i) = \mu.$$

En el caso continuo la expresión para la media es similar. Se define la media o esperanza matemática de una variable aleatoria continua X con función de densidad $f(x)$ como

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x)\ dx, \tag{6.9}$$

y su significado es el mismo. Cuando la variable aleatoria sólo tome valores en un intervalo (a, b), la media se puede escribir también como

$$\mu = E(X) = \int_{a}^{b} x f(x)\ dx.$$

El concepto de esperanza matemática se puede generalizar para una función $g(X)$ de la variable aleatoria X. Nótese que dicha función será una nueva variable aleatoria. La media de esa función vendrá dada entonces, en el caso discreto y continuo, por

$$\mu_{g(X)} = E(g(X)) = \begin{cases} \sum_i g(x_i) f(x_i) \\ \int_{-\infty}^{\infty} g(x) f(x)\ dx \end{cases} \tag{6.10}$$

En particular, si la función es de la forma $g(X) = aX + b$ donde a y b son constantes, se tiene

$$\mu_{aX+b} = E(aX + b) = a\mu_X + b, \tag{6.11}$$

ya que, aplicando (6.10) en el caso continuo

$$\mu_{aX+b} = \int_{-\infty}^{\infty} (ax + b) f(x)\ dx = a \int_{-\infty}^{\infty} x f(x)\ dx + b \int_{-\infty}^{\infty} f(x)\ dx = a\mu_X + b.$$

Particularizando a los casos especiales de $a = 0$ y $b = 0$ se obtienen dos propiedades importantes de la media

$$\mu_b = E(b) = b \qquad (a = 0); \qquad \mu_{aX} = E(aX) = a\mu_X \qquad (b = 0). \tag{6.12}$$

Ejemplo II–12

Calculemos la media en el lanzamiento de dos dados: $\mu = \sum_i x_i\ f(x_i) = \dfrac{252}{36} = 7$

6.2.2. Varianza y desviación típica

La media por sí sola no proporciona una adecuada descripción de la distribución de la variable aleatoria. Además de conocer en qué valor se centra esa distribución es importante determinar la dispersión o variación de los valores de la variable aleatoria en torno a la media. Para ello se define la **varianza**, representada por σ^2 ó $\mathrm{Var}(X)$, de una variable aleatoria discreta X como

$$\mathrm{Var}(X) = \sigma^2 = E\left((X-\mu)^2\right) = \sum_i (x_i-\mu)^2 f(x_i). \tag{6.13}$$

Es decir, es la esperanza matemática de las desviaciones al cuadrado de los valores de la variable respecto a su media. Es claro que cuanto mayor sea la varianza menos concentrados estarán los valores de X respecto a su media. Al igual que ocurría con la media, la definición anterior de la varianza está íntimamente ligada a la definición, ya vista, de varianza de una distribución de frecuencias

$$\lim_{N\to\infty} s^2 = \lim_{N\to\infty} \frac{\sum_{i=1}^k (x_i-\overline{x})^2 n_i}{N-1} = \lim_{N\to\infty} \frac{N}{N-1} \sum_{i=1}^k (x_i-\overline{x})^2 \frac{n_i}{N}.$$

Teniendo en cuenta que cuando N tiende a ∞, $N/(N-1)$ tiende a 1, $\overline{x}$ tiende a μ, y n_i/N tiende a la probabilidad de x_i

$$\lim_{N\to\infty} s^2 = \sum_{i=1}^k (x_i-\mu)^2 P(X=x_i) = \sigma^2.$$

Con el fin de obtener una medida de dispersión que tenga las mismas unidades que la variable aleatoria se define la **desviación típica** σ como la raíz cuadrada positiva de la varianza

$$\sigma = +\sqrt{\sigma^2} = \sqrt{\sum_i (x_i-\mu)^2 f(x_i)}. \tag{6.14}$$

Existe una expresión alternativa más útil en la práctica para calcular la varianza

$$\sigma^2 = \sum_i x_i^2 f(x_i) - \mu^2 = E(X^2) - \mu^2. \tag{6.15}$$

Para demostrar esta expresión desarrollamos el cuadrado en (6.13) y aplicamos la definición de media

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

$$= \sum_i x_i^2 f(x_i) + \mu^2 \sum_i f(x_i) - 2\mu \sum_i x_i f(x_i) = E(X^2) + \mu^2 - 2\mu\mu = E(X^2) - \mu^2.$$

De la misma manera se puede definir la varianza y desviación típica de una variable aleatoria continua X con función de densidad $f(x)$

$$\mathrm{Var}(X) = \sigma^2 = E\left((X-\mu)^2\right) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx, \tag{6.16}$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx}. \tag{6.17}$$

Cuando X sólo toma valores en un intervalo (a,b), la definición de la varianza se reduce a

$$\sigma^2 = \int_a^b (x-\mu)^2 f(x)\ dx.$$

También, al igual que en el caso discreto, existe una expresión más práctica para su cálculo

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\ dx\ -\ \mu^2 = E(X^2) - \mu^2. \tag{6.18}$$

Análogamente a la media, suponiendo una función $g(X)$ de la variable aleatoria X, su varianza será

$$\sigma^2_{g(X)} = E\left((g(X) - \mu_{g(X)})^2\right) = \begin{cases} \sum_i (g(x_i) - \mu_{g(X)})^2 f(x_i) \\ \int_{-\infty}^{\infty} (g(x) - \mu_{g(X)})^2 f(x)\ dx \end{cases} \tag{6.19}$$

y en el caso particular de que la función sea de la forma $g(X) = aX + b$, donde a y b son constantes

$$\sigma^2_{aX+b} = \mathrm{Var}(aX + b) = a^2\sigma^2_X. \tag{6.20}$$

La demostración es rápida ya que, aplicando la relación (6.11) para la media de $aX + b$

$$\sigma^2_{aX+b} = E\left((aX + b - \mu_{aX+b})^2\right) = E\left((aX + b - a\mu_X - b)^2\right) =$$

$$= E\left(a^2(X - \mu_X)^2\right) = a^2 E\left((X - \mu_X)^2\right) = a^2\sigma^2_X.$$

Particularizando a los casos $a = 0$ y $b = 0$ se obtienen las siguientes propiedades de la varianza

$$\sigma^2_b = \mathrm{Var}(b) = 0 \qquad ; \qquad \sigma^2_{aX} = \mathrm{Var}(aX) = a^2\sigma^2_X. \tag{6.21}$$

Es decir, la varianza de una constante es nula. Estas expresiones son muy útiles para realizar cambios de variables que simplifiquen los cálculos.

Ejemplo II–12

(Continuación.)

Calculemos la varianza en el lanzamiento de dos dados:

$$\sigma^2 = \sum_i x_i^2\ f(x_i) - \mu^2 = \frac{1974}{36} - 7^2 = 5.83 \qquad \Rightarrow \qquad \sigma = 2.42$$

6.2.3. Momentos

Media y varianza son en realidad casos particulares de la definición más general de momento. Dada una variable aleatoria X se define el **momento de orden r respecto al parámetro c** como la esperanza matemática de $(X - c)^r$

$$E\left((X - c)^r\right) = \begin{cases} \sum_i (x_i - c)^r f(x_i) \\ \int_{-\infty}^{\infty} (x - c)^r f(x)\ dx \end{cases} \tag{6.22}$$

Cuando $c = 0$ tenemos los momentos respecto al origen

$$\mu'_r = \begin{cases} \sum_i x_i^r f(x_i) \\ \int_{-\infty}^{\infty} x^r f(x)\ dx \end{cases}$$

Nótese que $\mu'_0 = 1$, $\mu'_1 = \mu$, y que $\mu'_2 - \mu = \sigma^2$.

Por otra parte, cuando c es la media μ, tenemos los momentos centrales

$$\mu_r = \begin{cases} \sum_i (x_i - \mu)^r f(x_i) \\ \int_{-\infty}^{\infty} (x - \mu)^r f(x)\ dx \end{cases}$$

y se tiene: $\mu_0 = 1$, $\mu_1 = 0$ (fácil de comprobar por la definición de media) y $\mu_2 = \sigma^2$.

Una definición importante es la de **función generatriz de momentos**. Dada una variable aleatoria X, esta función se define, para cualquier real t, como la esperanza matemática de e^{tX} y se denota por $M_X(t)$. Es decir, en el caso discreto y continuo, será

$$M_X(t) = E(e^{tX}) = \begin{cases} \sum_i e^{tx_i} f(x_i) \\ \int_{-\infty}^{\infty} e^{tx} f(x)\ dx \end{cases} \tag{6.23}$$

La utilidad de la función generatriz de momentos estriba en que puede utilizarse para generar (o calcular) todos los momentos respecto al origen de la variable X, ya que se cumple

$$\mu'_r = \left. \frac{d^r M_X(t)}{dt^r} \right|_{t=0} \tag{6.24}$$

Es decir, el momento de orden r respecto al origen es la r–ésima derivada de la función generatriz de momentos, evaluada en $t = 0$. La demostración, en el caso discreto, es

$$\left. \frac{d^r M_X(t)}{dt^r} \right|_{t=0} = \frac{d^r}{dt^r} \left(\sum_i e^{tx_i} f(x_i) \right) \Bigg|_{t=0} = \sum_i \left. \frac{d^r}{dt^r} (e^{tx_i}) \right|_{t=0} f(x_i) =$$

$$= \sum_i x_i^r \left. e^{tx_i} \right|_{t=0} f(x_i) = \sum_i x_i^r f(x_i) = \mu'_r$$

Una propiedad de la función generatriz de momentos que se usará con posterioridad es la siguiente: Si a y b son dos números reales, entonces

$$M_{(X+a)/b}(t) = e^{at/b} M_X \left(\frac{t}{b} \right), \tag{6.25}$$

y la demostración es

$$M_{(X+a)/b}(t) = E\left(e^{t(X+a)/b} \right) = E\left(e^{tX/b} e^{ta/b} \right) = e^{ta/b} E\left(e^{(t/b)X} \right) = e^{at/b} M_X \left(\frac{t}{b} \right).$$

6.3. Variable aleatoria bidimensional

A veces es interesante estudiar simultáneamente varios aspectos de un experimento aleatorio. Para ello se define la **variable aleatoria bidimensional** como una función que asigna un par de números reales a cada uno de los puntos, o resultados posibles, del espacio muestral (ej. peso y altura de una muestra de individuos). En general, denotaremos una variable aleatoria bidimensional de un experimento aleatorio por (X, Y), de forma que tomará valores (x, y) en un espacio bidimensional real. Diremos que una variable bidimensional es discreta cuando las dos variables que la componen lo sean. Asimismo será continua cuando tanto X como Y sean continuas. No es difícil generalizar el estudio de las variables aleatorias bidimensionales a las variables multidimensionales, aunque no se hará aquí.

6.3.1. Distribución de probabilidad conjunta y marginal

Sea una variable aleatoria bidimensional (X, Y) discreta asociada a un experimento aleatorio. Se define la **función de probabilidad conjunta** como la función

$$f(x, y) = P(X = x, Y = y). \tag{6.26}$$

En el caso de que la variable aleatoria bidimensional sea continua se define la **función de densidad conjunta** como la función $f(x,y)$ tal que

$$P(x_1 < X < x_2, y_1 < Y < y_2) = \int_{x_1}^{x_2} \int_{y_1}^{y_2} f(x,y)\ dx\ dy. \tag{6.27}$$

Para que estas definiciones sean completas hay que añadir la condición

$$f(x,y) \geq 0, \tag{6.28}$$

junto con (para el caso discreto y continuo respectivamente)

$$\sum_i \sum_j f(x_i, y_j) = 1 \qquad ; \qquad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y)\ dx\ dy = 1. \tag{6.29}$$

Gráficamente, la función de densidad conjunta $f(x,y)$ representa una superficie con volumen (entre ella y el plano xy) unidad. Así la probabilidad de que la variable (X,Y) tome valores en unos intervalos se evalúa calculando un volumen mediante (6.27).

Para el caso discreto la función de probabilidad se suele representar mediante una tabla de doble entrada. Si asumimos que X toma valores entre x_1 y x_n, e Y toma valores entre y_1 e y_m, dicha tabla tendrá la forma

$X \setminus Y$	y_1	y_2	$\cdots$	y_m	Total
x_1	$f(x_1,y_1)$	$f(x_1,y_2)$	$\cdots$	$f(x_1,y_m)$	$f_1(x_1)$
x_2	$f(x_2,y_1)$	$f(x_2,y_2)$	$\cdots$	$f(x_2,y_m)$	$f_1(x_2)$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_n	$f(x_n,y_1)$	$f(x_n,y_2)$	$\cdots$	$f(x_n,y_m)$	$f_1(x_n)$
Total	$f_2(y_1)$	$f_2(y_2)$	$\cdots$	$f_2(y_m)$	1

donde las funciones $f_1(x)$ y $f_2(y)$ son las **funciones de probabilidad marginal** de X e Y respectivamente. Representan la probabilidad de que X (ó Y) tome un determinado valor independientemente de los valores de Y (ó X) y se calculan por

$$f_1(x) = P(X = x) = \sum_j f(x, y_j) \qquad ; \qquad f_2(y) = P(Y = y) = \sum_i f(x_i, y). \tag{6.30}$$

Evidentemente, y como puede observarse en la tabla, cumplen la condición

$$\sum_i f_1(x_i) = 1 \qquad ; \qquad \sum_j f_2(y_j) = 1.$$

Análogamente, para variable aleatoria continua, se pueden definir las **funciones de densidad marginal** como

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\ dy \qquad ; \qquad f_2(y) = \int_{-\infty}^{\infty} f(x,y)\ dx. \tag{6.31}$$

Al igual que en caso unidimensional, se puede definir la **función de distribución conjunta** como la probabilidad de que X e Y sean inferiores a unos valores dados. Así, en el caso discreto y continuo

$$F(x,y) = P(X \leq x, Y \leq y) = \sum_{x_i \leq x} \sum_{y_j \leq y} f(x_i, y_j), \tag{6.32}$$

$$F(x,y) = P(X < x, Y < y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f(u,v)\ du\ dv, \tag{6.33}$$

cumpliéndose además

$$f(x,y) = \frac{\partial^2 F}{\partial x\ \partial y}.$$

También se pueden definir las **funciones de distribución marginal** $F_1(x)$ y $F_2(y)$ como

$$F_1(x) = P(X \leq x) = \sum_{x_i \leq x} f_1(x_i) \qquad ; \qquad F_2(y) = P(Y \leq y) = \sum_{y_j \leq y} f_2(y_j) \tag{6.34}$$

$$F_1(x) = \int_{-\infty}^{x} \int_{-\infty}^{\infty} f(u,v)\ du\ dv \qquad ; \qquad F_2(y) = \int_{-\infty}^{\infty} \int_{-\infty}^{y} f(u,v)\ du\ dv, \tag{6.35}$$

con propiedades similares a las ya vistas para el caso unidimensional.

6.3.2. Distribución condicionada e independencia estadística

Dada una variable aleatoria bidimensional se define la **distribución condicionada** de X cuando la variable Y toma un valor fijo ($Y = y$) a la distribución unidimensional de la variable X para los elementos de la población que tienen como valor de Y el valor fijado. Recordando la definición (5.11) de probabilidad condicionada se puede escribir

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{f(x,y)}{f_2(y)}$$

siempre que $P(Y = y) \neq 0$. Esto nos permite definir la **función de probabilidad condicionada**, en el caso discreto, o la **función de densidad condicionada**, en el caso continuo, de X dado Y (y, análogamente, de Y dado X) como el cociente entre la función de probabilidad conjunta y la función de probabilidad marginal de la variable cuyo valor se fija

$$f(x|y) = \frac{f(x,y)}{f_2(y)} \qquad ; \qquad f(y|x) = \frac{f(x,y)}{f_1(x)}, \tag{6.36}$$

por ejemplo

$$f(x_2|y_3) = \frac{f(x_2,y_3)}{f_2(y_3)} \qquad ; \qquad f(y_4|x_2) = \frac{f(x_2,y_4)}{f_1(x_2)}.$$

De esta forma, si se desea encontrar la probabilidad de que la variable aleatoria X tome valores entre a y b cuando la variable Y tiene un valor y, habrá que evaluar, en el caso discreto y continuo

$$P(a \leq X \leq b|Y = y) = \sum_{a \leq x_i \leq b} f(x_i|y),$$

$$P(a < X < b|Y = y) = \int_a^b f(x|y)\ dx.$$

Un concepto fundamental en el estudio de las variables aleatorias bidimensionales es el de **independencia** estadística. Diremos que dos variables X e Y son independientes cuando el conocimiento de los valores que toma una de ellas no aporta información sobre los valores que puede tomar la otra. En este caso es claro que las distribuciones condicionadas son iguales a las distribuciones marginales

$$f(x|y) = f_1(x) \qquad ; \qquad f(y|x) = f_2(y).$$

Esto puede demostrarse fácilmente, por ejemplo en el caso continuo, desarrollando la definición de la función

de densidad marginal dada en (6.31)

$$f_1(x) = \int_{-\infty}^{\infty} f(x,y)\ dy = \int_{-\infty}^{\infty} f(x|y) f_2(y)\ dy = f(x|y) \int_{-\infty}^{\infty} f_2(y)\ dy = f(x|y),$$

donde se ha aplicado que $f(x|y)$ no depende del valor de y. Utilizando entonces la definición de la función de probabilidad (o de densidad) condicionada, vista en (6.36), en el caso de que las variables sean independientes se cumplirá

$$f(x,y) = f_1(x) f_2(y). \tag{6.37}$$

Esto se suele tomar como la condición necesaria y suficiente para la condición de independencia, de forma que diremos que dos variables aleatorias X e Y son independientes si la función de probabilidad conjunta (o la función de densidad conjunta, en el caso continuo) puede expresarse como el producto de una función de X y una función de Y, las cuales coinciden con las funciones de probabilidad (o de densidad) marginales. Esta definición de variables aleatorias independientes es equivalente a la definición de sucesos independientes vista en (5.15). En el caso de independencia es evidente que la función de distribución conjunta también se puede expresar en función de las funciones de distribución marginales

$$F(x,y) = F_1(x) F_2(y).$$

6.3.3. Medias, varianzas y covarianza

Sea una variable aleatoria bidimensional (X,Y) con función de probabilidad, o función de densidad, conjunta $f(x,y)$. Al igual que en el caso unidimensional, se pueden definir las **medias, o esperanzas matemáticas**, de cada una de las dos variables como (en el caso discreto y continuo)

$$\mu_X = E(X) = \sum_i \sum_j x_i f(x_i, y_j)\ \ ;\ \ \mu_Y = E(Y) = \sum_i \sum_j y_j f(x_i, y_j),$$

$$\mu_X = E(X) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x,y)\ dx\ dy\ \ ;\ \ \mu_Y = E(Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x,y)\ dx\ dy.$$

En el caso de tener una variable aleatoria expresada como una función $g(X,Y)$ de las dos variables X e Y, su media vendrá dada por

$$\mu_{g(X,Y)} = E(g(X,Y)) = \begin{cases} \sum_i \sum_j g(x_i, y_j) f(x_i, y_j) \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y)\ dx\ dy \end{cases} \tag{6.38}$$

En particular, si la función es una combinación de lineal de las dos variables de la forma $g(X,Y) = aX + bY$ es inmediato que

$$\mu_{aX+bY} = a\mu_X + b\mu_Y \qquad \text{y en concreto :} \qquad \mu_{X+Y} = \mu_X + \mu_Y. \tag{6.39}$$

La esperanza matemática es entonces un operador lineal. Otra importante expresión puede deducirse suponiendo que $g(X,Y) = XY$. En este caso, si las dos variables son **independientes**, se cumple

$$\mu_{XY} = E(XY) = E(X)E(Y) = \mu_X \mu_Y. \tag{6.40}$$

Para demostrarlo se parte de la definición dada en (6.38) y se aplica la condición de independencia (6.37)

$$\mu_{XY} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x,y)\ dx\ dy = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_1(x) f_2(y)\ dx\ dy =$$

$$= \int_{-\infty}^{\infty} x f_1(x)\ dx \ \int_{-\infty}^{\infty} y f_2(y)\ dy = \mu_x \mu_y.$$

Por otra parte, se pueden definir las **varianzas** de X e Y, para variables aleatorias discretas y continuas, como (en este caso sólo escribimos las varianzas de X, para Y las expresiones son análogas)

$$\sigma_X^2 = \mathrm{Var}(X) = \sum_i \sum_j (x_i - \mu_X)^2 f(x_i, y_j),$$

$$\sigma_X^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)^2 f(x, y)\ dx\ dy.$$

Una cantidad importante en el caso bidimensional es la **covarianza**. Se define ésta como

$$\sigma_{XY}^2 = \mathrm{Cov}(X, Y) = E\left((X - \mu_X)(Y - \mu_Y)\right). \tag{6.41}$$

De manera que, en el caso discreto y continuo, es

$$\sigma_{XY}^2 = \sum_i \sum_j (x_i - \mu_X)(y_j - \mu_Y) f(x_i, y_j), \tag{6.42}$$

$$\sigma_{XY}^2 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\ dx\ dy. \tag{6.43}$$

Hay que indicar que en algunos textos no se incluye el cuadrado en la notación de la covarianza, representándose ésta por σ_{XY}. Otra forma, útil en la práctica, de expresar la covarianza es

$$\sigma_{XY}^2 = E(XY) - \mu_X \mu_Y = \mu_{XY} - \mu_X \mu_Y. \tag{6.44}$$

Se puede demostrar desarrollando la expresión (6.42)

$$\sigma_{XY}^2 = \sum_i \sum_j (x_i y_j - x_i \mu_Y - \mu_X y_j + \mu_X \mu_Y) f(x_i, y_j) =$$

$$= \sum_i \sum_j x_i y_j f(x_i, y_j) - \mu_Y \sum_i \sum_j x_i f(x_i, y_j) - \mu_X \sum_i \sum_j y_j f(x_i, y_j) +$$

$$+ \mu_X \mu_Y \sum_i \sum_j f(x_i, y_j).$$

Puesto que el primer término es la esperanza matemática del producto XY y el sumatorio del último término es la unidad

$$\sigma_{XY}^2 = E(XY) - \mu_Y \mu_X - \mu_X \mu_Y + \mu_X \mu_Y = \mu_{XY} - \mu_X \mu_Y,$$

como queríamos demostrar.

Si aplicamos la relación (6.40) a esta última expresión de la covarianza se obtiene que, para variables aleatorias independientes, la covarianza es nula ($\sigma_{XY} = 0$). Este resultado indica que la covarianza es una medida del grado de correlación, o asociación, entre las dos variables, al igual que ocurría con la covarianza de una variable estadística bidimensional. Un valor alto de la covarianza indicará una correlación (positiva o negativa, dependiendo del signo de la covarianza) importante (los valores de una variable tienden a aumentar al aumentar la otra, en el caso de covarianza positiva). Hay que indicar, sin embargo, que el que la covarianza sea nula no implica que las dos variables sean estadísticamente independientes.

Una expresión importante es la de la varianza de una combinación lineal de variables aleatorias, la cual

se puede expresar en función de las varianzas de ambas variables y la covarianza

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma^2_{XY}. \tag{6.45}$$

Para demostrarlo se parte de la definición de varianza y se aplica la expresión (6.39)

$$\sigma^2_{aX+bY} = E\left((aX+bY-\mu_{aX+bY})^2\right) = E\left((aX+bY-a\mu_X-b\mu_Y)^2\right) =$$

$$= E\left((a(X-\mu_X)+b(Y-\mu_Y))^2\right) =$$

$$= a^2E\left((X-\mu_X)^2\right) + b^2E\left((Y-\mu_Y)^2\right) + 2abE\left((X-\mu_X)(Y-\mu_Y)\right) =$$

$$= a^2\sigma^2_X + b^2\sigma^2_Y + 2ab\sigma^2_{XY}.$$

En el caso importante de variables aleatorias independientes la covarianza es nula y, por tanto, (6.45) se convierte en

$$\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y \qquad \text{y en particular :} \qquad \sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y. \tag{6.46}$$

Nótese que la expresión es la misma para la suma o resta de dos variables aleatorias.

6.4. Teorema de Chebyshev

Como ya se ha visto anteriormente, la varianza, o la desviación típica, de una variable aleatoria proporciona una medida de la dispersión, o variabilidad, de las observaciones respecto a su valor medio. Si la varianza es pequeña la mayoría de los valores de la variable se agrupan alrededor de la media. Por el contrario, si σ es grande existirá una gran dispersión de estos valores. En este sentido, el **teorema de Chebyshev** establece una relación entre la desviación típica y la probabilidad de que la variable tome un valor entre dos valores simétricos alrededor de la media. En particular, proporciona una estimación conservadora de la probabilidad de que una variable aleatoria asuma un valor dentro de k desviaciones típicas alrededor de la media.

El enunciado del teorema es el siguiente: Sea una variable aleatoria X con media μ y desviación típica σ. La probabilidad de que X tome un valor dentro de k desviaciones típicas de la media es al menos $1 - 1/k^2$. Es decir

$$P(\mu - k\sigma < X < \mu + k\sigma) \geq 1 - \frac{1}{k^2}. \tag{6.47}$$

Para demostrarlo, en el caso continuo, desarrollamos la definición de varianza

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

$$= \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\ dx + \int_{\mu-k\sigma}^{\mu+k\sigma} (x-\mu)^2 f(x)\ dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)\ dx,$$

entonces

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} (x-\mu)^2 f(x)\ dx + \int_{\mu+k\sigma}^{\infty} (x-\mu)^2 f(x)\ dx,$$

puesto que ninguna de las integrales es negativa. Puesto que en los intervalos que cubren las dos últimas integrales siempre se cumple

$$|x-\mu| \geq k\sigma \Rightarrow (x-\mu)^2 \geq k^2\sigma^2,$$

y por ello

$$\sigma^2 \geq \int_{-\infty}^{\mu-k\sigma} k^2\sigma^2 f(x)\ dx + \int_{\mu+k\sigma}^{\infty} k^2\sigma^2 f(x)\ dx$$

$$\Rightarrow \qquad \frac{1}{k^2} \geq \int_{-\infty}^{\mu-k\sigma} f(x)\ dx + \int_{\mu+k\sigma}^{\infty} f(x)\ dx = 1 - \int_{\mu-k\sigma}^{\mu+k\sigma} f(x)\ dx,$$

puesto que el segundo término es la probabilidad de que X tome un valor fuera del intervalo $(\mu - k\sigma, \mu + k\sigma)$. Por tanto

$$P(\mu - k\sigma < X < \mu + k\sigma) = \int_{\mu-k\sigma}^{\mu+k\sigma} f(x)\ dx \geq 1 - \frac{1}{k^2},$$

como queríamos demostrar.

Nótese que, por ejemplo, haciendo $k = 2$, el teorema nos dice que la probabilidad de que una variable, con cualquier distribución de probabilidad, tome un valor más cerca de 2σ de la media es al menos 0.75. Para calcular un valor exacto de estas probabilidades habrá que conocer cual es la forma de la distribución de probabilidad. Análogamente el intervalo $\mu \pm 3\sigma$ $(k = 3)$ contiene al menos el 89 % de la distribución y $\mu \pm 4\sigma$ $(k = 4)$ contiene al menos el 94 %.

Capítulo 7

Distribuciones discretas de probabilidad

"La vida merece la pena sólo por dos cosas: por descubrir las matemáticas y por enseñarlas."

Siméon Poisson (1781-1840)

Existen muchos fenómenos naturales que obedecen a distribuciones de probabilidad similares. En este tema vamos a conocer algunas de las más frecuentes e importantes.

El comportamiento de una variable aleatoria queda, en general, descrito por su **distribución de probabilidad**, o función de probabilidad $f(x)$, que, en el caso de que la variable sea discreta, indica la probabilidad de que se dé cada uno de los valores x posibles de la variable aleatoria ($f(x) = P(X = x)$). La práctica indica que muchos experimentos aleatorios tienen comportamientos similares, de forma que sus resultados siguen la misma distribución de probabilidad. En este capítulo se van a presentar las principales distribuciones discretas de probabilidad. Existen otras distribuciones discretas que no se abordarán aquí por brevedad.

7.1. Distribución discreta uniforme

La distribución uniforme es la más simple de todas las distribuciones discretas de probabilidad. Diremos que tenemos una **distribución discreta uniforme** cuando todos los posibles valores de la variable aleatoria sean igualmente probables. En este caso, si la variable aleatoria X puede tomar los valores $x_1, x_2, \ldots, x_n$ con probabilidades iguales, la función de probabilidad vendrá dada por

$$f(x; n) = \frac{1}{n}, \qquad \text{donde} \quad x = x_1, x_2, \ldots, x_n \tag{7.1}$$

por la condición de normalización (6.2) ($\sum f(x_i) = 1$). Se ha utilizado la notación $f(x; n)$ puesto que, en este caso, la distribución de probabilidad depende (únicamente) del parámetro n, o número de valores posibles.

Las expresiones para la media y varianza de esta distribución son, evidentemente

$$\mu = \sum_{i=1}^{n} x_i f(x_i, n) = \sum_{i=1}^{n} \frac{x_i}{n} = \frac{\sum_{i=1}^{n} x_i}{n},$$

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 f(x_i, n) = \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{n} = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}.$$

Ejemplo II–13

Lanzamiento de un dado (no trucado). Es una distribución discreta uniforme.

$$x = 1, 2, 3, 4, 5, 6 \qquad n = 6 \qquad f(x; 6) = \frac{1}{6}$$

$$\mu = \frac{\sum x_i}{n} = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{n} = \frac{\sum (x_i - 3.5)^2}{6} = 2.92 \qquad \Rightarrow \qquad \sigma = 1.71$$

7.2. Distribución binomial

Supongamos un experimento aleatorio consistente en realizar un número de ensayos o pruebas repetidas, cada una de ellas con únicamente dos posibles resultados mutuamente excluyentes, que denominaremos *éxito* o *fracaso*. Supongamos que la probabilidad de obtener un éxito en un ensayo es siempre constante y que los diferentes ensayos son independientes, en el sentido de que el resultado de un ensayo no afecta a los otros. En este caso diremos que tenemos un **proceso de Bernoulli**. En concreto, el proceso de Bernoulli debe tener las siguientes propiedades

1. El experimento consiste en n ensayos repetidos.
2. El resultado de cada uno de los ensayos puede clasificarse en *éxito* o *fracaso* (excluyentes).
3. La probabilidad de éxito, que denotaremos por p, es constante en todos los ensayos.
4. Los diferentes ensayos son independientes.

Ejemplos de procesos de Bernoulli son la prueba de artículos de una cadena de producción para determinar cuáles son defectuosos, la extracción de una carta para ver si es de un palo o no (siempre que se devuelva la carta extraída a la baraja) o la observación del sexo de recien nacidos.

Se define la **variable aleatoria binomial** como la función que da el número de éxitos en un proceso de Bernoulli. Evidentemente, la variable binomial X podrá tener valores en el rango $X = \{0, 1, 2, \ldots, n\}$, donde n es el número de veces que se repite el ensayo. La distribución de probabilidad asociada con esta variable aleatoria se denomina **distribución binomial** y vendrá representada por

$$f(x) = P(X = x) = b(x; n, p),$$

ya que depende del número de ensayos n y la probabilidad de éxito p en un solo ensayo. Para calcular una expresión para $b(x; n, p)$ consideremos la probabilidad de que se obtengan x éxitos y $n - x$ fracasos en un orden determinado. Llamando q a la probabilidad de fracaso (que será evidentemente $q = 1 - p$) y teniendo en cuenta que los n ensayos son independientes, la probabilidad de esa disposición de resultados particular será el producto de las probabilidades de cada ensayo, es decir

$$\overbrace{p \ldots p}^{x} \overbrace{q \ldots q}^{n-x} = p^x q^{n-x}.$$

Para calcular la probabilidad total de x éxitos, tenemos que sumar la probabilidad anterior para todas las disposiciones posibles de resultados en que se dan esos x éxitos. Ese número se puede calcular como las permutaciones con repetición de n elementos con x y $n - x$ elementos repetidos, que por (5.22) se puede

expresar como

$$P_n^{x,n-x} = \frac{n!}{x!(n-x)!} = \left(\begin{array}{c} n \\ x \end{array} \right).$$

De esta forma, la probabilidad de obtener x éxitos, o la distribución de probabilidad binomial, viene dada por

$$b(x;n,p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}, \qquad \text{donde} \qquad x = 0, 1, \ldots, n \tag{7.2}$$

El término de distribución binomial viene del hecho de que los diversos valores de $b(x;n,p)$ con $x = 0, 1, 2, \ldots, n$ corresponden a los $n+1$ términos de la expansión binomial de $(q+p)^n$ pues

$$(q+p)^n = \left(\begin{array}{c} n \\ 0 \end{array} \right) q^n + \left(\begin{array}{c} n \\ 1 \end{array} \right) pq^{n-1} + \left(\begin{array}{c} n \\ 2 \end{array} \right) p^2 q^{n-2} + \ldots + \left(\begin{array}{c} n \\ n \end{array} \right) p^n =$$

$$(q+p)^n = b(0;n,p) + b(1;n,p) + \ldots + b(n;n,p) = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x}.$$

Nótese además que, puesto que $(q+p) = 1$, la expresión anterior implica

$$\sum_{x=0}^{n} b(x;n,p) = \sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} = 1,$$

como debe cumplir cualquier función de probabilidad.

Dado que el cálculo de probabilidades binomiales por la expresión (7.2) es, generalmente, laborioso, en la **Tabla I** (Apéndice A) se presentan las probabilidades de que la variable aleatoria binomial X tome los diferentes posibles valores para diferentes n y p. Con frecuencia es necesario calcular la probabilidad de que X sea menor a un determinado valor, o esté en un intervalo dado. Para ello es necesario calcular la función de distribución de la variable aleatoria bidimensional

$$P(X \leq x) = B(x;n,p) = \sum_{r=0}^{x} b(r;n,p), \tag{7.3}$$

cuyos valores se encuentran tabulados en la **Tabla II** (Apéndice A) para diferentes valores de n y p. En realidad se tabula

$$P(X \geq r) = \sum_{x=r}^{n} b(x;n,p),$$

utilizando la notación de la tabla. Es decir se tabula la cola de la derecha.

Un caso particular importante de la distribución binomial es cuando $n = 1$, es decir, cuando sólo se hace un ensayo. En este caso llamaremos **variable de Bernoulli** a X, que sólo podrá tomar los valores 0 (fracaso) y 1 (éxito), y diremos que tenemos una **distribución de Bernoulli**. La función de probabilidad será

$$f(x) = \left(\begin{array}{c} 1 \\ x \end{array} \right) p^x q^{1-x} = p^x q^{1-x} = \left\{ \begin{array}{lll} q & ; & x = 0 \\ p & ; & x = 1 \end{array} \right. \tag{7.4}$$

Calculemos a continuación la media y la varianza de la distribución de Bernoulli

$$\mu = \sum_{x_i=0}^{1} x_i f(x_i) = 0q + 1p = p, \tag{7.5}$$

$$\sigma^2 = \sum_{x_i=0}^{1} x_i^2 f(x_i) - \mu^2 = 0^2 q + 1^2 p - p^2 = p - p^2 = p(1-p) = pq. \tag{7.6}$$

Estas relaciones pueden utilizarse para calcular la media y la varianza de la distribución binomial. Efectivamente, la variable binomial puede expresarse como la suma de n variables de Bernoulli (independientes) ($x = x_1 + x_2 + \ldots + x_n$) y, por tanto, la media de la distribución binomial, utilizando (6.39) ($\mu_{aX+bY} = a\mu_X + b\mu_Y$) vendrá dada por

$$\mu_X = \mu_{X_1+X_2+\ldots+X_n} = \mu_{X_1} + \mu_{X_2} + \ldots + \mu_{X_n} = \overbrace{p + p + \ldots + p}^{n}$$

$$\Rightarrow \qquad \mu = np. \tag{7.7}$$

Asimismo, podemos utilizar (6.45) para calcular la varianza de la distribución binomial, y puesto que las n variables son independientes ($\sigma^2_{aX+bY} = a^2\sigma^2_X + b^2\sigma^2_Y$)

$$\sigma^2_X = \sigma^2_{X_1+X_2+\ldots+X_n} = \sigma^2_{X_1} + \sigma^2_{X_2} + \ldots + \sigma^2_{X_n} = \overbrace{pq + pq + \ldots + pq}^{n}$$

$$\Rightarrow \qquad \sigma^2 = npq, \tag{7.8}$$

y, por tanto, la desviación típica será

$$\sigma = \sqrt{npq}. \tag{7.9}$$

Una propiedad importante de la distribución binomial es que será simétrica en el caso de $p = q$ y presentará asimetría a la derecha (serán más probables los valores bajos de x) cuando $p < q$ (y al contrario), como es lógico esperar. La distribución binomial es de gran utilidad en numerosos campos científicos, incluido el control de calidad y aplicaciones médicas.

Ejemplo II–14

Sea un jugador de baloncesto que tiene que tirar 3 tiros libres. Sabemos que su promedio de acierto es del 80 %. Determinemos las probabilidades de que enceste 0, 1, 2 ó 3 canastas.
Si llamamos: Canasta → S ; Fallo → N; x: número de canastas o puntos.
Podemos calcular la probabilidad de cada suceso como el producto de las probabilidades de cada tiro ya que son sucesos independientes.

	x	P
SSS	3	0.512
SSN	2	0.128
SNS	2	0.128
SNN	1	0.032
NSS	2	0.128
NSN	1	0.032
NNS	1	0.032
NNN	0	0.008
		1.000

$$P(S) = 0.8 \qquad P(N) = 0.2$$

$$P(SSS) = 0.8 \times 0.8 \times 0.8 = 0.512$$

$$P(SSN) = 0.8 \times 0.8 \times 0.2 = 0.128$$

$$P(SNN) = 0.8 \times 0.2 \times 0.2 = 0.032$$

$$P(NNN) = 0.2 \times 0.2 \times 0.2 = 0.008$$

La probabilidad de cada x se calcula sumando las probabilidades para cada disposición:

$P(x = 0) = 0.008$ $\qquad$ $P(x = 1) = 3 \times 0.032 = 0.096$

$P(x = 2) = 3 \times 0.128 = 0.384$ $\qquad$ $P(x = 3) = 0.512$

Ejemplo II–14

(Continuación.)

La prob. de 2 éxitos en 3 intentos:

$$p^2\, q^1 = 0.8^2 \times 0.2^1 = 0.128$$

El número de disposiciones para cada x:

$$(x=3)\quad P_3^{3,0} = \begin{pmatrix} 3 \\ 3 \end{pmatrix} = \frac{3!}{3!0!} = 1$$

$$(x=2)\quad P_3^{2,1} = \begin{pmatrix} 3 \\ 2 \end{pmatrix} = \frac{3!}{2!1!} = 3$$

$$(x=1)\quad P_3^{1,2} = \begin{pmatrix} 3 \\ 1 \end{pmatrix} = \frac{3!}{1!2!} = 3$$

$$(x=0)\quad P_3^{0,3} = \begin{pmatrix} 3 \\ 0 \end{pmatrix} = \frac{3!}{0!3!} = 1$$

También puede usarse:

$$b(x;n,p) = \begin{pmatrix} n \\ x \end{pmatrix} p^x q^{n-x}$$

$$b(x;3,0.8) = \begin{pmatrix} 3 \\ x \end{pmatrix} 0.8^x 0.2^{3-x}$$

$$b(0;3,0.8) = \begin{pmatrix} 3 \\ 0 \end{pmatrix} 0.8^0 0.2^3 = 0.008$$

$$b(1;3,0.8) = \begin{pmatrix} 3 \\ 1 \end{pmatrix} 0.8^1 0.2^2 = 0.096$$

$$b(2;3,0.8) = \begin{pmatrix} 3 \\ 2 \end{pmatrix} 0.8^2 0.2^1 = 0.384$$

$$b(3;3,0.8) = \begin{pmatrix} 3 \\ 3 \end{pmatrix} 0.8^3 0.2^0 = 0.512$$

También puede buscarse en las tablas. En este caso en la **Tabla I** con $n = 3$, $p = 0.80$ y $x = 0, 1, 2, 3$.
Si queremos calcular la probabilidad de que acierte 2 o más canastas, debemos calcular la función de distribución.

n	x	0.1 . . . . . 0.7	0.8	0.9 . . .
2	0			
	1			
3	0		0.008	
	1		0.096	
	2		0.384	
	3		0.512	
4	0			

.

$$P(X \geq 2) = \sum_{x=2}^{3} b(x;3,0.80) = 0.384 + 0.512 = 0.896$$

o buscar en la **Tabla II** con $n = 3$, $r = 2$, $p = 0.80$.

La media se obtiene como:

$$\mu = np = 3 \times 0.8 = 2.4$$

puede comprobarse haciendo,

$$\mu = \sum_{x=0}^{3} x\; b(x;n,p) = 2.4$$

La varianza y la desviación típica:

$$\sigma^2 = n\; p\; q = 3 \times 0.8 \times 0.2 = 0.48 \rightarrow \sigma = 0.69$$

puede comprobarse haciendo,

$$\sigma = \sqrt{\sum_{x=0}^{3} (x-\mu)^2\; b(x;n,p)} = 0.69$$

7.3. Distribución de Poisson

Consideremos un experimento aleatorio consistente en medir el número de resultados, o sucesos de un tipo dado, que se producen en un cierto intervalo continuo. Este intervalo puede ser un intervalo de tiempo, de espacio, una región dada, etc. Ejemplos de este experimento podrían ser: el número de partículas radiactivas

emitidas por un material en un tiempo dado, el número de fotones que llegan a un detector en un tiempo fijado, el número de días al año en que llueve en un cierto lugar, el número de estrellas que se observan en el cielo en cuadrículas del mismo tamaño, etc. Diremos que un experimento de este tipo sigue un **proceso de Poisson** cuando se cumplan las siguientes condiciones:

1. El número de resultados que ocurren en un intervalo es independiente del número que ocurre en otro intervalo disjunto. Es decir, los sucesos aparecen aleatoriamente de forma independiente. Se dice entonces que el proceso no tiene memoria.

2. La probabilidad de que un resultado sencillo ocurra en un intervalo pequeño es proporcional a la longitud de dicho intervalo. Además dicha probabilidad permanece constante, de forma que se puede definir un número medio de resultados por unidad de intervalo. Se dice que el proceso es estable.

3. La probabilidad de que ocurra más de un resultado en un intervalo suficientemente pequeño es despreciable.

Se define entonces la **variable aleatoria de Poisson** como el número de resultados que aparecen en un experimento que sigue el proceso de Poisson. Nótese que el campo de variabilidad de la variable de Poisson será: $X = \{0, 1, 2, \ldots\}$. La distribución de probabilidad asociada con esta variable se denomina **distribución de Poisson** y dependerá fundamentalmente del número medio de resultados (o sucesos) por intervalo, que denotaremos por λ. De esta forma, la distribución de Poisson se escribe

$$f(x) = P(X = x) = p(x; \lambda).$$

Para calcular una expresión para $p(x; \lambda)$ es importante relacionar la distribución de Poisson con la binomial. Efectivamente, la distribución de Poisson aparece como límite de la distribución binomial cuando el número de observaciones en ésta última es muy grande y la probabilidad de que en una observación se dé el suceso (se obtenga un éxito, en la nomenclatura de la distribución binomial) es muy pequeña. Para ello dividimos el intervalo de observación en n intervalos muy pequeños, con n suficientemente grande para que, por la tercera propiedad del proceso de Poisson, no se puedan dar dos sucesos en cada subintervalo, y la probabilidad p de que ocurra un suceso en un subintervalo sea muy pequeña. De esta forma, el experimento de observar cuantos sucesos aparecen en un intervalo se convierte en observar si ocurre o no un suceso en n subintervalos (proceso de Bernoulli). Podemos suponer entonces una distribución binomial con n ensayos y probabilidad de éxito en cada uno p, que podremos escribir

$$b(x; n, p) = \left(\begin{array}{c} n \\ x \end{array} \right) p^x q^{n-x} = \frac{n(n-1)\ldots(n-x+1)}{x!} p^x (1-p)^{n-x}.$$

Nótese que, aunque $n \to \infty$ y $p \to 0$, el número medio esperado de sucesos en el intervalo total ha de permanecer constante, e igual a λ, es decir: $\mu = np = \lambda$. Haciendo tender n a infinito y sustituyendo p por λ/n

$$\lim_{n\to\infty} b(x; n, p) = \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{x!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} =$$

$$= \lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^n \left(1 - \frac{\lambda}{n}\right)^{-x} = \frac{\lambda^x}{x!} e^{-\lambda},$$

donde se ha introducido el valor de los siguientes límites

$$\lim_{n\to\infty} \frac{n(n-1)\ldots(n-x+1)}{n^x} = \lim_{n\to\infty} 1 \left(1 - \frac{1}{n}\right) \ldots \left(1 - \frac{x-1}{n}\right) = 1$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-x}=1$$

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{n}=\left(\lim_{n\to\infty}\left(1+\frac{1}{n/(-\lambda)}\right)^{n/(-\lambda)}\right)^{-\lambda}=e^{-\lambda}$$

De esta forma, la distribución de probabilidad de Poisson, o probabilidad de que se den x sucesos en un proceso de Poisson con valor promedio λ, vendrá dada por

$$p(x;\lambda)=\frac{\lambda^x}{x!}e^{-\lambda},\qquad \text{donde}\qquad x=0,1,2,\ldots \tag{7.10}$$

Aunque el campo de variabilidad de X es infinito, las probabilidades disminuirán muy rápidamente al aumentar x (Nótese que $x\gg\lambda\Rightarrow\lambda^x\ll x!$). Es inmediato comprobar que esta función de probabilidad cumple la propiedad de que la suma para todos los valores de x de las probabilidades es la unidad, ya que

$$\sum_{x=0}^{\infty}p(x;\lambda)=\sum_{x=0}^{\infty}\frac{\lambda^x}{x!}e^{-\lambda}=e^{-\lambda}\left(1+\frac{\lambda}{1!}+\frac{\lambda^2}{2!}+\ldots\right)=e^{-\lambda}e^{\lambda}=1.$$

Para facilitar su cálculo, en la **Tabla III** (Apéndice A) se da la función de distribución de Poisson (o probabilidades acumuladas) para diferentes valores de λ y x, definida como

$$P(x;\lambda)=\sum_{r=0}^{x}p(r;\lambda)=\sum_{r=0}^{x}\frac{\lambda^r}{r!}e^{-\lambda}.$$

Es fácil demostrar que la media de la distribución de Poisson coincide con el parámetro λ, como cabría esperar

$$\mu=\sum_{x=0}^{\infty}xp(x;\lambda)=\sum_{x=0}^{\infty}x\frac{\lambda^x}{x!}e^{-\lambda}=\sum_{x=1}^{\infty}x\frac{\lambda^x}{x!}e^{-\lambda}=\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!}e^{-\lambda}.$$

Haciendo el cambio de variable $y=x-1$

$$\mu=\lambda\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}e^{-\lambda}=\lambda\sum_{y=0}^{\infty}p(y;\lambda)=\lambda\times 1\qquad\Rightarrow\qquad\mu=\lambda. \tag{7.11}$$

Para calcular la varianza σ^2 encontramos primero una expresión alternativa para dicho parámetro. En general

$$\sigma^2=E\left(X^2\right)-\mu^2=E\left(X^2\right)-E(X)+\mu-\mu^2=E\left(X(X-1)\right)+\mu-\mu^2. \tag{7.12}$$

En el caso particular del cálculo de la distribución de Poisson podemos entonces desarrollar la esperanza que aparece en el último término de la expresión anterior

$$E\left(X(X-1)\right)=\sum_{x=0}^{\infty}x(x-1)\frac{\lambda^x}{x!}e^{-\lambda}=\sum_{x=2}^{\infty}x(x-1)\frac{\lambda^x}{x!}e^{-\lambda}=\lambda^2\sum_{x=2}^{\infty}\frac{\lambda^{x-2}}{(x-2)!}e^{-\lambda}.$$

Haciendo el cambio de variable $y=x-2$

$$E\left(X(X-1)\right)=\lambda^2\sum_{y=0}^{\infty}\frac{\lambda^y}{y!}e^{-\lambda}=\lambda^2\sum_{y=0}^{\infty}p(y;\lambda)=\lambda^2,$$

$$\sigma^2=E\left(X(X-1)\right)+\mu-\mu^2=\lambda^2+\mu-\mu^2=\mu^2+\mu-\mu^2=\mu$$

$$\Rightarrow\qquad\sigma^2=\lambda\quad;\quad\sigma=\sqrt{\lambda} \tag{7.13}$$

Es decir, la varianza de la distribución de Poisson coincide con su valor medio y con el parámetro λ que

fija la función de probabilidad. La expresión para la desviación típica se suele expresar en teoría de la señal diciendo que el error (desviación típica) es la raíz cuadrada de la señal (valor medio).

Respecto a la forma de la distribución de Poisson se encuentra que presenta una asimetría a la derecha y tiende a hacerse simétrica cuando $n \to \infty$.

Ejemplo II–15

Sea un detector astronómico al que llegan una media de 3 fotones cada segundo. Calcular las probabilidades de que lleguen $0, 1, 2, 3, 4, \ldots$ fotones/s.
Es una distribución de Poisson con $\lambda = 3$.

$(x;\lambda)$	$p(x;\lambda)$
(0;3)	0.05
(1;3)	0.15
(2;3)	0.22
(3;3)	0.22
(4;3)	0.17
(5;3)	0.10
(6;3)	0.05
(7;3)	0.02
(8;3)	0.008
(9;3)	0.003
(10;3)	0.0008
(50;3)	1.2×10^{-42}

$$p(x;\lambda) = \frac{\lambda^x}{x!}e^{-\lambda} \to p(x;3) = \frac{3^x}{x!}e^{-3}$$

Probabilidades acumuladas:

$$P(x \leq 3) = \sum_{x=0}^{3} p(x;\lambda) = 0.05 + 0.15 + 0.22 + 0.22 = 0.64$$

o mirando en la **Tabla III** ($\lambda = 3$ y $x = 3$) que sale 0.647.
También usando las tablas se puede calcular la probabilidad de un valor concreto (ej: 5) haciendo:

$$p(5;3) = \sum_{x=0}^{5} p(x;3) - \sum_{x=0}^{4} p(x;3) = 0.916 - 0.815 = 0.101$$

La media se obtiene como:

$$\mu = \lambda = 3$$

y podemos comprobarlo haciendo,

$$\mu = \sum_{x=0}^{\infty} xp(x;3) \simeq \sum_{x=0}^{10} xp(x;3) = 2.97 \simeq 3$$

La desviación típica:

$$\sigma = \sqrt{\lambda} = \sqrt{3} = 1.73$$

Y se puede comprobar (saldría exacto si se sumaran todos los términos hasta infinito),

$$\sigma = \sqrt{\sum_{x=0}^{\infty} (x-\mu)^2 p(x;3)} = 1.72 \simeq 1.73$$

Las aplicaciones de la distribución de Poisson son numerosas, desde el control de calidad y el muestreo de aceptación hasta problemas físicos en los que se mide el número de sucesos que se dan en un tiempo dado, o el número de casos que aparecen en una superficie. Recuerdese además que es una buena aproximación aplicar esta distribución a distribuciones binomiales con un gran número de ensayos y probabilidades pequeñas.

Ejemplo II–16

Aproximación de la distribución binomial a la de Poisson.

Sea un experimento binomial donde se realizan $n = 17$ ensayos. La probabilidad de éxito en cada uno es $p = 0.05$. Calcular la probabilidad de obtener $x = 4$ éxitos.

Usando las tablas con $n = 17, p = 0.05$,

$$P(x = 4) = b(4; 17, 0.05) = 0.008$$

Si la aproximamos por una distribución de Poisson,

$$p = \frac{\lambda}{n} \qquad \rightarrow \qquad \lambda = p\, n = 0.85$$

$$P(x = 4) \simeq p(4; 0.85) = \frac{0.85^4}{4!}\, e^{-0.85} = 0.009$$

La aproximación es mejor si el número de ensayos aumenta.

Por ejemplo para $n = 1000$, $p = 0.001$ y $x = 2$,

$$P(x = 2) = \begin{cases} b(2; 1000, 0.001) = \begin{pmatrix} 1000 \\ 2 \end{pmatrix} \times 0.001^2 \times 0.999^{1000-2} = 0.184 \\ p(2; 1) = \frac{1^2}{2!}\, e^{-1} = 0.184 \end{cases}$$

Capítulo 8

Distribuciones continuas de probabilidad

"¿Cómo nos atrevemos a hablar de leyes del azar? ¿No es el azar la antítesis de toda ley?"

Bertrand Russell (1872-1970)

En este tema se presentan algunas de las distribuciones continuas de probabilidad más comunes y frecuentemente utilizadas en Física. También resultan fundamentales a la hora de tomar decisiones en inferencia estadística y al realizar contrastes de hipótesis, como se estudiará más adelante.

8.1. Distribución continua uniforme

Se dice que una variable aleatoria X sigue una **distribución continua uniforme** cuando su función de densidad $f(x)$ toma valores constantes en el intervalo $[a, b]$. Es decir, $f(x) = K$ en ese intervalo y, por tanto, la probabilidad de que tome un valor en cualquier incremento (de la misma anchura) dentro de ese intervalo es la misma. Para calcular esa constante aplicamos la condición de normalización de la función de densidad

$$1 = \int_{-\infty}^{\infty} f(x)\ dx = \int_a^b f(x)\ dx = \int_a^b K\ dx = K(b-a) \quad \Rightarrow \quad K = \frac{1}{b-a}.$$

Por lo tanto la función de densidad tiene la forma

$$f(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a < x < b \\ 0 & x > b \end{cases} \tag{8.1}$$

Podemos además calcular la función de distribución $F(x)$. Cuando x esté en el intervalo $[a, b]$

$$F(x) = P(X < x) = \int_{-\infty}^{x} f(t)\ dt = \int_a^x \frac{1}{b-a}\ dt = \frac{x-a}{b-a},$$

y, en general,

$$F(x) = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & a < x < b \\ 1 & x > b \end{cases} \tag{8.2}$$

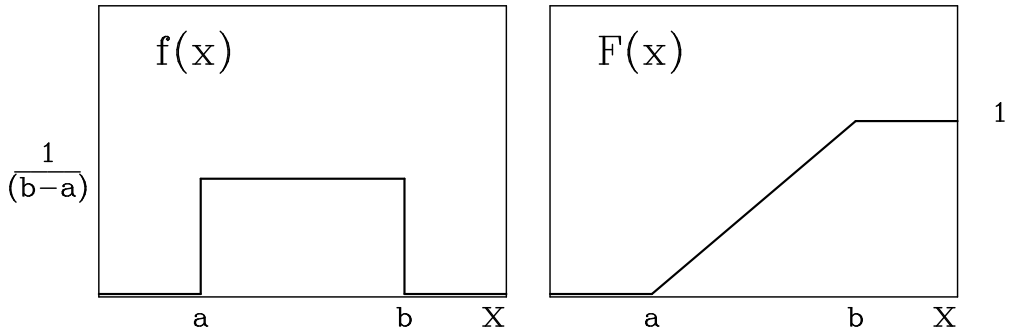


Figura 8.1: *Función de densidad, $f(x)$, y función de distribución, $F(x)$, para una distribución continua uniforme.*

La representación gráfica de la función de densidad y de la función de distribución será como la mostrada en la Figura 8.1.

La media, o esperanza matemática, de la distribución continua, se puede expresar como

$$\mu = \int_{-\infty}^{\infty} x f(x)\, dx = \int_a^b x \frac{dx}{b-a} = \frac{1}{b-a}\left[\frac{x^2}{2}\right]_a^b = \frac{b^2-a^2}{2(b-a)} = \frac{(a+b)(b-a)}{2(b-a)}$$

$$\Rightarrow \qquad \mu = \frac{a+b}{2}. \tag{8.3}$$

Por otra parte, la varianza puede calcularse como

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\, dx = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{dx}{b-a} =$$

$$\frac{1}{b-a}\left[\frac{x^3}{3} - \frac{a+b}{2}x^2 + \left(\frac{a+b}{2}\right)^2 x\right]_a^b.$$

Desarrollando se llega a la expresión para la varianza y la desviación típica

$$\sigma^2 = \frac{(b-a)^2}{12} \qquad ; \qquad \sigma = \frac{b-a}{\sqrt{12}}. \tag{8.4}$$

8.2. Distribución normal

La distribución continua de probabilidad más importante de toda la estadística es, sin duda alguna, la **distribución normal**. La importancia de esta distribución se debe a que describe con gran aproximación la distribución de las variables asociadas con muchos fenómenos de la naturaleza. En particular, las medidas de magnitudes físicas suelen distribuirse según una distribución normal. Por ejemplo, la distribución de alturas de un grupo de población, las medidas de calidad de procesos industriales, o la distribución de temperaturas de una población, se pueden aproximar por distribuciones normales. Además, los errores en las medidas también se aproximan con mucha exactitud a la distribución normal. Por otra parte, bajo ciertas condiciones, la distribución normal constituye una buena aproximación a otras distribuciones de probabilidad, como la binomial y la de Poisson. Frecuentemente, a la distribución normal se la denomina también distribución gaussiana.

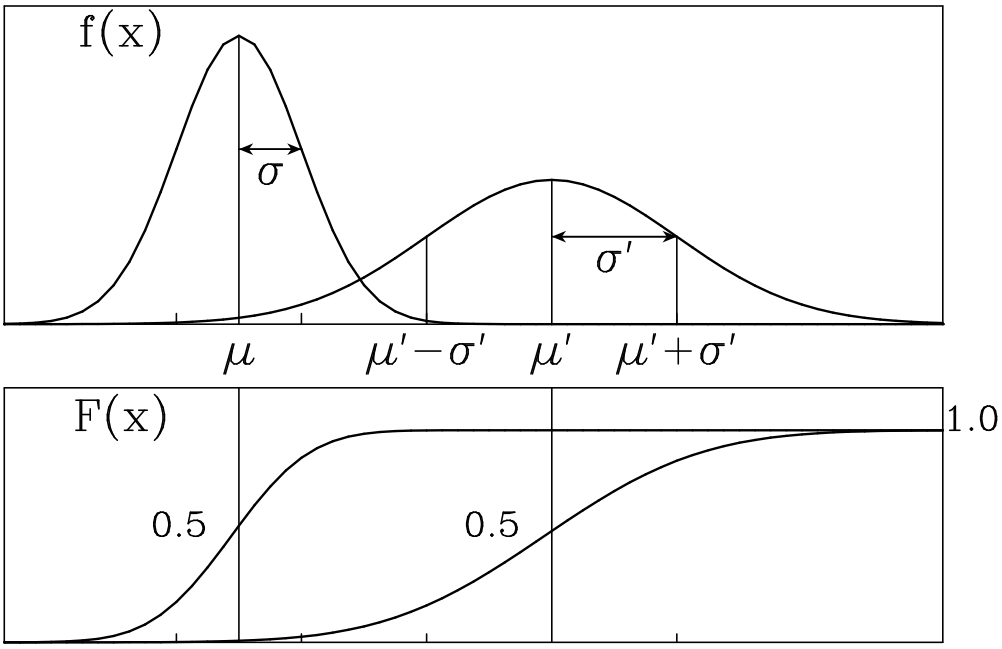


Figura 8.2: *Función de densidad, $f(x)$, y función de distribución, $F(x)$, para una distribución normal. Se muestran las representaciones correspondientes a dos valores de la media μ y la desviación típica σ.*

8.2.1. Definición y propiedades

Por definición, se dice que una variable aleatoria continua X sigue una **distribución normal** de media μ y desviación típica σ si su función de densidad es

$$f(x) = N(\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad ; \qquad -\infty < x < \infty \tag{8.5}$$

De esta forma, una vez que se especifican μ y σ la distribución queda determinada completamente. Puede comprobarse que esta distribución de probabilidad cumple la condición de normalización dada en (6.4), ya que

$$\int_{-\infty}^{\infty} f(x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\, dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\, dz = \frac{1}{\sqrt{2\pi}}\sqrt{2\pi} = 1, \tag{8.6}$$

donde se ha hecho el cambio de variable $z = (x-\mu)/\sigma$ (es decir $dx = \sigma\, dz$) y se ha aplicado el siguiente valor tabulado de la integral: $\int_{-\infty}^{\infty} e^{-ax^2}\, dx = \sqrt{\pi/a}$.

Gráficamente (Figura 8.2), la distribución de probabilidad normal tiene forma de campana (llamada campana de Gauss, o curva normal), simétrica (por depender de x a través del término $(x-\mu)^2$), centrada en μ y con anchura proporcional a σ (como es lógico esperar del significado de la desviación típica). Evidentemente, el máximo de la función de densidad ocurre para $x = \mu$ y, por tanto, media, mediana y moda coinciden en ese punto. Se puede demostrar que los puntos de inflexión de la curva normal están situados en $\mu-\sigma$ y $\mu+\sigma$. La curva tiende asintóticamente a cero al alejarse del valor medio. Además, por (8.6), el área entre la curva normal y el eje X es la unidad.

La función de distribución normal, útil para el cálculo de probabilidades, vendrá dada por

$$F(x) = P(X < x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt. \tag{8.7}$$

Es claro que la probabilidad de que X tome un valor entre x_1 y x_2 puede calcularse por

$$P(x_1 < X < x_2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx. \tag{8.8}$$

Se puede demostrar que, efectivamente, los parámetros μ y σ de la distribución normal coinciden con la media y la desviación típica de dicha distribución. Para el caso de la media

$$E(X) = \int_{-\infty}^{\infty} xf(x)\ dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} xe^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} (\mu + \sigma z)e^{-\frac{z^2}{2}}\ dz,$$

donde hemos aplicado el mismo cambio de variables que anteriormente ($z = (x-\mu)/\sigma$). Separando la integral en dos términos

$$E(X) = \frac{\mu}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\ dz + \frac{\sigma}{\sqrt{2\pi}} \int_{-\infty}^{\infty} ze^{-\frac{z^2}{2}}\ dz =$$

$$= \frac{\mu}{\sqrt{2\pi}}\sqrt{2\pi} + \frac{\sigma}{\sqrt{2\pi}} \left[-e^{-\frac{z^2}{2}}\right]_{-\infty}^{\infty} = \mu,$$

como queríamos demostrar. Para la varianza

$$\mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\ dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (x-\mu)^2 e^{-\frac{(x-\mu)^2}{2\sigma^2}}\ dx =$$

$$= \frac{\sigma^2}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^2 e^{-\frac{z^2}{2}}\ dz,$$

donde se ha hecho el mismo cambio de variable. Integrando ahora por partes haciendo $u = z$, $dv = ze^{-z^2/2}\,dz$, de forma que: $du = dz$ y $v = -e^{-z^2/2}$, se obtiene

$$\mathrm{Var}(X) = \frac{\sigma^2}{\sqrt{2\pi}} \left(-ze^{-\frac{z^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} e^{-\frac{z^2}{2}}\ dz \right) = \frac{\sigma^2}{\sqrt{2\pi}} \left(0 + \sqrt{2\pi}\right) = \sigma^2.$$

8.2.2. Distribución normal tipificada

La dificultad de integración de las ecuaciones (8.7) y (8.8) para calcular probabilidades de una distribución hace que sea sumamente útil presentar las áreas bajo la curva normal en forma tabular. Para no tener que presentar estas tablas para todos los posibles valores de μ y σ se define la **variable normal tipificada** Z a partir de una transformación lineal de la variable original X de la forma

$$Z = \frac{X - \mu}{\sigma}. \tag{8.9}$$

Haciendo esta sustitución en la función de densidad de X ($f(x)dx = f(z)dz$)

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \Rightarrow f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} = N(0,1). \tag{8.10}$$

Por lo tanto, la variable tipificada sigue una distribución normal con media 0 y desviación típica 1, llamada **función de densidad tipificada**, o estándar. Es claro que esta distribución no depende de ningún parámetro y su representación gráfica es una campana simétrica respecto al eje z=0, en el que alcanza el máximo valor.

El problema de calcular la probabilidad de que X se encuentre en un intervalo (x_1, x_2) se puede reducir entonces a calcular la probabilidad de que Z esté en un intervalo equivalente (z_1, z_2)

$$P(x_1 < X < x_2) = P(z_1 < Z < z_2), \qquad \text{con} \qquad z_1 = \frac{x_1 - \mu}{\sigma} \quad \text{y} \quad z_2 = \frac{x_2 - \mu}{\sigma}.$$

Por lo tanto, usando la variable tipificada sólo es necesario trabajar con una tabla de la distribución

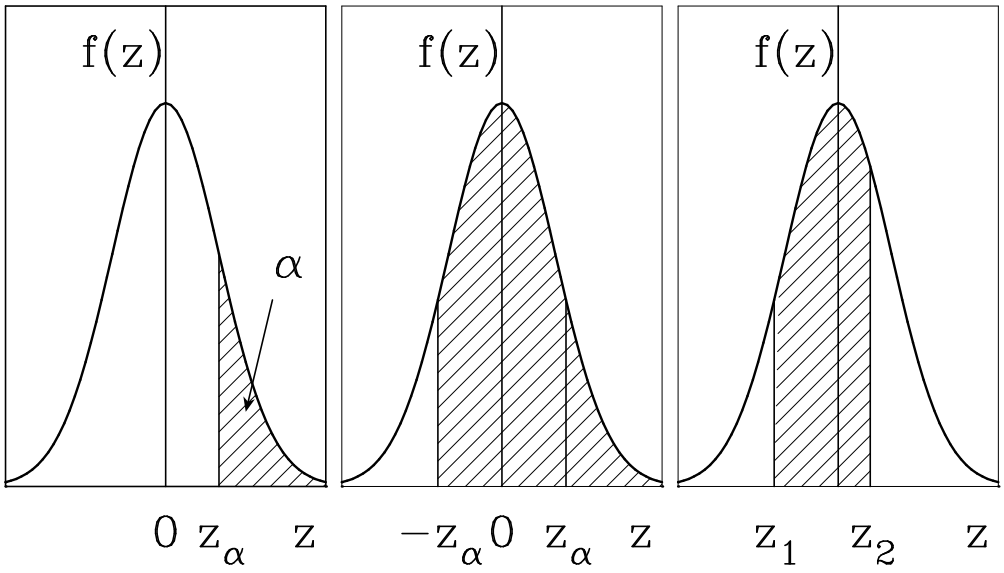


Figura 8.3: *Determinación de la probabilidad para la distribución normal tipificada.*

normal. En la **Tabla IV** (Apéndice A) se presentan las probabilidades de que Z tenga un valor mayor que un z_α dado. Se tabulan únicamente los valores de $z_\alpha \geq 0$. Es lo que se conoce como la áreas de la cola derecha de la distribución

$$P(Z > z_\alpha) = \alpha = \frac{1}{\sqrt{2\pi}} \int_{z_\alpha}^{\infty} e^{-\frac{z^2}{2}}\, dz$$

$$Ejemplo: \qquad P(Z > 1.75) = 0.0401$$

Para calcular la probabilidad de que Z esté por debajo de un determinado valor z_α se usará, por el condición de normalización

$$P(Z < z_\alpha) = 1 - P(Z > z_\alpha) = 1 - \alpha$$

$$Ejemplo: \qquad P(Z < 1.75) = 1 - 0.0401 = 0.9599$$

Asimismo, si z_α fuese negativo, por ser la curva simétrica

$$P(Z > (-z_\alpha)) = 1 - P(Z < (-z_\alpha)) = 1 - P(Z > z_\alpha) = 1 - \alpha$$

$$Ejemplo: \qquad P(Z > -1.75) = 0.9599$$

y la probabilidad de que Z esté entre dos valores se calcula por

$$P(z_1 < Z < z_2) = P(Z > z_1) - P(Z > z_2)$$

$$Ejemplo: \qquad P(-1 < Z < 0.5) = P(Z > -1) - P(Z > 0.5) =$$

$$= (1 - P(Z > 1)) - P(Z > 0.5) = 1 - 0.1587 - 0.3085 = 0.5328$$

como puede comprobarse en las gráficas (Figura 8.3).

En particular, puede calcularse la probabilidad de que Z se encuentre en el intervalo $(-1, 1)$, correspondiente a un intervalo $(\mu - \sigma, \mu + \sigma)$ para cualquier distribución normal

$$P(\mu - \sigma < X < \mu + \sigma) = P(-1 < Z < 1) = P(Z > -1) - P(Z > 1) =$$

$$= (1 - P(Z > 1)) - P(Z > 1) = 1 - 2P(Z > 1) = 1 - 2 \times 0.1587 = 0.6826$$

De manera análoga

$$P(\mu - 2\sigma < X < \mu + 2\sigma) = P(-2 < Z < 2) = 0.9544$$

$$P(\mu - 3\sigma < X < \mu + 3\sigma) = P(-3 < Z < 3) = 0.9973$$

Nótese que estas probabilidades son más precisas que las que daba el teorema de Chebyshev, que indicaba que las probabilidades eran, como mínimo 0.0, 0.75 y 0.89, para 1σ, 2σ y 3σ respectivamente.

8.2.3. Relación con otras distribuciones

Existe un teorema básico en estadística que explica porqué la distribución normal es tan frecuente. El teorema es el siguiente:

Teorema del límite central: Si $X_1, X_2, \ldots, X_n$ son variables aleatorias independientes con medias μ_i, desviaciones típicas σ_i, y distribuciones de probabilidad cualesquiera (y no necesariamente la misma), y definimos la variable suma $Y = X_1 + X_2 + \ldots + X_n$, entonces, cuando n crece, la variable

$$Z = \frac{Y - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}}$$

tiende hacia una distribución normal estándar $N(0, 1)$. Es decir, las probabilidades de Y las podremos calcular utilizando la distribución normal $N(\sum \mu_i, \sqrt{\sum \sigma_i^2})$. Esto explica por qué una medida de un fenómeno natural que está influenciado por un gran número de efectos (con cualquier distribución) ha de de seguir una distribución normal. Hay que indicar además que, cuando las variables X_i siguen distribuciones normales, no es necesario que n sea grande para que la variable suma siga una distribución normal. Este teorema es de gran utilidad en temas posteriores.

El teorema del límite central además nos permite relacionar otras distribuciones con la distribución normal. En particular, el cálculo de probabilidades de la distribución binomial puede efectuarse usando tablas, pero puede hacerse muy complicado cuando n (número de ensayos) se hace muy grande, superando los valores tabulados. Para estos casos, la distribución normal supone una buena aproximación a la distribución binomial. En particular, si X es una variable aleatoria binomial con media $\mu = np$ y desviación típica $\sigma = \sqrt{npq}$, la variable

$$Z = \frac{X - np}{\sqrt{npq}} \tag{8.11}$$

sigue la distribución normal tipificada (o estándar) cuando n tiende a infinito (teorema de Moivre). Esto es una consecuencia inmediata del teorema del límite central ya que la variable binomial puede considerarse, como ya vimos, como la suma de n variables de Bernoulli con media $\mu = p$ y varianza $\sigma^2 = pq$, de forma que

$$Z = \frac{X - \sum_{i=1}^{n} \mu_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} = \frac{X - \sum_{i=1}^{n} p}{\sqrt{\sum_{i=1}^{n} pq}} = \frac{X - np}{\sqrt{npq}}.$$

Esta importante propiedad se puede comprobar además empíricamente calculando probabilidades binomiales y normales. Como la distribución binomial se hace más simétrica cuando p es próximo a 0.5, la distribución tiende más rápidamente a la normal para esos valores de p. Para p próximos a 0 ó 1, habrá que aumentar mucho n para que la asimetría, clara para un número pequeño de ensayos, desaparezca. Como regla práctica podemos considerar que la distribución normal es una aproximación aceptable de la distribución binomial cuando tanto np como nq sean mayor que 5 ($n\,p > 5; n\,q > 5$). Esto quiere decir que si $p = 0.5$, bastará con que $n = 10$ para que la aproximación sea aceptable, pero para $p = 0.1$, será necesario que el número de

ensayos sea, al menos, 50.

De forma similar existe una relación entre la distribución normal y la de Poisson. En particular, si X es una variable aleatoria de Poisson con parámetro λ, la variable

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \tag{8.12}$$

sigue la distribución normal estándar cuando λ tiende a infinito. Es decir, la distribución de Poisson se puede aproximar a la normal con parámetros $\mu = \lambda$ y $\sigma = \sqrt{\lambda}$ (Recordemos que λ era la media y la varianza de la distribución de Poisson). Esta aproximación empieza a ser aceptable para $\lambda > 5$. Es también una consecuencia del teorema del límite central, ya que la variable de Poisson se puede considerar como la suma de muchas variables de Poisson subdiviendo el intervalo de medida.

La aplicación de la distribución normal es entonces muy útil para calcular probabilidades de la distribución binomial o de Poisson cuando n (ó λ) es grande. Hay que tener en cuenta que al pasar de una variable discreta X a una continua X' habrá que utilizar la, llamada, *corrección de continuidad*, que consiste en calcular las probabilidades como

$$P(x_1 \leq X \leq x_2) = P(x_1 - 0.5 < X' < x_2 + 0.5).$$

8.3. Distribución χ^2 de Pearson

Sean $X_1, X_2, \ldots, X_n$ n variables aleatorias normales con media 0 y varianza 1 independientes entre sí, entonces la variable

$$\chi^2_n = X_1^2 + X_2^2 + \ldots + X_n^2 \tag{8.13}$$

recibe el nombre de χ^2 (*chi–cuadrado*) con n grados de libertad. La función de densidad asociada es la **distribución χ^2 de Pearson**, que se puede expresar como

$$f(x) = \begin{cases} \frac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{8.14}$$

donde $\Gamma(\alpha)$ es la función gamma, definida, para cualquier real positivo α, como

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx \qquad \text{con} \qquad \alpha > 0. \tag{8.15}$$

Nótese que la variable χ^2 toma únicamente valores positivos, al ser una suma de cuadrados. Además su distribución depende únicamente del parámetro n, o número de grados de libertad. Gráficamente, su función de densidad es muy asimétrica (para $n = 1$ corresponde a elevar al cuadrado una curva normal tipificada), pero se va haciendo más simétrica a medida que n aumenta.

En particular, para $n \geq 30$, es una buena aproximación suponer que la variable $\sqrt{2\chi^2_n}$ se distribuye como una distribución normal con media $\sqrt{2n-1}$ y varianza 1 ($N(\sqrt{2n-1}, 1)$).

Una propiedad importante de la distribución χ^2 es que si $\chi^2_{n_1}$ y $\chi^2_{n_2}$ son dos variables χ^2 con grados de libertad n_1 y n_2 respectivamente, entonces la variable suma $\chi^2_n = \chi^2_{n_1} + \chi^2_{n_2}$ es una χ^2 con $n = n_1 + n_2$ grados de libertad. Esto es evidente a partir de la definición dada en (8.13).

La media y la varianza de la distribución χ^2_n están dadas por

$$\mu = n \qquad ; \qquad \sigma^2 = 2n. \tag{8.16}$$

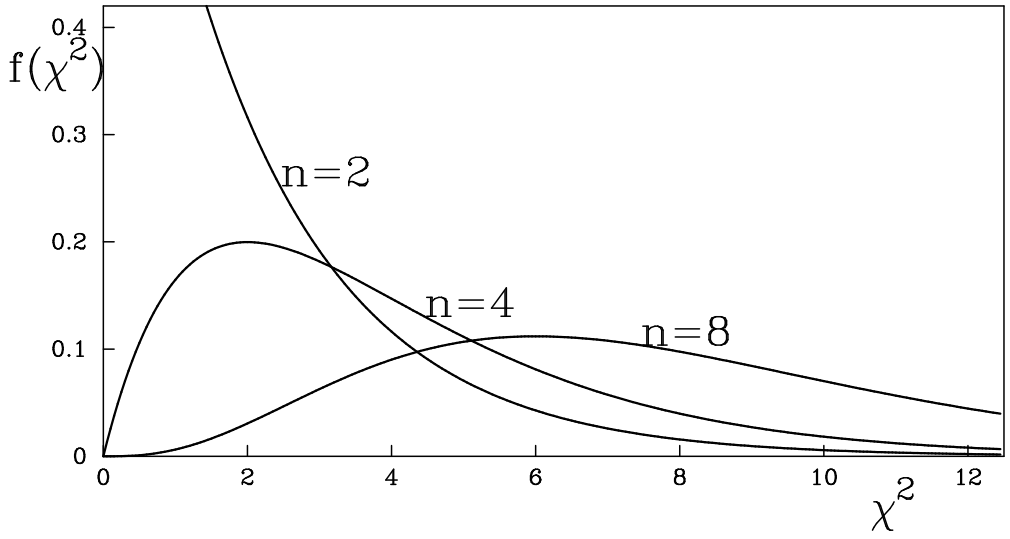


Figura 8.4: *Distribuciones* χ^2.

Para demostrar estas relaciones partimos de la definición de χ^2 (8.13) y utilizamos la propiedad de la media y varianza de una suma de variables independientes

$$\mu = E(\chi_n^2) = E\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} E\left(X_i^2\right),$$

$$\sigma^2 = \mathrm{Var}(\chi_n^2) = \mathrm{Var}\left(\sum_{i=1}^{n} X_i^2\right) = \sum_{i=1}^{n} \mathrm{Var}\left(X_i^2\right).$$

Es necesario entonces calcular la media y la varianza de un variable X_i^2. Puesto que X_i es normal con media 0 y varianza 1, se cumple

$$\sigma_{X_i}^2 = E\left(X_i^2\right) - \mu_{X_i}^2 \quad \Rightarrow \quad 1 = E\left(X_i^2\right) - 0 \quad \Rightarrow \quad E\left(X_i^2\right) = 1.$$

Para calcular la varianza de X_i^2 hacemos

$$\mathrm{Var}\left(X_i^2\right) = \sigma_{X_i^2}^2 = E\left(X_i^4\right) - \mu_{X_i^2}^2 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^4 e^{-\frac{x^2}{2}}\, dx - E\left(X_i^2\right)^2.$$

Integrando por partes con $u = x^3$ y $dv = xe^{-x^2/2}dx$ ($\Rightarrow du = 3x^2\, dx$, $v = -e^{-x^2/2}$)

$$\mathrm{Var}\left(X_i^2\right) = \frac{1}{\sqrt{2\pi}} \left[-x^3 e^{-\frac{x^2}{2}} \Big|_{-\infty}^{\infty} + \int_{-\infty}^{\infty} 3x^2 e^{-\frac{x^2}{2}}\, dx \right] - 1^2 =$$

$$= \frac{3}{\sqrt{2\pi}} \int_{-\infty}^{\infty} x^2 e^{-\frac{x^2}{2}}\, dx - 1 = 3E\left(X_i^2\right) - 1 = 2.$$

Y, por lo tanto,

$$\mu = \sum_{i=1}^{n} E\left(X_i^2\right) = \sum_{i=1}^{n} 1 = n,$$

$$\sigma^2 = \sum_{i=1}^{n} \mathrm{Var}\left(X_i^2\right) = \sum_{i=1}^{n} 2 = 2n.$$

Estas expresiones se pueden también demostrar integrando directamente en la definición de media y varianza usando (8.14).

Para calcular las probabilidades de que la variable χ^2 tome valores por encima o debajo de un determinado valor puede usarse la **Tabla V** (Apéndice A). En ésta se dan las abscisas, denotadas por $\chi_{\alpha,n}$, que dejan a su derecha un área (o probabilidad) bajo la función de densidad igual a cierto valor α, llamado *nivel de significación*. Es decir

$$P(\chi_n^2 > \chi_{\alpha,n}^2) = \alpha \qquad \text{y} \qquad P(\chi_n^2 < \chi_{\alpha,n}^2) = 1 - \alpha.$$

La importancia de la distribución χ^2 en estadística se basa en la siguiente propiedad: Sea σ^2 la varianza de una población normal y s^2 la varianza de una muestra de tamaño n extraída al azar de dicha población. Entonces la variable aleatoria que cambia de muestra a muestra y viene dada por

$$\chi_{n-1}^2 = (n-1)\frac{s^2}{\sigma^2}, \tag{8.17}$$

obedece a una distribución χ^2 con $(n-1)$ grados de libertad. Esta propiedad es sumamente importante para la estimación de la varianza y el contraste de hipótesis sobre la varianza σ^2.

8.4. Distribución t de Student

Sean $X_1, X_2, \ldots, X_n$ y X, $n+1$ variables aleatorias normales con media 0 y desviación típica σ independientes entre sí, entonces la variable

$$t_n = \frac{X}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} X_i^2}} \tag{8.18}$$

recibe el nombre de t **de Student** con n grados de libertad. Podemos llegar a una expresión más usual de la variable t dividiendo numerador y denominador por la desviación típica σ

$$t_n = \frac{\frac{X}{\sigma}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{X_i}{\sigma}\right)^2}} = \frac{Z}{\sqrt{\frac{1}{n}\chi_n^2}}, \tag{8.19}$$

donde Z es una variable que sigue una distribución normal estándar $N(0,1)$ y χ_n^2 es una χ^2 con n grados de libertad, siendo ambas independientes.

La función de densidad asociada es la **distribución t de Student** (introducida por W.S. Gosset), que se puede expresar como

$$f(x) = f(t) = \frac{1}{\sqrt{n}\beta\left(\frac{1}{2}, \frac{n}{2}\right)}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}} \qquad ; \qquad -\infty < t < \infty \tag{8.20}$$

donde $\beta(p,q)$ es la función beta, definida, para un par de reales p y q positivos, haciendo uso de la función gamma, como

$$\beta(p,q) = \frac{\Gamma(p)\Gamma(q)}{\Gamma(p+q)}. \tag{8.21}$$

La demostración de que la variable t definida en (8.19) sigue la función de densidad anterior está fuera del alcance de este libro.

El campo de variabilidad de la variable t de Student será de $-\infty$ a ∞ y su función de densidad dependerá únicamente del parámetro n (grados de libertad). Nótese que, al depender $f(t)$ de t a través de t^2, la función de densidad será simétrica alrededor de $t = 0$. Su forma será campaniforme, siendo más achatada para valores bajos de n.

Cuando n aumenta $f(t)$ se va haciendo cada vez más apuntada, tendiendo a la curva normal tipificada

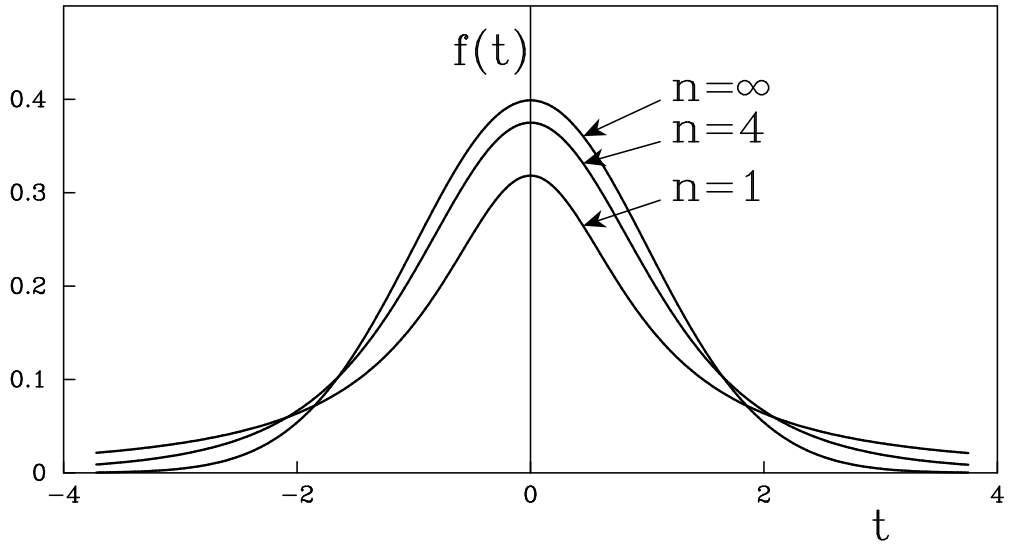


Figura 8.5: *Distribución t de Student.*

$(N(0,1))$ cuando $n \to \infty$. En general, la curva normal es una buena aproximación de la distribución t cuando $n \geq 30$.

La media y la varianza de la distribución t vienen dadas por

$$\mu = 0 \qquad ; \qquad \sigma^2 = \frac{n}{n-2} \qquad (\text{para } n > 2). \tag{8.22}$$

Es evidente que, al ser $f(t)$ simétrica respecto a $t = 0$, la media ha de ser nula. Respecto a la varianza, nótese que es mayor que 1 y depende del número de grados de libertad. Sólo al hacerse n muy grande, σ tiende a 1, y, por tanto, a la distribución normal estándar.

Para calcular las áreas debajo de la distribución t se puede usar la **Tabla VI** (Apéndice A). Al igual que con la distribución χ^2, ésta da las abscisas, denotadas por $t_{\alpha,n}$, que dejan a su derecha un área (o probabilidad) bajo la función de densidad igual a cierto valor α, llamado *nivel de significación*. Es decir

$$P(t_n > t_{\alpha,n}) = \alpha \qquad \text{y} \qquad P(t_n < t_{\alpha,n}) = 1 - \alpha.$$

Para valores de t negativos, al ser la distribución simétrica, se cumple

$$P(t_n > -t_{\alpha,n}) = 1 - P(t_n < -t_{\alpha,n}) = 1 - P(t_n > t_{\alpha,n}) = 1 - \alpha,$$

$$P(t_n < -t_{\alpha,n}) = \alpha,$$

además de

$$t_{\alpha,n} = -t_{1-\alpha,n},$$

relación muy útil para calcular valores de t que dan $\alpha > 0.5$, que no vienen tabulados en las tablas.

La distribución t de Student es sumamente importante para la estimación y el contraste de hipótesis sobre la media de una población, como se verá en temas posteriores. Si se tiene una población que sigue una distribución normal con media μ y desviación típica σ $(N(\mu,\sigma))$, y se extrae una muestra aleatoria de tamaño n sobre la que se calcula una media $\overline{x}$ y una desviación típica s, entonces la variable aleatoria dada por

$$t_{n-1} = \frac{\overline{x} - \mu}{s/\sqrt{n}} \tag{8.23}$$

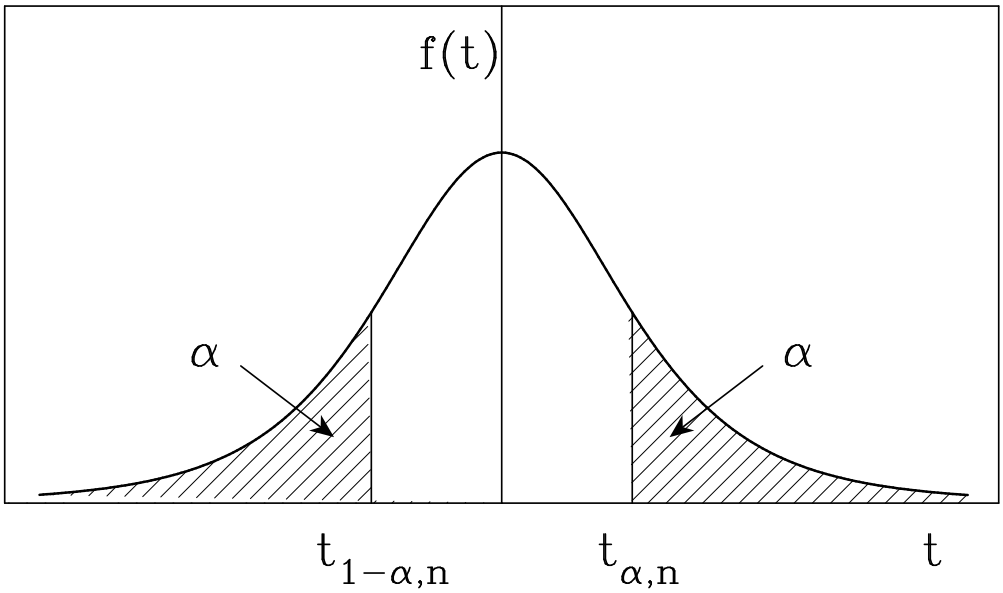


Figura 8.6: *Distribución t de Student. Simetría y* $P(t_n < -t_{\alpha,n}) = \alpha$ *y* $t_{\alpha,n} = -t_{1-\alpha,n}$.

obedece a una distribución t de Student con $(n-1)$ grados de libertad.

8.5. Distribución F de Fisher

Sean $\chi^2_{n_1}$ y $\chi^2_{n_2}$ dos variables χ^2 de Pearson con n_1 y n_2 grados de libertad e independientes entre sí. Entonces, la variable aleatoria definida como

$$F_{n_1,n_2} = \frac{\dfrac{\chi^2_{n_1}}{n_1}}{\dfrac{\chi^2_{n_2}}{n_2}} \tag{8.24}$$

recibe el nombre de F **de Fisher** con n_1 y n_2 grados de libertad.

La función de densidad asociada es la **distribución** F **de Fisher**, cuya expresión es la siguiente

$$f(x) = f_{n_1,n_2}(x) = \begin{cases} \dfrac{\Gamma\left(\frac{n_1+n_2}{2}\right)\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \dfrac{x^{(n_1/2)-1}}{\left(1+\frac{n_1}{n_2}x\right)^{(n_1+n_2)/2}} & x > 0 \\ 0 & x \leq 0 \end{cases} \tag{8.25}$$

Nótese que el campo de variabilidad de la variable F es entre 0 e ∞ (al ser un cociente de cuadrados) y que su función de densidad depende exclusivamente de los dos parámetros n_1 y n_2, aunque es importante el orden en el que se dan estos. En particular, por la definición de F dada en (8.24), se cumple

$$F_{n_1,n_2} = \frac{1}{F_{n_2,n_1}}. \tag{8.26}$$

La representación gráfica de la distribución F será de la forma representada en la figura y dependerá, lógicamente, de n_1 y n_2.

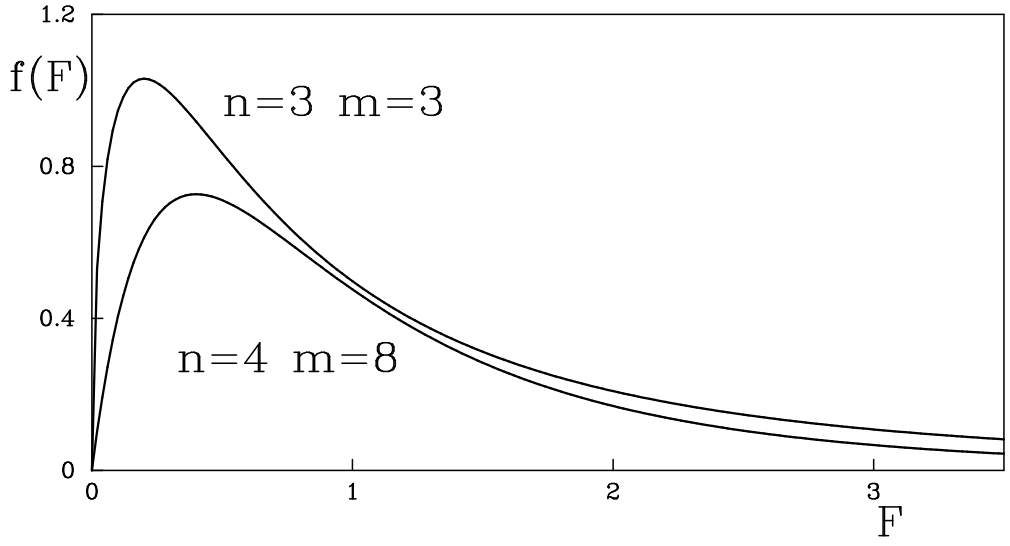


Figura 8.7: *Distribución F de Fisher.*

Se puede demostrar que la media y la varianza de la distribución F de Fisher vienen dadas por

$$\mu = \frac{n_2}{n_2 - 2} \quad (n_2 > 2) \qquad ; \qquad \sigma^2 = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2 - 4)(n_2 - 2)^2} \quad (n > 4), \tag{8.27}$$

y que la media sólo depende de n_2.

Las áreas bajo la curva de la distribución F se pueden calcular usando la **Tabla VII** (Apéndice A). Esta da, en función de n_1 y n_2, las abscisas, denotadas por $F_{\alpha;n_1,n_2}$, que dejan a su derecha un área (o probabilidad) bajo la función de densidad igual a cierto valor α, llamado *nivel de significación*. Por tanto

$$P(F_{n_1,n_2} > F_{\alpha;n_1,n_2}) = \alpha \qquad \text{y} \qquad P(F_{n_1,n_2} < F_{\alpha;n_1,n_2}) = 1 - \alpha$$

En dicha Tabla se tabulan los valores de $F_{\alpha;n_1,n_2}$ para valores de α próximos a 0. Para α cercano a 1, puede usarse la propiedad dada en (8.26), de forma que

$$F_{1-\alpha;n_2,n_1} = \frac{1}{F_{\alpha;n_1,n_2}}.$$

Es importante notar que las distribuciones χ^2 y t son en realidad casos particulares de la distribución F, ya que

$$F_{1,n} = t_n^2 \qquad ; \qquad F_{n,\infty} = \frac{\chi_n^2}{n},$$

como puede comprobarse fácilmente (Nótese que χ_1^2 es una variable que sigue una distribución normal tipificada).

La distribución F de Fisher es muy utilizada en el análisis de varianza y, en particular, es usada para comparar las varianzas de dos poblaciones normales. Efectivamente, sea X_1 una variable aletoria normal $N(\mu_1, \sigma_1)$ y X_2 una variable normal $N(\mu_2, \sigma_2)$, independientes entre sí. Si de la primera población se extrae una muestra aleatoria de tamaño n_1 en la cual se mide una desviación típica s_1, y de la segunda población se extrae una muestra de tamaño n_2, con desviación típica s_2, entonces, por la propiedad (8.17) se pueden definir las variables χ^2

$$\chi_{n_1-1}^2 = (n_1 - 1)\frac{s_1^2}{\sigma_1^2} \qquad ; \qquad \chi_{n_2-1}^2 = (n_2 - 1)\frac{s_2^2}{\sigma_2^2},$$

de forma que se puede construir la variable F dada por

$$F_{n_1-1,n_2-1} = \frac{\dfrac{\chi^2_{n_1-1}}{n_1-1}}{\dfrac{\chi^2_{n_2-1}}{n_2-1}}.$$

En otras palabras, si s_1^2 y s_2^2 son las varianzas de variables aleatorias independientes de tamaños n_1 y n_2 que se extraen de poblaciones normales con varianzas σ_1^2 y σ_2^2 respectivamente, entonces la variable

$$F_{n_1-1,n_2-1} = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{8.28}$$

sigue una distribución F de Fisher con $n_1 - 1$ y $n_2 - 1$ grados de libertad. En particular, si $\sigma_1 = \sigma_2$

$$F_{n_1-1,n_2-1} = \frac{s_1^2}{s_2^2}.$$

Tema III

INFERENCIA ESTADÍSTICA

Capítulo 9

Teoría elemental del muestreo

"Lo malo del infinito es que es muy muy largo, especialmente la última parte."

Woody Allen (1935-)

Uno de los objetivos principales de la estadística es extraer conclusiones e información sobre una determinada población. Recordemos que por **población** se denomina al conjunto completo de elementos, con alguna característica común, objeto de nuestro estudio (personas, objetos, experimentos, etc.). Evidentemente, la forma más directa de cumplir dicho objetivo sería estudiar todos y cada uno de los elementos de la población. Sin embargo, en numerosas ocasiones esto no es posible ya que, por ejemplo, el tamaño de la población puede ser demasiado grande (ej. estrellas del cielo) e incluso infinito (ej. tiradas posibles de un dado), o porque estudiar los elementos supone la destrucción de estos (ej. ensayos destructivos de control de calidad) o, simplemente, porque el coste económico es prohibitivo. En estos casos, es necesario trabajar con un subconjunto de elementos de la población, es decir una **muestra**. Al proceso de obtener muestras se le denomina **muestreo**.

La **inferencia estadística** se ocupa de estudiar los métodos necesarios para extraer, o inferir, conclusiones válidas e información sobre una población a partir del estudio experimental de una muestra de dicha población. Los métodos utilizados en la inferencia estadística dependen de la información previa que se tenga de la población a estudiar. Cuando se conoce la forma de la distribución de probabilidad que sigue la variable aleatoria a estudiar en la población, el problema consiste en determinar los diferentes parámetros de dicha distribución (ej. media y varianza para la distribución normal). Para ello se utilizan los **métodos paramétricos**, consistentes en procedimientos óptimos para encontrar dichos parámetros. Cuando la distribución de la población es desconocida, el problema principal es encontrar la forma y características de la distribución, lo cual se hace mediante los llamados **métodos no paramétricos**. En este capítulo y en los dos siguientes nos limitaremos a estudiar los principales métodos paramétricos de inferencia estadística.

9.1. Conceptos básicos

Para poder estudiar correctamente una población mediante la inferencia estadística es fundamental que la muestra esté bien escogida. La clave de un proceso de muestreo es que la muestra sea representativa de la población. Una forma de conseguir esto es haciendo que todos los elementos de la población tengan la misma probabilidad de ser elegidos para la muestra. Diremos en este caso que tenemos un **muestreo aleatorio**. Para realizar estos muestreos aleatorios se utilizan a menudo tablas de números aleatorios.

Por otra parte, cuando cada elemento de la población pueda seleccionarse más de una vez tendremos un **muestreo con reemplazamiento**, mientras que cuando cada elemento sólo se puede seleccionar una única vez será un **muestreo sin reemplazamiento**. Evidentemente, una población finita muestreada con reemplazamiento puede considerarse infinita. Si la población es infinita, o el tamaño de ésta (N) es muy grande comparado con el tamaño de la muestra (n), es prácticamente indiferente que el muestreo sea con o sin reemplazamiento. Como veremos, normalmente el análisis se simplifica cuando la población es infinita o el muestreo es con reemplazamiento.

Supongamos que tenemos una población de la cual conocemos la distribución de probabilidad $f(x)$ que sigue su variable aleatoria asociada X. Se dirá que tenemos una población normal, binomial, etc. cuando $f(x)$ corresponda a una distribución normal, binomial, etc. Para poder conocer la población objeto de nuestro estudio es necesario calcular los parámetros que definen su distribución de probabilidad, por ejemplo, la media μ y la desviación típica σ para una distribución normal, o la probabilidad de éxito p para una distribución binomial. Estas cantidades que definen la distribución de la población son los **parámetros poblacionales**.

El problema se concreta entonces en calcular, o estimar, los parámetros poblacionales. Para ello se toma una muestra aleatoria de la población. Para caracterizar una muestra aleatoria de tamaño n vamos a definir las variables aleatorias $X_i, i = 1, 2, \ldots, n$, que representan las medidas o valores muestrales que se observen. Así, en una muestra en particular, dichas variables aleatorias tomarán los valores numéricos $x_i, i = 1, 2, \ldots, n$. Nótese que cada una de las variables aleatorias X_i seguirá la misma distribución de probabilidad $f(x)$ de la población. En el caso de un muestreo con reemplazamiento las diferentes X_i serán independientes entre sí (el valor que tome una X_i particular no dependerá de los valores que se hayan obtenido anteriormente) y, por tanto, la distribución de probabilidad conjunta podrá expresarse como

$$P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n). \tag{9.1}$$

Para poder estimar los parámetros poblacionales se usan las medidas de las variables aleatorias X_i que definen la muestra. Por ejemplo, como veremos más adelante, para estimar la media de una población normal, se calcula la media aritmética de los diferentes valores x_i que se observan en la muestra. Dicha media aritmética es una función de las variables aleatorias X_i. En general, a cualquier función $g(X_1, X_2, \ldots, X_n)$ de las variables aleatorias que constituyen una muestra aleatoria se le llama **estadístico**. Es importante indicar que a cada parámetro poblacional le corresponderá un estadístico de la muestra, que constituirá una estimación del primero. Por ejemplo, para estimar el parámetro poblacional *media* calcularemos el estadístico muestral consistente en la media aritmética de los valores x_i. Para distinguir valores de la población de los valores medidos en la muestra, se denotarán por letras griegas (μ, σ, etc.) los parámetros poblacionales y por letras romanas ($\overline{X}$, S, etc.) los estadísticos de la muestra.

Al ser una función de variables aleatorias, una estadístico de la muestra se podrá considerar también como una variable aleatoria, es decir, podrá obtener diferentes valores dependiendo de la muestra en particular que se elija. Tendrá, por lo tanto, una distribución de probabilidad asociada. A ésta se le llama **distribución muestral** del estadístico. Dicho de otra forma, consideremos todas las muestras posibles que se pueden extraer de una población. Sin en cada una de estas muestras se midiese un estadístico, por ejemplo la media, éste tomaría valores diferentes, que se distribuirían en una determinada distribución muestral. Puesto que los estadísticos van a ser la base para la estimación de los parámetros poblacionales, es sumamente importante estudiar sus distribuciones, para así verificar su utilidad como estimadores. A continuación se estudian los principales estadísticos y sus distribuciones muestrales.

9.2. Media muestral

El primer estadístico importante es la media muestral. Si tenemos una muestra aleatoria de tamaño n representada por las variables aleatorias $X_i, i = 1, 2, \ldots, n$, se define la **media muestral**, o media de la muestra, como

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}. \tag{9.2}$$

Evidentemente, cuando las variables aleatorias X_i tomen, en una muestra, los valores particulares x_i, el valor que tendrá la media muestral vendrá dado por

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

9.2.1. Distribución muestral de la media

Al ser una combinación lineal de variables aleatorias, la media muestral es asimismo una nueva variable aleatoria y tendrá asociada una distribución de probabilidad. Es decir, consideremos una población de la que se toman diferentes muestras de tamaño n, calculando para cada muestra la media $\overline{x}$. Si tomamos k muestras distintas, obtendremos k valores, en general diferentes, de medias muestrales $\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k$. Si hacemos que k tienda a infinito, los valores $\overline{x}_i$ tendrán una distribución llamada **distribución muestral de la media**.

Vamos a calcular la media y la varianza de la distribución muestral de la media. Supongamos que tenemos una población con una distribución de probabilidad $f(x)$ caracterizada por los parámetros poblacionales media μ y varianza σ^2 y que tomamos una muestra de tamaño n representada por las variables aleatorias $X_i, i = 1, 2, \ldots, n$. Puesto que cada X_i sigue la misma distribución de probabilidad $f(x)$ de la población, con media μ, la media, o esperanza matemática, de cada X_i será

$$E(X_i) = \mu_{X_i} = \mu.$$

De forma que podemos calcular la media, o esperanza matemática, de la distribución muestral de la media, como

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) = \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right) = \frac{1}{n}(n\mu)$$

$$\Rightarrow \qquad \mu_{\overline{X}} = E(\overline{X}) = \mu. \tag{9.3}$$

Es decir, el valor esperado de la media muestral es la media de la población. Este resultado es sumamente importante.

De forma similar se puede calcular la varianza de la distribución muestral de la media. Puesto que la varianza de cada X_i coincide con la varianza de la población σ^2

$$\mathrm{Var}(X_i) = \sigma^2_{X_i} = \sigma^2,$$

podemos calcular la varianza de la distribución de la media utilizando la expresión para la varianza de una combinación lineal de variables aleatorias. Para ello vamos a suponer que el muestreo es con reemplazamiento o, equivalentemente, que la población es infinita. En este caso, las diferentes X_i son independientes y podemos hacer el siguiente desarrollo (Recuérdese que para variables aleatorias independientes se cumple $\sigma^2_{aX+bY} =$

$a^2\sigma_X^2 + b^2\sigma_Y^2$)

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \ldots + X_n}{n}\right) =$$

$$= \frac{1}{n^2}\mathrm{Var}(X_1) + \frac{1}{n^2}\mathrm{Var}(X_2) + \ldots + \frac{1}{n^2}\mathrm{Var}(X_n) = n\left(\frac{1}{n^2}\sigma^2\right)$$

$$\Rightarrow \qquad \sigma^2_{\overline{X}} = E\left((\overline{X} - \mu)^2\right) = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}. \tag{9.4}$$

Es decir, la desviación típica de la distribución de medias será la de la población original, dividido por un factor $\sqrt{n}$ que depende del tamaño de la muestra.

Ejemplo III–1

Consideremos una caja con tarjetas, cada una con un número. Suponemos que la población tiene $\mu = 10$ y $\sigma = 4$. Extraemos muestras de tamaño $n = 9$ (con reemplazamiento):

Primera muestra: 4, 13, 8, 12, 8, 15, 14, 7, 8. Media $\overline{X} = 9.9$.
Segunda muestra: 17, 14, 2, 12, 12, 6, 5, 11, 5. Media $\overline{X} = 9.3$.
...

Tras una serie de 10 muestras obtenemos $\overline{X}$ =9.9, 9.3, 9.9, 10.9, 9.6, 9.2, 10.2, 11.5, 9.0 y 11.8. Comprobamos que el valor medio de $\overline{X}$ es 10.13, y su desviación típica 0.97. Aplicando las fórmulas se obtiene

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}} = \frac{4}{\sqrt{9}} = 1.3333.$$

La expresión anterior es válida solo para el caso de población infinita o muestreo con reemplazamiento. Si tenemos una población finita en que se hace muestreo sin reemplazamiento, la expresión para la media de la distribución sigue siendo válida, pero la de la varianza hay que substituirla por

$$\sigma^2_{\overline{X}} = \mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right), \tag{9.5}$$

donde N es el tamaño de la población y n el tamaño de la muestra (Ver la demostración en ej. *Probabilidad y Estadística* de Schaum, pags. 186-187). Nótese que la expresión anterior se convierte en (9.4) cuando $N \to \infty$ ó N se hace mucho más grande que n.

Respecto a la forma de la distribución muestral de la media, ésta en principio depende de la distribución de la población de partida, pero, en virtud del **teorema del límite central**, se puede establecer que $\overline{X}$ seguirá una distribución asintóticamente normal. Es decir:

Si $\overline{X}$ es la media de una muestra aleatoria de tamaño n que se toma de una población con distribución cualquiera, media μ y varianza σ^2, entonces la variable tipificada

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \tag{9.6}$$

tiende a una distribución normal estándar $N(0, 1)$ cuando n tiende a infinito.

Efectivamente, el teorema del límite central establecía que, si se define una variable aleatoria $Y = X_1 + X_2 + \ldots + X_n$, suma de variables aleatorias independientes con medias μ_i y desviaciones típicas σ_i, entonces la variable tipificada

$$Z = \frac{Y - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}}$$

era asintóticamente normal. Por la definición de media muestral (9.2) podemos hacer $Y = n\overline{X}$, y por tanto, puesto que todas las X_i tienen la misma media μ y desviación típica σ de la población, Z se convierte en

$$Z = \frac{n\overline{X} - n\mu}{\sqrt{n\sigma^2}} = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}},$$

como queríamos demostrar. En resumen, $\overline{X}$ es asintóticamente normal, sea cual sea la forma de la distribución de la población de partida. Evidentemente, cuanto mayor sea el tamaño de la muestra, más se aproximará la distribución de $\overline{X}$ a la normal. En la práctica, la aproximación de distribución normal se utiliza cuando $n \geq 30$, y la bondad de ésta dependerá de la forma más o menos simétrica de la distribución de la población muestreada.

Un caso particular muy importante es cuando la distribución de la población de partida es normal. En este caso, no es necesario que el tamaño de la muestra sea grande para que la distribución muestral de $\overline{X}$ sea normal y podemos establecer que:

Si la población de la cual se toman muestras está distribuida normalmente con media μ y varianza σ^2, entonces la media muestral sigue una distribución normal con media μ y varianza σ^2/n, con independencia del tamaño de la muestra.

Esto es también consecuencia del teorema del límite central. Una combinación lineal, como $\overline{X}$, de variables aleatorias normales será también normal.

Para derivar estos últimos resultados hemos supuesto que la población era infinita o el muestreo con reemplazamiento (para que las diferentes X_i fuesen independientes). Si esto no se cumpliese y tuviésemos un muestreo sin reemplazamiento de una población finita, en (9.6) habría que substituir $\sigma/\sqrt{n}$ por la expresión dada en (9.5).

9.2.2. Distribución muestral de una proporción

Supongamos que tenemos una población sobre la que se experimenta un proceso de Bernoulli. Es decir, se llevan a cabo n ensayos y el resultado de cada uno de ellos es un éxito o un fracaso. Llamemos p a la probabilidad de éxito en cada ensayo y q $(= 1 - p)$ a la probabilidad de fracaso. Cada n ensayos se pueden considerar como una muestra de tamaño n. Para cada muestra vamos a definir el estadístico P como la proporción de éxitos, o número de éxitos dividido por el número de ensayos. Nótese que P puede considerarse como la media muestral de una variable de Bernoulli (o variable binomial con un único ensayo). P seguirá una distribución de probabilidad, llamada **distribución muestral de una proporción**, que es, entonces, un caso particular de la distribución muestral de una media.

Para calcular los parámetros poblacionales de esta distribución recordemos que la media y varianza de una variable de Bernoulli vienen dadas por

$$\mu = p \qquad ; \qquad \sigma^2 = pq.$$

Entonces, la media y varianza de la distribución de una proporción las podemos calcular aplicando (9.3) y (9.4) como

$$\mu_{\overline{P}} = E(\overline{P}) = \mu = p, \tag{9.7}$$

$$\sigma^2_{\overline{P}} = \mathrm{Var}(\overline{P}) = \frac{\sigma^2}{n} = \frac{pq}{n} = \frac{p(1-p)}{n}. \tag{9.8}$$

Al igual que antes, en el caso de un muestreo sin reemplazamiento de una muestra finita, la segunda ecuación hay que substituirla por

$$\sigma^2_{\overline{P}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right) = \frac{pq}{n}\left(\frac{N-n}{N-1}\right). \tag{9.9}$$

Al ser un caso particular de la distribución muestral de la media, la distribución muestral de una proporción puede aproximarse por una distribución normal para valores grandes del número de ensayos n. En la práctica esta aproximación se hace para $n \geq 30$.

Ejemplo III–2

Un jugador de baloncesto tiene un promedio de acierto en tiros libres del 80 %. Si tira tandas de 100 tiros libres y se calcula el promedio de aciertos, o la probabilidad de éxitos, la distribución tendrá una media $\mu_{\overline{P}} = p = 0.80$, y una desviación típica

$$\sigma_{\overline{P}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.80 \times 0.20}{100}} = 0.04.$$

Como $n \geq 30$, la aproximación a una distribución normal funcionará bien.

9.2.3. Distribución muestral de la diferencia de medias

Supongamos que tenemos dos poblaciones, la primera caracterizada por una media μ_1 y una varianza σ_1^2, y la segunda por μ_2 y σ_2^2. Supongamos que se extraen muestras aleatorias independientes de cada población, con tamaños n_1 y n_2 respectivamente. Siguiendo la misma notación, llamemos $\overline{X_1}$ al estadístico que representa la media muestral de la primera población y $\overline{X_2}$ a la media muestral de la segunda. Vamos a estudiar un nuevo estadístico, consistente en la diferencia de las medias muestrales $\overline{X_1} - \overline{X_2}$. Efectivamente, al ser una combinación lineal de dos variables aleatorias, será una nueva variable aleatoria, o estadístico, que tomará diferentes valores para todas las diferentes combinaciones de muestras extraídas de cada población. Su distribución vendrá dada por la **distribución muestral de la diferencia de medias**.

Para calcular la media y varianza de la distribución muestral de la diferencia de medias hacemos uso de las expresiones para la media y varianza de la diferencia de variables aleatorias independientes ($E(X \pm Y) = E(X) \pm E(Y)$ y $\mathrm{Var}(X \pm Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$) y de las expresiones (9.3) y (9.4) para la media y varianza de la distribución muestral de la media. Entonces

$$\mu_{\overline{X_1}-\overline{X_2}} = \mu_{\overline{X_1}} - \mu_{\overline{X_2}} = \mu_1 - \mu_2, \tag{9.10}$$

$$\sigma^2_{\overline{X_1}-\overline{X_2}} = \sigma^2_{\overline{X_1}} + \sigma^2_{\overline{X_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \tag{9.11}$$

Este último resultado solo será válido para poblaciones infinitas o en muestreos con reemplazamiento. En otro caso deberíamos usar la expresión (9.5) para llegar a una expresión equivalente.

Por otra parte, respecto a la forma de la distribución, por el teorema del límite central la variable tipificada definida por

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{9.12}$$

tenderá a la distribución normal estándar cuando tanto n_1 como n_2 tiendan a infinito. En la práctica se suele aplicar la aproximación normal si $n_1 + n_2 > 30$ (y $n_1 \simeq n_2$). Aún cuando n_1 y n_2 sean menores de 30, la aproximación normal puede ser razonablemente buena si las distribuciones originales no son muy asimétricas. Por supuesto, si ambas poblaciones fuesen normales, entonces $\overline{X_1} - \overline{X_2}$ tiene una distribución normal sin importar los tamaños de las muestras.

Ejemplo III–3

Se tienen dos poblaciones normales $N(20, 5)$ y $N(10, 6)$ y se extraen dos muestras de tamaños $n_1 = 25$ y $n_2 = 12$. ¿Cuál será la distribución muestral de la diferencia de medias?

$$\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2 = 20 - 10 = 10,$$

$$\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{5^2}{25} + \frac{6^2}{12}} = 2$$

$$\Rightarrow N(10, 2).$$

Ejemplo III–3

(Continuación) ¿Cuál será la probabilidad de obtener una diferencia de medias $\overline{X_1} - \overline{X_2} > 14$?
Para responder, utilizamos la distribución normal tipificada

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{14 - 10}{2} = 2,$$

por lo que (consultando en las tablas) obtenemos

$$P(\overline{X_1} - \overline{X_2} > 14) = P(Z > 2) = 0.0228.$$

De forma similar se puede deducir la distribución muestral de la diferencia de proporciones para dos poblaciones con distribuciones de Bernoulli y parámetros p_1, q_1 y p_2, q_2 respectivamente. En este caso, el estadístico diferencia de proporciones de éxitos $(\overline{P_1} - \overline{P_2})$ de muestras tomadas de cada población sigue una distribución con media y varianza dadas por

$$\mu_{\overline{P_1} - \overline{P_2}} = \mu_{\overline{P_1}} - \mu_{\overline{P_2}} = p_1 - p_2,$$

$$\sigma^2_{\overline{P_1} - \overline{P_2}} = \sigma^2_{\overline{P_1}} + \sigma^2_{\overline{P_2}} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}.$$

9.3. Varianza muestral

Otro estadístico importante es la varianza muestral. Si $X_i, i = 1, 2, \ldots, n$, representan las variables aleatorias para una muestra de tamaño n, entonces se define la **varianza muestral**, o varianza de la muestra, como

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}, \tag{9.13}$$

donde $\overline{X}$ es la media muestral. Se sigue entonces la misma definición que para la varianza de una tabla de frecuencias. En algunos textos se define la varianza muestral dividiendo por n en vez de $n - 1$. Más adelante veremos la razón de esta definición.

En una muestra particular, donde las variables aleatorias X_i toman los valores particulares x_i, el valor que tomará la varianza muestral vendrá dado, entonces, por

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n - 1}.$$

9.3.1. Distribución muestral de la varianza

Al igual que la media muestral, la varianza muestral es una variable aleatoria. Es decir, los valores que toma dependen de la muestra en particular que se tenga. Tiene por tanto una distribución de probabilidad asociada, llamada **distribución muestral de la varianza**. Para la media muestral vimos que la media, o esperanza matemática, de su distribución coincidía con la media poblacional. Para la varianza muestral sucede lo mismo: El valor esperado de la varianza muestral es la varianza poblacional, es decir

$$E(S^2) = \mu_{S^2} = \sigma^2. \tag{9.14}$$

Para demostrarlo empezamos desarrollando el numerador de (9.13)

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}\left((X_i-\mu)-(\overline{X}-\mu)\right)^2=\sum_{i=1}^{n}(X_i-\mu)^2-2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu)+n(\overline{X}-\mu)^2.$$

Ahora en el segundo término aplicamos: $\sum(X_i-\mu)=\sum X_i-n\mu=n(\overline{X}-\mu)$, resultando

$$\sum_{i=1}^{n}(X_i-\overline{X})^2=\sum_{i=1}^{n}(X_i-\mu)^2-n(\overline{X}-\mu)^2. \tag{9.15}$$

Introducimos esto en la definición de la varianza y tomamos esperanzas matemáticas

$$E(S^2)=E\left(\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n-1}\right)=\frac{1}{n-1}\left(\sum_{i=1}^{n}E\left((X_i-\mu)^2\right)-nE\left((\overline{X}-\mu)^2\right)\right).$$

Aplicando la definición de varianza de una variable aleatoria ($E\left((X-\mu)^2\right)=\sigma^2$), que la varianza de X_i es la varianza poblacional ($\sigma^2_{X_i}=\sigma^2$), y que la varianza de la media muestral es, por (9.4), $\sigma^2_{\overline{X}}=\sigma^2/n$

$$E(S^2)=\frac{1}{n-1}\left(\sum_{i=1}^{n}\sigma^2_{X_i}-n\sigma^2_{\overline{X}}\right)=\frac{1}{n-1}\left(n\sigma^2-n\frac{\sigma^2}{n}\right)=\frac{1}{n-1}(n-1)\sigma^2=\sigma^2,$$

como queríamos demostrar.

Nótese que si para la varianza muestral hubiésemos utilizado la definición alternativa

$$S'^2=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{n}, \tag{9.16}$$

hubiésemos obtenido

$$E(S'^2)=\frac{n-1}{n}\sigma^2,$$

y la varianza muestral hubiese subestimado la varianza poblacional. Este es el motivo por el que estamos trabajando con la definición (9.13) para la varianza. Como veremos más adelante, se dice que S^2 es un estimador insesgado de la varianza, mientras que S'^2 es un estimador sesgado. Evidentemente, cuando el tamaño n de la muestra sea grande apenas habrá diferencia de usar una definición u otra para la varianza muestral.

Los resultados anteriores son válidos si la población es infinita o el muestreo es con reemplazamiento. En el caso de tener un muestreo sin reemplazamiento de una población finita de tamaño N, la esperanza matemática de la varianza muestral estaría dada por

$$E(S^2)=\mu_{S^2}=\left(\frac{N}{N-1}\right)\sigma^2. \tag{9.17}$$

9.3.2. Distribución muestral de $(n-1)S^2/\sigma^2$

En vez de trabajar con la distribución muestral de la varianza S^2, es más cómodo utilizar la distribución muestral de la nueva variable aleatoria en el muestreo dada por

$$(n-1)\frac{S^2}{\sigma^2}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})^2}{\sigma^2}, \tag{9.18}$$

donde hemos usado la definición de varianza muestral dada en (9.13).

Para ver la importancia de esta distribución suponemos que tenemos una población normal y partimos

de la relación (9.15) escrita como

$$\sum_{i=1}^{n}(X_i-\mu)^2=\sum_{i=1}^{n}(X_i-\overline{X})^2+n(\overline{X}-\mu)^2. \tag{9.19}$$

Esta expresión tiene un importante significado pues descompone la variabilidad de los datos respecto a la media verdadera (o poblacional) en la suma de dos variabilidades: la de los datos respecto a la media muestral, y la de la media muestral respecto a la poblacional. Si en esta expresión dividimos en todos los miembros por σ^2, y se aplica la igualdad (9.18) se obtiene

$$\sum_{i=1}^{n}\left(\frac{X_i-\mu}{\sigma}\right)^2=\frac{(n-1)S^2}{\sigma^2}+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2. \tag{9.20}$$

Recordemos ahora que se definía una variable χ^2 con n grados de libertad como la suma de los cuadrados de n variables aleatorias normales X_i tipificadas ($N(0,1)$), es decir $\chi^2_n=X_1^2+\ldots+X_n^2$. El primer término de (9.20) es la suma de cuadrados de n variables aleatorias $N(0,1)$ (pues la media y desviación típica de cada X_i es μ y σ respectivamente) y, por lo tanto, es una χ^2 con n grados de libertad. Por otra parte, puesto que la media y desviación típica de la distribución muestral de la media $\overline{X}$ son respectivamente μ, por (9.3), y $\sigma/\sqrt{n}$, por (9.4), el último termino del segundo miembro es el cuadrado de una variable normal tipificada y, por tanto, puede considerarse como una χ^2 con 1 grado de libertad. Es decir, tenemos que una χ^2 con n grados de libertad es igual a la variable $(n-1)S^2/\sigma^2$ más una χ^2 con 1 grado de libertad. Por las propiedades de la distribución χ^2 puede deducirse entonces que $(n-1)S^2/\sigma^2$ es una χ^2 con $(n-1)$ grados de libertad. Estrictamente, para que esto se cumpla es necesario que el primer y último término de (9.20) sean independientes entre si. Aunque queda fuera del alcance de este libro, se puede demostrar que dicha condición se cumple. En resumen:

Si de una población con distribución normal y parámetros μ, σ, se toman muestras aleatorias de tamaño n, entonces la siguiente variable aleatoria obedece a una distribución χ^2 con $(n-1)$ grados de libertad

$$\chi^2_{n-1}=(n-1)\frac{S^2}{\sigma^2}. \tag{9.21}$$

Más adelante se verá cómo esta última propiedad es de importancia para la estimación de la varianza de una población normal.

Nótese que mientras que $\sum(X_i-\mu)^2/\sigma^2$ era una χ^2 con n grados de libertad, la variable $\sum(X_i-\overline{X})^2/\sigma^2$ es una χ^2 con $(n-1)$ grados de libertad. Es debido a que, al no conocer μ y estimarla a partir de $\overline{X}$, se pierde un grado de libertad pues esta media muestral se calcula a partir de los diferentes X_i. De esta forma, en general, cuando se quiere calcular un parámetro poblacional (ej. σ) y no se conoce el otro (ej. μ) la substitución de éste último por su parámetro muestral (ej. $\overline{X}$) hace que el sistema pierda un grado de libertad. Lo mismo ocurrirá en los dos siguientes apartados.

Ejemplo III–4

Un vendedor asegura que la pintura anticorrosiva de un automóvil dura 10 años, con una desviación típica de 3 años. Se pintan 6 coches y la pintura dura 12, 17, 3, 9, 5 y 13 años. ¿Podemos creer al vendedor cuando afirma que $\sigma = 3$?
Obtenemos la media muestral $\overline{X} = 9.83$ (que lógicamente debe ser próxima a μ). Calculamos ahora la varianza muestral

$$S^2 = \frac{\sum_{i=1}^{6}(X_i - \overline{X})^2}{n-1} = 27.4$$

y por tanto

$$\chi^2_{n-1} = (n-1)\frac{S^2}{\sigma^2} = 15.22,$$

que está muy lejos de lo esperado (recordemos que una distribución χ^2_{n-1} tiene $\mu = (n-1) = 5$ y $\sigma^2 = 2(n-1) = 10$).

9.3.3. El estadístico t

Al estudiar la distribución muestral de la media se vió que la variable aleatoria tipificada dada por

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

seguía una distribución normal si la población era normal, o tendía asintóticamente a la normal en otro caso. Como veremos, esta expresión se usa para estimar la media μ de la población. Sin embargo, en la mayoría de los casos no se conoce a priori la varianza σ^2 de la población. En ese caso, lo mejor que se puede hacer es reemplazar dicha varianza σ^2 por el valor de la varianza muestral S^2, definiéndose así el estadístico

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}}. \tag{9.22}$$

Este nuevo estadístico t toma valores diferentes de muestra a muestra. Si la muestras son pequeñas, los valores de S pueden fluctuar considerablemente de una a otra y la distribución de la variable aleatoria t puede desviarse apreciablemente de la distribución normal.

Para calcular la forma de la distribución de t, dividimos numerador y denominador de (9.22) por la desviación típica poblacional σ

$$t = \frac{(\overline{X} - \mu)/\sigma}{(S/\sigma)/\sqrt{n}} = \frac{(\overline{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{S^2/\sigma^2}}.$$

El numerador de esta última expresión representa, por (9.6), una variable normal tipificada que denotaremos por Z. Por otra parte, por (9.21), el denominador puede expresarse en función de una χ^2 con $(n-1)$ grados de libertad

$$t = \frac{Z}{\sqrt{\chi^2_{n-1}/(n-1)}}.$$

Esto es exactamente la definición de una variable t de Student con $(n-1)$ grados de libertad ($t_n = Z/\sqrt{\chi^2_n/n}$) ya que se cumple que numerador y denominador son independientes. Por tanto, podemos concluir que:

Si se toman muestras aleatorias de tamaño n de una población normalmente distribuida entonces el estadístico t, dado por (9.22), sigue una distribución t de Student con $(n-1)$ grados de libertad.

Este resultado, que se usa para la estimación de la media de una población, sigue siendo válido aún cuando la población no sea normal pero tenga una distribución en forma de campana similar a la normal.

Ejemplo III–5

Retomando el caso del ejemplo III–1 ($\mu = 10, \sigma = 4$), supongamos que no conocemos la desviación típica σ. Calculemos el valor del estadístico t.
Datos de la primera muestra ($n = 9$): 4, 13, 8, 12, 8, 15, 14, 7, 8 $\Rightarrow \overline{X} = 9.9$.

$$S^2 = \frac{\sum_i (X_i - \overline{X})^2}{n-1} \Rightarrow S = 3.72$$

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{9.9 - 10}{3.72/\sqrt{9}} = -0.08,$$

que resulta un valor muy centrado.

9.3.4. Distribución muestral de la razón de varianzas

Anteriormente hemos visto cómo para comparar dos poblaciones independientes se estudiaba la distribución muestral de la diferencia de medias. En el caso de las varianzas podría hacerse lo mismo y construir un estadístico de la diferencia de varianzas muestrales. Sin embargo, la distribución muestral de ese estadístico es demasiado complicada y, para poder comparar las varianzas de dos poblaciones, es mejor definir un estadístico basado en la razón de las varianzas muestrales, en vez de en su diferencia. Supongamos que tenemos dos poblaciones normales independientes con varianzas poblacionales σ_1^2 y σ_2^2 respectivamente. Sean S_1^2 y S_2^2 las varianzas muestrales medidas en una muestra aleatoria extraída de cada población. Se define entonces el estadístico F como

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}. \tag{9.23}$$

Evidentemente este estadístico será diferente para cada pareja de muestras. Es fácil ver cuál es su distribución ya que, suponiendo que las muestras tienen tamaños n_1 y n_2 respectivamente, usando (9.21), se pueden construir las variables χ^2

$$\chi^2_{n_1-1} = (n_1 - 1)\frac{S_1^2}{\sigma_1^2} \qquad ; \qquad \chi^2_{n_2-1} = (n_2 - 1)\frac{S_2^2}{\sigma_2^2}.$$

Sustituyendo en la definición (9.23) del estadístico F llegamos inmediatamente a

$$F = \frac{\chi^2_{n_1-1}/(n_1-1)}{\chi^2_{n_2-1}/(n_2-1)},$$

y esto es la definición de una variable F de Fisher con $(n_1 - 1)$ y $(n_2 - 1)$ grados de libertad (pues se define $F_{n_1,n_2} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}$). Es decir, si se extraen dos muestras aleatorias independientes de tamaños n_1 y n_2 de dos poblaciones normales con varianzas σ_1^2 y σ_2^2 respectivamente, y si las varianzas muestrales para cada muestra están dadas por S_1^2 y S_2^2, entonces el estadístico F, definido en (9.23), tiene una distribución F con $(n_1 - 1)$ y $(n_2 - 1)$ grados de libertad.

Este resultado sigue siendo válido aunque las poblaciones no sean normales pero su distribución tenga forma de campana.

Capítulo 10

Estimación puntual de parámetros

"No tenemos dinero, luego nos toca pensar."

Ernest Rutherford (1871-1937)

10.1. La estimación de parámetros

El objetivo de este tema es describir cómo se puede realizar la estimación de las características de una población a partir del estudio de una muestra aleatoria extraída de la misma. Vamos a suponer que se conoce la distribución de probabilidad que sigue la variable en estudio de la población, es decir, estamos en el caso de la estadística paramétrica. El problema se reduce entonces a estimar los valores de los **parámetros poblacionales** que definen dicha distribución. Sea α el parámetro poblacional a estimar. Supongamos que los posibles valores de la variable aleatoria en la muestra se representan por $X_1, X_2, \ldots, X_n$. El problema se resuelve definiendo una función $A = A(X_1, X_2, \ldots, X_n)$ de las medidas realizadas en la muestra tal que A constituya una estimación razonable del parámetro poblacional α. Evidentemente, para una muestra en particular A tomará un valor $a = a(x_1, x_2, \ldots, x_n)$ que variará de muestra a muestra. Es decir, al ser una función de variables aleatorias, A será asimismo una variable aleatoria, o un estadístico, con una distribución de probabilidad asociada. Al estadístico que sirve para realizar una estimación de un parámetro poblacional se le llama **estimador**. Por ejemplo, para estimar la media μ de una población normal se define el estimador $\overline{X}$ que tomará los valores particulares representados por $\overline{x}$.

Evidentemente queremos disponer de un buen estimador, en el sentido de que proporcione una estimación lo más precisa posible del parámetro poblacional. En general, la bondad de cada estimador dependerá de su distribución de probabilidad asociada. Por ejemplo, será conveniente que los diferentes valores que puede tomar el estimador para muestras de la misma población se distribuyan alrededor del valor del parámetro poblacional con una pequeña dispersión. En general, para cada parámetro poblacional se podrán definir varios estimadores, cada uno con sus características. Será importante elegir, de entre todos los estimadores posibles, el estimador óptimo para cada parámetro poblacional. Las propiedades que definen un buen estimador son las siguientes:

- Diremos que un estimador A de un parámetro poblacional α es **insesgado**, o centrado, si su media, o esperanza matemática, coincide con el parámetro poblacional. Es decir

$$E(A) = \mu_A = \alpha. \tag{10.1}$$

  Por ejemplo, la media aritmética $\overline{X}$ es un estimador insesgado de la media de una población (9.3)

y S^2 es un estimador insesgado de la varianza (9.14). Sin embargo, ${S'}^2$, definida como (9.16), es un estimador sesgado.

- Si se tienen dos estimadores A_1, A_2 de un parámetro poblacional, se dice que A_1 es más **eficiente** que A_2 si su varianza es menor. Es decir

$$\sigma^2_{A_1} < \sigma^2_{A_2}. \tag{10.2}$$

  Por ejemplo, para la estimación de la media poblacional, los estimadores media aritmética $\overline{X}$ y mediana M_e son insesgados, pero la media es más eficiente que la mediana (su varianza es menor). Evidentemente, entre dos estimadores insesgados siempre será preferible usar el más eficiente. Incluso en algunos casos será mejor usar un estimador algo sesgado pero más eficiente que otro insesgado.

- Se dice que un estimador es **consistente** cuando, al crecer el tamaño muestral, se aproxima asintóticamente al valor del parámetro poblacional y su varianza se hace nula. Es decir

$$\lim_{n\to\infty} A = \alpha \qquad ; \qquad \lim_{n\to\infty} \sigma^2_A = 0. \tag{10.3}$$

  Evidentemente, la media aritmética (por ejemplo) es un estimador consistente pues la varianza de su distribución muestral se puede expresar por $\sigma^2_{\overline{X}} = \sigma^2/n$ (9.4).

Un estimador ideal ha de ser insesgado y con una eficacia máxima. Sin embargo, en la práctica, a veces no es posible calcular dichos estimadores, y, por la comodidad con que se obtienen, se trabaja con estimadores sesgados o poco eficientes. De todas formas, un requisito mínimo que ha de cumplir cualquier estimador es que sea consistente.

Existen dos procedimientos para realizar la estimación de un parámetro poblacional. Cuando se determina un único valor de un estimador que se aproxime al parámetro poblacional desconocido se dice que se hace una **estimación puntual**. Cuando, alternativamente, se calculan dos valores entre los cuales se considera que, con cierta probabilidad, se encuentra el parámetro poblacional, el procedimiento se conoce como **estimación por intervalos de confianza**. En este tema veremos la estimación puntual y en el siguiente la estimación por intervalos.

10.2. Principales estimadores puntuales

Un **estimador puntual** de un parámetro poblacional es una función real de los n valores que la variable estadística toma en el muestreo. Es decir, es un estadístico (variable aleatoria) que cambia de muestra a muestra de forma aleatoria. Una **estimación puntual** es el valor concreto que toma el estimador puntual en una muestra en particular. Como ya se ha indicado, los estimadores puntuales se usan para realizar la estimación de parámetros poblacionales. En general, a cada parámetro poblacional se le pueden asociar diferentes estimadores puntuales aunque normalmente se elegirán aquellos que sean insesgados y más eficientes. Evidentemente, no se espera que un estimador puntual proporcione sin error el parámetro poblacional, sino que se pretende que las estimaciones puntuales no se alejen mucho del valor desconocido a calcular.

A continuación se dan los estimadores puntuales más usados asociados a las principales distribuciones de probabilidad que puede seguir la población a estudiar:

- Supongamos que la característica en estudio de la población sigue una **distribución normal** con media μ y varianza σ^2, es decir es $N(\mu, \sigma)$. Como estimadores puntuales de los parámetros poblaciones

μ y σ^2 normalmente se utilizan la media aritmética $\overline{X}$ y la varianza muestral S^2 respectivamente. Efectivamente, en (9.3) y (9.14) se demostró que ambos estimadores son insesgados pues

$$E(\overline{X}) = \mu \qquad ; \qquad E(S^2) = \sigma^2. \tag{10.4}$$

Además, puede demostrarse que ambos estimadores puntuales tienen una eficiencia máxima, es decir son de varianza mínima comparados con otros estimadores de los mismos parámetros poblacionales.

- Supongamos que la población obedece a una **distribución binomial** de parámetro p (probabilidad de éxito). Como estimador puntual de p se usa la proporción de éxitos $\overline{P}$, definida como el número de éxitos dividido por el número de ensayos (o frecuencia relativa de éxitos). En (9.7) se demostró que este estimador es insesgado. Es decir

$$E(\overline{P}) = p. \tag{10.5}$$

Además puede demostrarse que es de varianza mínima ($\sigma^2_{\overline{P}} = p(1-p)/n$).

- Consideremos ahora una población cuya característica en estudio siga una **distribución de Poisson**. Sea λ, o número medio de sucesos por intervalo, el parámetro poblacional a determinar. Sean $X_1, X_2, \ldots, X_n$ los números de resultados obtenidos en n experimentos (muestra de tamaño n). Entonces, un estimador puntual para λ es la media muestral, definida como

$$\overline{\lambda} = \frac{\sum_{i=1}^{n} X_i}{n}. \tag{10.6}$$

Este estimador es insesgado, es decir $E(\overline{\lambda}) = \lambda$, y además tiene varianza mínima (es el más eficiente).

10.3. El método de máxima verosimilitud

En la seccion anterior se ha visto como, con frecuencia, los estimadores puntuales mejores coinciden con los que se elegirían intuitivamente. Por ejemplo, es lógico que la media muestral $\overline{X}$ sea un estimador apropiado para la media poblacional μ. Sin embargo, en ocasiones, no es del todo obvio cual ha de ser el mejor estimador. Para ello, se presenta a continuación un metodo general muy potente para hallar estimadores puntuales. Se trata del **método de la máxima verosimilitud**.

Para ilustrar el método supongamos que la distribución de probabilidad de la población, caracterizada por una variable aleatoria X, contiene un único parámetro α a determinar. Sea $f(x, \alpha)$ la función de probabilidad, en el caso discreto, o función de densidad, en el caso continuo, de dicha variable aleatoria. Si de esta población se extrae una muestra de tamaño n representada por los valores $X_1, X_2, \ldots, X_n$, podemos expresar la distribución de probabilidad conjunta (9.1) por

$$L(X_1, X_2, \ldots, X_n; \alpha) = f(X_1, X_2, \ldots, X_n; \alpha) = f(X_1, \alpha) f(X_2, \alpha) \ldots f(X_n, \alpha), \tag{10.7}$$

donde hemos supuesto que las diferentes X_i son independientes (población infinita o muestreo con reemplazamiento). A esta función L se le llama **función de verosimilitud** y variará de muestra a muestra y con el parámetro α. Evidentemente, la función de verosimilitud para una muestra discreta en particular, da la probabilidad de que las variables tomen unos determinados valores. Se define entonces el estimador puntual de máxima verosimilitud como el valor de α que hace máxima dicha función de verosimilitud L. Es decir, es el parámetro α para el cual la probabilidad de haber obtenido la muestra en particular que se tiene es máxima.

Ejemplo III–6

Supongamos que se hace un experimento de Bernoulli (por ejemplo en el control de calidad de 3 artículos para ver sin son defectuosos) y encontramos dos éxitos y un fracaso. Queremos estimar el parámetro p (probabilidad de éxito) de la distribución binomial. Si consideramos $X = 1$ como éxito y $X = 0$ como fracaso, la función de verosimilitud podrá calcularse como

$$L(X_1, X_2, X_3; p) = f(X_1, p)\ f(X_2, p)\ f(X_3, p) =$$

$$= P(X_1 = 1; p)\ P(X_2 = 1; p)\ P(X_3 = 0; p) = p\ p\ q = p^2(1-p) = p^2 - p^3.$$

Como buscamos el máximo de esta función, tomamos derivadas e igualamos a cero, es decir

$$\frac{\mathrm{d}L}{\mathrm{d}p} = 2p - 3p^2 = 0 \Rightarrow (2-3p)p = 0,$$

cuyas soluciones son $p = 0$ (no nos vale) y $p = 2/3$. Así que $p = 2/3$ es la estimación de máxima verosimilitud de p y coincide, además, con lo que se esperaría de forma natural como probabilidad de éxito (número de éxitos dividido por el número de ensayos).

Por razones prácticas, se suele trabajar con el logarítmo neperiano de la función de verosimilitud. De esta forma para encontrar el valor de α que lo hace máximo se iguala la siguiente derivada a cero

$$\frac{d \ln L}{d\alpha} = \frac{1}{L}\frac{dL}{d\alpha} = 0, \tag{10.8}$$

y se resuelve esta ecuación para encontrar α. En el caso de que la distribución de probabilidad tenga más de un parámetro poblacional, se hacen las derivadas parciales respecto a cada parámetro y se resuelve el sistema de ecuaciones.

Como ejemplo del método a continuación se derivan los estimadores de máxima verosimilitud para las principales distribuciones:

- Supongamos que la población sigue una **distribución binomial**, consistiendo la muestra en n ensayos en los que, en cada uno, se obtiene un éxito, que representaremos por $X = 1$, o un fracaso, $X = 0$. La función de probabilidad para un único ensayo vendrá dada por

$$f(x, p) = p^x(1-p)^{1-x} = \begin{cases} 1-p & ; \quad x = 0 \\ p & ; \quad x = 1 \end{cases}$$

  donde p es la probabilidad de éxito, parámetro desconocido a determinar. Supongamos que en el experimento de n ensayos se obtienen f éxitos. Entonces, la función de verosimilitud, o función de probabilidad conjunta, será

$$L = \prod_{i=1}^{n} f(x_i, p) = p^f (1-p)^{n-f},$$

$$\ln L = f \ln p + (n-f) \ln (1-p).$$

  Derivando respecto al parámetro p, e igualando la derivada a cero

$$\frac{d \ln L}{dp} = \frac{f}{p} - \frac{n-f}{1-p} = 0.$$

  Despejando p

$$p(n-f) = f - f\,p \quad \Rightarrow \quad p(n-f+f) = f \quad \Rightarrow \quad p = \frac{f}{n}.$$

  Por lo tanto, el estimador de máxima verosimilitud del parámetro p es la frecuencia relativa de éxitos, como cabría esperar.

- Supongamos ahora que se tiene una distribución normal con parámetros μ y σ, es decir $N(\mu, \sigma)$, de la

que se extrae una muestra de tamaño n. La función de verosimilitud será en este caso

$$L = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}},$$

$$\ln L = \sum_{i=1}^{n} \left(-\ln\sqrt{2\pi} - \ln\sigma - \frac{(x_i-\mu)^2}{2\sigma^2} \right) =$$

$$= -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum (x_i-\mu)^2.$$

A continuación se hacen las derivadas parciales respecto a los dos parámetros poblacionales para calcular sus estimadores

$$\frac{\partial \ln L}{\partial \mu} = -\frac{1}{2\sigma^2} 2 \sum (x_i - \mu) = 0 \quad \Rightarrow$$

$$\sum (x_i - \mu) = 0 \quad \Rightarrow \quad \sum x_i - n\mu = 0 \quad \Rightarrow \quad \mu = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Por lo tanto, el estimador de máxima verosimilitud para μ coincide con la media muestra, es decir, con el estimador puntual usado hasta ahora.

Similarmente, para la varianza

$$\frac{\partial \ln L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum (x_i-\mu)^2 = 0.$$

Multiplicando por $2\sigma^4$

$$n\sigma^2 = \sum (x_i-\mu)^2 \quad \Rightarrow \quad \sigma^2 = \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}.$$

Luego, el estimador de máxima verosimilitud para la varianza es la varianza muestral en su definición de (9.16), o ${S'}^2$. Nótese que esta es la varianza sesgada y no coincide con el estimador puntual que hemos usado hasta ahora. En general, los estimadores de máxima verosimilitud no tienen porque ser insesgados, auque gozan de propiedades asintóticas muy importantes.

- Es fácil demostrar que el estimador de máxima verosimilitud para el parámetro λ de la **distribución de Poisson** es la media muestral definida en (10.6).

Ejemplo III–7

Calcular el estimador de máxima verosimilitud para el parámetro λ de la distribución de Poisson.

La función de probabilidad

$$f(x;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

La función de verosimilitud será entonces

$$L = \prod_{i=1}^{n} \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Tomando logaritmos, derivando y operando

$$\ln L = \sum_{i=1}^{n} (x_i \ln\lambda - \ln(x_i!) - \lambda) = \ln\lambda \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \ln(x_i!) - n\lambda.$$

$$\frac{\mathrm{d}\ln L}{\mathrm{d}\lambda} = \frac{1}{\lambda}\sum_{i=1}^{n} x_i - n = 0 \Rightarrow \sum_{i=1}^{n} x_i = \lambda\, n$$

$$\Rightarrow \lambda = \frac{\sum_{i=1}^{n} x_i}{n}, \quad \text{que es el número promedio de eventos/intervalo.}$$

Capítulo 11

Estimación por intervalos de confianza

> "No puedo juzgar mi trabajo mientras lo hago. He de hacer como los pintores, alejarme y mirarlo desde cierta distancia, aunque no demasiada. ¿Cuánta? Adivínelo."
>
> Blaise Pascal (1623-1662)

Generalmente, una estimación puntual no proporciona un valor exacto del parámetro poblacional a determinar. Es más, en la mayoría de los casos, no tendremos información sobre la precisión de tal estimación, de forma que su valor único no nos informa sobre la probabilidad de que se encuentre cerca o lejos del valor verdadero. En la práctica, interesa no solamente dar una estimación, sino precisar la incertidumbre de dicha estimación. Esto se consigue mediante la **estimación por intervalos de confianza**, en la cual se calcula un intervalo sobre el que podamos establecer que, con cierta probabilidad, está contenido el parámetro poblacional desconocido. De esta manera, en vez de calcular un único estimador, se determinan dos estimadores que serán los límites inferior (L_1) y superior (L_2) (o límites de confianza) de un **intervalo de confianza** $I = [L_1, L_2]$. A esta pareja de valores se le llama **estimador por intervalo**. Estos límites de confianza serán estadísticos que variarán de muestra a muestra, de forma que podrá considerarse al intervalo como una variable aleatoria bidimensional. Efectivamente, los límites del intervalo serán función de los valores que toma la variable aleatoria en el muestreo

$$L_1 = f_1(X_1, X_2, \ldots, X_n) \qquad ; \qquad L_2 = f_2(X_1, X_2, \ldots, X_n).$$

Al valor concreto que toma el intervalo aleatorio en una muestra en particular se le llama **estimación por intervalo**. Al ser el estimador por intervalo una variable aleatoria, podrá decirse que existe una cierta probabilidad de que el intervalo aleatorio cubra el verdadero valor del parámetro poblacional β. Es decir

$$P(L_1 < \beta < L_2) = 1 - \alpha, \tag{11.1}$$

donde, por definición, a $1-\alpha$ se le llama **nivel de confianza** y al intervalo $[L_1, L_2]$ se le denomina **intervalo de confianza del** $(1-\alpha)100\,\%$.

Nótese que, una vez tomada una muestra en particular, no tiene sentido decir que β estará dentro del intervalo con una cierta probabilidad, puesto que estará o no estará. La forma correcta de expresar esto es diciendo que $1-\alpha$ es la probabilidad de seleccionar una muestra concreta que conduzca a un intervalo que contenga al parámetro poblacional. En otras palabras, el $100(1-\alpha)\,\%$ de los intervalos correspondientes a todas las muestras posibles del mismo tamaño contienen a β y el $100\alpha\,\%$ no lo contienen.

Evidentemente, al aumentar el tamaño de la muestra ha de aumentar la precisión con que se conoce el parámetro poblacional, y por lo tanto, para un nivel de confianza fijo, el intervalo de confianza ha de hacerse

más pequeño. Es decir, la longitud del intervalo de confianza indica la precisión de la estimación.

Para ilustrar los conceptos anteriores, supongamos que para realizar la estimación por intervalos de confianza de un parámetro poblacional se calcula un estadístico B. Este estadístico tendrá un distribución muestral asociada, con media μ_B y desviación típica σ_B. Supongamos que la distribución muestral de B es aproximadamente normal (sabemos que esto es una buena aproximación si la muestra es suficientemente grande). En este caso, usando las propiedades de la curva normal, podemos establecer las siguientes probabilidades

$$P(\mu_B - \sigma_B < B < \mu_B + \sigma_B) = 0.6827$$

$$P(\mu_B - 2\sigma_B < B < \mu_B + 2\sigma_B) = 0.9544$$

$$P(\mu_B - 3\sigma_B < B < \mu_B + 3\sigma_B) = 0.9973$$

Es fácil ver que lo anterior es equivalente a

$$P(B - \sigma_B < \mu_B < B + \sigma_B) = 0.6827$$

$$P(B - 2\sigma_B < \mu_B < B + 2\sigma_B) = 0.9544$$

$$P(B - 3\sigma_B < \mu_B < B + 3\sigma_B) = 0.9973$$

Si B es insesgado, es decir si μ_B coincide con el parámetro poblacional β a determinar, las expresiones anteriores proporcionan intervalos de confianza del 68.27 %, 95.44 % y 99.73 % respectivamente para dicho parámetro poblacional. Normalmente, se suele trabajar con niveles de confianza de 0.95 ó 0.99. Para conseguir estas probabilidades hay que buscar en la tabla de la distribución normal las abscisas que dejan a su derecha un área igual a $(1 - 0.95)/2 = 0.05/2 = 0.025$ y $(1 - 0.99)/2 = 0.01/2 = 0.005$ respectivamente. Estas son aproximadamente $z_{0.025} = 1.96$ y $z_{0.005} = 2.58$. Por lo tanto, los intervalos de confianza del 95 % y 99 % serán respectivamente

$$P(B - 1.96\sigma_B < \mu_B < B + 1.96\sigma_B) = 0.95,$$

$$P(B - 2.58\sigma_B < \mu_B < B + 2.58\sigma_B) = 0.99.$$

En general, para un nivel de confianza $1-\alpha$ habrá que buscar las abscisas $z_{\alpha/2}$ de la distribución normal tipificada $N(0, 1)$ que dejan a su derecha un área igual a $\alpha/2$, expresándose entonces el intervalo de confianza del $(1 - \alpha)100\,\%$ como

$$P(B - z_{\alpha/2}\sigma_B < \mu_B < B + z_{\alpha/2}\sigma_B) = 1 - \alpha. \tag{11.2}$$

La expresión anterior es sumamente útil para calcular intervalos de confianza usando estadísticos con distribuciones muestrales normales. Lo único que habrá que hacer será substituir B por el estadístico insesgado correspondiente y μ_B y σ_B por la media y desviación típica de la distribución muestral.

En el caso de que la distribución muestral del estadístico no sea normal, se pueden hacer las modificaciones correspondientes. Así si B siguiera una distribución t de Student con n grados de libertad, el intervalo vendría dado por

$$P(B - t_{\alpha/2,n}\sigma_B < \mu_B < B + t_{\alpha/2,n}\sigma_B) = 1 - \alpha, \tag{11.3}$$

donde $t_{\alpha/2,n}$ representa el valor de la abscisa de la distribución t con n grados de libertad que deja a su derecha un área igual a $\alpha/2$. Así mismo, se pueden encontrar las expresiones correspondientes para las distribuciones χ^2 y F, introduciendo las abscisas $\chi^2_{\alpha/2,n}$ y $F_{\alpha/2;n_1,n_2}$.

11.1. Intervalos de confianza para la media

Supongamos en primer lugar que la población en estudio sigue una **distribución normal** $N(\mu, \sigma)$ y que como estimador puntual de la media poblacional μ se usa la media muestral $\overline{X}$. Distinguiremos tres casos principales:

- **Varianza poblacional σ^2 conocida:**

Ya se ha visto que si la población es normal, la media muestral sigue una distribución normal con media $\mu_{\overline{X}} = \mu$ (9.3) y varianza $\sigma^2_{\overline{X}} = \sigma^2/n$ (9.4). Entonces, aplicando (11.2), el intervalo de confianza del $(1-\alpha)100\,\%$ para la media puede expresarse como

$$P(\overline{X} - z_{\alpha/2}\sigma_{\overline{X}} < \mu_{\overline{X}} < \overline{X} + z_{\alpha/2}\sigma_{\overline{X}}) = 1-\alpha \qquad \Rightarrow$$

$$P\left(\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha. \tag{11.4}$$

Al mismo resultado puede llegarse teniendo en cuenta que, en este caso, la variable $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ es una normal tipificada $N(0,1)$. Entonces

$$P\left(-z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha,$$

que conduce inmediatamente a (11.4).

En resumen, el intervalo de confianza de nivel $(1-\alpha)$ para la media de una distribución normal de varianza conocida es

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]. \tag{11.5}$$

El resultado anterior es válido para un población infinita o en un muestreo con reemplazamiento. Si el muestreo es sin reemplazamiento en una población finita de tamaño N, habrá que usar la expresión (9.5) para la varianza de la distribución de medias, de forma que el intervalo de confianza es

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right] \tag{11.6}$$

Nótese que muestras diferentes darán lugar a valores diferentes de $\overline{X}$ y, por lo tanto, a intervalos diferentes. Sin embargo, la longitud de los intervalos será siempre la misma y dependerá únicamente (para muestras de igual tamaño) del nivel de confianza $1-\alpha$ que se haya fijado (a menor α mayor anchura del intervalo). Evidentemente, no todos los intervalos que se construyan de diferentes muestras contendrán al parámetro μ, aunque sabemos que esto se cumplirá para el $100(1-\alpha)\,\%$ de los intervalos posibles.

Ejemplo III–8

Retornando al ejemplo III–1, calculemos el intervalo de confianza para la media ($\sigma = 4$ es conocida) de las dos primeras muestras (usar nivel de confianza 0.95).

- Muestra i): 4, 13, 8, 12, 8, 15, 14, 7, 8. $\Rightarrow \overline{X} = 9.9$

$$1 - \alpha = 0.95 \Rightarrow \alpha = 0.05$$

$$z_{\alpha/2} = z_{0.025} = 1.96$$

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[9.9 \pm 1.96\frac{4}{\sqrt{9}}\right] = [9.9 \pm 2.6]$$

- Muestra ii): 17, 14, 2, 12, 12, 6, 5, 11, 5. $\Rightarrow \overline{X} = 9.3$

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[9.3 \pm 1.96\frac{4}{\sqrt{9}}\right] = [9.3 \pm 2.6]$$

De cada 100 muestras, en el 95 % de ellas el intervalo de confianza así calculado incluirá al valor real.

- **Varianza poblacional σ^2 desconocida y $n > 30$:**

  En general, la desviación típica σ de la población se desconoce a priori, de forma que, estrictamente, no se puede aplicar la expresión (11.5) para calcular el intervalo de confianza. Sin embargo, cuando la muestra es grande, la desviación típica muestral S suele ser un estimador muy preciso de σ, de forma que, en primera aproximación, el intervalo de confianza se puede construir sustituyendo σ por S en (11.5), obteniéndose

$$P\left(\overline{X} - z_{\alpha/2}\frac{S}{\sqrt{n}} < \mu < \overline{X} + z_{\alpha/2}\frac{S}{\sqrt{n}}\right) = 1 - \alpha, \tag{11.7}$$

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right]. \tag{11.8}$$

  En la práctica, esta aproximación se usa cuando el tamaño de la muestra n es mayor que 30.

- **Varianza poblacional σ^2 desconocida y $n < 30$:**

  Cuando las muestras son pequeñas la varianza muestral puede variar considerablemente de muestra a muestra, por lo que la aproximación anterior no se considera válida. En estos casos, el intervalo confianza se puede construir recordando que la variable

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

  sigue una distribución t de Student con $n - 1$ grados de libertad. Por lo tanto, al ser la distribución t también simétrica, se puede expresar que

$$P\left(-t_{\alpha/2,n-1} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) = 1 - \alpha.$$

  Por lo que, operando

$$P\left(\overline{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1 - \alpha. \tag{11.9}$$

  De manera que el intervalo de confianza de nivel $(1 - \alpha)$ para la media de una distribución normal de varianza desconocida y muestra pequeña es

$$I = \left[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right], \tag{11.10}$$

donde $t_{\alpha/2,n-1}$ es la abscisa de la distribución t que deja a su derecha un área igual a $\alpha/2$. Esta expresión será además exacta y podrá utilizarse para calcular el intervalo de confianza para muestras grandes ($n > 30$). Sin embargo, por las propiedades de la distribución t, esta distribución tiende a la normal al aumentar los grados de libertad, por lo que la expresión (11.8) es suficientemente buena si n es grande.

Ejemplo III–9

Calcular los intervalos de confianza para la media en el ejemplo anterior suponiendo que la varianza es desconocida.

- Muestra i): $\overline{X} = 9.9$, $S = 3.72$

$$\alpha = 0.05 \Rightarrow t_{\alpha/2,n-1} = t_{0.025,8} = 2.306$$

$$I = \left[\overline{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right] = \left[9.9 \pm 2.306\frac{3.72}{\sqrt{9}}\right] = [9.9 \pm 2.9]\,,$$

lo que nos conduce a un intervalo mayor que en el ejemplo anterior, (7.0,12.8), lo cual es lógico porque hemos introducido una nueva fuente de incertidumbre al haber tenido que estimar la varianza (al no ser ahora conocida).

- Muestra ii): en este caso se obtiene

$$I = [9.3 \pm 3.8]\,,$$

que también es un intervalo mayor (5.5,13.1).

Para calcular los intervalos de confianza para la media anteriores se ha supuesto que la población de partida sigue una distribución normal. Sin embargo, en virtud del teorema del límite central y según se vió en (9.6), la distribución muestral de la media tiende asintóticamente a la normal cualquiera que sea la población de partida. Esto quiere decir que, para muestras grandes de cualquier población, el intervalo de confianza para la media es aproximadamente

$$I = \left[\overline{X} \pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right], \tag{11.11}$$

donde se ha supuesto que S es un buen estimador de σ si la muestra es grande.

Dos casos particulares de esta propiedad son los siguientes:

- **Intervalo de confianza para una proporción (distribución binomial)**

  Supongamos que la población sigue una distribución binomial con parámetro desconocido p. Ya se ha visto como la proporción de éxitos $\overline{P}$ (número de éxitos dividido por el número de ensayos) constituye un buen estimador de p. Además la distribución muestral del estadístico $\overline{P}$ puede aproximarse a la distribución normal cuando la muestra (o número de ensayos) es grande. En (9.7) y (9.8) se demostró que la media y varianza de la distribución muestral de una proporción son respectivamente $\mu_{\overline{P}} = p$ y $\sigma^2_{\overline{P}} = p(1-p)/n$. Entonces, aproximando la distribución por una normal y aplicando (11.2), donde el estadístico es $\overline{P}$, se obtiene

$$P\left(\overline{P} - z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}} < p < \overline{P} + z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right) = 1-\alpha. \tag{11.12}$$

  Es decir, para una muestra grande, el intervalo de confianza de nivel $(1-\alpha)$ para el parámetro p de una distribución binomial es

$$I = \left[\overline{P} \pm z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right]. \tag{11.13}$$

Nótese que en la varianza muestral se ha substituido p por $\overline{P}$, lo cual es una buena aproximación si la muestra es grande. Para muestras pequeñas ($n < 30$) la aproximación realizada de substituir la binomial por una normal es posible que no sea buena, especialmente si p se acerca a 0 ó a 1. Como ya se explicó, cuando se cumpla que conjuntamente $np > 5$ y $n(1-p) > 5$, la aproximación anterior es válida incluso para muestras pequeñas.

Ejemplo III–10

Un jugador de baloncesto lanza 100 tiros libres y anota 85. Calcular el intervalo de confianza para la proporción de aciertos.
Como $n = 100$ es claramente mayor que 30, podemos aproximar por la distribución normal. La proporción de éxitos será entonces $\overline{P} = 85/100 = 0.85$. Usando un nivel de confianza $1 - \alpha = 0.95$,

$$I = \left[\overline{P} \pm z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right] = \left[0.85 \pm 1.96\sqrt{\frac{0.85 \times 0.15}{100}}\right] = [0.85 \pm 0.07]\,,$$

lo que nos conduce al intervalo (0.78,0.92).

- **Intervalo de confianza para el parámetro λ de una distribución de Poisson**

Consideremos ahora que la población sigue una distribución de Poisson con parámetro λ. Ya se ha visto como un estimador puntual de dicho parámetro poblacional es la media muestral $\overline{\lambda}$, definida en (10.6). Para calcular el intervalo de confianza vamos a suponer que la muestra es grande, por lo que se puede aproximar la distribución por una normal. Igualando la media y la desviación típica muestral respectivamente a $\overline{X} = \overline{\lambda}$ y $S = \sqrt{\overline{\lambda}}$ (por las propiedades de la distribución de Poison), y aplicando (11.2), se puede escribir

$$P\left(\overline{\lambda} - z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}} < \lambda < \overline{\lambda} + z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}}\right) = 1 - \alpha. \tag{11.14}$$

Es decir, para una muestra grande, el intervalo de confianza de nivel $(1-\alpha)$ para el parámetro λ de una distribución de Poisson es

$$I = \left[\overline{\lambda} \pm z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}}\right]. \tag{11.15}$$

También suele exigirse $\overline{\lambda} > 5$.

11.2. Intervalos de confianza para la diferencia de medias

Supongamos que se tienen dos poblaciones normales $N(\mu_1, \sigma_1)$ y $N(\mu_2, \sigma_2)$. Vamos a estudiar cómo se puede determinar un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ a partir de muestras aleatorias independientes de tamaños n_1 y n_2 extraídas de cada población respectivamente. Distinguiremos diferentes casos

- **Varianzas poblacionales σ_1^2 y σ_2^2 conocidas:**

Ya se ha visto que un buen estimador puntual para la diferencia de medias es la diferencia de medias muestrales $\overline{X_1} - \overline{X_2}$. Además se cumple que la distribución muestral de la diferencia de medias es normal con media $\mu_{\overline{X_1}-\overline{X_2}} = \mu_1 - \mu_2$ (9.10) y varianza $\sigma^2_{\overline{X_1}-\overline{X_2}} = \sigma_1^2/n_1 + \sigma_2^2/n_2$ (9.11). Por tanto,

aplicando (11.2), se puede escribir

$$P\left((\overline{X_1}-\overline{X_2})-z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\mu_1-\mu_2<(\overline{X_1}-\overline{X_2})+z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$$

$$=1-\alpha. \tag{11.16}$$

Es decir, el intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de dos distribuciones normales de varianzas conocidas es

$$I=\left[(\overline{X_1}-\overline{X_2})\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]. \tag{11.17}$$

Ejemplo III–11

Volviendo a utilizar los datos del ejemplo III–1, determinar el intervalo de confianza para la diferencia de medias de las dos primeras muestras. Suponer la varianza poblacional conocida.

$$\overline{X}_1=9.9 \quad n_1=9 \quad \sigma_1=4$$
$$\overline{X}_1=9.3 \quad n_2=9 \quad \sigma_2=4$$

$$I=\left[(\overline{X_1}-\overline{X_2})\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]=$$

$$=\left[(9.9-9.3)\pm 1.96\sqrt{\frac{16}{9}+\frac{16}{9}}\right]=[0.6\pm 3.7]$$

por lo que el intervalo de confianza es (−3.1,4.3).

- **Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas y $n_1+n_2>30$ (con $n_1\simeq n_2$):**

  Generalmente no se conocerán a priori los valores de las varianzas poblacionales. Sin embargo, cuando las muestras son grandes, ya se ha visto como las varianzas muestrales son generalmente una buena aproximación a las varianzas poblacionales. Por lo tanto, en este caso el intervalo de confianza para la diferencia de medias puede aproximarse por las expresiones (11.16) y (11.17) sustituyendo σ_1^2 y σ_2^2 por S_1^2 y S_2^2 respectivamente

$$P\left((\overline{X_1}-\overline{X_2})-z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}<\mu_1-\mu_2<(\overline{X_1}-\overline{X_2})+z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$$

$$=1-\alpha \tag{11.18}$$

$$\Rightarrow I=\left[(\overline{X_1}-\overline{X_2})\pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]. \tag{11.19}$$

  Las aproximaciones anteriores son entonces válidas para muestras grandes. Para esto se usan diferentes criterios. Algunos autores exigen que tanto $n_1>30$ como $n_2>30$. Aqui vamos a fijar el criterio de que $n_1+n_2>30$, con la condición adicional de que ambos tamaños muestrales sean similares ($n_1\simeq n_2$).

- **Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas con $\sigma_1=\sigma_2$ (muestras pequeñas):**

  Supongamos ahora el caso de que las muestras no son grandes, por lo que no se pueden aplicar las aproximaciones anteriores. Consideremos en primer lugar que se puede asegurar a priori que las dos varianzas poblacionales han de ser iguales ($\sigma_1^2=\sigma_2^2$), aunque con valor desconocido. En este caso, por

(9.12), puede construirse la siguiente variable normal tipificada

$$Z = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}. \tag{11.20}$$

Por otra parte, por (9.21), sabemos que $(n_1 - 1)S_1^2/\sigma^2$ y $(n_2 - 1)S_2^2/\sigma^2$ obedecen a distribuciones χ^2 con $n_1 - 1$ y $n_2 - 1$ grados de libertad respectivamente. Por tanto, se puede construir la siguiente variable χ^2 con $n_1 + n_2 - 2$ grados de libertad

$$\chi^2_{n_1+n_2-2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}.$$

Recordando que una variable t de Student con n grados de libertad se define como $t_n = Z/\sqrt{\chi^2_n/n}$, el siguiente estadístico seguirá una distribución t con $n_1 + n_2 - 2$ grados de libertad

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \Bigg/ \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2(n_1 + n_2 - 2)}} =$$

$$= \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{11.21}$$

donde se ha definido S_p como

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \tag{11.22}$$

Por lo tanto, para dicha variable T se puede escribir

$$P\left(-t_{\alpha/2,n_1+n_2-2} < \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} < t_{\alpha/2,n_1+n_2-2}\right) = 1 - \alpha$$

$$P\left((\overline{X_1} - \overline{X_2}) - t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

$$= 1 - \alpha. \tag{11.23}$$

Y el intervalo de confianza de nivel $(1 - \alpha)$ para la diferencia de medias de dos poblaciones normales de varianzas desconocidas pero iguales es

$$I = \left[(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]. \tag{11.24}$$

Al calcularse por (11.22), S_p^2 representa una estimación puntual de la varianza común σ^2, calculándose como una media ponderada, con el número de grados de libertad, de las dos varianzas observadas.

Hay que indicar que las relaciones anteriores siguen siendo una buena aproximación aún cuando existan algunas diferencias entre las varianzas poblacionales si los tamaños de las muestras son iguales. En general, para calcular intervalos de confianza para la diferencia de medias siempre será conveniente contar con muestras de tamaño lo más parecido posible.

Ejemplo III–12 Calcular el intervalo de confianza para la diferencia de medias en dos métodos distintos empleado por Michelson para determinar la velocidad de la luz (expresamos la velocidad como $c = x + 299000$ km/s).

- Método i): 850, 740, 900, 1070, 930, 850, 950, 980; $n_1 = 8$.
- Método ii): 883, 816, 778, 796, 682, 711, 611, 599, 1051, 781, 578, 796; $n_2 = 12$.

Tenemos $n_1 + n_2 < 30$. Supondremos $\sigma_1 = \sigma_2$.

$$\begin{array}{lll} \overline{X}_1 = 908.75 & S_1 = 99.1 & n_1 = 8 \\ \overline{X}_2 = 756.83 & S_2 = 133.5 & n_2 = 12 \end{array}$$

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2} = \frac{7 \times 99.1^2 + 11 \times 133.5^2}{18} = 14710.6$$

$$\Rightarrow S_p = 121.3$$

Por otro lado, si usamos $\alpha = 0.05$, tenemos $t_{0.025,18} = 2.101$ (tablas). El intervalo será entonces

$$I = \left[(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2, n_1+n_2-2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right] =$$

$$= \left[(908.8 - 756.8) \pm 2.101 \times 121.3 \times S_p \sqrt{\frac{1}{8} + \frac{1}{12}} \right] = [152 \pm 116]\,.$$

El intervalo de confianza solicitado es entonces (36,268) km/s (+299000).

- **Varianzas poblacionales σ_1^2 y σ_2^2 desconocidas con $\sigma_1 \neq \sigma_2$ (muestras pequeñas):**

Veamos ahora el caso general en el que no se conocen las varianzas poblacionales, no se puede asumir que sean iguales y las muestras no son grandes. En este caso se puede hacer un desarrollo similar al anterior y definir un estadístico equivalente a (11.21) de la forma

$$t = \frac{(\overline{X_1} - \overline{X_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}. \tag{11.25}$$

Se puede demostrar que la variable anterior sigue aproximadamente una distribución t de Student con f grados de libertad, donde f es el entero más próximo a la aproximación de Welch

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2.$$

Al igual que en el apartado anterior, la inclusión de esta nueva variable conduce a

$$P\left((\overline{X_1} - \overline{X_2}) - t_{\alpha/2,f} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{X_1} - \overline{X_2}) + t_{\alpha/2,f} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right)$$

$$= 1 - \alpha. \tag{11.26}$$

Por lo tanto, el intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de dos poblaciones normales de varianzas desconocidas es

$$I = \left[(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2,f} \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}} \right]. \tag{11.27}$$

Ejemplo III–13 Repetir el ejemplo anterior, suponiendo ahora que $\sigma_1 \neq \sigma_2$.

$$f = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1} + \frac{(S_2^2/n_2)^2}{n_2+1}} - 2. = 19.8 \simeq 20.$$

Consultando en las tablas, obtenemos $t_{0.025,20} = 2.086$. Entonces

$$I = \left[(\overline{X_1} - \overline{X_2}) \pm t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right] =$$

$$I = \left[(908.8 - 756.8) \pm 2.086\sqrt{\frac{99.1^2}{8} + \frac{133.5^2}{12}}\right] = [152 \pm 109]\,.$$

El intervalo de confianza es ahora (43,261) km/s (+299000).

Para calcular los intervalos de confianza anteriores se ha supuesto que las poblaciones de partida son normales. Como consecuencia del teorema del límite central, para cualesquiera distribuciones de partida la distribución muestral de la diferencia de medias puede aproximarse por una normal siempre que el tamaño de las muestras sea suficientemente grande. En consecuencia, la expresión (11.19) sigue siendo aplicable para distribuciones no normales y muestras grandes. Un caso particular de este resultado es el siguiente:

- **Intervalo de confianza para la diferencia de proporciones**

  Supongamos que se quiere encontrar un intervalo de confianza para la diferencia entre los parámetros p_1 y p_2 de dos distribuciones binomiales. Un buen estimador puntual de esta diferencia es la diferencia de proporciones $P_1 - P_2$, donde P_1 es la proporción de éxitos en una muestra de tamaño n_1 de la primera población, y lo mismo para P_2. Teniendo en cuenta que la varianza de la distribución muestral de una proporción puede escribirse como: $\sigma_p = p(1-p)/n$, la varianza de la distribución muestral de la diferencia de proporciones será

  $$\sigma^2_{p_1-p_2} = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

  Por tanto, suponiendo que las muestras son grandes, y que, por lo tanto, la distribución muestral de la diferencia de proporciones es aproximadamente normal, se puede escribir, por analogía con (11.19), que el intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de proporciones es

  $$I = \left[(\overline{P_1} - \overline{P_2}) \pm z_{\alpha/2}\sqrt{\frac{\overline{P_1}(1-\overline{P_1})}{n_1} + \frac{\overline{P_2}(1-\overline{P_2})}{n_2}}\right]. \tag{11.28}$$

11.3. Intervalos de confianza para la varianza

A continuación se estudia cómo se puede calcular un intervalo de confianza para la varianza de una distribución normal. Supongamos que se extrae una muestra de tamaño n sobre la que se calcula la varianza muestral S^2. Por (9.21) sabemos que el estadístico $(n-1)S^2/\sigma^2$ sigue una distribución χ^2 con $n-1$ grados de libertad. Por lo tanto se puede expresar

$$P\left(\chi^2_{1-\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2,n-1}\right) = 1-\alpha,$$

donde $\chi^2_{\alpha/2,n-1}$ es la abscisa de la distribución χ^2 con $n-1$ grados de libertad que deja a su derecha un área

igual a $\alpha/2$, y de manera similar para $\chi^2_{1-\alpha/2,n-1}$. Nótese que aunque la distribución de χ^2 no es simétrica, el intervalo se ha escogido para que el área de las dos colas sea igual a $\alpha/2$.

Dividiendo cada término de la desigualdad por $(n-1)S^2$ e invirtiendo las desigualdades, se obtiene

$$P\left(\frac{\chi^2_{1-\alpha/2,n-1}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{\alpha/2,n-1}}{(n-1)S^2}\right) = 1-\alpha \quad \Rightarrow$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right) = 1-\alpha. \tag{11.29}$$

Por lo tanto, el intervalo de confianza de nivel $(1-\alpha)$ para la varianza de una distribución normal con varianza muestral S^2 es

$$I = \left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]. \tag{11.30}$$

Este intervalo no tiene por qué ser simétrico en torno a la varianza muestral. De la misma manera, el intervalo de confianza para la desviación típica de una población normal puede escribirse como

$$I = \left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}\right]. \tag{11.31}$$

Ejemplo III–14

Calcular el intervalo de confianza para la desviación típica de la segunda muestra del ejemplo III–12.

Ya vimos que $S = 133.5$ y $n = 12$. Por otro lado, consultando las tablas vemos que, para $\alpha/2 = 0.025$ tenemos

$$\chi^2_{0.025,11} = 21.920 \qquad \text{y} \qquad \chi^2_{0.975,11} = 3.816.$$

El intervalo será entonces

$$I = \left[\sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}\right] = \left[\sqrt{\frac{11\times 133.5^2}{21.920}}, \sqrt{\frac{11\times 133.5^2}{3.816}}\right],$$

lo que nos conduce al intervalo (94.6,226.7) km/s (+299000).

11.4. Intervalos de confianza para la razón de varianzas

Supongamos que se tienen dos poblaciones normales con varianzas σ_1^2 y σ_2^2. Vamos a estudiar cómo construir un intervalo de confianza para la razón de dichas varianzas a partir de dos muestras independientes de tamaños n_1 y n_2 y varianzas muestrales S_1^2 y S_2^2 respectivamente. Anteriormente se ha demostrado que, en este caso, el estadístico $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ sigue una distribución F de Fisher con (n_1-1) y (n_2-1) grados de libertad (9.23). Por lo tanto, se puede escribir

$$P\left(F_{1-\alpha/2;n_1-1,n_2-1} < \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} < F_{\alpha/2;n_1-1,n_2-1}\right) = 1-\alpha,$$

donde $F_{1-\alpha/2;n_1-1,n_2-1}$ y $F_{\alpha/2;n_1-1,n_2-1}$ son los valores de la distribución F, con (n_1-1) y (n_2-1) grados de libertad, que dejan a su derecha áreas iguales a $1-\alpha/2$ y $\alpha/2$ respectivamente. Multiplicando las desigualdades anteriores por S_2^2/S_1^2 e invirtiendo los términos, se obtiene

$$P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2;n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{F_{1-\alpha/2;n_1-1,n_2-1}}\right) = 1-\alpha.$$

Aplicando ahora la propiedad de la distribución F según la cual $F_{1-\beta;\nu_1,\nu_2} = 1/F_{\beta;\nu_2,\nu_1}$, se llega a:

$$P\left(\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2;n_1-1,n_2-1}} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}F_{\alpha/2;n_2-1,n_1-1}\right) = 1-\alpha. \tag{11.32}$$

Por lo tanto, el intervalo de confianza $(1-\alpha)$ para el cociente de varianzas de dos poblaciones normales independientes puede expresarse como

$$I = \left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2;n_1-1,n_2-1}}, \frac{S_1^2}{S_2^2}F_{\alpha/2;n_2-1,n_1-1}\right]. \tag{11.33}$$

y el intervalo para la razón de desviaciones típicas se obtiene tomando raices cuadradas en la expresión anterior.

Ejemplo III–15

Calcular el intervalo de confianza para la razón de varianzas de las dos poblaciones del ejemplo III–12.

$$S_1 = 99.1 \quad n_1 = 8 \quad S_1^2 = 9820.81$$
$$S_2 = 133.5 \quad n_2 = 12 \quad S_2^2 = 17822.25$$

$$\Rightarrow \frac{S_1^2}{S_2^2} = 0.5510$$

y además

$$F_{\alpha/2;n_1-1,n_2-1} = F_{0.025;7,11} = 3.7586$$
$$F_{\alpha/2;n_2-1,n_1-1} = F_{0.025;11,7} = \frac{4.7611+4.6658}{2} = 4.71345$$

Y el intervalo se calcula finalmente como

$$I = \left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2;n_1-1,n_2-1}}, \frac{S_1^2}{S_2^2}F_{\alpha/2;n_2-1,n_1-1}\right] = \left[\frac{0.5510}{3.7586}, 0.5510\times 4.7135\right],$$

por lo que el intervalo buscado es (0.15,2.60). Vemos que este intervalo es compatible con que las varianzas sean iguales.

11.5. Intervalos de confianza para datos apareados

En los apartados anteriores siempre que se ha trabajado con dos poblaciones se ha supuesto que éstas eran independientes. Pero éste no es siempre el caso. Vamos a suponer ahora que se tienen dos poblaciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ de las que se extraen dos muestras que no son independientes. Nos vamos a restringir al caso en el cual los tamaños n de ambas muestras son iguales entre si. Típicamente consideraremos la situación en la cual las muestras no se extraen de forma independiente de cada población, sino que cada muestra consiste en la medida de una característica en los mismos elementos de una población. Por ejemplo, supongamos que sobre los elementos de una muestra se mide cierta variable, después se aplica un determinado tratamiento a la muestra y, sobre los mismos elementos, se vuelve a medir la misma variable (ej. temperatura antes y después de aplicar un tratamiento). A este tipo de experimentos se le llama de **observaciones pareadas**.

El objetivo en este caso es calcular un intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ en dichas muestras. Para ello se consideran las diferencias $d_i = x_{1_i} - x_{2_i}$ $(i = 1, 2, \ldots, n)$ entre los valores de las variables en cada uno de los elementos de la muestra. Para plantear el problema se asume que estas diferencias son los valores de una nueva variable aleatoria D. Si la muestra es suficientemente grande (en la práctica $n > 30$) puede considerarse que dicha variable se distribuye normalmente con media $\mu_D = \mu_1 - \mu_2$ y

varianza σ_D^2. Las estimaciones puntuales de estos parámetros serán respectivamente $\overline{D}$ y S_D^2, que tomarán, para una muestra en particular, los valores concretos

$$\overline{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{\sum_{i=1}^n (x_{1_i} - x_{2_i})}{n},$$

$$s_d^2 = \frac{\sum_{i=1}^n (d_i - \overline{d})^2}{n-1}.$$

El problema se reduce entonces a calcular un intervalo de confianza para la media μ_D de una distribución normal. Por analogía con (11.7) y aproximando la varianza σ_D^2 por S_D^2 por ser la muestra grande, puede escribirse entonces

$$P\left(\overline{D} - z_{\alpha/2}\frac{S_D}{\sqrt{n}} < \mu_1 - \mu_2 < \overline{D} + z_{\alpha/2}\frac{S_D}{\sqrt{n}}\right) = 1 - \alpha, \tag{11.34}$$

donde se ha igualado μ_D a $\mu_1 - \mu_2$. Por lo tanto, el intervalo de confianza de nivel $(1-\alpha)$ para la diferencia de medias de observaciones pareadas con $n > 30$ puede expresarse como

$$I = \left[\overline{D} \pm z_{\alpha/2}\frac{S_D}{\sqrt{n}}\right]. \tag{11.35}$$

En el caso de que la muestra fuera pequeña ($n < 30$) habría que substituir la distribución normal por una distribución t, siendo el intervalo de confianza

$$I = \left[\overline{D} \pm t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}\right]. \tag{11.36}$$

Ejemplo III–16

Se aplica un proceso para aumentar el rendimiento en 10 fábricas muy diferentes (no dejar tomarse el bocadillo a media mañana). Los rendimientos (en ciertas unidades, como toneladas/día) antes y después son:

antes	13	22	4	10	63	18	34	6	19	43	X_1
después	15	22	2	15	65	17	30	12	20	42	X_2

Calcular el intervalo de confianza para el aumento del rendimiento.
Si definimos las diferencias como

$$D_i = X_{2,i} - X_{1,i}$$

obtenemos: D_i = 2, 0, -2, 5, 2, -1, -4, 6, 1, -1. Con estos datos ya podemos calcular

$$\overline{D} = \frac{\sum_i D_i}{n} = \frac{8}{10} = 0.8$$

$$S_D = \sqrt{\frac{\sum_{i=1}^n (d_i - \overline{d})^2}{n-1}} = 3.08$$

Como el número de datos es menor que 30, usamos $t_{0.025,9} = 2.262$ (tablas). El intervalo que buscamos será entonces

$$I = \left[\overline{D} \pm t_{\alpha/2,n-1}\frac{S_D}{\sqrt{n}}\right] \left[0.8 \pm 2.262\frac{3.08}{\sqrt{10}}\right] = [0.8 \pm 2.2],$$

es decir, (−1.4,3.0).

11.6. Determinación del tamaño de la muestra

Hasta ahora siempre se ha supuesto conocido el tamaño de la muestra n. Sin embargo, y fundamentalmente en el diseño de experimentos, en ocasiones el problema principal es la determinación del tamaño muestral

requerido para obtener la estimación de los parámetros poblacionales con una determinada precisión. Nótese que una muestra demasiado grande puede traducirse en una perdida de tiempo y dinero, mientras que, si la muestra es demasiado pequeña, no se obtendrá la fiabilidad deseada y el experimento será un fracaso.

La precisión de una estimación por intervalos de confianza vendrá marcada por la longitud del intervalo (en ocasiones, llamada error). Para ilustrar el problema supongamos que tenemos una distribución normal y que queremos determinar la media poblacional μ a partir de la media muestral $\overline{X}$. El intervalo de confianza vendrá entonces dado por (11.5), de manera que la longitud l del intervalo es

$$l = 2z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

Es decir, la longitud del intervalo es inversamente proporcional al tamaño de la muestra y la precisión aumenta, por tanto, al aumentar n. El problema se plantea entonces en cómo calcular el tamaño de la muestra n para estimar la media poblacional con una cierta precisión, es decir, para que la diferencia entre la media poblacional y muestral sea, en valor absoluto y con un cierto nivel de confianza $(1-\alpha)$, menor que un cierto error, denotado por ϵ

$$P(\overline{X} - \epsilon < \mu < \overline{X} + \epsilon) = 1 - \alpha.$$

De esta forma, comparando la expresión anterior con (11.4), una vez fijado α puede calcularse n igualando el error ϵ a la semilongitud del intervalo $(l/2)$

$$\epsilon = z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \qquad \Rightarrow \qquad n = z_{\alpha/2}^2\frac{\sigma^2}{\epsilon^2}. \tag{11.37}$$

Es decir, si se utiliza $\overline{X}$ como una estimación de μ, puede tenerse una confianza del $(1-\alpha)100\,\%$ de que, en una muestra del tamaño anterior, el error no excederá a un valor ϵ.

Para poder aplicar la expresión anterior es necesario conocer previamente σ. Si éste no es el caso, en la práctica se toma una muestra piloto pequeña (aunque es deseable que $n > 30$) para poder estimar σ mediante la desviación típica muestral S.

Ejemplo III–17

En el ejemplo III–1, ¿cuál ha de ser el tamaño de la muestra para poder determinar la media con un error de 0.5?

$$n = z_{\alpha/2}^2\frac{\sigma^2}{\epsilon^2}$$

En este caso tenemos $z_{0.025} = 1.96$, $\sigma = 4$ y $\epsilon = 0.5$. Por tanto, $n = 245.86 \simeq 246$.

Tema IV

CONTRASTE DE HIPÓTESIS

Capítulo 12

Contrastes de hipótesis

"La primera condición de una hipótesis es que debe poder entenderse."

Thomas Henry Huxley (1825-1895)

Las aplicaciones de la estadística a la investigación científica van mucho más allá de la estimación de parámetros poblacionales vista en el tema anterior. Típicamente, el método científico se caracteriza por basarse en la construcción de hipótesis, o modelos, lo más simples posibles de cómo funciona cierto aspecto de la naturaleza, y la comprobación o refutación de tales hipótesis por medio de la experimentación. A través del **contraste de hipótesis**, la estadística proporciona procedimientos óptimos para decidir la aceptación o el rechazo de afirmaciones o hipótesis acerca de la población en estudio. Las hipótesis se contrastan comparando sus predicciones con los datos experimentales. Si coinciden dentro de un margen de error, la hipótesis se mantiene. En caso contrario se rechaza y hay que buscar hipótesis o modelos alternativos que expliquen la realidad. De esta manera, el contraste de hipótesis juega un papel fundamental en el avance de cualquier disciplina científica.

12.1. Ensayos de hipótesis

Una **hipótesis estadística** es una afirmación o conjetura que se hace sobre una, o varias, características de una población. Ejemplos de dichas afirmaciones incluyen el que la media de una población tenga un determinado valor, o que los valores de una variable presenten menor dispersión en torno a un valor medio en una población comparada con la dispersión en otra, etc. Evidentemente, la forma más directa de comprobar tales hipótesis sería estudiando todos y cada uno de los elementos de la población. Sin embargo, frecuentemente esto no es posible (la población podría ser incluso infinita), por lo que el contraste de la hipótesis ha de basarse en una muestra, que supondremos aleatoria, de la población en estudio. Al no estudiarse la población entera, nunca podremos estar completamente seguros de si la hipótesis realizada es verdadera o falsa. Es decir, siempre existe la probabilidad de llegar a una conclusión equivocada.

Los métodos de **ensayos de hipótesis** que se tratan en este tema permitirán estudiar si, en términos de probabilidad, la hipótesis de partida puede ser aceptada o debe ser rechazada. Debe quedar claro que el rechazo de una hipótesis implica que la evidencia de la muestra la refuta. Es decir, que existe una probabilidad muy pequeña de que, siendo la hipótesis verdadera, se haya obtenido una muestra como la estudiada. Por otro lado, una hipótesis se aceptará cuando la muestra no proporcione evidencias suficientes para refutarla, lo cual no quiere decir que la hipótesis sea verdadera. Por ejemplo, si se ha hecho la hipótesis de que la media de una población es cero, y se encuentra que los valores tomados tienen, por ejemplo, media 0.1 y desviación

típica 10, podremos llegar a la conclusión de aceptar la hipótesis, lo cual no descarta que la media real de la población sea, por ejemplo, 0.2.

El primer paso en un proceso de ensayo de hipótesis es la formulación de la hipótesis estadística que se quiere aceptar o rechazar. Comunmente, se formulan las hipótesis estadísticas con el propósito de rechazarlas para así probar el argumento deseado. Por ejemplo, para demostrar que un producto es mejor que otro, se hace la hipótesis de que son iguales, es decir, que cualquier diferencia observada es debida únicamente a fluctuaciones en el muestreo. O por ejemplo, si se quiere demostrar que una moneda está trucada (no existe la misma probabilidad de que salga cara o cruz) se hace la hipótesis de que no está trucada (es decir, la probabilidad p de cara o cruz es siempre 0.5) y a continuación se estudia si los datos de la muestra llevan a un rechazo de esa hipótesis. Por este motivo, a la hipótesis de partida que se quiere contrastar se la llama **hipótesis nula**, y se representa por H_0. La hipótesis nula es por tanto la hipótesis que se acepta o rechaza como consecuencia del contraste de hipótesis. Por otra parte, la hipótesis que se acepta cuando se rechaza H_0 es la **hipótesis alternativa**, denotada por H_1. Es decir, si se acepta H_0 se rechaza H_1 y al contrario. En el ejemplo de la moneda trucada la hipótesis nula sería $p = 0.5$ y la hipótesis alternativa $p \neq 0.5$. En muchas ocasiones una hipótesis nula referida a un parámetro poblacional especificará un valor exacto del parámetro, mientras que la hipótesis alternativa incluirá la posibilidad de varios valores. Por otra parte, cuando se trate de comparar dos poblaciones, la hipótesis nula suele ser que las dos poblaciones tienen el mismo parámetro (ejemplo, media) y la alternativa, que los parámetros son diferentes.

Es importante recordar que la hipótesis nula, aunque se acepte, nunca se considera probada (por ejemplo, para probar que exactamente la media de una población tiene un determinado valor, habría que estudiar todos los elementos de la población). Sin embargo, sí puede rechazarse. Así, si suponiendo que H_0 es cierta, se encuentra que los resultados observados en una muestra aleatoria difieren marcadamente de los que cabría esperar teniendo en cuenta la variación propia del muestreo, se dice que las diferencias son **significativas** y se rechaza H_0.

Para realizar un contraste de hipótesis se utiliza un **estadístico de prueba** (también llamado función de decisión del contraste) cuya distribución muestral se supone conocida si la hipótesis nula H_0 es verdadera. Así, por ejemplo, si H_0 es que en una población normal la media tiene un determinado valor μ, el estadístico de prueba será la media muestral $\overline{X}$, cuya distribución tendrá media μ y desviación típica $\sigma/\sqrt{n}$. Una vez elegida una muestra, se medirá el estadístico de prueba y se comprobará si el valor que toma es compatible con la distribución muestral esperada si H_0 fuese cierta. Si el valor medido difiere considerablemente de los valores esperados, la hipótesis nula se rechazará. Todos los posibles valores del estadístico que llevan a rechazar H_0 constituyen la **región crítica** del contraste. Por el contrario, todos los valores que llevan a una aceptación de H_0 determinan la **región de aceptación**. En el ejemplo anterior, los valores de $\overline{X}$ próximos a μ determinarán la región de aceptación, mientras que los alejados de μ constituirán la región crítica.

12.2. Tipos de errores y significación

Como ya se ha indicado, un ensayo de una hipótesis estadística nunca es infalible, en el sentido de que siempre existe una probabilidad de cometer un error en las conclusiones del contraste. Este error es básicamente debido a la limitación de información intrínseca a la muestra. Diferenciaremos entre dos tipos posibles de errores:

- Si se rechaza la hipótesis H_0 cuando es verdadera se dice que se comete un **error de tipo I**.
- Si se acepta la hipótesis H_0 cuando es falsa se dice que se comete un **error de tipo II**.

En cualquiera de los dos casos se comete un error al tomar una decisión equivocada. Estos dos tipos de errores se resumen en la siguiente tabla:

	H_0 verdadera	H_0 falsa
Se acepta H_0	Decisión correcta	Error tipo II
Se rechaza H_0	Error tipo I	Decisión correcta

Una definición importante es la siguiente: se define **nivel de significación** α de un contraste de hipótesis a la probabilidad de cometer un error de tipo I. Es decir, si se repitiera un gran número de veces un contraste de hipótesis y H_0 fuese verdadera, en el $100(1-\alpha)\,\%$ de los casos llegaríamos a la conclusión correcta de aceptar H_0 y el $100\alpha\,\%$ de las veces cometeríamos el error de rechazar H_0. Normalmente, el nivel de significación se fija antes de realizar el contraste. Nótese que el valor de α es el que determina los tamaños de la región crítica y la región de aceptación, de forma que a menor α mayor será el tamaño de la región de aceptación (o menor el de la región crítica), al ser menor la probabilidad de equivocarse y rechazar H_0 cuando es verdadera. Típicamente se suelen tomar niveles de significación fijos de 0.05 ó 0.01, aunque cualquier valor es en principio posible. Cuando, por ejemplo, se usa $\alpha = 0.05$ se dice que la hipótesis se acepta o se rechaza al nivel de significación 0.05. Evidentemente, interesa que dicho nivel de significación sea lo más pequeño posible. Sin embargo esto no puede hacerse sin tener también en cuenta los posibles errores de tipo II.

Ejemplo IV–1

Se quiere probar si una moneda está trucada. Para ello se lanza la moneda 10 veces y se anota el número de caras. El proceso seguirá una distribución binomial.

$$\text{Hipótesis nula } H_0\text{: } p = 0.5$$
$$\text{Hipótesis alternativa } H_1\text{: } p \neq 0.5$$

El estadístico de prueba es la proporción de éxitos

$$\overline{P} = \frac{\text{número de caras}}{\text{número de ensayos}}$$

Aceptando H_0 como hipótesis inicial, vamos a calcular las probabilidades de que el estadístico de prueba esté dentro de diferentes intervalos. Usamos la tabla de la distribución binomial.

$$P(0.4 \leq \overline{P} \leq 0.6) = \sum_{x=4}^{10} b(x;10,0.5) - \sum_{x=7}^{10} b(x;10,0.5) = 0.828 - 0.172 = 0.656.$$

Y, de la misma forma,

$$P(0.3 \leq \overline{P} \leq 0.7) = 0.890$$
$$P(0.2 \leq \overline{P} \leq 0.8) = 0.978$$
$$P(0.1 \leq \overline{P} \leq 0.9) = 0.998$$

Si nos fijamos, por ejemplo, en $P(0.2 \leq \overline{P} \leq 0.8) = 0.978$, vemos que entonces podemos también escribir $P(X = 0, 1, 9, 10) = 1 - 0.978 = 0.022$, donde X es el estadístico número de caras. En este caso definiríamos las regiones críticas y de aceptación como

$$\text{A: } \{x : 2 \leq x \leq 8\}$$
$$\text{C: } \{x : x < 2 \text{ o } x > 8\}$$

Según esto, la probabilidad de comer un error de tipo I (o rechazar la hipótesis nula cuando es verdadera) es 0.02. Es decir, $\alpha = 0.02$, donde α es el nivel de significación. En resumen, nos equivocaremos en un $2\,\%$ de los casos.

La probabilidad de cometer un error de tipo II, denotada por β, es típicamente imposible de calcular a no ser que se tenga una hipótesis alternativa específica. Por ejemplo, en el contraste de la media de una población, si la media real μ' fuese un valor muy cercano a la media que estamos suponiendo en la hipótesis H_0, la probabilidad de cometer un error de tipo II sería muy alta, pero no la podemos conocer a priori a

no ser que se supongan ciertos valores para μ'. En otras palabras, si la hipótesis nula es falsa, β aumenta cuando el valor verdadero del parámetro se acerca al valor hipotético establecido en H_0. Cuanto mayor es la diferencia entre dicho valor hipotético y el real, menor será β. Típicamente, los errores de tipo II han de acotarse imponiendo que, si hubiese una diferencia que se considere significativa entre el valor supuesto en H_0 y el valor real, la probabilidad β de cometer un error de tipo II (y aceptar H_0 cuando es falsa) no sea mayor que un determinado valor.

Es claro que los errores de tipo I y tipo II se relacionan entre sí. Desafortunadamente, para una muestra dada, una disminución en la probabilidad de uno se convierte en un aumento en la probabilidad del otro. De forma que normalmente no es posible reducir ambos errores simultáneamente. La única forma en que esto es posible es aumentando el tamaño de la muestra. Para cada caso particular, habrá que estudiar cuál de los dos tipos de errores es más importante controlar, y fijar las regiones de aceptación y crítica de forma que se acote el error menos deseable de los dos. Para disminuir α se disminuye el tamaño de la región crítica, y lo contrario para β. Esto nos lleva a un concepto importante en el contraste de hipótesis: se denomina **potencia de una prueba** a la probabilidad de rechazar la hipótesis nula H_0 cuando es falsa. Es decir, su valor es $1-\beta$ y, depende, por tanto, del verdadero valor del parámetro. La potencia de una prueba se puede considerar como una medida de la sensibilidad para detectar diferencias en los valores del parámetro. Si se fija de antemano el nivel de significación, se elegirá siempre el tipo de contraste que presente una potencia mayor para un determinado tamaño muestral.

Ejemplo IV–2

En el ejemplo anterior, para calcular la probabilidad de cometer un error de tipo II debemos suponer un valor conocido para la proporción de éxitos, p_{verd}.

a) Supongamos que $p_{\text{verd}} = 0.7$. Entonces

$$\beta = P(2 \leq X \leq 8, \text{ dado que } p_{\text{verd}} = 0.7) = \sum_{x=2}^{10} b(x; 10, 0.7) - \sum_{x=9}^{10} b(x; 10, 0.7) = 1.000 - 0.149 = 0.851.$$

b) Supongamos que $p_{\text{verd}} = 0.9$. Entonces

$$\beta = P(2 \leq X \leq 8, \text{ dado que } p_{\text{verd}} = 0.9) = \sum_{x=2}^{10} b(x; 10, 0.9) - \sum_{x=9}^{10} b(x; 10, 0.9) = 1.000 - 0.736 = 0.264.$$

La potencia de la prueba (probabilidad de rechazar H_0 cuando es falsa) sería
a) $1-\beta = 0.149$
b) $1-\beta = 0.736$
Sería necesario aumentar el tamaño de la muestra para obtener potencias mayores.

Con el fin de ilustrar los conceptos expuestos anteriormente supongamos que se quiere hacer un contraste sobre la media de una población normal. La hipótesis nula H_0 es en este caso $\mu = \mu_0$. Como estadístico de prueba se utiliza la media muestral, que como sabemos, si H_0 es cierta, seguirá un distribución $N(\mu_0, \sigma/\sqrt{n})$. Es decir, la variable dada por $Z = (\overline{X} - \mu_0)/(\sigma/\sqrt{n})$ sigue una distribución normal tipificada.

Por las propiedades de la distribución normal, sabemos que, si H_0 es cierta, el 95 % de las veces el estadístico Z se situaría entre los valores -1.96 y 1.96 mientras que sólo un 5 % de las veces obtendríamos valores mayores que 1.96 o menores que -1.96. Esto quiere decir que, para un nivel de significación de $\alpha = 0.05$ la región de aceptación estaría definida por los valores del intervalo $(-1.96, 1.96)$ mientras que la región crítica estaría dada por $(-\infty, -1.96)$ y $(1.96, \infty)$. Es decir, la probabilidad de que cometer un error de tipo I (o el nivel de significación) ha de coincidir con el área de la región crítica. De esta manera, cuando se obtuviese un valor de $\overline{X}$ situado en la región crítica rechazaríamos la hipótesis nula al nivel de significación 0.05, mientras que la aceptaríamos en caso contrario. Nótese que si H_0 fuese falsa pero el valor verdadero de μ estuviese muy próximo a μ_0 tendríamos una probabilidad muy alta de aceptar H_0, y por lo tanto de cometer un error de tipo II.

El ejemplo anterior nos permite ver cómo el contraste de hipótesis está íntimamente relacionado con la

estimación por intervalos de confianza vista en el tema anterior. Efectivamente, en dicho ejemplo, el intervalo de confianza del $(1-\alpha)\,\%$ para la media μ_0 viene dado por

$$P\left(\overline{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu_0 < \overline{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha \qquad \Rightarrow$$

$$P\left(-z_{\alpha/2} < \frac{\overline{x}-\mu_0}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1-\alpha.$$

y esto coincide con la región de aceptación para un nivel de significación α. Es decir, el contraste de la hipótesis H_0 (en este caso, $\mu = \mu_0$) con un nivel de significación α es equivalente a calcular un intervalo de nivel de confianza $1-\alpha$ y rechazar H_0 si la media muestral no está dentro del intervalo. De esta forma, generalmente se puede emplear el intervalo de confianza para realizar el contraste de hipótesis. Este resultado se puede extender a los intervalos de confianza de varianzas, diferencia de medias, etc.

Ejemplo IV–3

Supongamos que tiramos una moneda 100 veces. Como n es grande, bajo la hipótesis nula $H_0 : p = 0.5$, tenemos que $\overline{p}$ sigue una distribución normal de media 0.5 y desviación típica $\sigma = \sqrt{\overline{p}(1-\overline{p})/n}$, es decir

$$N\left(\overline{p}, \sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right) = N(0.5, 0.05).$$

Podemos construir una distribución normal tipificada utilizando

$$z = \frac{\overline{p}-0.5}{0.05}$$

Para buscar la región de aceptación y la región crítica tomamos como nivel de significación $\alpha = 0.05$. En ese caso, $z_{\alpha/2} = 1.96$. Es decir

$$+1.96 = \frac{\overline{p}-0.5}{0.05} \Rightarrow \overline{p} = 0.598 \Rightarrow x = \overline{p}\times n = 59.8 \text{ caras}$$

$$-1.96 = \frac{\overline{p}-0.5}{0.05} \Rightarrow \overline{p} = 0.402 \Rightarrow x = \overline{p}\times n = 40.2 \text{ caras}$$

Entonces podemos decir que, con un nivel de confianza del 95 %,

A: $\{40 < x < 60\}$
C: $\{x \le 40 \text{ y } x \ge 60\}$

Dicho de otra forma, si obtenemos un número de caras comprendido entre 40 y 60, no podemos rechazar H_0 (al nivel de significación elegido).

Calculemos ahora la probabilidad de cometer un error de tipo II.
a) Si $p_{\text{verd}} = 0.7 \Rightarrow N\left(0.7, \sqrt{\frac{0.7\times 0.3}{100}}\right) = N(0.7, 0.0458)$. Usando $z = (\overline{p}-0.7)/0.0458$,

$$\beta = P(40 < x < 60) = P(0.4 < \overline{p} < 0.6) = P(-6.55 < z < -2.18) = 0.0146.$$

La potencia será $1-\beta = 0.9854$ (probabilidad de rechazar H_0 siendo falsa). Es la probabilidad de que si $p_{\text{verd}} = 0.7$ nuestro experimento detecte esa diferencia.
b) Si $p_{\text{verd}} = 0.9 \Rightarrow N\left(0.9, \sqrt{\frac{0.9\times 0.1}{100}}\right) = N(0.9, 0.03)$. Usando $z = (\overline{p}-0.9)/0.03$,

$$\beta = P(40 < x < 60) = P(0.4 < \overline{p} < 0.6) = P(-16.67 < z < -10.) \simeq 0.0.$$

La potencia será $1-\beta \simeq 1.0$ (seguro que lo detectamos; la moneda es "muy falsa" y hemos realizado muchos lanzamientos).

12.3. Contrastes bilaterales y unilaterales

En el ejemplo anterior se ha visto como la región crítica se dividía en dos intervalos de la recta representada por los valores posible del estadístico. En general, a un contraste de hipótesis en el que la región crítica se divide en dos partes se le llama **bilateral** y se dice que se hace un ensayo de **dos colas** (ver Fig. 12.1). Generalmente, aunque no siempre, el área de cada cola suele coincidir con la mitad del nivel de significación. Por ejemplo, si el contraste se hace sobre el valor de la media poblacional μ las hipótesis nula y alternativa tendrán típicamente la siguiente forma

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases} \tag{12.1}$$

Es decir, se intenta probar si el parámetro puede tomar un determinado valor o si, por el contrario, ha de ser diferente (sin importar que sea mayor o menor). Otro ejemplo sería el contraste sobre la igualdad de medias de dos poblaciones. En este caso la hipótesis nula es que las dos medias coinciden y la alternativa es que son diferentes

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} \tag{12.2}$$

A veces interesa comprobar si un parámetro es mayor (o menor) que un determinado valor. Es decir, no sólo interesa que sea diferente sino que hay que comprobar la hipótesis de que la diferencia vaya en un cierto sentido. En estos casos se define un contraste **unilateral**, o un ensayo de **una cola**, como aquel en el que la región crítica está formada por un único conjunto de puntos de la recta real. En este caso, el área de la única región crítica ha de coincidir con el nivel de significación (ver Fig. 12.1). Por ejemplo, si se quiere comprobar que la media de una población es mayor que un cierto valor se plantearán las siguientes hipótesis

$$\begin{cases} H_0: & \mu \leq \mu_0 \\ H_1: & \mu > \mu_0 \end{cases} \tag{12.3}$$

En este caso la región crítica cae en la cola derecha del estadístico de prueba, mientras que la cola izquierda forma parte de la región de aceptación. Otro ejemplo es aquel en el que interesa comprobar si la media de una población es mayor que la de otra. En este caso

$$\begin{cases} H_0: & \mu_1 \leq \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases} \tag{12.4}$$

Nótese que, para un mismo nivel de significación que en el caso bilateral, en el contraste unilateral la abscisa en la que comienza la región crítica (llamada **valor crítico**) ha de disminuir para que se conserve el área total (comparar gráficas izquierda y derecha en la Fig. 12.1).

En la siguiente tabla se dan los valores críticos para ensayos de una y dos colas y diferentes niveles de significación en el caso de que el estadístico siga una distribución normal:

Nivel de significación α	0.10	0.05	0.01	0.005	0.001
$\|z\|$ crítico (unilateral)	1.282	1.645	2.326	2.576	3.090
$\|z\|$ crítico (bilateral)	1.645	1.960	2.576	2.807	3.291

Es importante hacer notar que el hecho de hacer un contraste unilateral o bilateral depende de la conclusión que se quiera extraer y es algo que, en general, hay que decidir a priori, es decir, antes de realizar las medidas y los cálculos.

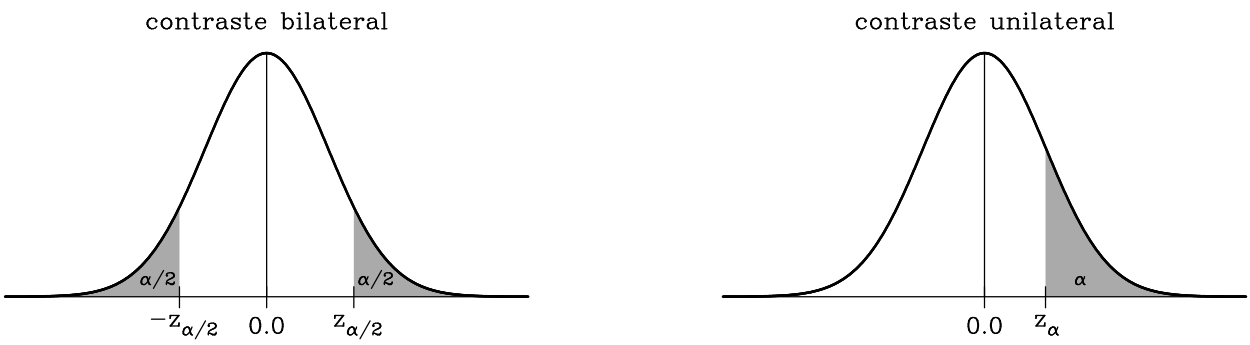


Figura 12.1: *Contrastes bilaterales y unilaterales: en la figura de la izquierda se muestran sombreadas las dos regiones críticas de un contraste bilateral, en el que el área de cada cola es $\alpha/2$, es decir, la mitad del nivel de significación. En la figura de la derecha se muestra la única región crítica de un contraste unilateral, cuya área ha de coincidir en este caso con el nivel de significación.*

Ejemplo IV–4

Necesitamos utilizar un contraste unilateral para probar que una moneda está cargada para sacar más caras:

$$H_0\text{: } p \leq 0.5$$
$$H_1\text{: } p > 0.5$$

Si, como en el ejemplo anterior, suponemos $n = 100$, tenemos $z_{0.05} = 1.645$ y

$$z = \frac{\overline{p} - 0.5}{0.05}.$$

Es decir

$$1.645 = \frac{\overline{p} - 0.5}{0.05} \Rightarrow \overline{p} = 0.582.$$

Las regiones crítica y de aceptación será entonces

$$\text{A: } \{x : x \leq 58\}$$
$$\text{C: } \{x : x > 58\}$$

Si $x \in A$ no podemos rechazar H_0 (incluso con 58 caras).

12.4. Fases de un contraste de hipótesis

Como resumen de los conceptos vistos hasta ahora, a continuación se especifican los procedimientos que hay que seguir para realizar un contraste de hipótesis:

1. Establecer cuáles son las hipótesis nula H_0 y alternativa H_1. En este momento habrá que decidir si el contraste va a ser unilateral o bilateral para así elegir entre las formulaciones (12.1) y (12.2) o (12.3) y (12.4).

2. Elegir un nivel de significación α.

3. Especificar el tamaño muestral n. En ocasiones, dicho tamaño viene dado antes de hacer el contraste. Sin embargo, cuando se está diseñando un experimento habrá que elegir un tamaño muestral óptimo. Normalmente esto se hace, para un α fijo, acotando los errores de tipo II que nos podemos permitir.

4. Seleccionar el estadístico de prueba apropiado. Nótese que la distribución muestral de este estadístico se supone conocida bajo la hipótesis de que H_0 es verdadera.

5. Determinar la región crítica a partir del tipo de estadístico de prueba y el nivel de significación deseado.

6. Calcular el valor del estadístico a partir de los datos de la muestra particular que se tenga.

7. Tomar la decisión estadística apropiada. Es decir, rechazar H_0 si el estadístico toma un valor en la región crítica, o aceptarla (o como mínimo, no rechazarla) en caso contrario.

Capítulo 13

Contrastes de hipótesis para una población

"Los grandes conocimientos engendran las grandes dudas."

Aristóteles (384-322 a.C.)

En este tema se presentan los contrastes de hipótesis para diferentes parámetros poblacionales de una única población. Debido a la íntima relación existente entre los contrastes de hipótesis y los intervalos de confianza, utilizaremos las expresiones vistas en temas anteriores para estos últimos para describir los contrastes. En todo lo siguiente se supone que se tiene un muestreo con reemplazamiento o en una población infinita. En otro caso habrá que hacer las modificaciones necesarias en las expresiones ya vistas.

13.1. Contraste de la media de una población normal

Supongamos que se tiene una población normal de la cual se extrae una muestra aleatoria descrita por $X_1, X_2, \ldots, X_n$. Como estimador de la media poblacional se usará la media muestral $\overline{X} = \sum_{i=1}^{n} X_i/n$, que, en una muestra en particular tomará el valor $\overline{x}$. A continuación se describen los contrastes de hipótesis para la media de la población. Al igual que para calcular los intervalos de confianza, se distinguirán varios casos:

13.1.1. Varianza σ^2 conocida

a) Constraste bilateral

En este caso, las hipótesis nula y alternativa serán respectivamente

$$\begin{cases} H_0: & \mu = \mu_0 \\ H_1: & \mu \neq \mu_0 \end{cases} \tag{13.1}$$

Es decir, se intenta contrastar si la media de la población tiene un determinado valor μ_0, o si, por el contrario, la media ha de ser distinta. En este caso, si se supone H_0 verdadera sabemos que la distribución muestral de medias será normal con media $\mu_{\overline{X}} = \mu_0$ y $\sigma^2_{\overline{X}} = \sigma^2/n$. Por lo tanto, se puede definir el siguiente estadístico que seguirá una normal tipificada (en el caso de que $\mu = \mu_0$) y tomará valores

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}. \tag{13.2}$$

Además, podemos establecer que, en el caso de que H_0 fuese cierta, z se distribuiría de forma que

$$P\left(-z_{\alpha/2} < z < z_{\alpha/2}\right) = 1 - \alpha,$$

donde $z_{\alpha/2}$ es la abscisa de la normal $N(0,1)$ que deja a su derecha un área de probabilidad igual a $\alpha/2$.

Es decir, existiría una probabilidad α (nivel de significación) de encontrar $\overline{x}$ fuera de ese intervalo. Esto nos define entonces la región de aceptación A y crítica C del contraste como

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}. \tag{13.3}$$

En otras palabras, si se encuentra que

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} \leq z_{\alpha/2}, \tag{13.4}$$

se acepta H_0. Por el contrario, si

$$\frac{|\overline{x} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2},$$

la hipótesis nula se rechaza al nivel de significación α.

Ejemplo IV–5

Se hacen 50 medidas de la acelaración de la gravedad, g, y se obtienen valores que conducen a $\overline{x} = 9.9$ m/s^2. Se sabe que, por el error en el método, $\sigma = 0.4$ m/s^2. ¿Es el valor medio significativamente diferente del valor esperado de g ($\mu_0 = 9.8$ m/s^2)?

Seguimos los pasos del contraste de hipótesis:

1. Establecemos las hipótesis nula y alternativa
$$\begin{cases} H_0 : \mu = 9.8 \\ H_1 : \mu \neq 9.8 \end{cases}$$
2. Fijamos el nivel de significación: $\alpha = 0.05$.
3. Especificamos el tamaño muestral: $n = 50$.
4. Seleccionamos el estadístico de prueba adecuado: si H_0 es correcta, entonces $z = (\overline{x} - 9.8)/(\sigma/\sqrt{n})$ sigue una distribución normal tipificada.
5. La región crítica será entonces: $C = \{z : |z| > z_{\alpha/2}\}$, donde $z_{\alpha/2} = z_{0.025} = 1.96$.
6. Calculamos el valor del estadístico:
$$|z| = \frac{|9.9 - 9.8|}{0.4/\sqrt{50}} = 1.77 < 1.96$$
7. Como $|z| < z_{\alpha/2} \Rightarrow$ no se rechaza H_0.

b) Contraste unilateral

En este caso las hipótesis nula y alternativa serían del tipo

$$\begin{cases} H_0 : & \mu \leq \mu_0 \\ H_1 : & \mu > \mu_0 \end{cases} \tag{13.5}$$

donde estamos contrastando si la media de la población puede o no ser mayor que un determinado valor. También podrían invertirse las desigualdades y hacer el contraste de una cola contrario. Se define aquí el mismo estadístico z (13.2) que para el contraste bilateral.

La región crítica se sitúa en este caso en la cola derecha de la distribución, de forma que podemos establecer que

$$A = \{z : z \leq z_\alpha\} \qquad ; \qquad C = \{z : z > z_\alpha\}, \tag{13.6}$$

donde z_α es la abscisa de la normal $N(0,1)$ que deja a su derecha un área de probabilidad igual a α. Es decir, solo se rechaza H_0 si la media muestral toma un valor mucho mayor que el supuesto en la hipótesis nula.

En otras palabras, si se encuentra que

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_\alpha, \tag{13.7}$$

se acepta H_0. Por el contrario, si

$$\frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_\alpha,$$

la hipótesis nula se rechaza al nivel de significación α.

Ejemplo IV–6

Con los datos del ejemplo anterior, queremos probar si el valor obtenido es significativamente mayor que $\mu_0 = 9.8$ m/s^2.

Es un contraste unilateral

$$\begin{cases} H_0 : \mu \leq 9.8 \\ H_1 : \mu > 9.8 \end{cases}$$

Usamos el mismo nivel de significación ($\alpha = 0.05$), $\overline{x}$ y n. La región crítica será ahora $C = \{z : z > z_\alpha\}$, donde $z_\alpha = z_{0.05} = 1.645$.

Calculamos el estadístico

$$z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} = 1.77$$

Como $z > z_\alpha$, rechazamos H_0 al nivel de significación $\alpha = 0.05$.

13.1.2. Varianza σ^2 desconocida y $n > 30$

En el caso común de desconocer la varianza poblacional, no puede aplicarse estrictamente el estadístico z dado en (13.2) para hacer el contraste de hipótesis. Sin embargo, si la muestra es grande, la varianza muestral definida como $S^2 = \sum_{i=1}^n (X_i - \overline{X})^2/(n-1)$ puede considerarse como un estimador preciso de la varianza poblacional. Por lo tanto, y de forma aproximada (en la práctica para $n > 30$) el contraste de hipótesis sobre la media se puede realizar igual que en el caso anterior sustituyendo σ por s en el estadístico z

$$z = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}, \tag{13.8}$$

y los contrastes, con las mismas hipótesis nulas y alternativas expresadas en (13.1) y (13.5), quedan:

a) Constraste bilateral

Las regiones de aceptación y crítica son

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}$$

Es decir, si

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq z_{\alpha/2}, \tag{13.9}$$

se acepta H_0. Por el contrario, H_0 se rechaza al nivel de significación α si

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} > z_{\alpha/2}.$$

b) Contraste unilateral

En este caso las regiones de aceptación y crítica se expresan como

$$A = \{z : z \leq z_\alpha\} \qquad ; \qquad C = \{z : z > z_\alpha\}.$$

Por tanto si se encuentra que

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq z_\alpha, \tag{13.10}$$

se acepta H_0. Por el contrario, si

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} > z_\alpha,$$

la hipótesis nula se rechaza al nivel de significación α.

13.1.3. Varianza σ^2 desconocida y $n \leq 30$

En el caso de que la varianza poblacional sea desconocida y la muestra sea pequeña no se considera válido suponer que el estadístico (13.8) sigue una distribución normal. En este caso, el contraste de hipótesis sobre la media puede hacerse definiendo un nuevo estadístico t

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} \tag{13.11}$$

y utilizando que, como se estudió en el tema anterior, esta nueva variable sigue una distribución t de Student con $n - 1$ grados de libertad.

Entonces, los contrastes para la media, con las mismas hipótesis nulas y alternativas expresadas en (13.1) y (13.5), son iguales que para el caso de varianza conocida pero sustituyendo σ por la desviación típica muestral s y la distribución normal por la distribución t. Es decir:

a) Constraste bilateral

Al ser la distribución t una distribución simétrica se puede expresar que, si H_0 se cumple (es decir, si $\mu = \mu_0$), entonces

$$P\left(-t_{\alpha/2,n-1} < t < t_{\alpha/2,n-1}\right) = 1 - \alpha,$$

donde $t_{\alpha/2,n-1}$ es la abscisa de la distribución t de Student con $n - 1$ grados de libertad que deja a su derecha un área de probabilidad igual a $\alpha/2$. Por lo tanto, las regiones de aceptación A y crítica C del contraste son

$$A = \{t : |t| \leq t_{\alpha/2,n-1}\} \qquad ; \qquad C = \{t : |t| > t_{\alpha/2,n-1}\}, \tag{13.12}$$

donde la variable t se define en (13.11) Entonces, si se encuentra que

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}, \tag{13.13}$$

se acepta H_0. Por el contrario, si

$$\frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} > t_{\alpha/2,n-1},$$

la hipótesis nula se rechaza al nivel de significación α.

b) Contraste unilateral

De forma similar, las regiones de aceptación A y crítica C para un contraste bilateral son

$$A = \{t : t \leq t_{\alpha,n-1}\} \quad ; \quad C = \{t : |t| > t_{\alpha,n-1}\}. \tag{13.14}$$

Por lo que H_0 se acepta si

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha,n-1}, \tag{13.15}$$

y se rechaza al nivel de significación α si

$$\frac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha,n-1}.$$

Hay que indicar que todas las expresiones anteriores sólo son estrictamente válidas si se puede asegurar que la población en estudio sigue una distribución normal. Sin embargo, siempre que las muestras sean grandes no se comete un error excesivo si se supone normalidad y se aplican las relaciones anteriores (sobre todo si la distribución tiene forma de campana).

Ejemplo IV–7

Considerando la siguiente serie de medidas de la velocidad de la luz por Michelson (299000+): 850, 740, 900, 1070, 930, 850, 950, 980 (km/s) se quiere saber si la media es significativamente diferente de 1000.

De la muestra anterior deducimos de forma inmediata $n = 8$, $\overline{x} = 908.8$ km/s y $s = 99.1$ km/s. El valor de σ es desconocido y el número de datos $n \leq 30$. Las hipótesis nula y alternativa son:

$$\begin{cases} H_0 : \mu = 1000 \\ H_1 : \mu \neq 1000 \end{cases}$$

Aceptaremos H_0 si

$$t = \frac{|\overline{x} - \mu_0|}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}.$$

Usando $\alpha = 0.10 \Rightarrow t_{0.05,7} = 1.895$. Por tanto

$$t = \frac{|908.8 - 1000.0|}{99.1/\sqrt{8}} = 2.60 > t_{0.05,7},$$

por lo que rechazamos la hipótesis nula.

13.2. Contraste de una proporción

Supongamos que se quiere hacer un contraste de hipótesis para el parámetro p de una distribución binomial. Ya se ha visto cómo la proporción de éxitos (o número de éxitos dividido por el número de ensayos) constituye un estimador puntual de p. Supongamos que $\overline{p}$ es el valor de dicha proporción en una muestra en particular. Para realizar el contraste de hipótesis vamos a suponer que la muestra es suficientemente grande para aproximar la distribución muestral de $\overline{p}$ por una normal con media p y varianza $p(1-p)/n$. Si la muestra no fuese grande, las aproximaciones siguientes no son válidas y habría que utilizar las propiedades de la distribución binomial para realizar el contraste.

a) Constraste bilateral

La hipótesis nula en este caso es que el parámetro p toma un determinado valor p_0. Es decir

$$\begin{cases} H_0: & p = p_0 \\ H_1: & p \neq p_0 \end{cases} \tag{13.16}$$

Al ser la muestra grande, el siguiente estadístico seguirá una distribución normal tipificada

$$z = \frac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}, \tag{13.17}$$

donde $\overline{p}$ es la proporción de éxitos observada en la muestra y donde se ha aproximado la varianza poblacional por la varianza muestral. Es decir, si H_0 es cierta se cumplirá

$$P\left(-z_{\alpha/2} < \frac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} < z_{\alpha/2}\right) = 1 - \alpha$$

y, por lo tanto, las regiones de aceptación y crítica serán:

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}$$

y, H_0 se aceptará si

$$\frac{|\overline{p} - p_0|}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \leq z_{\alpha/2}, \tag{13.18}$$

mientras que se rechazará al nivel de significación α si

$$\frac{|\overline{p} - p_0|}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} > z_{\alpha/2}.$$

b) Contraste unilateral

De manera similar puede establecerse el contraste unilateral, con hipótesis

$$\begin{cases} H_0: & p \leq p_0 \\ H_1: & p > p_0 \end{cases} \tag{13.19}$$

Las regiones de aceptación y crítica serían:

$$A = \{z : z \leq z_\alpha\} \qquad ; \qquad C = \{z : z > z_\alpha\}.$$

aceptándose H_0 si

$$\frac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \leq z_\alpha \tag{13.20}$$

y rechazándose al nivel de significación α si

$$\frac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} > z_\alpha.$$

Ejemplo IV–8

Un amigo nos dice que tiene un porcentaje de acierto en tiros libres del 90 %. Para probarlo tira 100 lanzamientos y encesta sólo 85. ¿Le podemos creer?

Usaremos un nivel de significación $\alpha = 0.05$. Estamos ante un ensayo unilateral de una proporción:

$$\begin{cases} H_0 : p \geq 0.90 \\ H_1 : p < 0.90 \end{cases}$$

Se aceptará H_0 si

$$\frac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \leq z_\alpha.$$

En nuestro caso, $z_\alpha = z_{0.05} = 1.645$ y $\overline{p} = 0.85$, es decir

$$\frac{0.90 - 0.85}{\sqrt{\frac{0.85(1-0.85)}{100}}} = 1.40 \leq z_\alpha,$$

por lo que no rechazamos H_0 (creemos a nuestro amigo).

13.3. Contraste de varianza de una población normal

A continuación se plantea el contraste de hipótesis sobre la varianza, o la desviación típica, de una población normal. Para ello se utilizará la propiedad vista en el tema anterior de que la variable $(n-1)S^2/\sigma^2$, donde S^2 es la varianza muestral y σ^2 la poblacional, sigue una distribución χ^2 con $n-1$ grados de libertad.

a) Contraste bilateral

En este caso, la hipótesis nula y alternativa vendrán dadas por

$$\begin{cases} H_0 : & \sigma^2 = \sigma_0^2 \\ H_1 : & \sigma^2 \neq \sigma_0^2 \end{cases} \tag{13.21}$$

Es decir, se quiere comprobar si la varianza de una población puede coincidir con un determinado valor σ_0^2. Para ello se define el estadístico

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}. \tag{13.22}$$

Sabemos que, si se cumple H_0, el estadístico anterior sigue una distribución χ^2 con $n-1$ grados de libertad. Es decir

$$P(\chi^2_{1-\alpha/2,n-1} < \chi^2 < \chi^2_{\alpha/2,n-1}) = 1 - \alpha,$$

donde $\chi^2_{\alpha/2,n-1}$ es la abscisa de la distribución χ^2 con $n-1$ grados de libertad que deja a su derecha un área de probabilidad igual a $\alpha/2$, y lo mismo para $\chi^2_{1-\alpha/2,n-1}$. Por lo tanto, las regiones de aceptación y rechazo de la hipótesis nula serán

$$A = \{\chi^2 : \chi^2_{1-\alpha/2,n-1} \leq \chi^2 \leq \chi^2_{\alpha/2,n-1}\},$$

$$C = \{\chi^2 : \chi^2 < \chi^2_{1-\alpha/2,n-1} \text{ o } \chi^2 > \chi^2_{\alpha/2,n-1}\}. \tag{13.23}$$

Nótese que en este caso la distribución no es simétrica, y región de confianza se escoge para tener áreas

iguales en ambas colas. En resumen, se aceptará la hipótesis nula si

$$\frac{(n-1)s^2}{\sigma_0^2} \in [\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}] \tag{13.24}$$

y se rechazará al nivel de significación α en caso contrario.

b) Contraste unilateral

El contraste unilateral para la varianza de una población normal puede plantearse de manera similar a partir de las hipótesis

$$\begin{cases} H_0: & \sigma^2 \leq \sigma_0^2 \\ H_1: & \sigma^2 > \sigma_0^2 \end{cases} \tag{13.25}$$

Se define entonces el estadístico χ^2 como en (13.22). La región crítica se sitúa ahora sólo en la cola derecha de la distribución de forma que se tienen las regiones

$$A = \{\chi^2 : \chi^2 \leq \chi^2_{\alpha,n-1}\} \quad ; \quad C = \{\chi^2 : \chi^2 > \chi^2_{\alpha,n-1}\} \tag{13.26}$$

y la hipótesis H_0 se acepta si

$$\frac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{\alpha,n-1} \tag{13.27}$$

rechazándose al nivel de significación α en caso contrario.

Ejemplo IV–9

¿Puede ser la desviación típica del ejemplo IV–7 igual a 200?

Usaremos $\alpha = 0.05$. Tenemos un ensayo bilateral:

$$\begin{cases} H_0 : \sigma^2 = 200^2 \\ H_1 : \sigma^2 \neq 200^2 \end{cases}$$

Aceptaremos H_0 si

$$\frac{(n-1)s^2}{\sigma_0^2} \in [\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}].$$

Consultando las tablas, vemos que ($n = 8$, $n - 1 = 7$)

$$\begin{aligned} \chi^2_{1-\alpha/2,n-1} &= \chi^2_{0.975,7} = 1.690 \\ \chi^2_{\alpha/2,n-1} &= \chi^2_{0.025,7} = 16.013 \end{aligned}$$

mientras que

$$\frac{(n-1)s^2}{\sigma_0^2} = \frac{7 \times 99.1^2}{200^2} = 1.72,$$

que se encuentra dentro del intervalo requerido. Por tanto, no rechazamos H_0 (la muestra es demasiado pequeña).

Capítulo 14

Contrastes de hipótesis para dos poblaciones

"Utilizo la palabra *prueba* no en el sentido de los abogados, para quienes dos medias verdades equivalen a una verdad, sino en el sentido de los matemáticos, para quienes media verdad es igual a nada."

Karl Friedrich Gauss (1777-1855)

En este capítulo se presentan los contrastes de hipótesis para diferentes parámetros poblacionales de dos poblaciones. Debido a la íntima relación existente entre los contrastes de hipótesis y los intervalos de confianza, utilizaremos las expresiones vistas en capítulos anteriores para estos últimos para describir los contrastes. En todo lo siguiente se supone que se tiene un muestreo con reemplazamiento o en una población infinita. En otro caso habría que hacer las modificaciones necesarias usando las expresiones presentadas en capítulos anteriores.

14.1. Contraste de la igualdad de medias de poblaciones normales

A continuación se describen los procedimientos de contraste de hipótesis para comparar las medias de dos poblaciones normales. Se supone que se cuenta con muestras aleatorias independientes de tamaños n_1 y n_2 para cada población. Se representará por μ_1 y μ_2 la media de cada población respectivamente, y por $\overline{x_1}$ y $\overline{x_2}$ los valores que tomen las medias muestrales para muestras particulares de ambas poblaciones. Los contrastes de hipótesis tendrán como finalidad en general verificar si ambas medias poblacionales pueden ser iguales o si hay evidencias a favor de que una puede ser mayor que la otra. Distinguiremos diferentes casos:

14.1.1. Varianzas conocidas

En este caso, los contrastes de hipótesis se desarrollan utilizando que, según se demostró en el tema anterior, el siguiente estadístico sigue una distribución normal tipificada (siempre que ambas poblaciones sean normales)

$$z = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{14.1}$$

a) Contraste bilateral

Para este contraste la hipótesis nula será que ambas medias son iguales, de forma que

$$\begin{cases} H_0: & \mu_1 = \mu_2 \\ H_1: & \mu_1 \neq \mu_2 \end{cases} \tag{14.2}$$

Es decir, H_0 implica que $\mu_1 - \mu_2 = 0$ y, por lo tanto, el estadístico dado en (14.1) se convierte, si H_0 se cumple, en

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}. \tag{14.3}$$

Este estadístico es similar al utilizado en (13.2), siguiendo una distribución normal tipificada, por lo que las regiones de aceptación y crítica para H_0 son

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}.$$

y la hipótesis nula de igualdad de medias se aceptará si se cumple

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2} \tag{14.4}$$

y se rechazará al nivel de significación α si

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$$

b) Contraste unilateral

La hipótesis nula y alternativa son este caso

$$\begin{cases} H_0: & \mu_1 \leq \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases} \tag{14.5}$$

Como estadístico de contraste se utiliza el especificado en (14.3) de forma que se tienen las regiones

$$A = \{z : z \leq z_\alpha\} \qquad ; \qquad C = \{z : z > z_\alpha\}.$$

y H_0 se acepta si

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_\alpha, \tag{14.6}$$

rechazándose al nivel de significación α en caso contrario.

14.1.2. Varianzas desconocidas y $n_1 + n_2 > 30$ $(n_1 \simeq n_2)$

Generalmente las varianzas poblacionales σ_1^2 y σ_2^2 serán desconocidas. Sin embargo, si las muestras son grandes, las varianzas muestrales son, en principio, una buena aproximación de las poblacionales. De esta forma el contraste de hipótesis para la diferencia de medias se puede realizar igual que en el caso anterior, sustituyendo σ_1 y σ_2 por s_1 y s_2 respectivamente, y asumiendo que el nuevo estadístico

$$z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \tag{14.7}$$

sigue una distribución normal tipificada. Las hipótesis nulas y alternativas son las mismas que las establecidas en (14.2) y (14.5), siendo los criterios de aceptación y rechazo los siguientes.

a) Contraste bilateral

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}$$

Y la hipótesis H_0 se acepta si

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq z_{\alpha/2}, \tag{14.8}$$

rechazándose al nivel α en caso contrario.

b) Contraste unilateral

$$A = \{z : z \leq z_{\alpha}\} \qquad ; \qquad C = \{z : z > z_{\alpha}\}$$

Y la hipótesis H_0 se acepta si

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq z_{\alpha}, \tag{14.9}$$

rechazándose al nivel α en caso contrario.

Ejemplo IV–10

La temperatura media durante el mes de julio en 2 ciudades diferentes es

$$\begin{array}{llll} \text{Ciudad 1} & \overline{x}_1 = 36^\circ & s_1 = 5^\circ & n_1 = 31 \\ \text{Ciudad 2} & \overline{x}_2 = 34^\circ & s_2 = 4^\circ & n_2 = 25 \end{array}$$

¿Es la ciudad 1 más calurosa que la ciudad 2?

Tenemos un ensayo unilateral

$$\begin{cases} H_0 : \mu_1 \leq \mu_2 \\ H_1 : \mu_1 > \mu_2 \end{cases}$$

Se aceptará H_0 si

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq z_{\alpha}.$$

Usamos $\alpha = 0.05 \Rightarrow z_\alpha = z_{0.05} = 1.645$. Es decir

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{36 - 34}{\sqrt{\frac{5^2}{31} + \frac{4^2}{25}}} = 1.66,$$

por lo que rechazamos H_0 y se puede considerar (al nivel de significación α) que la ciudad 1 es más calurosa que la ciudad 2.

14.1.3. Varianzas desconocidas y $\sigma_1 = \sigma_2$ $(n_1 + n_2 \leq 30)$

Cuando los tamaños muestrales no son grandes no se pueden hacer las aproximaciones anteriores. Supongamos en primer lugar que se puede suponer a priori que las dos varianzas poblacionales son iguales (en la práctica se debe hacer antes un contraste de igualdad de varianzas para poder aplicar esto). En este caso, en el tema anterior se comprobó que el siguiente estadístico sigue una distribución t de Student con $n_1 + n_2 - 2$ grados de libertad

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, \tag{14.10}$$

donde s_p es la varianza ponderada definida como

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

Los contrastes de hipótesis se basan en este estadístico. Nótese que cuando se hace la hipótesis nula de que las medias poblacionales son iguales, t se convierte en nuestro estadístico de prueba

$$t = \frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}. \tag{14.11}$$

Por lo tanto, los criterios de aceptación y rechazo para los contrastes, con las hipótesis establecidas en (14.2) y (14.5), son

a) Contraste bilateral

$$A = \{t : |t| \leq t_{\alpha/2,n_1+n_2-2}\} \qquad ; \qquad C = \{t : |t| > t_{\alpha/2,n_1+n_2-2}\} \tag{14.12}$$

La hipótesis nula ($\mu_1 = \mu_2$) se acepta si

$$\frac{|\overline{x_1} - \overline{x_2}|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2} \tag{14.13}$$

y se rechaza al nivel de significación α en caso contrario.

b) Contraste unilateral

$$A = \{t : t \leq t_{\alpha,n_1+n_2-2}\} \qquad ; \qquad C = \{t : t > t_{\alpha,n_1+n_2-2}\} \tag{14.14}$$

Y la hipótesis H_0 se acepta si

$$\frac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha,n_1+n_2-2} \tag{14.15}$$

rechazándose al nivel α en caso contrario.

14.1.4. Varianzas desconocidas con $\sigma_1 \neq \sigma_2$ $(n_1 + n_2 \leq 30)$

En un caso general no se podrá hacer a priori la suposición de que las dos varianzas poblacionales son iguales. Para hacer el contraste de hipótesis sobre la igualdad de medias en este caso se utiliza que, según se demostró en el tema anterior, se puede suponer que el siguiente estadístico sigue una distribución t de Student con f grados de libertad

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}, \tag{14.16}$$

donde f viene dado por (aproximación de Welch)

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}} - 2.$$

Al hacer la hipótesis nula el estadístico anterior se convierte en el estadístico a usar en este contraste de hipótesis

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}. \tag{14.17}$$

Entonces, se puede establecer que los criterios de aceptación y rechazo para los contrastes, con las hipótesis (14.2) y (14.5) son los siguientes:

a) Contraste bilateral

$$A = \{t : |t| \leq t_{\alpha/2,f}\} \qquad ; \qquad C = \{t : |t| > t_{\alpha/2,f}\} \tag{14.18}$$

La hipótesis nula de igualdad de medias se aceptará cuando

$$\frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha/2,f} \tag{14.19}$$

y se rechazará al nivel de significación α en caso contrario.

b) Contraste unilateral

$$A = \{t : t \leq t_{\alpha,f}\} \qquad ; \qquad C = \{t : t > t_{\alpha,f}\} \tag{14.20}$$

Y la hipótesis H_0 se acepta cuando

$$\frac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha,f} \tag{14.21}$$

rechazándose al nivel α en otro caso.

Ejemplo IV–11

Las notas de 2 alumnos, en las 9 asignaturas del primer curso, son

Alumno 1	5, 7, 7, 6, 5, 5, 8, 6, 8
Alumno 2	5, 6, 8, 9, 7, 6, 5, 8, 10

¿Son significativamente diferentes?

A partir de los datos deducimos de manera sencilla que

Alumno 1	$\overline{x_1} = 6.33$	$s_1 = 1.22$
Alumno 2	$\overline{x_2} = 7.11$	$s_2 = 1.76$

Tenemos

$$\begin{cases} H_0 : \mu_1 = \mu_2 \\ H_1 : \mu_1 \neq \mu_2 \end{cases}$$

Vamos a considerar dos casos

i) Varianzas desconocidas, y $\sigma_1 \neq \sigma_2$. En este caso, se aceptará H_0 si

$$t = \frac{|\overline{x_1} - \overline{x_2}|}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha/2,f}.$$

Calculamos primero f mediante

$$f = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1+1} + \frac{(s_2^2/n_2)^2}{n_2+1}} - 2 = 15.81 \simeq 16.$$

De esta forma,

$$t_{\alpha/2,f} = t_{0.025,16} = 2.120,$$

mientras que el valor del estadístico viene dado por

$$t = \frac{|6.33 - 7.11|}{\sqrt{\frac{1.22^2}{9} + \frac{1.76^2}{9}}} = 1.09 < t_{\alpha/2,f},$$

por lo que no rechazamos H_0 (no hay evidencias de que sean diferentes, al nivel de significación elegido).

Ejemplo IV–11

(Continuación)

ii) Varianzas desconocidas, y $\sigma_1 = \sigma_2$. Bajo estas suposiciones, se aceptará H_0 si

$$t = \frac{|\overline{x_1} - \overline{x_2}|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2, n_1+n_2-2}.$$

El valor de s_p se determina mediante

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 2.293 \Rightarrow s_p = 1.51,$$

por lo que finalmente se obtiene

$$t = \frac{|6.33 - 7.11|}{1.51\sqrt{\frac{1}{9} + \frac{1}{9}}} = \frac{0.78}{0.71} = 1.10.$$

Como $t_{\alpha/2,n_1+n_2-2} = t_{0.025,16} = 2.120$, tampoco se rechaza H_0.

14.2. Contraste de la igualdad entre dos proporciones

Supongamos ahora que se quiere hacer un contraste de hipótesis sobre la igualdad de los parámetros p_1 y p_2 de dos distribuciones binomiales. Denotaremos por $\overline{p_1}$ y $\overline{p_2}$ a las proporciones observadas en muestras de tamaños n_1 y n_2 extraídas de cada población. En la determinación del intervalo de confianza para la diferencia de p_1 y p_2 se demostró que, para muestras grandes, la distribución muestral de $\overline{p_1} - \overline{p_2}$ tiende a una distribución normal con media $p_1 - p_2$ y varianza

$$\sigma^2 = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

De esta manera, por analogía con (14.3), y en el caso de que se cumpla la hipótesis nula $p_1 = p_2$, el estadístico de prueba

$$z = \frac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \tag{14.22}$$

seguirá una distribución normal tipificada. Nótese que, puesto que estamos suponiendo muestras grandes, estamos sustituyendo la varianza poblacional por la varianza muestral. Los contrastes quedan entonces como sigue:

a) Contraste bilateral

Las hipótesis nula y alternativa son las siguientes

$$\begin{cases} H_0: & p_1 = p_2 \\ H_1: & p_1 \neq p_2 \end{cases} \tag{14.23}$$

Puesto que el estadístico dado en (14.22) sigue una distribución normal si H_0 es cierta, las regiones de aceptación y crítica serán

$$A = \{z : |z| \leq z_{\alpha/2}\} \qquad ; \qquad C = \{z : |z| > z_{\alpha/2}\}$$

y, por tanto, se acepta H_0 si se cumple

$$\frac{|\overline{p_1} - \overline{p_2}|}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq z_{\alpha/2}, \tag{14.24}$$

rechazándose al nivel de significación α en caso contrario.

b) Contraste unilateral

En este contraste las hipótesis nula y alternativa son:

$$\begin{cases} H_0: & p_1 \leq p_2 \\ H_1: & p_1 > p_2 \end{cases} \tag{14.25}$$

Utilizando el estadístico (14.22) se definen las regiones de aceptación y crítica

$$A = \{z : z \leq z_\alpha\} \qquad ; \qquad C = \{z : z > z_\alpha\},$$

por lo que se acepta la hipótesis nula si se cumple

$$\frac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq z_\alpha \tag{14.26}$$

y se rechaza al nivel α en caso contrario.

14.3. Contraste de la igualdad de varianzas de poblaciones normales

A continuación se describe el contraste de hipótesis para la comparación de varianzas de dos poblaciones normales independientes. Sean σ_1^2 y σ_2^2 las varianzas poblacionales, mientras que por s_1^2 y s_2^2 se representan los valores que toman las varianzas muestrales en muestras de tamaños n_1 y n_2 extraídas de cada población. En el tema anterior se demostró que, si ambas poblaciones son normales, el estadístico

$$F = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} \tag{14.27}$$

sigue una distribución F de Fisher con $(n_1 - 1)$ y $(n_2 - 1)$ grados de libertad. Aprovechando esta propiedad, los contrastes serán:

a) Contraste bilateral

Para este contraste la hipótesis nula será que las dos medias poblacionales son iguales, es decir

$$\begin{cases} H_0: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 \neq \sigma_2^2 \end{cases} \tag{14.28}$$

El estadístico de prueba será el descrito en (14.27) cuando se cumple la hipótesis nula. Es decir

$$F = \frac{s_1^2}{s_2^2}. \tag{14.29}$$

Al seguir este estadístico una distribución F, se puede escribir (igualando el área de las dos colas de la distribución)

$$P(F_{1-\alpha/2,n_1-1,n_2-1} < F < F_{\alpha/2,n_1-1,n_2-1}) = 1 - \alpha,$$

donde $F_{\alpha/2,n_1-1,n_2-1}$ es la abscisa de la distribución F con $n_1 - 1$ y $n_2 - 1$ grados de libertad que deja

a su derecha un área de probabilidad igual a $\alpha/2$, y lo mismo para $F_{1-\alpha/2,n_1-1,n_2-1}$. Por lo tanto, las regiones de aceptación y rechazo de la hipótesis nula serán

$$\begin{cases} A = \{F : F_{1-\alpha/2,n_1-1,n_2-1} \leq F \leq F_{\alpha/2,n_1-1,n_2-1}\} \\ C = \{F : F < F_{1-\alpha/2,n_1-1,n_2-1} \text{ o } F > F_{\alpha/2,n_1-1,n_2-1}\} \end{cases} \tag{14.30}$$

En resumen, la hipótesis nula se acepta cuando

$$\frac{s_1^2}{s_2^2} \in [F_{1-\alpha/2,n_1-1,n_2-1}, F_{\alpha/2,n_1-1,n_2-1}] \tag{14.31}$$

y se rechaza al nivel de significación α en caso contrario.

b) Contraste unilateral

En este contraste las hipótesis son:

$$\begin{cases} H_0 : & \sigma_1^2 \leq \sigma_2^2 \\ H_1 : & \sigma_1^2 > \sigma_2^2 \end{cases} \tag{14.32}$$

Como estadístico de prueba se usa el especificado en (14.29), situándose la región crítica en la cola derecha de la distribución F

$$A = \{F : F \leq F_{\alpha,n_1-1,n_2-1}\} \quad ; \quad C = \{F : F > F_{\alpha,n_1-1,n_2-1}\} \tag{14.33}$$

Por lo que la hipótesis H_0 se acepta cuando

$$\frac{s_1^2}{s_2^2} \leq F_{\alpha,n_1-1,n_2-1}, \tag{14.34}$$

rechazándose al nivel de significación α en caso contrario.

Ejemplo IV–12

¿Son las varianzas del ejemplo IV–10 diferentes? ¿Y las del ejemplo IV–11?

Las hipótesis son en este caso:

$$\begin{cases} H_0 : \sigma_1^2 = \sigma_2^2 \\ H_1 : \sigma_1^2 \neq \sigma_2^2 \end{cases}$$

Se aceptará H_0 si

$$F = \frac{s_1^2}{s_2^2} \in [F_{1-\alpha/2,n_1-1,n_2-1}, F_{\alpha/2,n_1-1,n_2-1}]$$

Ejemplo IV–10: supongamos $\alpha = 0.10$.

$$F_{1-\alpha/2,n_1-1,n_2-1} = F_{0.95,30,24} = \frac{1}{F_{0.05,24,30}} = \frac{1}{1.8874} = 0.5298$$

$$F_{\alpha/2,n_1-1,n_2-1} = F_{0.05,30,24} = 1.9390$$

Por lo que el estadístico será $F = s_1^2/s_2^2 = 5^2/4^2 = 1.56 \in [0.53, 1.94] \Rightarrow$ no se rechaza H_0.

Ejemplo IV–11: supongamos ahora que $\alpha = 0.05$. De formar similar a como hemos trabajado anteriormente

$$F_{1-\alpha/2,n_1-1,n_2-1} = F_{0.975,8,8} = \frac{1}{F_{0.025,8,8}} = \frac{1}{4.4332} = 0.2256$$

$$F_{\alpha/2,n_1-1,n_2-1} = F_{0.025,8,8} = 4.4332$$

Como $F = s_1^2/s_2^2 = 1.22^2/1.76^6 = 0.48 \in [0.23, 4.43] \Rightarrow$ se acepta también H_0.

14.4. Contraste de la igualdad de medias para datos apareados

Supongamos ahora que se tiene un experimento de observaciones pareadas. Es decir, se extraen dos muestras no independientes con el mismo tamaño n de dos poblaciones normales. En el tema anterior se vió cómo este problema se simplificaba definiendo una nueva variable aleatoria D consistente en las diferencias entre cada par de observaciones. De forma que para una muestra en particular se tenían n valores de $d_i = x_{1_i} - x_{2_i}$, pudiendo definirse una media y una varianza muestral de esta variable como $\overline{d} = \sum d_i/n$ y $s_d^2 = \sum(d_i - \overline{d})^2/(n-1)$. Entonces el contraste de hipótesis para la diferencia de medias se convierte en un contraste sobre el valor poblacional de $d = \mu_1 - \mu_2$. El problema es equivalente entonces al del contraste de la media de una población, por lo que se tiene que el estadístico

$$t = \frac{\overline{d} - d}{s_d/\sqrt{n}} \tag{14.35}$$

sigue una distribución t de Student con $n-1$ grados de libertad. Aquí se ha supuesto que la muestra no es demasiado grande, por lo que hay que utilizar la distribución t. Para muestras grandes de poblaciones normales ($n > 30$) se podría substituir la distribución t por una normal sin cometer un excesivo error.

a) Contraste bilateral

El contraste bilateral consiste en comprobar si la diferencia entre las dos medias es nula. Esto es equivalente a contrastar los siguientes valores de d

$$\begin{cases} H_0: & d = 0 & ; & \mu_1 = \mu_2 \\ H_1: & d \neq 0 & ; & \mu_1 \neq \mu_2 \end{cases} \tag{14.36}$$

Bajo la hipótesis H_0 el estadístico de prueba, dado en (14.35), se convierte en

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \tag{14.37}$$

Y las regiones de aceptación y crítica son

$$A = \{t : |t| \leq t_{\alpha/2,n-1}\} \qquad ; \qquad C = \{t : |t| > t_{\alpha/2,n-1}\}$$

Por lo tanto, la hipótesis nula se acepta si

$$\frac{|\overline{d}|}{s_d/\sqrt{n}} \leq t_{\alpha/2,n-1} \tag{14.38}$$

y se rechaza al nivel α en caso contrario.

b) Contraste unilateral

Para el contraste unilateral las hipótesis son:

$$\begin{cases} H_0: & d \leq 0 & ; & \mu_1 \leq \mu_2 \\ H_1: & d > 0 & ; & \mu_1 > \mu_2 \end{cases} \tag{14.39}$$

Evidentemente, el estadístico de prueba es el dado en (14.37), con las regiones

$$A = \{t : t \leq t_{\alpha,n-1}\} \qquad ; \qquad C = \{t : t > t_{\alpha,n-1}\}$$

y la hipótesis H_0 se acepta cuando

$$\frac{\overline{d}}{s_d/\sqrt{n}} \leq t_{\alpha,n-1} \tag{14.40}$$

rechazándose al nivel de significación α en caso contrario.

Ejemplo IV–13

En el ejemplo III–16, ¿aumenta la producción al no permitir el bocadillo a media mañana? Utilizar $\alpha = 0.05$.

Las hipótesis son

$$\begin{cases} H_0 : d \leq 0 \\ H_1 : d > 0 \end{cases}$$

Se aceptará H_0 si

$$t = \frac{\overline{d}}{s_d/\sqrt{n}} \leq t_{\alpha,n-1}.$$

Teníamos $\overline{d} = 0.8$, $s_d = 3.08$ y $n = 10$. Por tanto

$$t_{\alpha,n-1} = t_{0.05,9} = 1.833$$

$$t = \frac{0.8}{3.08/\sqrt{10}} = 0.82 \leq t_{0.05,9} \Rightarrow \text{ se acepta } H_0$$

y no se considera probado que aumente la producción.

Capítulo 15

Aplicaciones de la distribución χ^2

"Ninguna ciencia, en cuanto ciencia, engaña; el engaño está en quien no sabe."

Miguel de Cervantes (1547-1616)

En los temas anteriores nos hemos ocupado de los contrastes de hipótesis sobre los parámetros poblacionales. Para poderlos hacer hemos supuesto ciertas condiciones sobre la muestra, como que era aleatoria y provenía de una población que seguía una determinada distribución. Ahora se presentan algunos métodos para comprobar que una muestra dada cumple estas suposiciones. En particular, se estudiarán tests de hipótesis para comprobar si la distribución supuesta es consistente con la muestra, si diferentes muestras pueden considerarse homogéneas y si las observaciones de dos factores o parámetros de una misma población son independientes. Todos estos tests se basan en un procedimiento común consistente en la aplicación de la distribución χ^2.

15.1. Prueba de la bondad del ajuste

Los intervalos de confianza y los contrastes de hipótesis sobre parámetros poblacionales se basan en suponer que la población sigue una determinada distribución de probabilidad (normal, en muchos casos). Puesto que las conclusiones de dichos contrastes dependen de la elección de la distribución teórica, es importante determinar si dicha hipótesis puede ser correcta. Evidentemente, al trabajar con una muestra de una población, siempre existirán diferencias entre la distribución teórica y la observada. Sin embargo, habrá que comprobar si dichas desviaciones pueden ser debidas al azar o, por el contrario, proporcionan evidencias de que la distribución supuesta es incorrecta. Con este fin, en esta sección se presenta una prueba para, a partir de una muestra, determinar si una población sigue una distribución teórica específica.

La prueba aquí presentada, llamada de la **bondad del ajuste**, se basa en comparar las frecuencias observadas para una muestra concreta (es decir, el número de elementos de la muestra en los que la variable toma un valor concreto, o en un intervalo determinado) con las frecuencias esperadas si la muestra siguiese la distribución teórica hipotética.

Supongamos que tenemos una muestra de tamaño n y que la variable aleatoria X puede tomar los valores $X_1, X_2, \ldots, X_k$ excluyentes. Esto en principio sólo sería válido para una variable discreta, sin embargo se puede aplicar también a una variable continua realizando un agrupamiento en intervalos. Sean o_i las frecuencias observadas para cada X_i, es decir, el número de elementos de la muestra con $X = X_i$. Si se supone una distribución de probabilidad teórica, existirá una probabilidad p_i de que X tome un determinado valor X_i. Por lo tanto, las frecuencias esperadas para cada X_i serán $e_i = np_i$. Nótese que ha de cumplirse

que $\sum_{i=1}^k o_i = \sum_{i=1}^k e_i = n$ y $\sum_{i=1}^k p_i = 1$. Se puede escribir entonces la tabla:

X	X_1	X_2	...	X_i	...	X_k
Frecuencias observadas	o_1	o_2	...	o_i	...	o_k
Frecuencias esperadas	e_1	e_2	...	e_i	...	e_k

A continuación se hace la hipótesis nula H_0 consistente en suponer que la muestra sigue la distribución teórica elegida y, por tanto, las desviaciones encontradas respecto a ésta son debidas al azar. Para realizar el contraste de esta hipótesis se define el estadístico

$$\chi^2_{k-1} = \sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i}. \tag{15.1}$$

Se puede demostrar que, en el caso de que se cumpla H_0, el estadístico anterior sigue una distribución χ^2 con $k-1$ grados de libertad. Una demostración rigurosa de esto está fuera del alcance de este libro. Sin embargo, una justificación intuitiva es la siguiente:

Consideremos como variable el número de elementos de la muestra con valores $X = X_i$, es decir o_i. Si la muestra es grande, puede suponerse que esta variable sigue una distribución de Poisson, con parámetro $\lambda = np_i$ (valor esperado de o_i). Sabemos que si $\lambda > 5$, el siguiente estadístico sigue una normal tipificada

$$Z = \frac{o_i - \lambda}{\sqrt{\lambda}} = \frac{o_i - np_i}{\sqrt{np_i}} \simeq N(0,1)$$

y, por tanto, teniendo en cuenta que $e_i = np_i$, los términos de la expresión (15.1) son los cuadrados de variables aleatorias normales $N(0,1)$ y su suma constituye una χ^2. Puesto que, de las diferentes variables o_i, sólo $k-1$ son independientes (ya que $\sum o_i = n$), (15.1) será una χ^2 con $k-1$ grados de libertad.

Evidentemente, si las frecuencias observadas se acercan a las esperadas se obtendrá un valor bajo de χ^2 y la hipótesis nula (la muestra sigue la distribución teórica) se debe aceptar. Por el contrario, cuando existan considerables diferencias entre las frecuencias observadas y esperadas, el valor de χ^2 será grande y el ajuste será pobre, rechazándose H_0. La región crítica cae entonces en la cola derecha de la distribución y, para un nivel de significación α, se acepta H_0 si

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} \leq \chi^2_{\alpha,k-1} \tag{15.2}$$

y se rechaza si

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} > \chi^2_{\alpha,k-1}. \tag{15.3}$$

Para calcular el valor del estadístico χ^2 puede usarse la expresión alternativa

$$\sum_{i=1}^k \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^k \frac{o_i^2 - 2o_ie_i + e_i^2}{e_i} =$$

$$= \sum_{i=1}^k \frac{o_i^2}{e_i} - 2\sum_{i=1}^k o_i + \sum_{i=1}^k e_i = \sum_{i=1}^k \frac{o_i^2}{e_i} - 2n + n = \sum_{i=1}^k \frac{o_i^2}{e_i} - n$$

Para poder aplicar este método correctamente es necesario que el tamaño de la muestra sea suficientemente grande (típicamente $n > 30$). En particular, se suele poner la restricción de que las frecuencias esperadas para cada X_i (o intervalo) no sean inferiores a 5 ($e_i \geq 5$). Cuando no se cumpla esto habrá que agrupar diferentes valores de X_i (o intervalos) para que se verifique la condición. Evidentemente, ello reduce el número de grados de libertad.

Otra consideración importante es que, si para calcular las frecuencias esperadas hay que usar parámetros poblacionales estimados a partir de la propia muestra (ej. media y varianza para la distribución normal), el número de grados de libertad de la χ^2 hay que reducirlo a $k-p-1$, donde p es el número de parámetros poblacionales que se estiman (nótese que esto no se aplica si los parámetros poblacionales se conocen, o se suponen, a priori, sin estimarlos a partir de los datos muestrales).

Esta prueba de la bondad del ajuste es una herramienta muy importante debido, fundamentalmente, a que muchos procedimientos estadísticos dependen de la suposición de una determinada distribución de probabilidad. En particular, es importante para comprobar la suposición de normalidad para la población, aunque puede aplicarse en general para cualquier distribución.

Ejemplo IV–14

Consideremos el lanzamiento de un dado. Queremos saber si el dado está cargado. Es decir,
H_0: la población sigue una distribución uniforme.

Se lanza el dado 600 veces y se obtiene

x_i:	1	2	3	4	5	6
o_i:	92	85	102	94	117	110
e_i:	100	100	100	100	100	100

$$p_i = \frac{1}{6} \Rightarrow e_i = np_i = 600 \times \frac{1}{6} = 100$$

El número de grados de libertad será $k-1 = 6-1 = 5$. Calculemos el estadístico

$$\chi^2_{k-1} = \sum_{i=1}^{6} \frac{(o_i - e_i)^2}{e_i} = 7.18.$$

Tomando como nivel de significación $\alpha = 0.05$

$$\chi^2_{\alpha,k-1} = \chi^2_{0.05,5} = 11.070.$$

Como $\chi^2_{k-1} < \chi^2_{\alpha,k-1}$ $\Rightarrow$ no podemos rechazar H_0 (las diferencias observadas son compatibles con el azar).

15.2. Contraste de la independencia de caracteres

Un problema usual en las ciencias experimentales es el estudio de la dependencia o independencia entre dos caracteres o factores medidos sobre los elementos de una población (ej. entre peso y altura para una muestra de individuos). Además, a menudo hemos hecho la hipótesis de independencia para derivar expresiones simplificadas respecto a la estimación de parámetros poblacionales. Es importante contar con un método para contrastar dicha hipótesis. Para ello se puede seguir un procedimiento similar al de la prueba de la bondad del ajuste, basado en la distribución χ^2.

Supongamos que sobre una muestra de tamaño n de una población se miden dos caracteres dados por las variables aleatorias X e Y, que pueden tomar los valores $x_1, x_2, \ldots, x_k$ e $y_1, y_2, \ldots, y_m$. Estos valores particulares pueden representar a una variable cualitativa, discreta o continua agrupada en intervalos. Denotaremos por o_{ij} a la frecuencia o número de elementos de la muestra que tienen conjuntamente $X = x_i$ e $Y = y_j$. Las frecuencias observadas se presentan usualmente en una tabla, llamada **tabla de contingencia**. Para el caso de k valores posibles para X y m valores posibles para Y, la tabla de contingencia $k \times m$ será:

$x \setminus y$	y_1	y_2	$\cdots$	y_j	$\cdots$	y_m	
x_1	o_{11} (e_{11})	o_{12} (e_{12})	$\cdots$	o_{1j} (e_{1j})	$\cdots$	o_{1m} (e_{1m})	o_{x_1}
x_2	o_{21} (e_{21})	o_{22} (e_{22})	$\cdots$	o_{2j} (e_{2j})	$\cdots$	o_{2m} (e_{2m})	o_{x_2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_i	o_{i1} (e_{i1})	o_{i2} (e_{i2})	$\cdots$	o_{ij} (e_{ij})	$\cdots$	o_{im} (e_{im})	o_{x_i}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_k	o_{k1} (e_{k1})	o_{k2} (e_{k2})	$\cdots$	o_{kj} (e_{kj})	$\cdots$	n_{km} (e_{km})	o_{x_k}
	o_{y_1}	o_{y_2}	$\cdots$	o_{y_j}	$\cdots$	o_{y_m}	n

La última columna y fila muestran las frecuencias marginales de X e Y respectivamente, es decir, el número de elementos de la muestra que tienen un cierto valor de X (o Y) sin importar los valores que tome la otra variable. Nótese que se cumple que $\sum_{i=1}^{k}\sum_{j=1}^{m} o_{ij} = n$ y además $\sum_{i=1}^{k} o_{x_i} = \sum_{j=1}^{m} o_{y_j} = n$.

Se hace entonces la hipótesis nula H_0 de que los dos caracteres son independientes, es decir, que para cualquier valor fijo de Y las distribuciones para las diferentes X son las mismas, y viceversa. El contraste de esta hipótesis se basa en comparar las frecuencias observadas con las que se esperarían si realmente los dos caracteres fuesen independientes. Las frecuencias esperadas, representadas por e_{ij}, se pueden calcular a partir de las probabilidades p_{ij} de que ambas variables tomen conjuntamente unos determinados valores, que, bajo la hipótesis de independencia, serán

$$p_{ij} = P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j) \simeq \frac{o_{x_i}}{n}\frac{o_{y_j}}{n}.$$

Por tanto

$$e_{ij} = np_{ij} = \frac{o_{x_i}o_{y_j}}{n}. \tag{15.4}$$

Es decir, las frecuencias esperadas se calculan multiplicando los totales de la fila y columna correspondiente y dividendo por n. Estos valores se incluyen en la tabla de contingencia escribiéndolos entre paréntesis.

Para el contraste de la hipótesis de independencia se utiliza, igual que en la prueba de la bondad del ajuste, el estadístico

$$\chi^2_\nu = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{o_{ij}^2}{e_{ij}} - n. \tag{15.5}$$

En el caso de ser H_0 cierta, este estadístico sigue una distribución χ^2 con ν grados de libertad. Para calcular dicho número de grados de libertad hay que tener en cuenta que las sumas de las frecuencias esperadas de cada fila o columna deben dar las frecuencias marginales, de forma que, para cada fila o columna, sólo es necesario calcular $k-1$ o $m-1$ valores independientes. Así, por ejemplo, para una tabla 2×3 sólo hace falta calcular las frecuencias e_{11} y e_{12} por lo que el número de grados de libertad es 2. De la misma manera, una tabla de contingencia 2×2 tiene un único grado de libertad. De forma general, el número de grados de libertad se calcula como

$$\nu = (k-1)(m-1).$$

Para tablas de contingencia de dimensiones determinadas existen fórmulas para calcular el valor de χ^2 a partir únicamente de las frecuencias observadas. Así, para una tabla 2×2, la expresión (15.5) es equivalente a

$$\chi^2_\nu = \frac{n(o_{11}o_{22} - o_{12}o_{21})^2}{o_{x_1}o_{x_2}o_{y_1}o_{y_2}}, \tag{15.6}$$

mientras que para una tabla de contingencia 2×3 se puede demostrar que

$$\chi^2_\nu = \frac{n}{o_{x_1}}\left(\frac{o^2_{11}}{o_{y_1}} + \frac{o^2_{12}}{o_{y_2}} + \frac{o^2_{13}}{o_{y_3}}\right) + \frac{n}{o_{x_2}}\left(\frac{o^2_{21}}{o_{y_1}} + \frac{o^2_{22}}{o_{y_2}} + \frac{o^2_{23}}{o_{y_3}}\right) - n. \tag{15.7}$$

Al igual que ocurría en la prueba de la bondad del ajuste, el método sólo es fiable si el número de elementos es suficientemente grande. En particular, si alguna de las frecuencias esperadas es menor que 5 habrá que agrupar filas o columnas.

En resumen, puede establecerse que, para un nivel de significación α, la hipótesis H_0 de independencia de caracteres se acepta si

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \leq \chi^2_{\alpha,(k-1)(m-1)} \tag{15.8}$$

y se rechaza en caso contrario.

Conviene hacer notar que el estadístico χ^2 definido en (15.5) toma valores discretos, ya que las frecuencias observadas son discretas. Sin embargo, en el contraste de hipótesis estamos aproximando su distribución a una distribución de probabilidad continua como la χ^2. Para solucionar esto se suele aplicar una corrección de continuidad consistente en disminuir las diferencias entre las frecuencias observadas y esperadas en una cantidad 0.5. Es decir, si la frecuencia esperada es mayor que la observada se le resta 0.5 y al contrario. Esta corrección, llamada **corrección de continuidad de Yates** conduce a la siguiente expresión modificada para el estadístico

$$\chi^{2\prime}_\nu = \sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(|o_{ij} - e_{ij}| - 0.5)^2}{e_{ij}}. \tag{15.9}$$

La corrección es normalmente despreciable si el número de grados de libertad es mayor que 1. Es decir, en la práctica, sólo se aplica para tablas de contingencia 2×2. En este caso, la expresión dada en (15.6) se convierte en

$$\chi^{2\prime}_\nu = \frac{n\left(|o_{11}o_{22} - o_{12}o_{21}| - \frac{n}{2}\right)^2}{o_{x_1}o_{x_2}o_{y_1}o_{y_2}}. \tag{15.10}$$

Lógicamente, si las frecuencias esperadas son grandes la corrección es muy pequeña. En la práctica, sólo se aplica la corrección de Yates cuando las frecuencias esperadas están entre 5 y 10.

15.3. Contraste de la homogeneidad de muestras

Un problema similar al anterior es el contraste de la homogeneidad de varias muestras. Mientras que en el contraste de independencia se medían dos características de una misma muestra, ahora se elijen k muestras de tamaños predeterminados (y no necesariamente iguales) y se quiere comprobar si todas ellas pueden provenir de la misma población. Es decir, el objetivo es contrastar si la variable X se distribuye de igual manera dentro de cada muestra. La hipótesis nula H_0 es entonces que las k muestras son homogéneas y la forma de operar es la misma que la vista para el contraste de la independencia. Es decir se puede construir una tabla de contingencia y definir un estadístico χ^2 como el dado en (15.5). Ahora k es el número de muestras y m el número de valores posibles, o intervalos, de la variable. Entonces, la hipótesis H_0 de homogeneidad se acepta con un nivel de significación α cuando

$$\sum_{i=1}^{k}\sum_{j=1}^{m}\frac{(o_{ij} - e_{ij})^2}{e_{ij}} \leq \chi^2_{\alpha,(k-1)(m-1)}.$$

Ejemplo IV–15

Comparemos las notas de 4 grupos de primero en la asignatura de estadística

	Grupos				
Notas	A	B	C	D	o_{x_i}
NT–SB	14	5	13	5	37
AP	26	31	23	10	90
SS	29	30	25	26	110
o_{y_j}	69	66	61	41	237

Estudiar la homogeneidad de las calificaciones al comparar los distintos grupos.

Podemos calcular las frecuencias esperadas utilizando

$$e_{11} = \frac{o_{x_1} o_{y_1}}{n} = \frac{37 \times 69}{237} = 10.8$$

$$e_{12} = \frac{o_{x_1} o_{y_2}}{n} = \frac{37 \times 66}{237} = 10.3$$

$$\dots$$

$$\dots$$

De tal forma que podemos añadir a la tabla las frecuencias esperadas así calculadas (números entre paréntesis):

	Grupos				
Notas	A	B	C	D	o_{x_i}
NT–SB	14 (10.8)	5 (10.3)	13 (9.5)	5 (6.4)	37
AP	26 (26.2)	31 (25.1)	23 (23.2)	10 (15.6)	90
SS	29 (32.0)	30 (30.6)	25 (28.3)	26 (19.0)	110
o_{y_j}	69	66	61	41	237

El estadítico para el contraste se calcula mediante

$$\chi^2_\nu = \sum_{i=1}^{3} \sum_{j=1}^{4} \frac{o_{ij}^2}{e_{ij}} - n = 248.93 - 237 = 11.93.$$

El número de grados de libertad es $\nu = (k-1)(m-1) = 2 \times 3 = 6$. Con un nivel de significación $\alpha = 0.05$, se acepta H_0 (las muestras son homogéneas) si $\chi^2_\nu \leq \chi^2_{\alpha,\nu}$. Como $\chi^2_{0.05,6} = 12.592$, que es mayor que el estadístico calculado arriba, no rechazamos H_0.

Un caso particular interesante del contraste de homogeneidad es cuando se realiza un experimento de Bernouilli, cuyo resultado es éxito o fracaso, sobre una serie de muestras, y se quiere comprobar si la probabilidad de éxito p puede ser la misma en todas las muestras. Supongamos que se tienen k muestras de tamaños $n_1, n_2, \dots, n_k$. Representemos los números de éxitos en cada muestra por $a_1, a_2, \dots, a_k$. Por tanto los números de fracasos en las muestras serán $n_1 - a_1, n_2 - a_2, \dots, n_k - a_k$. Se puede construir entonces una tabla de contingencia $k \times 2$ como sigue:

Muestra:	éxitos	fracasos	
1	a_1 $(n_1 p)$	$n_1 - a_1$ $(n_1 - n_1 p)$	n_1
2	a_2 $(n_2 p)$	$n_2 - a_2$ $(n_2 - n_2 p)$	n_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	a_i $(n_i p)$	$n_i - a_i$ $(n_i - n_i p)$	n_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	a_k $(n_k p)$	$n_k - a_k$ $(n_k - n_k p)$	n_k

La probabilidad de éxito p se puede estimar a partir del conjunto de todas las muestras como

$$p = \frac{\sum_{i=1}^{k} a_i}{\sum_{i=1}^{k} n_i}.$$

De esta forma, se pueden calcular las frecuencias esperadas de éxitos como $n_1p, n_2p, \ldots, n_kp$ y las de fracasos como $n_1 - n_1p, n_2 - n_2p, \ldots, n_k - n_kp$. Estos valores esperados se muestran entre paréntesis en la tabla de contingencia.

La hipótesis nula H_0 es que las muestras son homogéneas, y por tanto no hay diferencias significativas entre las frecuencias observadas y esperadas. A partir de la tabla de contingencia, el estadístico en este caso se puede escribir como

$$\chi^2_{k-1} = \sum_{i=1}^{k}\sum_{j=1}^{2} \frac{(o_{ij} - e_{ij})^2}{e_{ij}} = \sum_{i=1}^{k} \frac{(a_i - n_ip)^2}{n_ip} + \sum_{i=1}^{k} \frac{\left((n_i - a_i) - (n_i - n_ip)\right)^2}{n_i - n_ip} =$$

$$= \sum_{i=1}^{k} \left(\frac{(a_i - n_ip)^2}{n_ip} + \frac{(a_i - n_ip)^2}{n_i(1-p)} \right) = \sum_{i=1}^{k} \frac{(1-p)(a_i - n_ip)^2 + p(a_i - n_ip)^2}{n_ip(1-p)}$$

$$\Rightarrow \qquad \chi^2_{k-1} = \frac{1}{p(1-p)} \sum_{i=1}^{k} \frac{(a_i - n_ip)^2}{n_i}, \tag{15.11}$$

y sigue una distribución χ^2 con un número de grados de libertad dado por $\nu = (k-1)(m-1) = k-1$ (puesto que p se ha calculado a partir de los datos muestrales, sólo $k-1$ de ellos son realmente independientes). Por lo tanto, la hipótesis H_0 de homogeneidad de las muestras puede aceptarse con un nivel de significación α cuando

$$\frac{1}{p(1-p)} \sum_{i=1}^{k} \frac{(a_i - n_ip)^2}{n_i} \leq \chi^2_{\alpha,k-1}. \tag{15.12}$$

Un caso similar es cuando se quiere contrastar que k muestras pertenecen a una población binomial con un parámetro p determinado. El análisis es el mismo con la diferencia de que, al no calcularse p a partir de los datos muestrales y estar determinado a priori, el número de grados de libertad de la χ^2 es k en vez de $k-1$ (los k números de éxitos esperados son ahora independientes).

Otro caso importante de aplicación del contraste de homogeneidad de muestras es cuando se quiere contrastar si para k muestras supuestamente extraídas de una población de Poisson, el parámetro λ, o número medio de sucesos, es el mismo. Representemos por $a_1, a_2, \ldots, a_k$ el número de sucesos observados en cada muestra. A partir de estos datos, asumiendo la hipótesis nula H_0 de homogeneidad, se puede realizar una estimación del parámetro λ como

$$\lambda = \frac{\sum_{i=1}^{k} a_i}{k}$$

Por lo tanto, el número de sucesos esperados en cada muestra ha de ser $e_i = \lambda$, para todas las muestras. De esta forma, el estadístico χ^2 del contraste de homogeneidad se puede escribir como

$$\chi^2_{k-1} = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(a_i - \lambda)^2}{\lambda} = \sum_{i=1}^{k} \frac{a_i^2}{\lambda} - 2\sum_{i=1}^{k} \frac{a_i\lambda}{\lambda} + \sum_{i=1}^{k} \lambda =$$

$$= \frac{1}{\lambda} \sum_{i=1}^{k} a_i^2 - 2\sum_{i=1}^{k} a_i + \lambda k = \frac{1}{\lambda} \sum_{i=1}^{k} a_i^2 - \sum_{i=1}^{k} a_i \tag{15.13}$$

y este estadístico seguirá una distribución χ^2 con $k-1$ grados de libertad. Por lo tanto, la hipótesis nula de

que el número de sucesos es constante, se aceptará, a un nivel de significación α, cuando

$$\frac{1}{\lambda}\sum_{i=1}^{k} a_i^2 - \sum_{i=1}^{k} a_i \leq \chi^2_{\alpha,k-1}. \tag{15.14}$$

Capítulo 16

Análisis de varianza

"No es el conocimiento, sino el acto de aprender, no la posesión, sino el acto de llegar allí, lo que brinda el mayor placer."

Carl Friedrich Gauss (1777-1855)

En el Capítulo 14 se estudiaron los contrastes de hipótesis para la comparación de dos poblaciones. En particular se presentó el contraste de igualdad de medias entre dos poblaciones, estudiándose el caso particular de que las varianzas poblacionales fuesen iguales. A veces es necesario ensayar la hipótesis de igualdad de medias cuando se tienen más de dos poblaciones con la misma varianza. Esto se puede conseguir utilizando la técnica del análisis de varianza. Este importante método de análisis estadístico se basa en el estudio de la variación total entre los datos y la descomposición de ésta en diversos factores. De esta manera se puede contestar a la pregunta de si existen diferencias significativas entre las medias de las poblaciones o si, por el contrario, las diferencias encontradas pueden deberse a las limitaciones del muestreo. Se distinguirán dos casos principales, dependiendo de que exista uno o dos factores de variación entre las poblaciones.

16.1. Análisis con un factor de variación

Supongamos que se tienen p poblaciones independientes de las que se extraen p muestras aleatorias de tamaños no necesariamente iguales y representados por $n_1, n_2, \ldots, n_p$. En el análisis de varianza se emplea normalmente el término **tratamiento** para hablar de la característica que diferencia a las p poblaciones. Típicamente dicho tratamiento será, por ejemplo, un diferente abono (en agricultura), un diferente medicamento (en medicina) o, en general, un proceso diferente que se ha aplicado a cada una de las poblaciones y sobre el que se quiere medir su efectividad. De esta forma diremos que se tienen p tratamientos diferentes. Representaremos por x_{ij} al valor que toma la variable aleatoria en estudio para el elemento i–esimo del tratamiento (o muestra) j. Los valores de la variable aleatoria obtenidos en el muestreo se pueden representar entonces en una tabla de la siguiente forma:

Tratamientos	1	2	...	j	...	p
	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}
	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}
Datos	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
muestrales	$x_{n_1 1}$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
		$\vdots$	$\vdots$	$x_{n_j j}$	$\vdots$	$\vdots$
		$\vdots$	$\vdots$		$\vdots$	$x_{n_p p}$
		$x_{n_2 2}$	$\vdots$		$\vdots$	
			$\vdots$		$\vdots$	
Tamaños muestrales	n_1	n_2	...	n_j	...	n_p
Sumas muestrales	T_1	T_2	...	T_j	...	T_p
Medias muestrales	$\overline{x_1}$	$\overline{x_2}$	...	$\overline{x_j}$	...	$\overline{x_p}$

La tabla lista además las sumas muestrales T_j y los tamaños de cada muestra, en los que se verifica

$$\sum_{j=1}^{p} n_j = n,$$

donde n es el número total de elementos observados. Se incluyen también las medias muestrales definidas como

$$\overline{x_j} = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \tag{16.1}$$

Se puede definir además una media total que se puede escribir como

$$\overline{x} = \frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n_j} x_{ij} = \frac{1}{n} \sum_{j=1}^{p} n_j \overline{x_j} \tag{16.2}$$

Para poder aplicar correctamente el análisis de varianza es necesario que las p poblaciones de partida cumplan las siguientes condiciones:

1. Las p poblaciones de partida han de seguir una distribución normal.
2. La varianza poblacional σ^2 de las p poblaciones ha de ser la misma.
3. Las p muestras han de ser elegidas aleatoriamente.

Bajo estas condiciones, el objetivo del análisis de varianza es comprobar si las p medias poblacionales pueden ser las mismas. Es decir, se trata de probar si los efectos producidos por los tratamientos son significativamente diferentes entre si o no (ej. abono o medicamento más eficiente). En otras palabras, las hipótesis nula y alternativa del análisis de varianza de un solo factor son:

$$\begin{cases} H_0: & \mu_1 = \mu_2 = \ldots = \mu_j = \ldots = \mu_p \\ H_1: & \text{Al menos dos de las medias son diferentes} \end{cases} \tag{16.3}$$

El método del análisis de varianza se basa en estudiar las variaciones que siempre existirán entre los datos x_{ij} de la tabla. En principio se supone que dichas variaciones se pueden separar en dos tipos de variaciones diferentes:

a) **Variación dentro de los tratamientos (VDT)**, es decir variaciones entre los elementos de cada columna. Estas variaciones se suponen debidas al azar, es decir intrínsecas al proceso aleatorio de elección de la muestra.

b) **Variación entre los tratamientos (VET)**, o variaciones entre los valores medios $\overline{x_j}$ de cada tratamiento. Estas serán debidas, por una parte a efectos aleatorios, y podrán incluir posibles variaciones sistemáticas entre las medias poblacionales de cada tratamiento.

De esta manera, el objetivo del método es estudiar si la variación entre tratamientos es consistente con lo que podría esperarse de las variaciones aleatorias, o si, por el contrario, existen evidencias de variaciones sistemáticas entre los diferentes tratamientos. En otras palabras se trata de contrastar si la variación entre tratamientos es significativamente mayor que la variación dentro de los tratamientos.

Para desarrollar este método matemáticamente, se define la **variación total (VT)** de los datos de la tabla como

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2. \tag{16.4}$$

Esta variación total se puede desarrollar de la siguiente forma

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}\left((x_{ij} - \overline{x_j}) + (\overline{x_j} - \overline{x})\right)^2 =$$

$$= \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2 + \sum_{j=1}^{p}\sum_{i=1}^{n_j}(\overline{x_j} - \overline{x})^2 + 2\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})(\overline{x_j} - \overline{x}).$$

Además se demuestra que el último término de esta expresión es nulo pues

$$\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})(\overline{x_j} - \overline{x}) = \sum_{j=1}^{p}\left(\sum_{i=1}^{n_j}x_{ij}(\overline{x_j} - \overline{x}) - \sum_{i=1}^{n_j}\overline{x_j}(\overline{x_j} - \overline{x})\right) =$$

$$= \sum_{j=1}^{p}\left((\overline{x_j} - \overline{x})\sum_{i=1}^{n_j}x_{ij} - n_j\overline{x_j}(\overline{x_j} - \overline{x})\right) = \sum_{j=1}^{p}\left((\overline{x_j} - \overline{x})n_j\overline{x_j} - n_j\overline{x_j}(\overline{x_j} - \overline{x})\right) = 0.$$

Por lo tanto, la variación total queda de la siguiente forma

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2 + \sum_{j=1}^{p}n_j(\overline{x_j} - \overline{x})^2. \tag{16.5}$$

Esta última expresión, considerada como la ecuación fundamental del análisis de varianza, implica que la variación total de los datos puede escribirse como una suma de dos variaciones. La primera coincide con la variación dentro de los tratamientos, denotada por VDT

$$VDT = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2, \tag{16.6}$$

mientras que la segunda es la variación entre tratamientos VET

$$VET = \sum_{j=1}^{p}n_j(\overline{x_j} - \overline{x})^2. \tag{16.7}$$

Es decir, se puede expresar

$$VT = VDT + VET. \tag{16.8}$$

Es importante hacer notar que ambas variaciones, VET y VDT, pueden servir para hacer una estimación de la varianza poblacional común σ^2 en el caso de que H_0 sea cierta (es decir, si no existe diferencia entre las medias para cada tratamiento). Sin embargo, VET y VDT no son exactamente estimadores de la varianza pues constituyen suma de cuadrados de desviaciones, sin dividir aún por el número de puntos usados en cada estimación.

En particular, a partir de la variación dentro de los tratamientos VDT puede estimarse σ^2. Por una parte, usando un único tratamiento, un estimador puntual de la varianza del tratamiento j será la varianza muestral

$$s_j^2 = \frac{\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2}{n_j - 1}$$

Como todas las columnas han de tener la misma varianza poblacional σ^2, una buena estimación de ésta puede conseguirse haciendo la media ponderada de las varianzas muestrales pesando con el número de grados de libertad (o número de puntos menos 1) de cada muestra. Llamemos s_{VDT}^2 a esta estimación de σ^2

$$s_{VDT}^2 = \frac{\sum_{j=1}^{p}(n_j - 1)s_j^2}{\sum_{j=1}^{p}(n_j - 1)} = \frac{\sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij} - \overline{x_j})^2}{n - p}.$$

Introduciendo la definición (16.6) de VDT:

$$\Rightarrow \qquad ME \equiv s_{VDT}^2 = \frac{VDT}{n - p}, \tag{16.9}$$

donde se ha denotado esta estimación de σ^2 por ME, llamado **cuadrado medio del azar**, ya que representa la varianza esperada únicamente por los efectos aleatorios. Es importante indicar que, se cumpla o no la hipótesis nula de igualdad de medias, ME constituye siempre una estimación insesgada de la varianza poblacional. El número de grados de libertad de esta estimación es lógicamente $n - p$ pues se han usado p medias muestrales para su cálculo (sólo $n - p$ valores son independientes).

Por otra parte, si la hipótesis H_0 fuese cierta, la varianza poblacional también podría estimarse a partir de la variación entre tratamientos VET. Supongamos por simplicidad que todas las muestras tienen el mismo tamaño, que denotaremos por n_0. Las diferentes $\overline{x_j}$ son estimaciones de la media muestral (que suponemos constante). De forma que la varianza de la distribución muestral de medias se puede expresar como $\sigma_{\overline{x}}^2 = \sigma^2/n_0$. Por lo tanto, una estimación, denotada por s_{VET}^2, de la varianza poblacional σ^2 puede obtenerse a partir de la varianza de la distribución muestral de medias como

$$s_{VET}^2 = n_0 s_{\overline{x}}^2 = n_0 \frac{\sum_{j=1}^{p}(\overline{x_j} - \overline{x})^2}{p - 1} = \frac{\sum_{j=1}^{p} n_0(\overline{x_j} - \overline{x})^2}{p - 1}.$$

Con un desarrollo algo más largo se puede también demostrar que, en el caso de muestras de tamaños desiguales, una estimación de σ^2 viene dada, como cabría esperarse, por

$$s_{VET}^2 = \frac{\sum_{j=1}^{p} n_j(\overline{x_j} - \overline{x})^2}{p - 1}.$$

Si ahora se introduce la definición de la variación entre tratamientos (16.7) se obtiene

$$\Rightarrow \qquad MT \equiv s_{VET}^2 = \frac{VET}{p - 1}, \tag{16.10}$$

donde esta estimación de σ^2 se ha denotado por MT, llamado **cuadrado medio de los tratamientos**, representando la varianza esperada tanto por efectos aleatorios como por posibles diferencias entre las medias de cada tratamiento. Es decir, MT es una estimación insesgada de la varianza poblacional únicamente en el

caso de que se cumpla H_0. En otro caso, se esperarían valores mayores de MT pues los efectos sistemáticos, debidos a las diferencias entre las distintas medias, se sumarían a los aleatorios. Lógicamente, el número de grados de libertad de esta varianza es $p-1$, pues se han usado $p-1$ datos independientes.

En resumen, si se cumple H_0, tanto ME como MT constituirán estimaciones insesgadas de σ^2. Por el contrario, si hay variaciones sistemáticas entre poblaciones, esperaríamos tener un valor de MT mayor que ME, que sigue constituyendo una estimación de σ^2. De esta manera, el problema se convierte en una comparación de varianzas y las hipótesis establecidas en (16.3) son equivalentes a

$$\begin{cases} H_0: & \sigma^2_{VET} \leq \sigma^2_{VDT} \\ H_1: & \sigma^2_{VET} > \sigma^2_{VDT} \end{cases} \tag{16.11}$$

Es, entonces, un contraste unilateral sobre la igualdad de varianzas. Solo se rechazará la hipótesis nula cuando la varianza calculada a partir de la variación entre tratamientos sea mayor que la varianza estimada a partir de la variación dentro de los tratamientos. Según se explicó en la sección 2.2.3, este contraste se resuelve definiendo el estadístico

$$F = \frac{s^2_{VET}}{s^2_{VDT}} = \frac{MT}{ME} \tag{16.12}$$

y aceptando la hipótesis nula de no diferencia entre todas las medias poblacionales, a un nivel de significación α, cuando

$$\frac{MT}{ME} \leq F_{\alpha,p-1,n-p}, \tag{16.13}$$

donde $F_{\alpha,p-1,n-p}$ es la abscisa de la distribución F de Fisher con $p-1$ y $n-p$ grados de libertad que deja a su derecha un área igual a α.

Como resumen, los cálculos que se han de realizar para llevar a cabo el análisis de varianza se pueden mostrar en la siguiente tabla de análisis de varianza:

Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
entre tratamientos	VET	$p-1$	$MT = VET/(p-1)$
dentro de los tratamientos	VDT	$n-p$	$ME = VDT/(n-p)$
total	VT	$n-1$	$F = MT/ME$

(Nótese cómo el número de grados de libertad de la variación total es la suma de los grados de libertad de VET y VDT)

En la práctica existen fórmulas sencillas para el cálculo de las diferentes variaciones necesarias para el análisis. Por una parte, se puede desarrollar la expresión (16.4) para la variación total como sigue

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{n_j}(x_{ij}-\overline{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}x_{ij}^2 - 2\overline{x}\sum_{j=1}^{p}\sum_{i=1}^{n_j}x_{ij} + \sum_{j=1}^{p}\sum_{i=1}^{n_j}\overline{x}^2 =$$

$$= \sum_{j=1}^{p}\sum_{i=1}^{n_j}x_{ij}^2 - 2\overline{x}n\overline{x} + n\overline{x}^2 = \sum_{j=1}^{p}\sum_{i=1}^{n_j}x_{ij}^2 - n\overline{x}^2$$

Definiendo ahora un factor C como

$$C \equiv n\overline{x}^2 = \frac{1}{n}\left(\sum_{j=1}^{p}\sum_{i=1}^{n_j}x_{ij}\right)^2 \tag{16.14}$$

se llega a la expresión para la variación total VT

$$VT = \sum_{j=1}^{p} \sum_{i=1}^{n_j} x_{ij}^2 - C. \tag{16.15}$$

Por otra parte, la variación entre tratamientos VET se puede calcular desarrollando (16.7)

$$VET = \sum_{j=1}^{p} n_j(\overline{x_j} - \overline{x})^2 = \sum_{j=1}^{p} n_j \overline{x_j}^2 - 2\overline{x} \sum_{j=1}^{p} n_j \overline{x_j} + \overline{x}^2 \sum_{j=1}^{p} n_j.$$

Definiendo ahora las sumas muestrales T_j como

$$T_j \equiv n_j \overline{x_j} = \sum_{i=1}^{n_j} x_{ij}, \tag{16.16}$$

se puede expresar VET como

$$VET = \sum_{j=1}^{p} n_j \left(\frac{T_j}{n_j}\right)^2 - 2\overline{x} n \overline{x} + \overline{x}^2 n = \sum_{j=1}^{p} \frac{T_j^2}{n_j} - n\overline{x}^2$$

$$\Rightarrow \qquad VET = \sum_{j=1}^{p} \frac{T_j^2}{n_j} - C. \tag{16.17}$$

Por último, la variación dentro de los tratamientos VDT se puede calcular a partir de VT y VET usando (16.8). Es decir

$$VDT = VT - VET. \tag{16.18}$$

A partir de aquí se calculan los cuadrados medios ME y MT usando (16.9) y (16.10), y el cociente $F = MT/ME$, que se comparará con el valor crítico $F_{1-\alpha,p-1,n-p}$ para aceptar o rechazar la hipótesis nula de igualdad de medias entre las poblaciones.

16.2. Análisis con dos factores de variación

El análisis de varianza con un sólo factor de variación puede generalizarse al caso en que se tengan más factores de variación entre las poblaciones. En el caso particular de dos factores de variación se supone que además de tener p poblaciones con distintos tratamientos, en las muestras que se extraen de éstas, cada elemento corresponde a un valor de un segundo factor. Es decir cada muestra se divide en b elementos diferenciados por un factor. A cada conjunto de elementos con este segundo factor igual pero variando el primer factor, o tratamiento, se le llama **bloque**. Un ejemplo claro es cuando se quiere probar la eficiencia de p máquinas distintas (aquí las diferentes máquinas serían los tratamientos). Para ello se prueba el rendimiento de cada máquina cuando en ella trabajan b diferentes operarios (cada operario sería un bloque). En realidad es como si se tuvieran $p \times b$ poblaciones diferentes y se tomase un único dato de cada una de ellas. Evidentemente, además de las esperables variaciones aleatorias podría haber diferencias significativas debidas a los distintos tratamientos (eficiencia de las máquinas en el ejemplo) o a los distintos bloques (eficiencia de los operarios en el ejemplo). El análisis de varianza con dos factores de variación es la herramienta adecuada para contrastar simultáneamente si pueden existir variaciones sistemáticas entre tratamientos o entre bloques.

En general se representará por x_{ij} al valor que toma la variable aleatoria en estudio para el bloque i y el tratamiento j. De esta forma, si se tienen p tratamientos y b bloques los valores de la variable aleatoria

obtenidos en el muestreo se pueden representar en la siguiente tabla (suponemos que hay un único dato para cada tratamiento y bloque):

Bloques \ Tratamientos	1	2	...	j	...	p	Sumas	Medias
1	x_{11}	x_{12}	...	x_{1j}	...	x_{1p}	T_{B_1}	$\overline{x_{B_1}}$
2	x_{21}	x_{22}	...	x_{2j}	...	x_{2p}	T_{B_2}	$\overline{x_{B_2}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	x_{i1}	x_{i2}	...	x_{ij}	...	x_{ip}	T_{B_i}	$\overline{x_{B_i}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
b	x_{b1}	x_{b2}	...	x_{bj}	...	x_{bp}	T_{B_b}	$\overline{x_{B_b}}$
Sumas	T_{T_1}	T_{T_2}	...	T_{T_j}	...	T_{T_p}	T	
Medias	$\overline{x_{T_1}}$	$\overline{x_{T_2}}$	...	$\overline{x_{T_j}}$	...	$\overline{x_{T_p}}$		$\overline{x}$

La tabla lista además las sumas muestrales para cada bloque (T_{B_i}) y tratamiento (T_{T_j}), junto con las medias muestrales, definidas para el bloque i y el tratamiento j como

$$\overline{x_{B_i}} = \frac{1}{p}\sum_{j=1}^{p} x_{ij} \qquad ; \qquad \overline{x_{T_j}} = \frac{1}{b}\sum_{i=1}^{b} x_{ij}. \tag{16.19}$$

La media total $\overline{x}$ se puede escribir entonces como

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{p}\sum_{i=1}^{b} x_{ij} = \frac{1}{b}\sum_{i=1}^{b}\overline{x_{B_i}} = \frac{1}{p}\sum_{j=1}^{p}\overline{x_{T_j}}, \tag{16.20}$$

donde se cumple que el número de elementos n es igual a bp.

Al igual que en el caso de un único factor de variación se hace la hipótesis de que las pb poblaciones de partida son normales y tienen la misma varianza poblacional σ^2. Bajo estas condiciones, el objetivo del análisis de varianza es comprobar simultáneamente la hipótesis de igualdad de medias para los diferentes tratamientos, por un lado, y para los diferentes bloques, por otro. Es decir, para comprobar si hay diferencias entre los tratamientos y diferencias entre los bloques se plantean las siguientes hipótesis nula y alternativa:

$$\begin{cases} H_0: & \mu_{T_1} = \mu_{T_2} = \ldots = \mu_{T_j} = \ldots = \mu_{T_p} \\ H_1: & \text{Al menos dos de las medias } \mu_{\mathrm{T_j}} \text{ son diferentes} \end{cases} \tag{16.21}$$

$$\begin{cases} H_0': & \mu_{B_1} = \mu_{B_2} = \ldots = \mu_{B_i} = \ldots = \mu_{B_b} \\ H_1': & \text{Al menos dos de las medias } \mu_{\mathrm{B_i}} \text{ son diferentes} \end{cases} \tag{16.22}$$

El método del análisis de varianza se basa entonces en estudiar las variaciones entre los datos. Dichas variaciones se suponen de tres tipos diferentes:

a) **Variación debida al azar**. Son las variaciones dentro de cada columna o fila de la tabla. Es decir, son similares a las variaciones dentro de los tratamientos en el análisis con un sólo factor.

b) **Variación entre los tratamientos**, o variaciones entre los valores medios $\overline{x_{T_j}}$ de cada tratamiento. Estas serán debidas a los efectos aleatorios más las posibles variaciones sistemáticas entre los tratamientos.

c) **Variación entre los bloques**, debidas a los efectos aleatorios más las posibles variaciones sistemáticas entre los bloques.

El objetivo del método es entonces comprobar si la variaciones dadas en b) y c) son significativamente mayores que las variaciones debidas al azar. Para estudiar estas variaciones se comienza desarrollando la variación total, dada en (16.4), como

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{b}(x_{ij} - \overline{x})^2 = \sum_{j=1}^{p}\sum_{i=1}^{b}\left((x_{ij} - \overline{x_{T_j}} - \overline{x_{B_i}} + \overline{x}) + (\overline{x_{T_j}} - \overline{x}) + (\overline{x_{B_i}} - \overline{x})\right)^2.$$

Se puede comprobar que, al igual que en el caso del análisis con un sólo factor, los términos cruzados de la expresión anterior se anulan, quedando la variación total como

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{b}(x_{ij} - \overline{x_{T_j}} - \overline{x_{B_i}} + \overline{x})^2 + \sum_{j=1}^{p}\sum_{i=1}^{b}(\overline{x_{T_j}} - \overline{x})^2 + \sum_{j=1}^{p}\sum_{i=1}^{b}(\overline{x_{B_i}} - \overline{x})^2 \tag{16.23}$$

Por lo tanto se puede descomponer la variación total en tres términos correspondientes a la variación debida al azar (denotada por VDT pues es similar a la variación dentro de los tratamientos para el caso de un factor), la variación entre tratamientos (VET) y la variación entre bloques (VEB). Es decir

$$VT = VDT + VET + VEB, \tag{16.24}$$

donde

$$VDT = \sum_{j=1}^{p}\sum_{i=1}^{b}(x_{ij} - \overline{x_{T_j}} - \overline{x_{B_i}} + \overline{x})^2, \tag{16.25}$$

$$VET = b\sum_{j=1}^{p}(\overline{x_{T_j}} - \overline{x})^2, \tag{16.26}$$

$$VEB = p\sum_{i=1}^{b}(\overline{x_{B_i}} - \overline{x})^2. \tag{16.27}$$

Estas tres variaciones, VDT, VET y VEB, pueden servir para hacer una estimación de la varianza poblacional común σ^2 en el caso de que H_0 y H_0' sean ciertas. Por analogía con el caso de un factor, estas estimaciones se pueden escribir como los siguientes **cuadrados medios** del azar (ME), tratamientos (MT) y bloques (MB)

$$ME \equiv s^2_{VDT} = \frac{VDT}{(p-1)(b-1)}, \tag{16.28}$$

$$MT \equiv s^2_{VET} = \frac{VET}{p-1}, \tag{16.29}$$

$$MB \equiv s^2_{VEB} = \frac{VEB}{b-1}, \tag{16.30}$$

donde se ha dividido cada suma de cuadrados por los grados de libertad, o número de datos independientes para calcular dichas sumas. Nótese que en el caso de ME, al usarse p medias de tratamientos y b medias de bloques, el número de grados de libertad ha de ser $(p-1)(b-1)$.

Es importante indicar que ME constituye siempre una estimación insesgada de σ^2, se cumplan o no las hipótesis nulas. Sin embargo, MT y MB sólo serán estimadores insegados cuando se cumplan, respectivamente, H_0 y H_0'. En otros casos, es decir cuando existan diferencias sistemáticas entre tratamientos o bloques, dichos cuadrados tomarían valores mayores que σ^2, y por tanto que ME. Por lo tanto, el problema se plantea como dos contrastes unilaterales de igualdad de varianzas donde las hipótesis son

$$\begin{cases} H_0: & \sigma^2_{VET} \leq \sigma^2_{VDT} \\ H_1: & \sigma^2_{VET} > \sigma^2_{VDT} \end{cases} \tag{16.31}$$

$$\begin{cases} H_0': & \sigma^2_{VEB} \leq \sigma^2_{VDT} \\ H_1': & \sigma^2_{VEB} > \sigma^2_{VDT} \end{cases} \tag{16.32}$$

Para realizar este contraste se definen entonces los estadísticos

$$F = \frac{s^2_{VET}}{s^2_{VDT}} = \frac{MT}{ME} \quad ; \quad F' = \frac{s^2_{VEB}}{s^2_{VDT}} = \frac{MB}{ME}, \tag{16.33}$$

aceptándose la hipótesis nula H_0 de no diferencia entre los tratamientos, a un nivel de significación α, cuando

$$\frac{MT}{ME} \leq F_{\alpha,p-1,(p-1)(b-1)} \tag{16.34}$$

y aceptándose la hipótesis nula H_0' de no diferencia entre los bloques cuando

$$\frac{MB}{ME} \leq F_{\alpha,b-1,(p-1)(b-1)}. \tag{16.35}$$

Al igual que antes, se puede escribir una tabla resumen con todos los factores necesarios para realizar este análisis de varianza como:

Variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
entre tratamientos	VET	$p-1$	$MT = VET/(p-1)$
entre bloques	VEB	$b-1$	$MB = VEB/(b-1)$
debida al azar	VDT	$(p-1)(b-1)$	$ME = VDT/(p-1)(b-1)$
total	VT	$pb-1$	$F = MT/ME \quad ; \quad F' = MB/ME$

(El número de grados de libertad de la variación total es $n-1$ $(= pb-1)$ y coincide con la suma de los grados de libertad de VET, VEB y VDT)

Las fórmulas para el cálculo de las diferentes variaciones necesarias para el análisis son similares a las presentadas para el caso de un único factor. Así la variación total puede calcularse como

$$VT = \sum_{j=1}^{p}\sum_{i=1}^{b} x_{ij}^2 - C \qquad \text{donde} \qquad C = \frac{1}{n}\left(\sum_{j=1}^{p}\sum_{i=1}^{b} x_{ij}\right)^2. \tag{16.36}$$

Por otra parte, las variaciones entre tratamientos VET y entre bloques VEB se pueden expresar como

$$VET = \sum_{j=1}^{p} \frac{T_{T_j}^2}{b} - C \qquad \text{donde} \qquad T_{T_j} = \sum_{i=1}^{b} x_{ij} \tag{16.37}$$

$$VEB = \sum_{i=1}^{b} \frac{T_{B_i}^2}{p} - C \qquad \text{donde} \qquad T_{B_i} = \sum_{j=1}^{p} x_{ij} \tag{16.38}$$

Por último, la variación debida al azar VDT se puede calcular, usando (16.24), como

$$VDT = VT - VET - VEB. \tag{16.39}$$

Hay que indicar que en el análisis anterior se ha supuesto que hay un único dato para cada bloque y tratamiento dado. Se pueden hacer modificaciones a los desarrollos anteriores para realizar el análisis de varianza con dos factores cuando para cada tratamiento y bloque (es decir, para cada celda de la tabla de datos) se tienen toda una serie de medidas.

Tema V

REGRESIÓN LINEAL

Capítulo 17

Regresión lineal

"Afirmaciones extraordinarias requieren pruebas extraordinarias."

David Hume (1711-1776)

17.1. Regresión lineal simple

Dentro del estudio de las variables estadísticas bidimensionales vamos a abordar el análisis de la existencia de relaciones o dependencias entre las dos variables x e y que forman la variable bidimensional. Básicamente, **la relación** entre las dos variables podrá ser de dos tipos: **funcional**, cuando exista una relación matemática exacta que ligue ambas variables (ej. el radio y el área de un círculo), o **aleatoria**, cuando, aunque no exista entre las variables una relación exacta, se puede observar (aunque no siempre es el caso) una cierta tendencia entre los comportamientos de ambas (ej. el peso y la altura de un individuo).

El primer paso para el estudio de la relación entre las variables consiste en la construcción y observación de un diagrama de dispersión (Figura 17.1). El problema de la regresión se concreta entonces en ajustar una función a la nube de puntos representada en dicho diagrama. Esta función permitirá entonces obtener, al menos de forma aproximada, una estimación del valor de una de las variables a partir del valor que tome la otra. Cuando la función sea del tipo $y = f(x)$, hablaremos de **regresión de** y **sobre** x (a partir de los valores de x se pueden estimar los de y). Al contrario, la **regresión de** x **sobre** y se basará en una función del tipo $x = f(y)$.

Se conoce como **línea de regresión** a la representación gráfica de la función que se ajusta a la nube de puntos del diagrama de dispersión. Un primer problema para el estudio de la regresión es la elección del tipo de línea de regresión. Efectivamente, ésta podrá adoptar diferentes formas funcionales, y el tipo de línea se elegirá a partir de la forma de la nube de puntos. Cuando dicha nube se distribuya aproximadamente a lo largo de una línea recta ajustaremos una **recta de regresión**. Será el caso particular de la **regresión lineal**. En este caso importante, la regresión de y sobre x vendrá dada entonces por

$$y = a + bx, \tag{17.1}$$

donde a y b son dos parámetros que habremos de determinar. Gráficamente a será la ordenada de la recta en el origen (es decir el valor de y para $x = 0$) y b la pendiente de ésta.

Aunque aquí nos concentraremos, por simplicidad, en la regresión lineal, la línea de regresión puede responder a otras formas funcionales como, por ejemplo, es el caso de la regresión parabólica ($y = a+bx+cx^2$) y exponencial ($y = ab^x$).

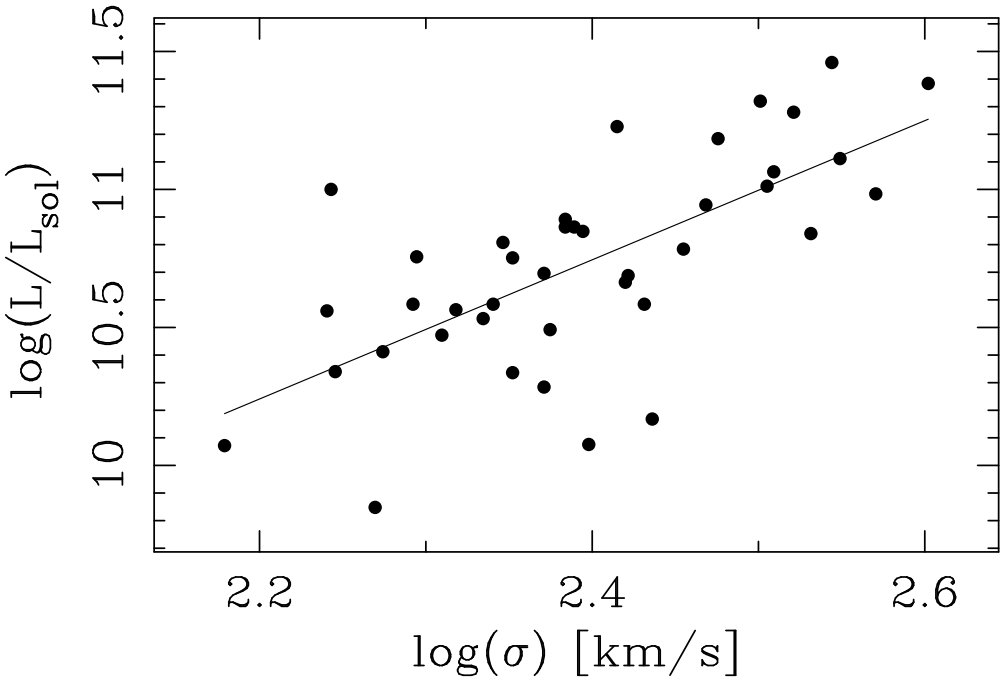


Figura 17.1: *Ejemplo de diagrama de dispersión. Los datos corresponden a las medidas de dispersión de velocidades y luminosidad en una muestra de 40 galaxias elípticas realizadas por Schechter (1980).*

17.2. Ajuste de una recta de regresión

Dentro del estudio de la regresión lineal vamos a analizar cómo se pueden determinar los parámetros a y b de la recta de regresión dada por (17.1), es decir, en el caso de la regresión de y sobre x (el caso contrario es similar). Como ya se ha indicado dicha recta de regresión nos permitirá obtener valores aproximados de y conocidos los de x.

Para calcular la recta que mejor se ajusta a la nube de puntos observada se usa el **método de mínimos cuadrados**. Veamos a continuación en qué consiste.

Sea una muestra de tamaño n en que la variable estadística bidimensional toma los valores

$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n).$$

A cada valor x_i de la variable x le corresponde entonces un valor y_i de la variable y, pudiendo además asociársele un valor y_i^*, que sería el dado por la recta que queremos calcular. Es decir

$$y_i^* = a + bx_i.$$

Llamemos d_i a la diferencia entre los dos valores, observado y dado por la recta, de la variable y en cada punto (ver Figura 17.2)

$$d_i = y_i^* - y_i.$$

Para que la recta a determinar sea la que mejor se ajuste a la nube de puntos de entre todas las rectas posibles, dichas distancias d_i deberán ser lo más pequeñas posible. Es decir, hay que minimizar los d_i. Para ello es conveniente tomar los cuadrados de las distancias, para que así no se anulen desviaciones positivas y

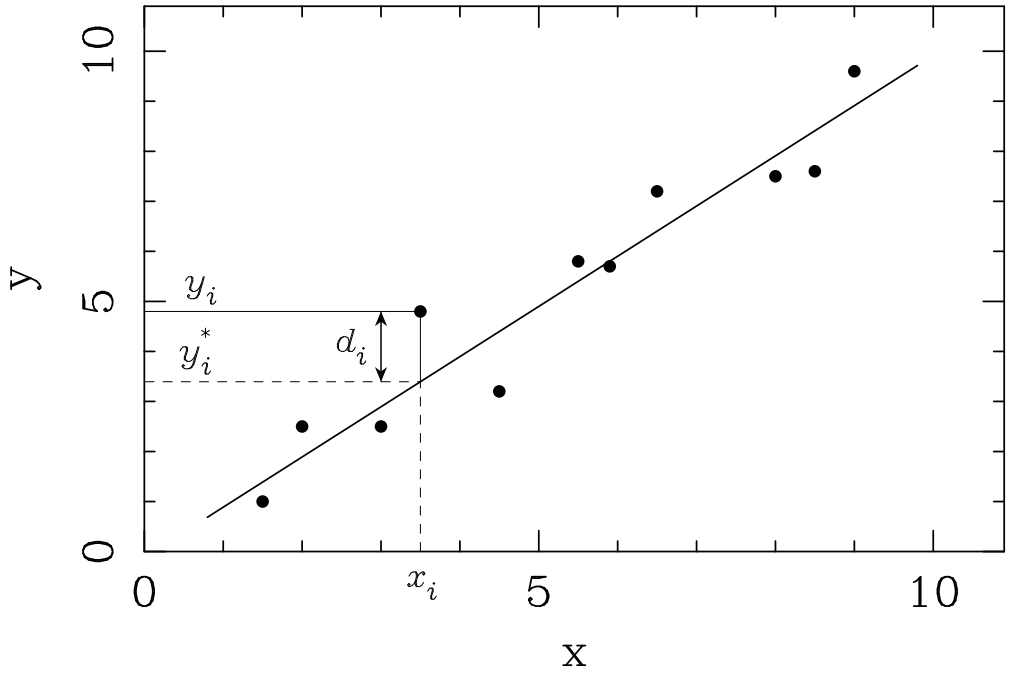


Figura 17.2: *Diferencia entre el valor observado y_i y el valor ajustado y_i^*.*

negativas. De esta forma, el problema se reduce a minimizar la expresión

$$M = \sum_{i=1}^{n} d_i^2 = \sum_{i=1}^{n} (y_i^* - y_i)^2,$$

o, utilizando la expresión para y_i^*

$$M = \sum_{i=1}^{n} (a + bx_i - y_i)^2.$$

Para encontrar los valores de a y b que hacen mínima esa expresión se deriva M respecto a esos dos parámetros y se igualan las derivadas a 0 (a partir de aquí se simplifica la notación de los sumatorios y no se indica que el índice va desde $i = 1$ hasta n)

$$\begin{cases} \dfrac{\partial M}{\partial a} = \sum 2(a + bx_i - y_i) = 0 \\ \\ \dfrac{\partial M}{\partial b} = \sum 2(a + bx_i - y_i)x_i = 0 \end{cases}$$

Desarrollando los sumatorios y recordando que $\sum_{i=1}^{n} a = an$

$$\Rightarrow \begin{cases} \sum(a + bx_i - y_i) = 0 \\ \\ \sum(ax_i + bx_i^2 - x_i y_i) = 0 \end{cases} \tag{17.2}$$

$$\Rightarrow \begin{cases} an + b\sum x_i = \sum y_i \\ \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i \end{cases} \tag{17.3}$$

Este sistema sencillo de ecuaciones, conocidas como **ecuaciones normales**, se puede resolver por el método

de Cramer, calculando en primer lugar el determinante

$$\Delta = \begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix} = n\sum x_i^2 - \Big(\sum x_i\Big)^2,$$

y cada uno de los parámetros por

$$a = \frac{1}{\Delta}\begin{vmatrix} \sum y_i & \sum x_i \\ \sum x_i y_i & \sum x_i^2 \end{vmatrix} = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b = \frac{1}{\Delta}\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Estas expresiones para los parámetros de la recta se pueden simplificar introduciendo las definiciones de media

$$\overline{x} = \frac{\sum x_i}{n} \qquad \text{y} \qquad \overline{y} = \frac{\sum y_i}{n}.$$

Dividiendo por n^2 en el numerador y denominador de la expresión para b, ésta queda

$$b = \frac{\frac{1}{n}\sum x_i y_i - \overline{x}\ \overline{y}}{\frac{1}{n}\sum x_i^2 - \overline{x}^2}. \tag{17.4}$$

Por otra parte, dividiendo por n en la primera expresión de (17.3)

$$\overline{y} = a + b\overline{x}. \tag{17.5}$$

Es decir, una vez calculado b, a se puede calcular de forma inmediata por

$$a = \overline{y} - b\overline{x}. \tag{17.6}$$

La expresión (17.5) es además interesante ya que indica que la recta de regresión debe pasar por $(\overline{x}, \overline{y})$, es decir, por el centro de la nube de puntos.

El desarrollo anterior puede generalizarse para calcular expresiones similares para la regresión parabólica y, en general, polinómica ($y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$). En el caso de la regresión exponencial el problema de la regresión se puede simplificar al de la regresión lineal ya que, tomando logaritmos

$$y = ab^x \qquad \Rightarrow \qquad \log y = \log a + x \log b.$$

17.3. Covarianza y coeficientes de regresión

Las expresiones para los parámetros de la recta de regresión se pueden simplificar más introduciendo una importante definición. Se define la **covarianza** de una muestra bidimensional a

$$\text{Cov} \equiv s_{xy}^2 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{n-1}. \tag{17.7}$$

Es decir, es una definición muy similar a la de la varianza s^2, pero mezclando las desviaciones de ambas variables. Al igual que ocurría con la varianza, en muchas ocasiones en el denominador se utiliza n en vez de $n-1$. Aquí usaremos esta segunda definición.

En el caso general de que haya valores repetidos, o agrupamiento en intervalos, la definición de la covarianza sería

$$\text{Cov} \equiv s^2_{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}(x_i-\overline{x})(y_j-\overline{y})n_{ij}}{n-1}. \tag{17.8}$$

Más adelante se profundizará más en el significado de la covarianza. Desarrollando la expresión (17.7) de la covarianza se puede llegar a una fórmula simplificada para calcularla

$$\begin{aligned} s^2_{xy} &= \frac{\sum(x_i-\overline{x})(y_i-\overline{y})}{n-1} = \frac{\sum(x_iy_i-\overline{x}y_i-x_i\overline{y}+\overline{x}\ \overline{y})}{n-1} = \\ &= \frac{\sum x_iy_i-\overline{x}\sum y_i-\overline{y}\sum x_i+n\overline{x}\ \overline{y}}{n-1} = \\ &= \frac{\sum x_iy_i-\overline{x}n\overline{y}-\overline{y}n\overline{x}+n\overline{x}\ \overline{y}}{n-1} = \frac{\sum x_iy_i-n\overline{x}\ \overline{y}}{n-1}. \end{aligned} \tag{17.9}$$

De la misma forma se puede desarrollar la expresión para la varianza de x

$$\begin{aligned} s^2_x &= \frac{\sum(x_i-\overline{x})^2}{n-1} = \frac{\sum(x_i^2-2x_i\overline{x}+\overline{x}^2)}{n-1} = \frac{\sum x_i^2-2\overline{x}\sum x_i+n\overline{x}^2}{n-1} = \\ &= \frac{\sum x_i^2-2n\overline{x}^2+n\overline{x}^2}{n-1} = \frac{\sum x_i^2-n\overline{x}^2}{n-1}. \end{aligned} \tag{17.10}$$

Nótese además que estas dos expresiones desarrolladas para la covarianza y varianza son similares al numerador y denominador, respectivamente, de la fórmula (17.4) para calcular el parámetro b (ordenada en el origen) de la recta de regresión. La similitud es más clara si escribimos dichas expresiones como

$$s^2_{xy} = \frac{n}{n-1}\left(\frac{1}{n}\sum x_iy_i-\overline{x}\ \overline{y}\right) \qquad ; \qquad s^2_x = \frac{n}{n-1}\left(\frac{1}{n}\sum x_i^2-\overline{x}^2\right).$$

De forma que la expresión para el coeficiente b de la recta de regresión de y sobre x puede escribirse como la razón entre la covarianza y la varianza de x. A dicho coeficiente se le llama **coeficiente de regresión de y sobre x** y se denota por b_{y_x}

$$b_{y_x} = \frac{s^2_{xy}}{s^2_x} = \frac{\text{Cov}}{s^2_x}. \tag{17.11}$$

Esto nos permite además, utilizando (17.6), poder escribir la ecuación de la recta de regresión como

$$\begin{aligned} y = a+bx &= (\overline{y}-b\overline{x})+\frac{\text{Cov}}{s^2_x}x = \overline{y}-\frac{\text{Cov}}{s^2_x}\overline{x}+\frac{\text{Cov}}{s^2_x}x \\ \Rightarrow \qquad y-\overline{y} &= \frac{\text{Cov}}{s^2_x}(x-\overline{x}). \end{aligned} \tag{17.12}$$

De igual manera se puede obtener la recta de regresión de x sobre y ($x = a+by$), minimizando en este caso las distancias horizontales $(x_i^*-x_i)$ a la recta. El resultado es que el **coeficiente de regresión de x sobre y** (denotado por b_{x_y}) y la recta resultante se pueden escribir

$$b_{x_y} = \frac{\text{Cov}}{s^2_y} \qquad ; \qquad x-\overline{x} = \frac{\text{Cov}}{s^2_y}(y-\overline{y}). \tag{17.13}$$

Nótese que ambas rectas de regresión (17.12) y (17.13) no coinciden en general y que ambas se cortan en el punto $(\overline{x},\overline{y})$ (ver Figura 17.3). Hay que indicar que la regresión de x sobre y es igualmente importante a la

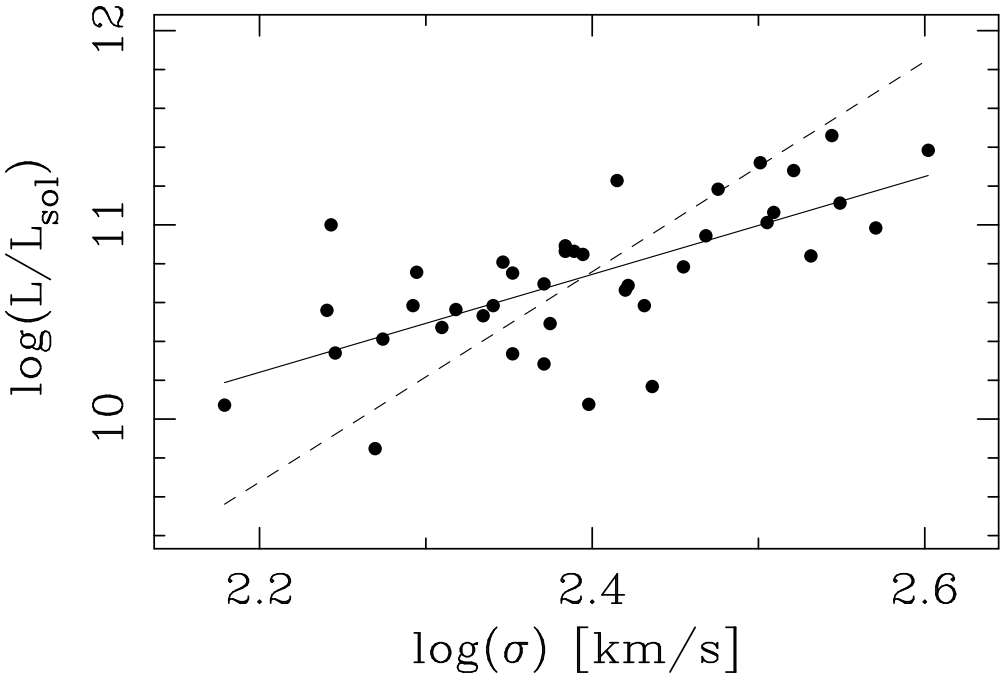


Figura 17.3: *Usando los mismos datos de la Figura 17.1 se comprueba que la recta de regresión de y sobre x (línea continua) no coincide con la recta de regresión de x sobre y (línea de trazos). Ambas rectas se cruzan en el punto* $(\overline{x}, \overline{y})$.

de y sobre x. En general, a no ser que se quiera estudiar en particular la dependencia de y con x, habrá que calcular ambas rectas.

El significado de los coeficientes de regresión es que b_{y_x} es, como ya se ha indicado, la pendiente de la recta de y sobre x, de forma que cuando sea positivo la recta será creciente y al contrario. En el caso de que $b_{y_x} = 0$ la recta será horizontal. De la misma manera, b_{x_y} representa la pendiente de la recta respecto al eje de ordenadas Y, y cuando sea nulo la recta será vertical. Se puede observar además que ambos coeficientes de regresión tienen el mismo signo (el signo de la covarianza, ya que las varianzas siempre son positivas). Esto implica que las dos rectas de regresión serán a la vez ascendentes o descendentes.

17.4. Correlación lineal

Después de haber considerado el tema de la regresión, cuyo objetivo era la estimación de una variable a partir de la otra, nos planteamos el problema de la **correlación**, el cual estudia el grado de asociación o dependencia entre las dos variables. Es decir, estudiar la correlación significa analizar hasta qué punto es significativa la dependencia de una variable con la otra. De esta manera, por ejemplo, cuando exista una dependencia funcional entre ambas variables diremos que tenemos una correlación perfecta (ej. radio y área de un círculo). Cuando, por el contrario, no exista ninguna dependencia entre las variables diremos que no hay correlación (ej. primera letra del apellido y altura de un individuo). El caso más interesante es el intermedio, cuando es posible que exista alguna correlación, aunque no perfecta, que habrá que cuantificar.

Nos vamos a concentrar aquí en un tipo particular de correlación que es la **correlación lineal**. Esta estudiará el grado en que la nube de puntos representada en el diagrama de dispersión se acerca a una recta. Cuanto mejor se aproxime dicha nube a una recta, mayor será el grado de correlación lineal. De esta forma, el estudio de la correlación lineal está íntimamente ligado al de la regresión lineal. Distinguiremos dos tipos

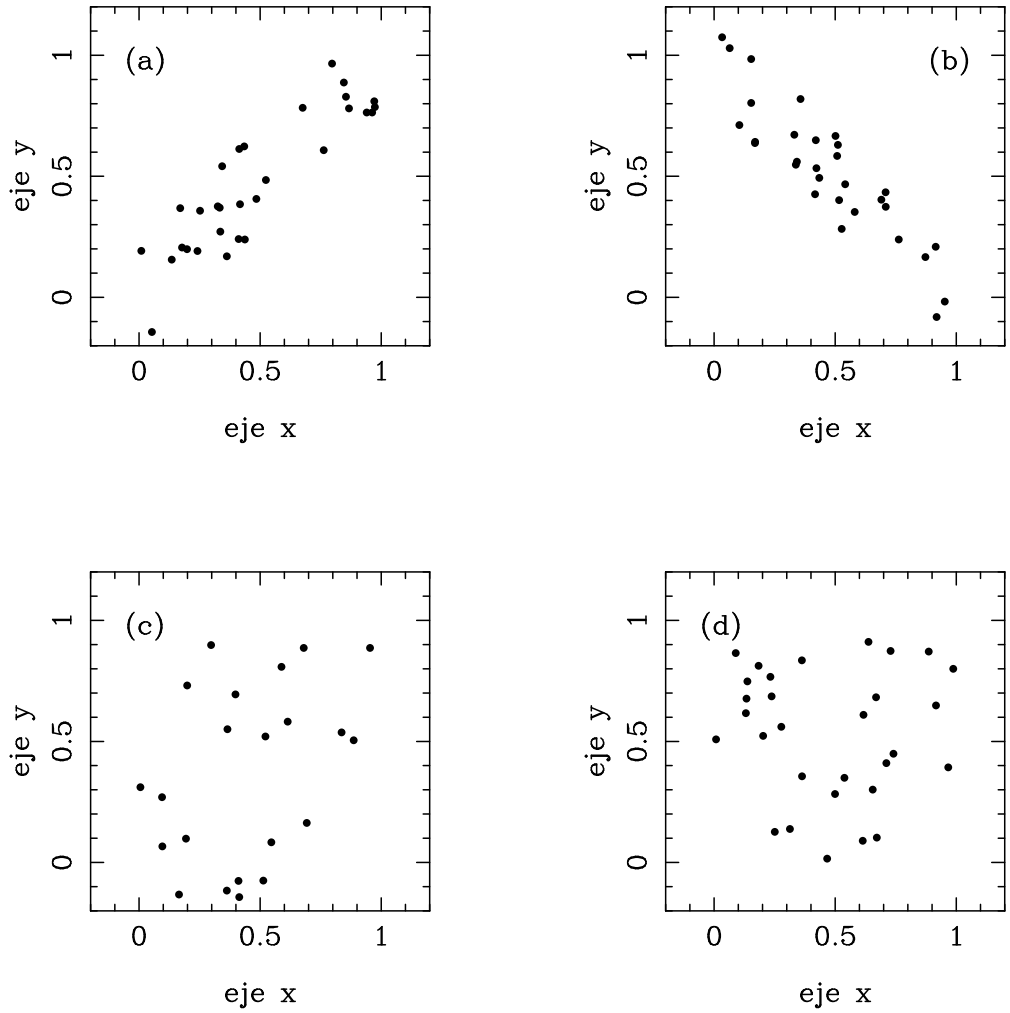


Figura 17.4: *Distintos ejemplos sencillos de correlaciones: (a) claramente positiva; (b) claramente negativa; (c) débilmente positiva; y (d) sin sin correlación.*

de correlación lineal. Cuando al crecer la variable x, la variable y tienda también a aumentar (pendiente positiva de la recta de regresión) diremos que tenemos una correlación **positiva o directa**. Cuando ocurra lo contrario, la correlación será **negativa o inversa**.

Evidentemente, la simple observación del diagrama de dispersión proporciona una idea cualitativa del grado de correlación. Sin embargo, es claramente más útil disponer de una medida cuantitativa de dicha correlación. Una primera cuantificación de la correlación se puede obtener a partir de la covarianza. Efectivamente, en la Figura 17.4 puede observarse que, en el caso de una clara correlación lineal positiva, la mayor parte de los puntos estarán en el segundo y tercer cuadrante, de forma que, en la definición de covarianza dada en (17.7) cuando x_i sea mayor que $\overline{x}$, también y_i tenderá a ser mayor que $\overline{y}$, y al revés. Por tanto, la mayoría de los términos del sumatorio serán positivos y la covarianza alcanzará un valor alto. Por el mismo argumento, si existe correlación lineal negativa, la mayoría de los términos del sumatorio serán negativos y la covarianza tendrá un valor alto y negativo. En el caso de que no hubiese correlación y los puntos estuviesen repartidos en los cuatro cuadrantes, en el numerador de (17.7) aparecerían por igual términos positivos y negativos, que se anularían dando un valor muy bajo, en valor absoluto, de la covarianza. En resumen, la covarianza es una medida de la correlación lineal entre las dos variables.

17.5. Coeficiente de correlación lineal y varianza residual

La utilidad de la covarianza como medida de correlación está limitada por el hecho de que depende de las unidades de medida en que se trabaje. Para construir una medida adimensional de la correlación habrá que dividir la varianza por un término con sus mismas dimensiones. De esta forma, se define el **coeficiente de correlación lineal** r como el cociente entre la covarianza y las desviaciones típicas (o raices cuadradas de las varianzas) de x e y

$$r = \frac{s_{xy}^2}{s_x s_y} = \frac{\text{Cov}}{s_x s_y}. \tag{17.14}$$

Desarrollando esta expresión mediante la aplicación de (17.9) y (17.10) se puede llegar a una fórmula más fácil de aplicar para el cálculo del coeficiente de correlación lineal

$$r = \frac{s_{xy}^2}{s_x s_y} = \frac{\frac{1}{n-1}\left(\sum x_i y_i - n\overline{x}\ \overline{y}\right)}{\sqrt{\frac{1}{n-1}\left(\sum x_i^2 - n\overline{x}^2\right)}\sqrt{\frac{1}{n-1}\left(\sum y_i^2 - n\overline{y}^2\right)}} =$$

$$= \frac{\sum x_i y_i - n\overline{x}\ \overline{y}}{\sqrt{\left(\sum x_i^2 - n\overline{x}^2\right)\left(\sum y_i^2 - n\overline{y}^2\right)}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2 - (\sum x_i)^2\right)\left(n\sum y_i^2 - (\sum y_i)^2\right)}}$$

Es importante resaltar que el coeficiente de correlación no depende de las unidades en que se midan las variables, al contrario que la varianza o la covarianza.

Es posible establecer una relación entre el coeficiente de correlación lineal (r) y los coeficientes de regresión (b_{y_x} y b_{x_y}). Usando las definiciones de ambos coeficientes

$$\left.\begin{array}{rcl} b_{y_x} = \dfrac{\text{Cov}}{s_x^2} & \Rightarrow & \text{Cov} = b_{y_x} s_x^2 \\ r = \dfrac{\text{Cov}}{s_x s_y} & \Rightarrow & \text{Cov} = r s_x s_y \end{array}\right\} \Rightarrow b_{y_x} s_x^2 = r s_x s_y \Rightarrow b_{y_x} = r\frac{s_y}{s_x}. \tag{17.15}$$

De la misma forma se puede encontrar una expresión para el coeficiente de regresión de x sobre y en función del coeficiente de correlación

$$b_{x_y} = r\frac{s_x}{s_y}. \tag{17.16}$$

Además se puede demostrar que el coeficiente de correlación es la media geométrica de los dos coeficientes de regresión, ya que

$$r = \frac{\text{Cov}}{s_x s_y} = \sqrt{\frac{\text{Cov}}{s_x^2}\frac{\text{Cov}}{s_y^2}} = \pm\sqrt{b_{y_x} b_{x_y}}.$$

Un concepto relacionado con el coeficiente de correlación es el de la **varianza residual**. Esta se introduce para proporcionar una estimación de la variación de los datos originales respecto a la recta de regresión que se ha ajustado. Su definición es la siguiente

$$s_r^2 = \frac{\sum_{i=1}^n (y_i - y_i^*)^2}{n-2} = \frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{n-2}. \tag{17.17}$$

Es decir, al igual que la varianza de una variable es una medida de la dispersión respecto al valor medio de ésta, la varianza residual mide la dispersión de los puntos respecto a la recta ajustada. Algunos autores definen la varianza residual utilizando n en vez de $n-2$. La definición aquí usada da una mejor estimación de la dispersión del ajuste. Nótese que, de forma similar a lo que ocurría en la definición de la varianza, solo

existen $n-2$ desviaciones independientes respecto a la recta (el sistema tiene $n-2$ grados de libertad), ya que si sólo tuviésemos 2 puntos conoceríamos sus desviaciones pues ambas serían 0, de aquí el sentido de promediar las desviaciones al cuadrado dividendo por ese número.

A partir de la varianza residual se puede definir la desviación típica residual como

$$s_r = \sqrt{\frac{\sum_{i=1}^{n}(y_i - a - bx_i)^2}{n-2}}. \tag{17.18}$$

También se puede encontrar una relación entre esta varianza residual y el coeficiente de correlación. Partiendo de la definición de varianza residual e introduciendo (17.6)

$$s_r^2 = \frac{\sum(y_i - a - bx_i)^2}{n-2} = \frac{\sum(y_i - \overline{y} + b\overline{x} - bx_i)^2}{n-2} = \frac{\sum\left((y_i - \overline{y}) - b(x_i - \overline{x})\right)^2}{n-2} =$$

$$= \frac{\sum(y_i - \overline{y})^2 + b^2\sum(x_i - \overline{x})^2 - 2b\sum(y_i - \overline{y})(x_i - \overline{x})}{n-2}.$$

Introducimos ahora las definiciones de varianza y covarianza (17.7)

$$s_r^2 = \frac{n-1}{n-2}\,(s_y^2 + b^2 s_x^2 - 2b\mathrm{Cov}).$$

Sustituyendo b por su expresión en (17.15) (nótese que el coeficiente de regresión que estamos usando es b_{y_x}) y poniendo la covarianza en función del coeficiente de correlación, usando (17.14)

$$s_r^2 = \frac{n-1}{n-2}\,\left(s_y^2 + r^2\frac{s_y^2}{s_x^2}s_x^2 - 2r\frac{s_y}{s_x}\mathrm{Cov}\right) = \frac{n-1}{n-2}\,\left(s_y^2 + r^2 s_y^2 - 2r\frac{s_y}{s_x}rs_xs_y\right) =$$

$$= \frac{n-1}{n-2}\,(s_y^2 + r^2 s_y^2 - 2r^2 s_y^2) = \frac{n-1}{n-2}\,(s_y^2 - r^2 s_y^2)$$

$$\Rightarrow s_r^2 = \frac{n-1}{n-2}\,s_y^2(1-r^2). \tag{17.19}$$

17.6. Interpretación del coeficiente de correlación

Usando las relaciones derivadas en el apartado anterior se puede hacer una interpretación del coeficiente de correlación. En primer lugar, a partir de (17.19) podemos acotar sus posibles valores. Efectivamente, dado que, por sus definiciones, tanto la varianza residual s_r^2 como la varianza s_y^2 han de ser positivas, podemos deducir que el coeficiente de correlación ha de estar acotado entre los valores -1 y $+1$

$$(1-r^2) \geq 0 \quad \Rightarrow \quad r^2 \leq 1 \quad \Rightarrow \quad -1 \leq r \leq 1.$$

Además, a partir de la relaciones (17.15) y (17.16), junto con la definición (17.14) del coeficiente de correlación, puede observarse que dicho coeficiente de correlación, los coeficientes de regresión y la covarianza han de tener el mismo signo

$$r \geq 0 \Longleftrightarrow b_{y_x} \geq 0 \Longleftrightarrow b_{x_y} \geq 0 \Longleftrightarrow \mathrm{Cov} \geq 0.$$

Es decir, cuando el coeficiente de correlación sea positivo, la pendiente de la recta será positiva (al igual que la varianza) y tendremos una correlación directa o positiva. Asimismo, cuando r sea negativo, nos

indicará que la correlación es inversa o negativa.

Respecto a los valores concretos del coeficiente de correlación podemos establecer los siguientes casos:

1. $r = 0$. En este caso, por las relaciones vistas en el apartado anterior, es claro que se cumple

$$r = 0 \quad \Rightarrow \quad \text{Cov} = 0 \quad ; \quad b_{y_x} = b_{x_y} = 0 \quad ; \quad s_r^2 \simeq s_y^2.$$

   Es decir, en este caso, al ser la covarianza nula no existirá correlación. Además las pendientes de la rectas de regresión de y sobre x y de x sobre y serán nulas, es decir sus orientaciones serán horizontal y vertical respectivamente. Por otra parte, al ser la varianza residual aproximadamente igual a la varianza de y, la dispersión de la variable y no se verá reducida al ajustar la recta de regresión.

2. $r = 1$. Es claro que en este caso se cumple que la varianza residual es nula ($s_r^2 = 0$), por lo que no habrá dispersión de los puntos respecto a la recta y todos se situaran sobre ella. En este caso tendremos una dependencia funcional entre ambas variables y una correlación positiva, o directa, perfecta. Además las dos rectas de regresión (de y sobre x y de x sobre y) coincidirán.

3. $r = -1$. Al igual que en el caso anterior todos los puntos se situarán sobre la recta y la correlación será negativa, o inversa, perfecta.

4. $0 < r < 1$. En este caso, la correlación será positiva pero no perfecta. Evidentemente la correlación (y la covarianza) será mejor cuanto más se acerque r a 1.

5. $-1 < r < 0$. De la misma manera tendremos una correlación negativa tanto mejor cuanto más próximo esté r a -1.

Para examinar más profundamente el significado del coeficiente de correlación, despejemos éste de la relación (17.19)

$$r^2 = 1 - \frac{(n-2)s_r^2}{(n-1)s_y^2} = 1 - \frac{\sum_i^n (y_i - y_i^*)^2}{\sum_i^n (y_i - \overline{y})^2}, \tag{17.20}$$

donde se han aplicado las definiciones de varianza de y y varianza residual (17.17). Además se puede desarrollar el término del denominador como

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n \left((y_i - y_i^*) + (y_i^* - \overline{y})\right)^2 =$$

$$\sum_{i=1}^n (y_i - y_i^*)^2 + \sum_{i=1}^n (y_i^* - \overline{y})^2 + 2\sum_{i=1}^n (y_i - y_i^*)(y_i^* - \overline{y}).$$

El término cruzado de la relación anterior es nulo ya que

$$\sum_{i=1}^n (y_i - y_i^*)(y_i^* - \overline{y}) = \sum_{i=1}^n (y_i - a - bx_i)(a + bx_i - \overline{y}) =$$

$$= a\sum_{i=1}^n (y_i - a - bx_i) + b\sum_{i=1}^n x_i(y_i - a - bx_i) - \overline{y}\sum_{i=1}^n (y_i - a - bx_i) = 0,$$

puesto que todos los sumatorios se anulan por (17.2). Por lo tanto, hemos demostrado que

$$\sum_{i=1}^n (y_i - \overline{y})^2 = \sum_{i=1}^n (y_i - y_i^*)^2 + \sum_{i=1}^n (y_i^* - \overline{y})^2. \tag{17.21}$$

Esta última expresión puede interpretarse usando la terminología del análisis de varianza. Efectivamente la suma de cuadrados del primer término representa la **variación total** (VT) de la variable dependiente

respecto a su valor medio $\overline{y}$. Por otra parte, el primer sumando del segundo término es la **variación no explicada** (VNE) por la recta de regresión, representando la variación de los datos, o residuos, alrededor de dicha recta. Al último sumando se le llama **variación explicada** (VE), ya que es la parte de la variación total que se explica por la recta ajustada. De esta forma, la variación total se descompone en dos variaciones, no explicada y explicada por la recta de regresión

$$VT = VNE + VE \tag{17.22}$$

Introduciendo la expresión (17.21) en la relación (17.20) para el coeficiente de correlación, se llega a

$$r^2 = \frac{\sum_{i=1}^n (y_i - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} - \frac{\sum_{i=1}^n (y_i - \overline{y})^2 - \sum_{i=1}^n (y_i^* - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

$$\Rightarrow \qquad r^2 = \frac{\sum_{i=1}^n (y_i^* - \overline{y})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{VE}{VT} = \frac{\text{Variación explicada}}{\text{Variación total}}. \tag{17.23}$$

Es decir, r^2, conocido como **coeficiente de determinación**, puede interpretarse como la fracción de la variación total que se explica por la recta de regresión. Así, un coeficiente de correlación próximo a ± 1 indica que casi todas las variaciones encontradas en y son explicadas por la recta (teniéndose una buena correlación), mientras que si r es 0, la recta de regresión apenas sirve para explicar las variaciones y la correlación lineal será pobre. Como ejemplo, si $r = 0.95$, podemos deducir que aproximadamente el 90 % de las variaciones de y son debidas a la regresión lineal.

Aunque el análisis de la regresión lineal y la derivación del coeficiente de correlación parecen un método muy adecuado para estudiar la relación entre dos variables, hay que indicar que tiene importantes debilidades. En particular:

- Tanto la recta de regresión como el coeficiente de correlación no son robustos, en el sentido de que resultan muy afectados por medidas particulares que se alejen mucho de la tendencia general.
- No hay que olvidar que el coeficiente de correlación no es más que una medida resumen. En ningún caso puede substituir al diagrama de dispersión, que siempre habrá que construir para extraer más información. Formas muy diferentes de la nube de puntos pueden conducir al mismo coeficiente de correlación.
- El que en un caso se obtenga un coeficiente de correlación bajo no significa que no pueda existir correlación entre las variables. De lo único que nos informa es de que la correlación no es lineal (no se ajusta a una recta), pero es posible que pueda existir una buena correlación de otro tipo.
- Un coeficiente de correlación alto no significa que exista una dependencia directa entre las variables. Es decir, no se puede extraer una conclusión de causa y efecto basándose únicamente en el coeficiente de correlación. En general hay que tener en cuenta que puede existir una tercera variable escondida que puede producir una correlación que, en muchos casos, puede no tener sentido.

Capítulo 18

Inferencia estadística sobre la regresión

"La predicción es difícil, especialmente si se trata del futuro."

Niels Bohr (1885-1962)

En este tema se van a utilizar los conceptos básicos de la teoría muestral y el contraste de hipótesis, ya estudiados en los temas anteriores, para elaborar un modelo estadístico de la regresión lineal simple. Esto nos permitirá estudiar desde un punto de vista probabilístico los parámetros de la recta de regresión y el concepto de correlación.

18.1. Fundamentos

En primer lugar es importante hacer la distinción entre las dos variables x e y que intervienen en la regresión lineal. Por una parte, y se considera como la **variable dependiente** (o respuesta), que tomará diferentes valores dependiendo del valor de x, o **variable independiente** (o de regresión). Supongamos que en el experimento se toma una muestra aleatoria representada por los pares (x_i, y_i), donde $i = 1, 2, \ldots, n$. Normalmente, los valores de x_i se fijan *a priori* (antes de realizar el experimento) y por tanto serán los mismos para las diferentes muestras que se puedan tomar. Se consideran entonces que tienen asociado un error despreciable y no son variables aleatorias. Por el contrario, para un valor de x fijo, el y_i particular medido podrá variar de una muestra a otra, de forma que, para cada x_i, la variable Y_i, que engloba a todos los posibles valores de y que se pueden obtener para $x = x_i$, se considerará una variable aleatoria en el muestreo. Tendrá, por lo tanto, una distribución de probabilidad asociada y se podrán definir su valor medio y varianza. Llamaremos $\mu_{Y|x}$ al valor medio de la variable Y para un valor fijo de x y $\sigma^2_{Y|x}$ a su varianza. Dichos valores medios dependerán entonces del valor concreto de x que se considere.

La hipótesis básica de la regresión lineal es que $\mu_{Y|x}$ está linealmente relacionado con x por la ecuación

$$\mu_{Y|x} = \alpha + \beta x. \tag{18.1}$$

Esta es la ecuación de regresión lineal poblacional. α y β serán los parámetros poblacionales correspondientes que tendrán que estimarse a partir de una muestra. Como se demostrará posteriormente, los coeficientes de la recta a y b se usarán como los estimadores de dichos parámetros poblacionales. De esta forma, $\mu_{Y|x}$ se

estimará por

$$y^* = a + bx, \tag{18.2}$$

que será la ecuación de regresión lineal ajustada o de la muestra. Es importante destacar que para diferentes muestras se obtendrán diferentes valores concretos de a y b, y por lo tanto diferentes rectas de regresión ajustadas, que en general no coincidirán con la recta poblacional dada en (18.1). A y B serán entonces también variables aleatorias en el muestreo.

El modelo estadístico para la regresión se basa entonces en suponer que todas las $\mu_{Y|x}$ caen sobre la recta poblacional y las diferencias encontradas se basan en la limitación del muestreo. En particular, para cada valor fijo de $x = x_i$, un valor concreto de Y_i (denotado por y_i) pofrá expresarse como

$$y_i = \mu_{Y|x_i} + \varepsilon_i = \alpha + \beta x_i + \varepsilon_i, \tag{18.3}$$

donde ε_i es el error aleatorio que tiene en cuenta la diferencia entre el valor observado y el valor medio esperado. Lógicamente se cumplirá que $\mu_{\varepsilon_i} = 0$.

Por otra parte, al usar la recta ajustada (18.2), los valores y_i medidos se podrán expresar como

$$y_i = y_i^* + e_i = a + bx_i + e_i, \tag{18.4}$$

donde e_i es el residuo y representa el error en el ajuste.

Una suposición adicional que se debe hacer para simplificar el estudio estadístico de la regresión lineal es que los errores ε_i para cada x_i tienen todos la misma varianza, denotada por σ^2. Esto quiere decir que para cada x_i los valores muestrales de Y_i se distribuyen todos alrededor de su correspondiente $\mu_{Y|x_i}$ con la misma dispersión. Es decir, los errores en la medida no han de depender del valor concreto de la variable independiente x. Bajo estas condiciones se puede expresar entonces que

$$\sigma^2_{Y_i} = \sigma^2_{\varepsilon_i} = \sigma^2. \tag{18.5}$$

σ^2 es por tanto la varianza de las diferentes variables aleatorias Y_i. Otra suposición importante es considerar que las variables aleatorias Y_i, para cada $x = x_i$, siguen una distribución normal, es decir, sus errores se distribuyen normalmente alrededor del valor medio. Por tanto, cada Y_i tendrá una distribución $N(\alpha+\beta x_i, \sigma)$.

18.2. Coeficientes de la recta

Como ya se ha indicado, para estimar los parámetros poblacionales α y β de la recta poblacional se usan los valores a y b deducidos a partir del método de los mínimos cuadrados. Diferentes muestras conducen a diferentes valores de dichos estimadores y, por lo tanto, A y B son variables aleatorias en el muestreo, con distribuciones de probabilidad asociadas. Para poder realizar contrastes de hipótesis sobre los parámetros de la recta es necesario entonces estudiar en primer lugar las características de dichas distribuciones muestrales.

18.2.1. Distribuciones de probabilidad

Estudiemos en primer lugar la distribución de probabilidad para el estimador B del coeficiente de regresión (pendiente del ajuste). Desarrollando la expresión (17.11) para b

$$b = \frac{s^2_{xy}}{s^2_x} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{(n-1)s^2_x} = \frac{\sum_{i=1}^n (x_i - \overline{x})y_i - \overline{y}\sum_{i=1}^n (x_i - \overline{x})}{(n-1)s^2_x}$$

$$b = \frac{\sum_{i=1}^{n}(x_i - \overline{x})y_i}{(n-1)s_x^2} = \sum_{i=1}^{n} w_i y_i \qquad \text{donde} \qquad w_i = \frac{x_i - \overline{x}}{(n-1)s_x^2}$$

De esta forma podemos expresar el coeficiente de regresión como una combinación lineal de las variables aleatorias Y_i. Nótese que cada w_i depende únicamente de los valores de las x y, por tanto, no cambia de muestra a muestra. Puesto que cada Y_i es normal, por las propiedades de dicha distribución el estadístico B seguirá también una distribución normal. El valor esperado (o medio) de B puede calcularse tomando esperanzas matemáticas en la expresión anterior

$$\mu_B = E(B) = \sum_{i=1}^{n} w_i E(Y_i) = \sum_{i=1}^{n} w_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} w_i + \beta \sum_{i=1}^{n} w_i x_i$$

Los sumatorios que aparecen en esta expresión pueden desarrollarse para demostrar que

$$\sum_i w_i = \frac{\sum_{i=1}^{n}(x_i - \overline{x})}{(n-1)s_x^2} = 0$$

$$\sum_{i=1}^{n} w_i x_i = \frac{\sum_{i=1}^{n}(x_i - \overline{x})x_i}{(n-1)s_x^2} = \frac{\sum_{i=1}^{n} x_i^2 - \overline{x}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} = 1$$

Por lo tanto

$$\mu_B = E(B) = \beta. \tag{18.6}$$

y B es un estimador insesgado de la pendiente β de la recta poblacional.

De forma similar se puede llegar a una expresión para la varianza de B, utilizando (18.5)

$$\sigma_B^2 = Var(B) = \sum_{i=1}^{n} w_i^2 \sigma_{Y_i}^2 = \sigma^2 \sum_{i=1}^{n} w_i^2 = \sigma^2 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{(n-1)^2 s_x^4}$$

$$\Rightarrow \qquad \sigma_B^2 = \sigma^2 \frac{s_x^2}{(n-1)s_x^4} = \frac{\sigma^2}{(n-1)s_x^2}.$$

Esta expresión tiene un importante significado intuitivo. El error en la determinación de la pendiente de la recta ha de ser inversamente proporcional al rango cubierto por las x, puesto que un rango pequeño conducirá a una pendiente muy indeterminada. En general, el error en la pendiente: (i) disminuirá al aumentar la dispersión de los valores de x; (ii) aumentará con σ^2, o el error intrínseco para las medidas de Y_i, y (iii) disminuirá al aumentar el número de puntos.

En resumen, hemos demostrado que B seguirá una distribución normal de parámetros

$$N\left(\beta, \frac{\sigma}{\sqrt{n-1}s_x}\right). \tag{18.7}$$

De forma similar se puede estudiar la distribución muestral del estadístico A que representa la ordenada en el origen. Desarrollando la expresión (17.6) para a se puede demostrar también que ésta puede expresarse como una combinación lineal de las variables aleatorias Y_i

$$a = \overline{y} - b\overline{x} = \frac{\sum_{i=1}^{n} y_i}{n} - \overline{x}\sum_{i=1}^{n} w_i y_i = \sum_{i=1}^{n}\left(\frac{1}{n} - \overline{x}w_i\right) y_i$$

$$a = \sum_{i=1}^{n} r_i y_i \qquad \text{donde} \qquad r_i = \frac{1}{n} - \overline{x}w_i$$

Al ser entonces una combinación lineal de variables normales independientes, A seguirá también una

distribución normal. Su valor medio se puede encontrar desarrollando la expresión anterior

$$\mu_A = E(A) = \sum_{i=1}^{n} r_i E(Y_i) = \sum_{i=1}^{n} r_i(\alpha + \beta x_i) = \alpha \sum_{i=1}^{n} r_i + \beta \sum_{i=1}^{n} r_i x_i,$$

donde los diferentes sumatorios tienen los siguientes valores

$$\sum_{i=1}^{n} r_i = \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} w_i \right) = 1 - \overline{x} \sum_{i=1}^{n} w_i = 1$$

$$\sum_{i=1}^{n} r_i x_i = \sum_{i=1}^{n} \left(\frac{x_i}{n} - \overline{x} w_i x_i \right) = \frac{\sum_{i=1}^{n} x_i}{n} - \overline{x} \sum_{i=1}^{n} w_i x_i = \overline{x} - \overline{x} = 0.$$

Por lo tanto

$$\mu_A = E(A) = \alpha \tag{18.8}$$

y A es un estimador insesgado del parámetro poblacional α. Respecto a su varianza

$$\sigma_A^2 = Var(A) = \sum_{i=1}^{n} r_i^2 \sigma_{Y_i}^2 = \sigma^2 \sum_{i=1}^{n} r_i^2 = \sigma^2 \sum_{i=1}^{n} \left(\frac{1}{n} - \overline{x} w_i \right)^2$$

$$\Rightarrow \qquad \sigma_A^2 = \sigma^2 \left(\sum_{i=1}^{n} \frac{1}{n^2} + \overline{x}^2 \sum_{i=1}^{n} w_i^2 - \frac{2\overline{x}}{n} \sum_{i=1}^{n} w_i \right) = \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2} \right).$$

Esta expresión también tiene un significado claro. El error en la ordenada en el origen es suma de dos términos: el primero es el error en la ordenada media $\overline{Y}$ y el segundo tiene en cuenta que el error será mayor cuanto más alejados estén los datos del origen $x = 0$. Es fácil comprobar que la expresión anterior es equivalente a la siguiente

$$\sigma_A^2 = \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \overline{x})^2}. \tag{18.9}$$

En definitiva el estimador A de la ordenada en el origen sigue una distribución normal del tipo

$$N\left(\alpha, \sigma \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} \right). \tag{18.10}$$

Para realizar contrastes sobre los coeficientes de la recta usando las expresiones anteriores es necesario conocer la varianza σ^2, es decir, la varianza de cada una de las Y_i, conocida como varianza del error del modelo. Se puede demostrar que, como cabría esperarse, la varianza residual de la muestra, definida en (17.17) como

$$s_r^2 = \frac{\sum_{i=1}^{n} (y_i - y^*)^2}{n-2} = \frac{\sum_{i=1}^{n} (y_i - a - b x_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

es un estimador insesgado de σ^2. Nótese que mientras que s_r^2 mide las desviaciones de los datos respecto a la recta ajustada ($y = a + bx$), σ^2 mide las desviaciones de cada Y_i respecto a su valor medio $\mu_{Y|x_i}$, lo que es equivalente a las desviaciones respecto a la recta poblacional ($y = \alpha + \beta x$) (puesto que los valores medios se han de situar sobre ésta). Por tanto, es lógico que la varianza residual sea el estimador insesgado de σ^2. Es decir

$$E(s_r^2) = \sigma^2. \tag{18.11}$$

De forma similar a lo que ocurría con la varianza muestral y poblacional de una variable, lo anterior

implica que se puede construir la siguiente variable χ^2

$$\chi^2_{n-2} = (n-2)\frac{s_r^2}{\sigma^2}, \tag{18.12}$$

lo cual puede servir para construir intervalos de confianza para la varianza σ^2.

18.2.2. Intervalos de confianza y contraste de hipótesis

Las propiedades anteriores de las distribuciones muestrales para los coeficientes de la recta de regresión pueden usarse para construir intervalos de confianza sobre los parámetros poblacionales de la recta. En el caso del coeficiente de regresión es claro que se puede construir la siguiente variable normal tipificada

$$z = \frac{b-\beta}{\sigma/(\sqrt{n-1}s_x)}$$

Entonces, por la definición de la distribución t de Student, el siguiente estadístico

$$t_{n-2} = \frac{z}{\sqrt{\chi^2_{n-2}/(n-2)}} = \frac{\frac{b-\beta}{\sigma/(\sqrt{n-1}s_x)}}{\sqrt{\frac{(n-2)s_r^2}{\sigma^2}\Big/(n-2)}} = \frac{b-\beta}{s_r/(\sqrt{n-1}s_x)} \tag{18.13}$$

seguirá una distribución t con $n-2$ grados de libertad. Por tanto, para un nivel de confianza $1-\alpha$ se puede expresar

$$P\left(-t_{\alpha/2,n-2} < \frac{b-\beta}{s_r/(\sqrt{n-1}s_x)} < t_{\alpha/2,n-2}\right) = 1-\alpha,$$

que conduce al siguiente intervalo de confianza para el parámetro poblacional β

$$P(\left(b - t_{\alpha/2,n-2}\frac{s_r}{\sqrt{n-1}s_x} < \beta < b + t_{\alpha/2,n-2}\frac{s_r}{\sqrt{n-1}s_x}\right) = 1-\alpha \tag{18.14}$$

$$I = \left[b \pm t_{\alpha/2,n-2}\frac{s_r}{\sqrt{n-1}s_x}\right]. \tag{18.15}$$

Por otra parte, lo anterior se puede usar para realizar constrastes de hipótesis sobre β. Si suponemos un contraste bilateral del tipo

$$\text{Hipótesis :} \quad \begin{cases} H_0: & \beta = \beta_0 \\ H_1: & \beta \neq \beta_0 \end{cases}$$

la hipótesis nula H_0 se aceptará, con un nivel de significación α, cuando

$$\frac{|b-\beta_0|}{s_r/(\sqrt{n-1}s_x)} \leq t_{\alpha/2,n-2}. \tag{18.16}$$

De la misma forma, a partir de la distribución muestral para la ordenada en el origen A, el estadístico

$$t = \frac{a-\alpha}{s_r\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}} \tag{18.17}$$

seguirá una distribución t con $n-2$ grados de libertad. Esto conduce al siguiente intervalo de confianza para α

$$P\left(a - t_{\alpha/2,n-2}s_r\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} < \alpha < a + t_{\alpha/2,n-2}s_r\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}\right) =$$

$$= 1-\alpha \tag{18.18}$$

$$I = \left[a \pm t_{\alpha/2,n-2} s_r \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}} \right], \tag{18.19}$$

para el que se puede dar también la siguiente expresión alternativa

$$I = \left[a \pm t_{\alpha/2,n-2} s_r \sqrt{\frac{\sum_{i=1}^n x_i^2}{n\sum_{i=1}^n (x_i - \overline{x})^2}} \right].$$

Esto implica que, en el contraste de hipótesis bilateral siguiente

$$\text{Hipótesis :} \quad \begin{cases} H_0: & \alpha = \alpha_0 \\ H_1: & \alpha \neq \alpha_0 \end{cases}$$

la hipótesis nula H_0 se acepta, a un nivel de significación α, cuando

$$\frac{|a - \alpha_0|}{s_r \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}} \leq t_{\alpha/2,n-2}. \tag{18.20}$$

Nótese que en estas expresiones el símbolo "α" se utiliza con dos sentidos diferentes: nivel de significación y ordenada en el origen de la recta poblacional.

18.3. Predicción

Aunque los intervalos de confianza para los parámetros poblacionales de la recta son importantes, en general el científico necesita calcular el intervalo de confianza para futuras evaluaciones de la recta, obtenidas para un valor concreto de la abscisa x_0, o lo que normalmente se conoce como intervalo de confianza para la predicción. En general dicho valor x_0 no coincidirá con ninguno de los valores x_i utilizados en para el cálculo de la recta de regresión. Vamos a distinguir dos situaciones diferentes.

18.3.1. Intervalo de confianza para el valor medio $\mu_{Y|x_0}$ en $x = x_0$

Para su cálculo utilizamos como estimador

$$Y_0^* = A + Bx_0 = (\overline{Y} - B\overline{x}) + Bx_0 = \overline{Y} + B(x_0 - \overline{x}),$$

que es un estadístico que tendrá una determinada distribución muestral. En concreto

$$\mu_{Y_0^*} = E(Y_0^*) = E(A + Bx_0) = \alpha + \beta x_0 = \mu_{Y|x_0}$$

$$\sigma^2_{Y_0^*} = \sigma^2_{A+Bx_0} = \sigma^2_{\overline{Y}+B(x_o-\overline{x})} = \sigma^2_{\overline{Y}} + (x_0 - \overline{x})^2\sigma_B^2 + 2(x_o - \overline{x})\text{cov}(\overline{Y}, B) =$$

$$= \frac{\sigma^2}{n} + (x_o - \overline{x})^2 \frac{\sigma^2}{(n-1)s_x^2} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2} \right)$$

El siguiente estadístico

$$t_{n-2} = \frac{Y_0^* - \mu_{Y|x_0}}{s_r \sqrt{\frac{1}{n} + \frac{(x_0-\overline{x})^2}{(n-1)s_x^2}}} \tag{18.21}$$

sigue una distribución t de Student con $n-2$ grados de libertad. El intervalo de confianza buscado vendrá dado por

$$I = \left[y_0^* \pm t_{\alpha/2,n-2} s_r \sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2}} \right] \tag{18.22}$$

18.3.2. Intervalo de confianza para un valor individual y_0 en $x = x_0$

En este caso estamos interesados en el intervalo de confianza para un único valor individual y_0. Sabemos que el valor real vendrá dado por

$$Y_0 = \alpha + \beta x_0 + \varepsilon_0$$

El estadístico $Y_0^* - Y_0$ seguirá entonces una determinada distribución muestral. En concreto

$$\mu_{Y_0^* - Y_0} = E(Y_0^* - Y_0) = E(A + Bx_0 - \alpha - \beta x_0 - \varepsilon_0) = 0$$

$$\sigma^2_{Y_0^* - Y_0} = \sigma^2_{Y_0^*} + \sigma^2_{Y_0} = \sigma^2_{Y_0^*} + \sigma^2_{\varepsilon_0} = \sigma^2 \left(\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2} \right) + \sigma^2 = \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2} \right)$$

El siguiente estadístico

$$t_{n-2} = \frac{Y_0^* - Y_0}{s_r \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2}}} \tag{18.23}$$

sigue una distribución t de Student con $n-2$ grados de libertad. Por tanto, el intervalo de confianza para Y_0 puede finalmente calcularse mediante

$$I = \left[y_0^* \pm t_{\alpha/2,n-2} s_r \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{(n-1)s_x^2}} \right] \tag{18.24}$$

18.4. Correlación

Hasta ahora hemos supuesto que la variable de regresión independiente x es una variable física o científica, pero no una variable aleatoria. De hecho, en este contexto, x frecuentemente recibe el nombre de **variable matemática**, la cual, en el proceso de muestreo, se mide con un error despreciable. Sin embargo, resulta mucho más realista suponer que tanto X como Y son variables aleatorias.

El **análisis de correlación** intenta cuantificar las relaciones entre dos variables por medio de un simple número que recibe el nombre de **coeficiente de correlación**.

Para ello vamos a considerar que el conjunto de medidas (x_i, y_i), con $i = 1, \ldots, n$, son observaciones de una población que tiene una función de densidad conjunta $f(x, y)$. No es difícil mostrar que en ese caso (ver libro de Walpole y Myers, Sección 9.10) la función de densidad conjunta de X e Y puede escribirse como una distribución normal bivariada

$$f(x,y) = \frac{1}{2\,\pi\,\sigma_X\,\sigma_Y\,\sqrt{1-\rho^2}} \times$$

$$\exp\left\{ \frac{-1}{2(1-\rho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X} \right)^2 - 2\rho \left(\frac{x-\mu_X}{\sigma_X} \right) \left(\frac{y-\mu_Y}{\sigma_Y} \right) + \left(\frac{y-\mu_Y}{\sigma_Y} \right)^2 \right] \right\}, \tag{18.25}$$

donde la constante ρ, definida como

$$\rho^2 = \beta^2\,\frac{\sigma_X^2}{\sigma_Y^2} = \frac{\sigma_{XY}^4}{\sigma_X^2\,\sigma_Y^2}$$

$$\Rightarrow \rho = \frac{\sigma_{XY}^2}{\sigma_X\ \sigma_Y} \tag{18.26}$$

recibe el nombre de **coeficiente de correlación poblacional** y juega un papel importante en muchos problemas de análisis de datos de dos variables.

De entrada, si hacemos $\rho = 0$ en (18.25) obtenemos

$$f(x,y) = \frac{1}{2\ \pi\ \sigma_X\ \sigma_Y}\ \exp\left\{-\frac{1}{2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} =$$

$$= \frac{1}{\sqrt{2\ \pi}\ \sigma_X}\ \exp\left\{-\frac{1}{2}\left(\frac{x-\mu_X}{\sigma_X}\right)^2\right\} \times \frac{1}{\sqrt{2\ \pi}\ \sigma_Y}\ \exp\left\{-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right\} =$$

$$f(x)\ f(y),$$

es decir, la función de distribución conjunta se puede expresar como producto de dos funciones independientes de X e Y. En otras palabras, si $\rho = 0$ las variables aleatorias X e Y son independientes. Por otra parte, si $\rho \neq 0$, no podemos separar las dos funciones y las variables no serán independientes.

Por otro lado, recordando que $\rho^2 = \beta^2\ \sigma_X^2/\sigma_Y^2$, vemos que estudiar la presencia de correlación se convertirá en estudiar si $\rho \neq 0$ o si $\beta \neq 0$. Dicho de otra forma

$$\text{No correlación} \qquad \Longleftrightarrow \qquad \rho = 0 \qquad \Longleftrightarrow \qquad \beta = 0$$

Finalmente estudiemos los contrastes para $\rho = 0$ y $\rho = \rho_0$:

- Contraste de la hipótesis $\rho = 0$

$$\text{Hipótesis :} \qquad \begin{cases} H_0: & \rho = 0 \\ H_1: & \rho \neq 0 \end{cases}$$

$$\beta = 0 \quad \rightarrow \quad t = \frac{b}{s_r/(\sqrt{n-1}s_x)} = \frac{rs_y/s_x}{s_r/(\sqrt{n-1}s_x)} = \frac{rs_y}{s_r/\sqrt{n-1}} =$$

$$\frac{rs_y}{\sqrt{\frac{n-1}{n-2}}s_y\sqrt{1-r^2}/\sqrt{n-1}}$$

$$\Rightarrow t_{n-2} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Se acepta H_0 si

$$\frac{|r|\sqrt{n-2}}{\sqrt{1-r^2}} \leq t_{\alpha/2,n-2}. \tag{18.27}$$

El que un valor de r sea o no indicativo de correlación dependerá también del número de puntos. Si n es grande, será fácil rechazar H_0 y existirá correlación.

- Contraste de la hipótesis $\rho = \rho_0$

$$\text{Hipótesis :} \qquad \begin{cases} H_0: & \rho = \rho_0 \\ H_1: & \rho \neq \rho_0 \end{cases}$$

Se puede demostrar que si X e Y siguen una distribución normal bivariada, la cantidad

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \quad \text{es aprox. normal con} \quad \begin{cases} \mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \\ \sigma^2 = \frac{1}{n-3} \end{cases}$$

Es decir

$$Z = \frac{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) - \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right)}{\sqrt{\frac{1}{n-3}}} = \frac{\sqrt{n-3}}{2}\ln\left(\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right) \qquad \text{es } N(0,1).$$

Se acepta H_0 si

$$\frac{\sqrt{n-3}}{2}\left|\ln\left(\frac{(1+r)(1-\rho_0)}{(1-r)(1+\rho_0)}\right)\right| \leq z_{\alpha/2}. \tag{18.28}$$

Vemos que si n crece es más fácil rechazar H_0. Por otro lado, si ρ es muy parecido a ρ_0, la cantidad dentro del logaritmo tiende a uno y el logaritmo a cero.

APÉNDICES

Capítulo 19

Apéndice A: Distribuciones de Probabilidad

En este apéndice aparecen tabuladas las siguientes funciones:

- Tabla I: probabilidades binomiales individuales.
- Tabla II: probabilidades binomiales acumuladas.
- Tabla III: probabilidades acumuladas de Poisson.
- Tabla IV: distribución normal tipificada.
- Tabla V: distribución χ^2 de Pearson.
- Tabla VI: distribución t de Student.
- Tabla VII: distribución F de Fisher.

Los datos que aparecen en las tablas han sido calculados utilizando funciones de *Numerical Recipes in Fortran 77* (Press et al. 1992) y programas propios de los autores de este libro.

Tabla I
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

		p																					
n	x	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
2	0	0.980	0.902	0.810	0.722	0.640	0.562	0.490	0.423	0.360	0.303	0.250	0.202	0.160	0.123	0.090	0.062	0.040	0.022	0.010	0.003	0.0+	0
	1	0.020	0.095	0.180	0.255	0.320	0.375	0.420	0.455	0.480	0.495	0.500	0.495	0.480	0.455	0.420	0.375	0.320	0.255	0.180	0.095	0.020	1
	2	0.0+	0.003	0.010	0.023	0.040	0.062	0.090	0.122	0.160	0.202	0.250	0.303	0.360	0.422	0.490	0.562	0.640	0.723	0.810	0.902	0.980	2
3	0	0.970	0.857	0.729	0.614	0.512	0.422	0.343	0.275	0.216	0.166	0.125	0.091	0.064	0.043	0.027	0.016	0.008	0.003	0.001	0.0+	0.0+	0
	1	0.029	0.135	0.243	0.325	0.384	0.422	0.441	0.444	0.432	0.408	0.375	0.334	0.288	0.239	0.189	0.141	0.096	0.057	0.027	0.007	0.0+	1
	2	0.0+	0.007	0.027	0.057	0.096	0.141	0.189	0.239	0.288	0.334	0.375	0.408	0.432	0.444	0.441	0.422	0.384	0.325	0.243	0.135	0.029	2
	3	0.0+	0.0+	0.001	0.003	0.008	0.016	0.027	0.043	0.064	0.091	0.125	0.166	0.216	0.275	0.343	0.422	0.512	0.614	0.729	0.857	0.970	3
4	0	0.961	0.815	0.656	0.522	0.410	0.316	0.240	0.179	0.130	0.092	0.062	0.041	0.026	0.015	0.008	0.004	0.002	0.001	0.0+	0.0+	0.0+	0
	1	0.039	0.171	0.292	0.368	0.410	0.422	0.412	0.384	0.346	0.299	0.250	0.200	0.154	0.111	0.076	0.047	0.026	0.011	0.004	0.0+	0.0+	1
	2	0.001	0.014	0.049	0.098	0.154	0.211	0.265	0.311	0.346	0.368	0.375	0.368	0.346	0.311	0.265	0.211	0.154	0.098	0.049	0.014	0.001	2
	3	0.0+	0.0+	0.004	0.011	0.026	0.047	0.076	0.111	0.154	0.200	0.250	0.299	0.346	0.384	0.412	0.422	0.410	0.368	0.292	0.171	0.039	3
	4	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.015	0.026	0.041	0.062	0.092	0.130	0.179	0.240	0.316	0.410	0.522	0.656	0.815	0.961	4
5	0	0.951	0.774	0.590	0.444	0.328	0.237	0.168	0.116	0.078	0.050	0.031	0.018	0.010	0.005	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.048	0.204	0.328	0.392	0.410	0.396	0.360	0.312	0.259	0.206	0.156	0.113	0.077	0.049	0.028	0.015	0.006	0.002	0.0+	0.0+	0.0+	1
	2	0.001	0.021	0.073	0.138	0.205	0.264	0.309	0.336	0.346	0.337	0.312	0.276	0.230	0.181	0.132	0.088	0.051	0.024	0.008	0.001	0.0+	2
	3	0.0+	0.001	0.008	0.024	0.051	0.088	0.132	0.181	0.230	0.276	0.312	0.337	0.346	0.336	0.309	0.264	0.205	0.138	0.073	0.021	0.001	3
	4	0.0+	0.0+	0.0+	0.002	0.006	0.015	0.028	0.049	0.077	0.113	0.156	0.206	0.259	0.312	0.360	0.396	0.410	0.392	0.328	0.204	0.048	4
	5	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.010	0.018	0.031	0.050	0.078	0.116	0.168	0.237	0.328	0.444	0.590	0.774	0.951	5
6	0	0.941	0.735	0.531	0.377	0.262	0.178	0.118	0.075	0.047	0.028	0.016	0.008	0.004	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.057	0.232	0.354	0.399	0.393	0.356	0.303	0.244	0.187	0.136	0.094	0.061	0.037	0.020	0.010	0.004	0.002	0.0+	0.0+	0.0+	0.0+	1
	2	0.001	0.031	0.098	0.176	0.246	0.297	0.324	0.328	0.311	0.278	0.234	0.186	0.138	0.095	0.060	0.033	0.015	0.005	0.001	0.0+	0.0+	2
	3	0.0+	0.002	0.015	0.041	0.082	0.132	0.185	0.235	0.276	0.303	0.312	0.303	0.276	0.235	0.185	0.132	0.082	0.041	0.015	0.002	0.0+	3
	4	0.0+	0.0+	0.001	0.005	0.015	0.033	0.060	0.095	0.138	0.186	0.234	0.278	0.311	0.328	0.324	0.297	0.246	0.176	0.098	0.031	0.001	4
	5	0.0+	0.0+	0.0+	0.0+	0.002	0.004	0.010	0.020	0.037	0.061	0.094	0.136	0.187	0.244	0.303	0.356	0.393	0.399	0.354	0.232	0.057	5
	6	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.016	0.028	0.047	0.075	0.118	0.178	0.262	0.377	0.531	0.735	0.941	6
7	0	0.932	0.698	0.478	0.321	0.210	0.133	0.082	0.049	0.028	0.015	0.008	0.004	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.066	0.257	0.372	0.396	0.367	0.311	0.247	0.185	0.131	0.087	0.055	0.032	0.017	0.008	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.002	0.041	0.124	0.210	0.275	0.311	0.318	0.298	0.261	0.214	0.164	0.117	0.077	0.047	0.025	0.012	0.004	0.001	0.0+	0.0+	0.0+	2
	3	0.0+	0.004	0.023	0.062	0.115	0.173	0.227	0.268	0.290	0.292	0.273	0.239	0.194	0.144	0.097	0.058	0.029	0.011	0.003	0.0+	0.0+	3
	4	0.0+	0.0+	0.003	0.011	0.029	0.058	0.097	0.144	0.194	0.239	0.273	0.292	0.290	0.268	0.227	0.173	0.115	0.062	0.023	0.004	0.0+	4
	5	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.025	0.047	0.077	0.117	0.164	0.214	0.261	0.298	0.318	0.311	0.275	0.210	0.124	0.041	0.002	5
	6	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.008	0.017	0.032	0.055	0.087	0.131	0.185	0.247	0.311	0.367	0.396	0.372	0.257	0.066	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.015	0.028	0.049	0.082	0.133	0.210	0.321	0.478	0.698	0.932	7

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
8	0	0.923	0.663	0.430	0.272	0.168	0.100	0.058	0.032	0.017	0.008	0.004	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.075	0.279	0.383	0.385	0.336	0.267	0.198	0.137	0.090	0.055	0.031	0.016	0.008	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.003	0.051	0.149	0.238	0.294	0.311	0.296	0.259	0.209	0.157	0.109	0.070	0.041	0.022	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.005	0.033	0.084	0.147	0.208	0.254	0.279	0.279	0.257	0.219	0.172	0.124	0.081	0.047	0.023	0.009	0.003	0.0+	0.0+	0.0+	3
	4	0.0+	0.0+	0.005	0.018	0.046	0.087	0.136	0.188	0.232	0.263	0.273	0.263	0.232	0.188	0.136	0.087	0.046	0.018	0.005	0.0+	0.0+	4
	5	0.0+	0.0+	0.0+	0.003	0.009	0.023	0.047	0.081	0.124	0.172	0.219	0.257	0.279	0.279	0.254	0.208	0.147	0.084	0.033	0.005	0.0+	5
	6	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.022	0.041	0.070	0.109	0.157	0.209	0.259	0.296	0.311	0.294	0.238	0.149	0.051	0.003	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.016	0.031	0.055	0.090	0.137	0.198	0.267	0.336	0.385	0.383	0.279	0.075	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.017	0.032	0.058	0.100	0.168	0.272	0.430	0.663	0.923	8
9	0	0.914	0.630	0.387	0.232	0.134	0.075	0.040	0.021	0.010	0.005	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.083	0.299	0.387	0.368	0.302	0.225	0.156	0.100	0.060	0.034	0.018	0.008	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.003	0.063	0.172	0.260	0.302	0.300	0.267	0.216	0.161	0.111	0.070	0.041	0.021	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.008	0.045	0.107	0.176	0.234	0.267	0.272	0.251	0.212	0.164	0.116	0.074	0.042	0.021	0.009	0.003	0.001	0.0+	0.0+	0.0+	3
	4	0.0+	0.001	0.007	0.028	0.066	0.117	0.172	0.219	0.251	0.260	0.246	0.213	0.167	0.118	0.074	0.039	0.017	0.005	0.001	0.0+	0.0+	4
	5	0.0+	0.0+	0.001	0.005	0.017	0.039	0.074	0.118	0.167	0.213	0.246	0.260	0.251	0.219	0.172	0.117	0.066	0.028	0.007	0.001	0.0+	5
	6	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.021	0.042	0.074	0.116	0.164	0.212	0.251	0.272	0.267	0.234	0.176	0.107	0.045	0.008	0.0+	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.021	0.041	0.070	0.111	0.161	0.216	0.267	0.300	0.302	0.260	0.172	0.063	0.003	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.008	0.018	0.034	0.060	0.100	0.156	0.225	0.302	0.368	0.387	0.299	0.083	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.010	0.021	0.040	0.075	0.134	0.232	0.387	0.630	0.914	9
10	0	0.904	0.599	0.349	0.197	0.107	0.056	0.028	0.013	0.006	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.091	0.315	0.387	0.347	0.268	0.188	0.121	0.072	0.040	0.021	0.010	0.004	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.004	0.075	0.194	0.276	0.302	0.282	0.233	0.176	0.121	0.076	0.044	0.023	0.011	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.010	0.057	0.130	0.201	0.250	0.267	0.252	0.215	0.166	0.117	0.075	0.042	0.021	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.001	0.011	0.040	0.088	0.146	0.200	0.238	0.251	0.238	0.205	0.160	0.111	0.069	0.037	0.016	0.006	0.001	0.0+	0.0+	0.0+	4
	5	0.0+	0.0+	0.001	0.008	0.026	0.058	0.103	0.154	0.201	0.234	0.246	0.234	0.201	0.154	0.103	0.058	0.026	0.008	0.001	0.0+	0.0+	5
	6	0.0+	0.0+	0.0+	0.001	0.006	0.016	0.037	0.069	0.111	0.160	0.205	0.238	0.251	0.238	0.200	0.146	0.088	0.040	0.011	0.001	0.0+	6
	7	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.021	0.042	0.075	0.117	0.166	0.215	0.252	0.267	0.250	0.201	0.130	0.057	0.010	0.0+	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.023	0.044	0.076	0.121	0.176	0.233	0.282	0.302	0.276	0.194	0.075	0.004	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.010	0.021	0.040	0.072	0.121	0.188	0.268	0.347	0.387	0.315	0.091	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.006	0.013	0.028	0.056	0.107	0.197	0.349	0.599	0.904	10
11	0	0.895	0.569	0.314	0.167	0.086	0.042	0.020	0.009	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.099	0.329	0.384	0.325	0.236	0.155	0.093	0.052	0.027	0.013	0.005	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.005	0.087	0.213	0.287	0.295	0.258	0.200	0.140	0.089	0.051	0.027	0.013	0.005	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.014	0.071	0.152	0.221	0.258	0.257	0.225	0.177	0.126	0.081	0.046	0.023	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.001	0.016	0.054	0.111	0.172	0.220	0.243	0.236	0.206	0.161	0.113	0.070	0.038	0.017	0.006	0.002	0.0+	0.0+	0.0+	0.0+	4

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
11	5	0.0+	0.0+	0.002	0.013	0.039	0.080	0.132	0.183	0.221	0.236	0.226	0.193	0.147	0.099	0.057	0.027	0.010	0.002	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.0+	0.002	0.010	0.027	0.057	0.099	0.147	0.193	0.226	0.236	0.221	0.183	0.132	0.080	0.039	0.013	0.002	0.0+	0.0+	6
	7	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.017	0.038	0.070	0.113	0.161	0.206	0.236	0.243	0.220	0.172	0.111	0.054	0.016	0.001	0.0+	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.023	0.046	0.081	0.126	0.177	0.225	0.257	0.258	0.221	0.152	0.071	0.014	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.013	0.027	0.051	0.089	0.140	0.200	0.258	0.295	0.287	0.213	0.087	0.005	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.013	0.027	0.052	0.093	0.155	0.236	0.325	0.384	0.329	0.099	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.009	0.020	0.042	0.086	0.167	0.314	0.569	0.895	11
12	0	0.886	0.540	0.282	0.142	0.069	0.032	0.014	0.006	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.107	0.341	0.377	0.301	0.206	0.127	0.071	0.037	0.017	0.008	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.006	0.099	0.230	0.292	0.283	0.232	0.168	0.109	0.064	0.034	0.016	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.017	0.085	0.172	0.236	0.258	0.240	0.195	0.142	0.092	0.054	0.028	0.012	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.002	0.021	0.068	0.133	0.194	0.231	0.237	0.213	0.170	0.121	0.076	0.042	0.020	0.008	0.002	0.001	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.0+	0.004	0.019	0.053	0.103	0.158	0.204	0.227	0.222	0.193	0.149	0.101	0.059	0.029	0.011	0.003	0.001	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.0+	0.004	0.016	0.040	0.079	0.128	0.177	0.212	0.226	0.212	0.177	0.128	0.079	0.040	0.016	0.004	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.0+	0.001	0.003	0.011	0.029	0.059	0.101	0.149	0.193	0.222	0.227	0.204	0.158	0.103	0.053	0.019	0.004	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.020	0.042	0.076	0.121	0.170	0.213	0.237	0.231	0.194	0.133	0.068	0.021	0.002	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.012	0.028	0.054	0.092	0.142	0.195	0.240	0.258	0.236	0.172	0.085	0.017	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.016	0.034	0.064	0.109	0.168	0.232	0.283	0.292	0.230	0.099	0.006	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.017	0.037	0.071	0.127	0.206	0.301	0.377	0.341	0.107	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.014	0.032	0.069	0.142	0.282	0.540	0.886	12
13	0	0.878	0.513	0.254	0.121	0.055	0.024	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.115	0.351	0.367	0.277	0.179	0.103	0.054	0.026	0.011	0.004	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.007	0.111	0.245	0.294	0.268	0.206	0.139	0.084	0.045	0.022	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.021	0.100	0.190	0.246	0.252	0.218	0.165	0.111	0.066	0.035	0.016	0.006	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.003	0.028	0.084	0.154	0.210	0.234	0.222	0.184	0.135	0.087	0.050	0.024	0.010	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.0+	0.006	0.027	0.069	0.126	0.180	0.215	0.221	0.199	0.157	0.109	0.066	0.034	0.014	0.005	0.001	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.001	0.006	0.023	0.056	0.103	0.155	0.197	0.217	0.209	0.177	0.131	0.083	0.044	0.019	0.006	0.001	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.0+	0.001	0.006	0.019	0.044	0.083	0.131	0.177	0.209	0.217	0.197	0.155	0.103	0.056	0.023	0.006	0.001	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.034	0.066	0.109	0.157	0.199	0.221	0.215	0.180	0.126	0.069	0.027	0.006	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.024	0.050	0.087	0.135	0.184	0.222	0.234	0.210	0.154	0.084	0.028	0.003	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.016	0.035	0.066	0.111	0.165	0.218	0.252	0.246	0.190	0.100	0.021	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.022	0.045	0.084	0.139	0.206	0.268	0.294	0.245	0.111	0.007	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.004	0.011	0.026	0.054	0.103	0.179	0.277	0.367	0.351	0.115	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.024	0.055	0.121	0.254	0.513	0.878	13
14	0	0.869	0.488	0.229	0.103	0.044	0.018	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.123	0.359	0.356	0.254	0.154	0.083	0.041	0.018	0.007	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
14	2	0.008	0.123	0.257	0.291	0.250	0.180	0.113	0.063	0.032	0.014	0.006	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.026	0.114	0.206	0.250	0.240	0.194	0.137	0.085	0.046	0.022	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.004	0.035	0.100	0.172	0.220	0.229	0.202	0.155	0.104	0.061	0.031	0.014	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.0+	0.008	0.035	0.086	0.147	0.196	0.218	0.207	0.170	0.122	0.076	0.041	0.018	0.007	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.001	0.009	0.032	0.073	0.126	0.176	0.207	0.209	0.183	0.140	0.092	0.051	0.023	0.008	0.002	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.0+	0.002	0.009	0.028	0.062	0.108	0.157	0.195	0.209	0.195	0.157	0.108	0.062	0.028	0.009	0.002	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.023	0.051	0.092	0.140	0.183	0.209	0.207	0.176	0.126	0.073	0.032	0.009	0.001	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.007	0.018	0.041	0.076	0.122	0.170	0.207	0.218	0.196	0.147	0.086	0.035	0.008	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.031	0.061	0.104	0.155	0.202	0.229	0.220	0.172	0.100	0.035	0.004	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.022	0.046	0.085	0.137	0.194	0.240	0.250	0.206	0.114	0.026	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.014	0.032	0.063	0.113	0.180	0.250	0.291	0.257	0.123	0.008	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.007	0.018	0.041	0.083	0.154	0.254	0.356	0.359	0.123	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.018	0.044	0.103	0.229	0.488	0.869	14
15	0	0.860	0.463	0.206	0.087	0.035	0.013	0.005	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.130	0.366	0.343	0.231	0.132	0.067	0.031	0.013	0.005	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.009	0.135	0.267	0.286	0.231	0.156	0.092	0.048	0.022	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.031	0.129	0.218	0.250	0.225	0.170	0.111	0.063	0.032	0.014	0.005	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.005	0.043	0.116	0.188	0.225	0.219	0.179	0.127	0.078	0.042	0.019	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.001	0.010	0.045	0.103	0.165	0.206	0.212	0.186	0.140	0.092	0.051	0.024	0.010	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.002	0.013	0.043	0.092	0.147	0.191	0.207	0.191	0.153	0.105	0.061	0.030	0.012	0.003	0.001	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.0+	0.003	0.014	0.039	0.081	0.132	0.177	0.201	0.196	0.165	0.118	0.071	0.035	0.013	0.003	0.001	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.001	0.003	0.013	0.035	0.071	0.118	0.165	0.196	0.201	0.177	0.132	0.081	0.039	0.014	0.003	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.030	0.061	0.105	0.153	0.191	0.207	0.191	0.147	0.092	0.043	0.013	0.002	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.024	0.051	0.092	0.140	0.186	0.212	0.206	0.165	0.103	0.045	0.010	0.001	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.019	0.042	0.078	0.127	0.179	0.219	0.225	0.188	0.116	0.043	0.005	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.014	0.032	0.063	0.111	0.170	0.225	0.250	0.218	0.129	0.031	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.022	0.048	0.092	0.156	0.231	0.286	0.267	0.135	0.009	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.013	0.031	0.067	0.132	0.231	0.343	0.366	0.130	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.013	0.035	0.087	0.206	0.463	0.860	15
16	0	0.851	0.440	0.185	0.074	0.028	0.010	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.138	0.371	0.329	0.210	0.113	0.053	0.023	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.010	0.146	0.275	0.277	0.211	0.134	0.073	0.035	0.015	0.006	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.0+	0.036	0.142	0.229	0.246	0.208	0.146	0.089	0.047	0.022	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.006	0.051	0.131	0.200	0.225	0.204	0.155	0.101	0.057	0.028	0.011	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.001	0.014	0.056	0.120	0.180	0.210	0.201	0.162	0.112	0.067	0.034	0.014	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.003	0.018	0.055	0.110	0.165	0.198	0.198	0.168	0.122	0.075	0.039	0.017	0.006	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	6

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
16	7	0.0+	0.0+	0.0+	0.005	0.020	0.052	0.101	0.152	0.189	0.197	0.175	0.132	0.084	0.044	0.019	0.006	0.001	0.0+	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.001	0.006	0.020	0.049	0.092	0.142	0.181	0.196	0.181	0.142	0.092	0.049	0.020	0.006	0.001	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.019	0.044	0.084	0.132	0.175	0.197	0.189	0.152	0.101	0.052	0.020	0.005	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.017	0.039	0.075	0.122	0.168	0.198	0.198	0.165	0.110	0.055	0.018	0.003	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.034	0.067	0.112	0.162	0.201	0.210	0.180	0.120	0.056	0.014	0.001	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.028	0.057	0.101	0.155	0.204	0.225	0.200	0.131	0.051	0.006	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.022	0.047	0.089	0.146	0.208	0.246	0.229	0.142	0.036	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.015	0.035	0.073	0.134	0.211	0.277	0.275	0.146	0.010	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.023	0.053	0.113	0.210	0.329	0.371	0.138	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.028	0.074	0.185	0.440	0.851	16
17	0	0.843	0.418	0.167	0.063	0.023	0.008	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.145	0.374	0.315	0.189	0.096	0.043	0.017	0.006	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.012	0.158	0.280	0.267	0.191	0.114	0.058	0.026	0.010	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.001	0.041	0.156	0.236	0.239	0.189	0.125	0.070	0.034	0.014	0.005	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.008	0.060	0.146	0.209	0.221	0.187	0.132	0.080	0.041	0.018	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.001	0.017	0.067	0.136	0.191	0.208	0.185	0.138	0.087	0.047	0.021	0.008	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.004	0.024	0.068	0.128	0.178	0.199	0.184	0.143	0.094	0.052	0.024	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.001	0.007	0.027	0.067	0.120	0.168	0.193	0.184	0.148	0.101	0.057	0.026	0.009	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.001	0.008	0.028	0.064	0.113	0.161	0.188	0.185	0.154	0.107	0.061	0.028	0.009	0.002	0.0+	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.002	0.009	0.028	0.061	0.107	0.154	0.185	0.188	0.161	0.113	0.064	0.028	0.008	0.001	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.009	0.026	0.057	0.101	0.148	0.184	0.193	0.168	0.120	0.067	0.027	0.007	0.001	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.024	0.052	0.094	0.143	0.184	0.199	0.178	0.128	0.068	0.024	0.004	0.0+	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.021	0.047	0.087	0.138	0.185	0.208	0.191	0.136	0.067	0.017	0.001	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.018	0.041	0.080	0.132	0.187	0.221	0.209	0.146	0.060	0.008	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.014	0.034	0.070	0.125	0.189	0.239	0.236	0.156	0.041	0.001	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.026	0.058	0.114	0.191	0.267	0.280	0.158	0.012	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.017	0.043	0.096	0.189	0.315	0.374	0.145	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.023	0.063	0.167	0.418	0.843	17
18	0	0.835	0.397	0.150	0.054	0.018	0.006	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.152	0.376	0.300	0.170	0.081	0.034	0.013	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.013	0.168	0.284	0.256	0.172	0.096	0.046	0.019	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.001	0.047	0.168	0.241	0.230	0.170	0.105	0.055	0.025	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.009	0.070	0.159	0.215	0.213	0.168	0.110	0.061	0.029	0.012	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.001	0.022	0.079	0.151	0.199	0.202	0.166	0.115	0.067	0.033	0.013	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.005	0.030	0.082	0.144	0.187	0.194	0.166	0.118	0.071	0.035	0.015	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.001	0.009	0.035	0.082	0.138	0.179	0.189	0.166	0.121	0.074	0.037	0.015	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	7

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
18	8	0.0+	0.0+	0.0+	0.002	0.012	0.038	0.081	0.133	0.173	0.186	0.167	0.125	0.077	0.038	0.015	0.004	0.001	0.0+	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.0+	0.003	0.014	0.039	0.079	0.128	0.169	0.185	0.169	0.128	0.079	0.039	0.014	0.003	0.0+	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.015	0.038	0.077	0.125	0.167	0.186	0.173	0.133	0.081	0.038	0.012	0.002	0.0+	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.037	0.074	0.121	0.166	0.189	0.179	0.138	0.082	0.035	0.009	0.001	0.0+	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.035	0.071	0.118	0.166	0.194	0.187	0.144	0.082	0.030	0.005	0.0+	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.033	0.067	0.115	0.166	0.202	0.199	0.151	0.079	0.022	0.001	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.029	0.061	0.110	0.168	0.213	0.215	0.159	0.070	0.009	0.0+	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.025	0.055	0.105	0.170	0.230	0.241	0.168	0.047	0.001	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.019	0.046	0.096	0.172	0.256	0.284	0.168	0.013	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.034	0.081	0.170	0.300	0.376	0.152	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.018	0.054	0.150	0.397	0.835	18
19	0	0.826	0.377	0.135	0.046	0.014	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.159	0.377	0.285	0.153	0.068	0.027	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.014	0.179	0.285	0.243	0.154	0.080	0.036	0.014	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.001	0.053	0.180	0.243	0.218	0.152	0.087	0.042	0.017	0.006	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.011	0.080	0.171	0.218	0.202	0.149	0.091	0.047	0.020	0.007	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.002	0.027	0.091	0.164	0.202	0.192	0.147	0.093	0.050	0.022	0.008	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.007	0.037	0.095	0.157	0.192	0.184	0.145	0.095	0.052	0.023	0.008	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.001	0.012	0.044	0.097	0.153	0.184	0.180	0.144	0.096	0.053	0.024	0.008	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.003	0.017	0.049	0.098	0.149	0.180	0.177	0.144	0.097	0.053	0.023	0.008	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.001	0.005	0.020	0.051	0.098	0.146	0.177	0.176	0.145	0.098	0.053	0.022	0.007	0.001	0.0+	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.001	0.007	0.022	0.053	0.098	0.145	0.176	0.177	0.146	0.098	0.051	0.020	0.005	0.001	0.0+	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.023	0.053	0.097	0.144	0.177	0.180	0.149	0.098	0.049	0.017	0.003	0.0+	0.0+	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.024	0.053	0.096	0.144	0.180	0.184	0.153	0.097	0.044	0.012	0.001	0.0+	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.023	0.052	0.095	0.145	0.184	0.192	0.157	0.095	0.037	0.007	0.0+	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.022	0.050	0.093	0.147	0.192	0.202	0.164	0.091	0.027	0.002	0.0+	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.020	0.047	0.091	0.149	0.202	0.218	0.171	0.080	0.011	0.0+	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.017	0.042	0.087	0.152	0.218	0.243	0.180	0.053	0.001	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.036	0.080	0.154	0.243	0.285	0.179	0.014	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.027	0.068	0.153	0.285	0.377	0.159	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.014	0.046	0.135	0.377	0.826	19
20	0	0.818	0.358	0.122	0.039	0.012	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.165	0.377	0.270	0.137	0.058	0.021	0.007	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.016	0.189	0.285	0.229	0.137	0.067	0.028	0.010	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.001	0.060	0.190	0.243	0.205	0.134	0.072	0.032	0.012	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.013	0.090	0.182	0.218	0.190	0.130	0.074	0.035	0.014	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4

Tabla I (*Continuación*)
PROBABILIDADES BINOMIALES INDIVIDUALES

$$b(x;n,p) = \binom{n}{x} p^x q^{n-x}$$

n	x	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	x
20	5	0.0+	0.002	0.032	0.103	0.175	0.202	0.179	0.127	0.075	0.036	0.015	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.009	0.045	0.109	0.169	0.192	0.171	0.124	0.075	0.037	0.015	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.002	0.016	0.055	0.112	0.164	0.184	0.166	0.122	0.074	0.037	0.015	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.0+	0.005	0.022	0.061	0.114	0.161	0.180	0.162	0.120	0.073	0.035	0.014	0.004	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.001	0.007	0.027	0.065	0.116	0.160	0.177	0.160	0.119	0.071	0.034	0.012	0.003	0.0+	0.0+	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.002	0.010	0.031	0.069	0.117	0.159	0.176	0.159	0.117	0.069	0.031	0.010	0.002	0.0+	0.0+	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.003	0.012	0.034	0.071	0.119	0.160	0.177	0.160	0.116	0.065	0.027	0.007	0.001	0.0+	0.0+	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.014	0.035	0.073	0.120	0.162	0.180	0.161	0.114	0.061	0.022	0.005	0.0+	0.0+	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.015	0.037	0.074	0.122	0.166	0.184	0.164	0.112	0.055	0.016	0.002	0.0+	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.037	0.075	0.124	0.171	0.192	0.169	0.109	0.045	0.009	0.0+	0.0+	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.036	0.075	0.127	0.179	0.202	0.175	0.103	0.032	0.002	0.0+	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.035	0.074	0.130	0.190	0.218	0.182	0.090	0.013	0.0+	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.032	0.072	0.134	0.205	0.243	0.190	0.060	0.001	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.028	0.067	0.137	0.229	0.285	0.189	0.016	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.007	0.021	0.058	0.137	0.270	0.377	0.165	19
	20	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.039	0.122	0.358	0.818	20
21	0	0.810	0.341	0.109	0.033	0.009	0.002	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0
	1	0.172	0.376	0.255	0.122	0.048	0.017	0.005	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	1
	2	0.017	0.198	0.284	0.215	0.121	0.055	0.022	0.007	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	2
	3	0.001	0.066	0.200	0.241	0.192	0.117	0.058	0.024	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	3
	4	0.0+	0.016	0.100	0.191	0.216	0.176	0.113	0.059	0.026	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	4
	5	0.0+	0.003	0.038	0.115	0.183	0.199	0.164	0.109	0.059	0.026	0.010	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	5
	6	0.0+	0.0+	0.011	0.054	0.122	0.177	0.188	0.156	0.105	0.057	0.026	0.009	0.003	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	6
	7	0.0+	0.0+	0.003	0.020	0.065	0.126	0.172	0.180	0.149	0.101	0.055	0.025	0.009	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	7
	8	0.0+	0.0+	0.001	0.006	0.029	0.074	0.129	0.169	0.174	0.144	0.097	0.053	0.023	0.008	0.002	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	8
	9	0.0+	0.0+	0.0+	0.002	0.010	0.036	0.080	0.132	0.168	0.170	0.140	0.093	0.050	0.021	0.006	0.001	0.0+	0.0+	0.0+	0.0+	0.0+	9
	10	0.0+	0.0+	0.0+	0.0+	0.003	0.014	0.041	0.085	0.134	0.167	0.168	0.137	0.089	0.046	0.018	0.005	0.001	0.0+	0.0+	0.0+	0.0+	10
	11	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.018	0.046	0.089	0.137	0.168	0.167	0.134	0.085	0.041	0.014	0.003	0.0+	0.0+	0.0+	0.0+	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.021	0.050	0.093	0.140	0.170	0.168	0.132	0.080	0.036	0.010	0.002	0.0+	0.0+	0.0+	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.023	0.053	0.097	0.144	0.174	0.169	0.129	0.074	0.029	0.006	0.001	0.0+	0.0+	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.009	0.025	0.055	0.101	0.149	0.180	0.172	0.126	0.065	0.020	0.003	0.0+	0.0+	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.026	0.057	0.105	0.156	0.188	0.177	0.122	0.054	0.011	0.0+	0.0+	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.026	0.059	0.109	0.164	0.199	0.183	0.115	0.038	0.003	0.0+	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.026	0.059	0.113	0.176	0.216	0.191	0.100	0.016	0.0+	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.024	0.058	0.117	0.192	0.241	0.200	0.066	0.001	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.007	0.022	0.055	0.121	0.215	0.284	0.198	0.017	19
	20	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.017	0.048	0.122	0.255	0.376	0.172	20
	21	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.009	0.033	0.109	0.341	0.810	21

Tabla II
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
2	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.020	0.098	0.190	0.278	0.360	0.438	0.510	0.577	0.640	0.697	0.750	0.798	0.840	0.877	0.910	0.938	0.960	0.978	0.990	0.997	1−	1
	2	0.0+	0.003	0.010	0.023	0.040	0.062	0.090	0.122	0.160	0.202	0.250	0.303	0.360	0.422	0.490	0.562	0.640	0.723	0.810	0.902	0.980	2
3	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.030	0.143	0.271	0.386	0.488	0.578	0.657	0.725	0.784	0.834	0.875	0.909	0.936	0.957	0.973	0.984	0.992	0.997	0.999	1−	1−	1
	2	0.0+	0.007	0.028	0.061	0.104	0.156	0.216	0.282	0.352	0.425	0.500	0.575	0.648	0.718	0.784	0.844	0.896	0.939	0.972	0.993	1−	2
	3	0.0+	0.0+	0.001	0.003	0.008	0.016	0.027	0.043	0.064	0.091	0.125	0.166	0.216	0.275	0.343	0.422	0.512	0.614	0.729	0.857	0.970	3
4	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.039	0.185	0.344	0.478	0.590	0.684	0.760	0.821	0.870	0.908	0.938	0.959	0.974	0.985	0.992	0.996	0.998	1−	1−	1−	1−	1
	2	0.001	0.014	0.052	0.110	0.181	0.262	0.348	0.437	0.525	0.609	0.688	0.759	0.821	0.874	0.916	0.949	0.973	0.988	0.996	1−	1−	2
	3	0.0+	0.0+	0.004	0.012	0.027	0.051	0.084	0.126	0.179	0.241	0.312	0.391	0.475	0.563	0.652	0.738	0.819	0.890	0.948	0.986	1−	3
	4	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.015	0.026	0.041	0.062	0.092	0.130	0.179	0.240	0.316	0.410	0.522	0.656	0.815	0.961	4
5	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.049	0.226	0.410	0.556	0.672	0.763	0.832	0.884	0.922	0.950	0.969	0.982	0.990	0.995	0.998	1−	1−	1−	1−	1−	1−	1
	2	0.001	0.023	0.081	0.165	0.263	0.367	0.472	0.572	0.663	0.744	0.812	0.869	0.913	0.946	0.969	0.984	0.993	0.998	1−	1−	1−	2
	3	0.0+	0.001	0.009	0.027	0.058	0.104	0.163	0.235	0.317	0.407	0.500	0.593	0.683	0.765	0.837	0.896	0.942	0.973	0.991	0.999	1−	3
	4	0.0+	0.0+	0.0+	0.002	0.007	0.016	0.031	0.054	0.087	0.131	0.188	0.256	0.337	0.428	0.528	0.633	0.737	0.835	0.919	0.977	1−	4
	5	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.010	0.018	0.031	0.050	0.078	0.116	0.168	0.237	0.328	0.444	0.590	0.774	0.951	5
6	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.059	0.265	0.469	0.623	0.738	0.822	0.882	0.925	0.953	0.972	0.984	0.992	0.996	0.998	1−	1−	1−	1−	1−	1−	1−	1
	2	0.001	0.033	0.114	0.224	0.345	0.466	0.580	0.681	0.767	0.836	0.891	0.931	0.959	0.978	0.989	0.995	0.998	1−	1−	1−	1−	2
	3	0.0+	0.002	0.016	0.047	0.099	0.169	0.256	0.353	0.456	0.558	0.656	0.745	0.821	0.883	0.930	0.962	0.983	0.994	0.999	1−	1−	3
	4	0.0+	0.0+	0.001	0.006	0.017	0.038	0.070	0.117	0.179	0.255	0.344	0.442	0.544	0.647	0.744	0.831	0.901	0.953	0.984	0.998	1−	4
	5	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.011	0.022	0.041	0.069	0.109	0.164	0.233	0.319	0.420	0.534	0.655	0.776	0.886	0.967	0.999	5
	6	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.016	0.028	0.047	0.075	0.118	0.178	0.262	0.377	0.531	0.735	0.941	6
7	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.068	0.302	0.522	0.679	0.790	0.867	0.918	0.951	0.972	0.985	0.992	0.996	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.002	0.044	0.150	0.283	0.423	0.555	0.671	0.766	0.841	0.898	0.938	0.964	0.981	0.991	0.996	0.999	1−	1−	1−	1−	1−	2
	3	0.0+	0.004	0.026	0.074	0.148	0.244	0.353	0.468	0.580	0.684	0.773	0.847	0.904	0.944	0.971	0.987	0.995	0.999	1−	1−	1−	3
	4	0.0+	0.0+	0.003	0.012	0.033	0.071	0.126	0.200	0.290	0.392	0.500	0.608	0.710	0.800	0.874	0.929	0.967	0.988	0.997	1−	1−	4
	5	0.0+	0.0+	0.0+	0.001	0.005	0.013	0.029	0.056	0.096	0.153	0.227	0.316	0.420	0.532	0.647	0.756	0.852	0.926	0.974	0.996	1−	5
	6	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.009	0.019	0.036	0.062	0.102	0.159	0.234	0.329	0.445	0.577	0.717	0.850	0.956	0.998	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.015	0.028	0.049	0.082	0.133	0.210	0.321	0.478	0.698	0.932	7

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

		p																					
n	r	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
8	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.077	0.337	0.570	0.728	0.832	0.900	0.942	0.968	0.983	0.992	0.996	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.003	0.057	0.187	0.343	0.497	0.633	0.745	0.831	0.894	0.937	0.965	0.982	0.991	0.996	0.999	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.006	0.038	0.105	0.203	0.321	0.448	0.572	0.685	0.780	0.855	0.912	0.950	0.975	0.989	0.996	0.999	1−	1−	1−	1−	3
	4	0.0+	0.0+	0.005	0.021	0.056	0.114	0.194	0.294	0.406	0.523	0.637	0.740	0.826	0.894	0.942	0.973	0.990	0.997	1−	1−	1−	4
	5	0.0+	0.0+	0.0+	0.003	0.010	0.027	0.058	0.106	0.174	0.260	0.363	0.477	0.594	0.706	0.806	0.886	0.944	0.979	0.995	1−	1−	5
	6	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.025	0.050	0.088	0.145	0.220	0.315	0.428	0.552	0.679	0.797	0.895	0.962	0.994	1−	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.009	0.018	0.035	0.063	0.106	0.169	0.255	0.367	0.503	0.657	0.813	0.943	0.997	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.004	0.008	0.017	0.032	0.058	0.100	0.168	0.272	0.430	0.663	0.923	8
9	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.086	0.370	0.613	0.768	0.866	0.925	0.960	0.979	0.990	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.003	0.071	0.225	0.401	0.564	0.700	0.804	0.879	0.929	0.961	0.980	0.991	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.008	0.053	0.141	0.262	0.399	0.537	0.663	0.768	0.850	0.910	0.950	0.975	0.989	0.996	0.999	1−	1−	1−	1−	1−	3
	4	0.0+	0.001	0.008	0.034	0.086	0.166	0.270	0.391	0.517	0.639	0.746	0.834	0.901	0.946	0.975	0.990	0.997	1−	1−	1−	1−	4
	5	0.0+	0.0+	0.001	0.006	0.020	0.049	0.099	0.172	0.267	0.379	0.500	0.621	0.733	0.828	0.901	0.951	0.980	0.994	1−	1−	1−	5
	6	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.025	0.054	0.099	0.166	0.254	0.361	0.483	0.609	0.730	0.834	0.914	0.966	0.992	1−	1−	6
	7	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.025	0.050	0.090	0.150	0.232	0.337	0.463	0.601	0.738	0.859	0.947	0.992	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.009	0.020	0.039	0.071	0.121	0.196	0.300	0.436	0.599	0.775	0.929	0.997	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.010	0.021	0.040	0.075	0.134	0.232	0.387	0.630	0.914	9
10	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.096	0.401	0.651	0.803	0.893	0.944	0.972	0.987	0.994	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.004	0.086	0.264	0.456	0.624	0.756	0.851	0.914	0.954	0.977	0.989	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.012	0.070	0.180	0.322	0.474	0.617	0.738	0.833	0.900	0.945	0.973	0.988	0.995	0.998	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.001	0.013	0.050	0.121	0.224	0.350	0.486	0.618	0.734	0.828	0.898	0.945	0.974	0.989	0.996	1−	1−	1−	1−	1−	4
	5	0.0+	0.0+	0.002	0.010	0.033	0.078	0.150	0.249	0.367	0.496	0.623	0.738	0.834	0.905	0.953	0.980	0.994	0.999	1−	1−	1−	5
	6	0.0+	0.0+	0.0+	0.001	0.006	0.020	0.047	0.095	0.166	0.262	0.377	0.504	0.633	0.751	0.850	0.922	0.967	0.990	0.998	1−	1−	6
	7	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.026	0.055	0.102	0.172	0.266	0.382	0.514	0.650	0.776	0.879	0.950	0.987	0.999	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.012	0.027	0.055	0.100	0.167	0.262	0.383	0.526	0.678	0.820	0.930	0.988	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.011	0.023	0.046	0.086	0.149	0.244	0.376	0.544	0.736	0.914	0.996	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.006	0.013	0.028	0.056	0.107	0.197	0.349	0.599	0.904	10
11	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.105	0.431	0.686	0.833	0.914	0.958	0.980	0.991	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.005	0.102	0.303	0.508	0.678	0.803	0.887	0.939	0.970	0.986	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.015	0.090	0.221	0.383	0.545	0.687	0.800	0.881	0.935	0.967	0.985	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.002	0.019	0.069	0.161	0.287	0.430	0.574	0.704	0.809	0.887	0.939	0.971	0.988	0.996	0.999	1−	1−	1−	1−	1−	4

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
11	5	0.0+	0.0+	0.003	0.016	0.050	0.115	0.210	0.332	0.467	0.603	0.726	0.826	0.901	0.950	0.978	0.992	0.998	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.0+	0.003	0.012	0.034	0.078	0.149	0.247	0.367	0.500	0.633	0.753	0.851	0.922	0.966	0.988	0.997	1−	1−	1−	6
	7	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.022	0.050	0.099	0.174	0.274	0.397	0.533	0.668	0.790	0.885	0.950	0.984	0.997	1−	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.029	0.061	0.113	0.191	0.296	0.426	0.570	0.713	0.839	0.931	0.981	0.998	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.015	0.033	0.065	0.119	0.200	0.313	0.455	0.617	0.779	0.910	0.985	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.014	0.030	0.061	0.113	0.197	0.322	0.492	0.697	0.898	0.995	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.009	0.020	0.042	0.086	0.167	0.314	0.569	0.895	11
12	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.114	0.460	0.718	0.858	0.931	0.968	0.986	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.006	0.118	0.341	0.557	0.725	0.842	0.915	0.958	0.980	0.992	0.997	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.020	0.111	0.264	0.442	0.609	0.747	0.849	0.917	0.958	0.981	0.992	0.997	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.002	0.026	0.092	0.205	0.351	0.507	0.653	0.775	0.866	0.927	0.964	0.985	0.994	0.998	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.0+	0.004	0.024	0.073	0.158	0.276	0.417	0.562	0.696	0.806	0.888	0.943	0.974	0.991	0.997	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.001	0.005	0.019	0.054	0.118	0.213	0.335	0.473	0.613	0.739	0.842	0.915	0.961	0.986	0.996	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.0+	0.001	0.004	0.014	0.039	0.085	0.158	0.261	0.387	0.527	0.665	0.787	0.882	0.946	0.981	0.995	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.026	0.057	0.112	0.194	0.304	0.438	0.583	0.724	0.842	0.927	0.976	0.996	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.015	0.036	0.073	0.134	0.225	0.347	0.493	0.649	0.795	0.908	0.974	0.998	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.019	0.042	0.083	0.151	0.253	0.391	0.558	0.736	0.889	0.980	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.020	0.042	0.085	0.158	0.275	0.443	0.659	0.882	0.994	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.014	0.032	0.069	0.142	0.282	0.540	0.886	12
13	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.122	0.487	0.746	0.879	0.945	0.976	0.990	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.007	0.135	0.379	0.602	0.766	0.873	0.936	0.970	0.987	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.025	0.134	0.308	0.498	0.667	0.798	0.887	0.942	0.973	0.989	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.003	0.034	0.118	0.253	0.416	0.579	0.722	0.831	0.907	0.954	0.980	0.992	0.997	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.0+	0.006	0.034	0.099	0.206	0.346	0.499	0.647	0.772	0.867	0.930	0.968	0.987	0.996	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.001	0.008	0.030	0.080	0.165	0.284	0.426	0.573	0.709	0.821	0.902	0.954	0.982	0.994	0.999	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.0+	0.001	0.007	0.024	0.062	0.129	0.229	0.356	0.500	0.644	0.771	0.871	0.938	0.976	0.993	0.999	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.018	0.046	0.098	0.179	0.291	0.427	0.574	0.716	0.835	0.920	0.970	0.992	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.032	0.070	0.133	0.228	0.353	0.501	0.654	0.794	0.901	0.966	0.994	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.020	0.046	0.093	0.169	0.278	0.421	0.584	0.747	0.882	0.966	0.997	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.027	0.058	0.113	0.202	0.333	0.502	0.692	0.866	0.975	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.005	0.013	0.030	0.064	0.127	0.234	0.398	0.621	0.865	0.993	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.010	0.024	0.055	0.121	0.254	0.513	0.878	13
14	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.131	0.512	0.771	0.897	0.956	0.982	0.993	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
14	2	0.008	0.153	0.415	0.643	0.802	0.899	0.953	0.979	0.992	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.030	0.158	0.352	0.552	0.719	0.839	0.916	0.960	0.983	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.004	0.044	0.147	0.302	0.479	0.645	0.780	0.876	0.937	0.971	0.989	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.0+	0.009	0.047	0.130	0.258	0.416	0.577	0.721	0.833	0.910	0.957	0.982	0.994	0.998	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.001	0.012	0.044	0.112	0.219	0.359	0.514	0.663	0.788	0.881	0.942	0.976	0.992	0.998	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.0+	0.002	0.012	0.038	0.093	0.184	0.308	0.454	0.605	0.741	0.850	0.925	0.969	0.990	0.998	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.0+	0.002	0.010	0.031	0.075	0.150	0.259	0.395	0.546	0.692	0.816	0.907	0.962	0.988	0.998	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.008	0.024	0.058	0.119	0.212	0.337	0.486	0.641	0.781	0.888	0.956	0.988	0.999	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.018	0.043	0.090	0.167	0.279	0.423	0.584	0.742	0.870	0.953	0.991	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.029	0.063	0.124	0.220	0.355	0.521	0.698	0.853	0.956	0.996	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.006	0.017	0.040	0.084	0.161	0.281	0.448	0.648	0.842	0.970	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.021	0.047	0.101	0.198	0.357	0.585	0.847	0.992	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.018	0.044	0.103	0.229	0.488	0.869	14
15	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.140	0.537	0.794	0.913	0.965	0.987	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.010	0.171	0.451	0.681	0.833	0.920	0.965	0.986	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.0+	0.036	0.184	0.396	0.602	0.764	0.873	0.938	0.973	0.989	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.005	0.056	0.177	0.352	0.539	0.703	0.827	0.909	0.958	0.982	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.001	0.013	0.062	0.164	0.314	0.485	0.648	0.783	0.880	0.941	0.975	0.991	0.997	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.002	0.017	0.061	0.148	0.278	0.436	0.597	0.739	0.849	0.923	0.966	0.988	0.996	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.0+	0.004	0.018	0.057	0.131	0.245	0.390	0.548	0.696	0.818	0.905	0.958	0.985	0.996	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.001	0.004	0.017	0.050	0.113	0.213	0.346	0.500	0.654	0.787	0.887	0.950	0.983	0.996	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.015	0.042	0.095	0.182	0.304	0.452	0.610	0.755	0.869	0.943	0.982	0.996	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.034	0.077	0.151	0.261	0.403	0.564	0.722	0.852	0.939	0.983	0.998	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.025	0.059	0.120	0.217	0.352	0.515	0.686	0.836	0.938	0.987	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.018	0.042	0.091	0.173	0.297	0.461	0.648	0.823	0.944	0.995	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.011	0.027	0.062	0.127	0.236	0.398	0.604	0.816	0.964	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.014	0.035	0.080	0.167	0.319	0.549	0.829	0.990	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.013	0.035	0.087	0.206	0.463	0.860	15
16	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.149	0.560	0.815	0.926	0.972	0.990	0.997	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.011	0.189	0.485	0.716	0.859	0.937	0.974	0.990	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.043	0.211	0.439	0.648	0.803	0.901	0.955	0.982	0.993	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.007	0.068	0.210	0.402	0.595	0.754	0.866	0.935	0.972	0.989	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.001	0.017	0.079	0.202	0.370	0.550	0.711	0.833	0.915	0.962	0.985	0.995	0.999	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.003	0.024	0.082	0.190	0.340	0.510	0.671	0.802	0.895	0.951	0.981	0.994	0.998	1−	1−	1−	1−	1−	1−	6

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
16	7	0.0+	0.0+	0.001	0.006	0.027	0.080	0.175	0.312	0.473	0.634	0.773	0.876	0.942	0.977	0.993	0.998	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.001	0.007	0.027	0.074	0.159	0.284	0.437	0.598	0.744	0.858	0.933	0.974	0.993	0.999	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.001	0.007	0.026	0.067	0.142	0.256	0.402	0.563	0.716	0.841	0.926	0.973	0.993	0.999	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.007	0.023	0.058	0.124	0.227	0.366	0.527	0.688	0.825	0.920	0.973	0.994	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.019	0.049	0.105	0.198	0.329	0.490	0.660	0.810	0.918	0.976	0.997	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.038	0.085	0.167	0.289	0.450	0.630	0.798	0.921	0.983	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.011	0.028	0.065	0.134	0.246	0.405	0.598	0.790	0.932	0.993	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.018	0.045	0.099	0.197	0.352	0.561	0.789	0.957	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.026	0.063	0.141	0.284	0.515	0.811	0.989	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.028	0.074	0.185	0.440	0.851	16
17	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.157	0.582	0.833	0.937	0.977	0.992	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.012	0.208	0.518	0.748	0.882	0.950	0.981	0.993	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.050	0.238	0.480	0.690	0.836	0.923	0.967	0.988	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.009	0.083	0.244	0.451	0.647	0.798	0.897	0.954	0.982	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.001	0.022	0.099	0.242	0.426	0.611	0.765	0.874	0.940	0.975	0.991	0.997	1−	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.005	0.032	0.106	0.235	0.403	0.580	0.736	0.853	0.928	0.970	0.989	0.997	1−	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.001	0.008	0.038	0.107	0.225	0.381	0.552	0.710	0.834	0.917	0.965	0.988	0.997	1−	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.002	0.011	0.040	0.105	0.213	0.359	0.526	0.685	0.817	0.908	0.962	0.987	0.997	1−	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.0+	0.003	0.012	0.040	0.099	0.199	0.337	0.500	0.663	0.801	0.901	0.960	0.988	0.997	1−	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.0+	0.003	0.013	0.038	0.092	0.183	0.315	0.474	0.641	0.787	0.895	0.960	0.989	0.998	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.035	0.083	0.166	0.290	0.448	0.619	0.775	0.893	0.962	0.992	1−	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.011	0.030	0.072	0.147	0.264	0.420	0.597	0.765	0.894	0.968	0.995	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.009	0.025	0.060	0.126	0.235	0.389	0.574	0.758	0.901	0.978	0.999	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.018	0.046	0.103	0.202	0.353	0.549	0.756	0.917	0.991	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.033	0.077	0.164	0.310	0.520	0.762	0.950	1−	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.007	0.019	0.050	0.118	0.252	0.482	0.792	0.988	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.023	0.063	0.167	0.418	0.843	17
18	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.165	0.603	0.850	0.946	0.982	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.014	0.226	0.550	0.776	0.901	0.961	0.986	0.995	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.058	0.266	0.520	0.729	0.865	0.940	0.976	0.992	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.011	0.098	0.280	0.499	0.694	0.835	0.922	0.967	0.988	0.996	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.002	0.028	0.121	0.284	0.481	0.667	0.811	0.906	0.959	0.985	0.995	0.999	1−	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.006	0.042	0.133	0.283	0.466	0.645	0.791	0.892	0.952	0.982	0.994	0.999	1−	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.001	0.012	0.051	0.139	0.278	0.451	0.626	0.774	0.881	0.946	0.980	0.994	0.999	1−	1−	1−	1−	1−	1−	7

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
18	8	0.0+	0.0+	0.0+	0.003	0.016	0.057	0.141	0.272	0.437	0.609	0.760	0.872	0.942	0.979	0.994	0.999	1−	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.001	0.004	0.019	0.060	0.139	0.263	0.422	0.593	0.747	0.865	0.940	0.979	0.995	1−	1−	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.021	0.060	0.135	0.253	0.407	0.578	0.737	0.861	0.940	0.981	0.996	1−	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.021	0.058	0.128	0.240	0.391	0.563	0.728	0.859	0.943	0.984	0.997	1−	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.020	0.054	0.119	0.226	0.374	0.549	0.722	0.861	0.949	0.988	0.999	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.018	0.048	0.108	0.209	0.355	0.534	0.717	0.867	0.958	0.994	1−	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.015	0.041	0.094	0.189	0.333	0.519	0.716	0.879	0.972	0.998	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.033	0.078	0.165	0.306	0.501	0.720	0.902	0.989	1−	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.008	0.024	0.060	0.135	0.271	0.480	0.734	0.942	1−	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.014	0.039	0.099	0.224	0.450	0.774	0.986	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.018	0.054	0.150	0.397	0.835	18
19	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.174	0.623	0.865	0.954	0.986	0.996	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.015	0.245	0.580	0.802	0.917	0.969	0.990	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.067	0.295	0.559	0.763	0.889	0.954	0.983	0.995	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.013	0.115	0.316	0.545	0.737	0.867	0.941	0.977	0.992	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.002	0.035	0.144	0.327	0.535	0.718	0.850	0.930	0.972	0.990	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.009	0.054	0.163	0.332	0.526	0.703	0.837	0.922	0.968	0.989	0.997	1−	1−	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.002	0.016	0.068	0.175	0.334	0.519	0.692	0.827	0.916	0.966	0.988	0.997	1−	1−	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.004	0.023	0.077	0.182	0.334	0.512	0.683	0.820	0.913	0.965	0.989	0.997	1−	1−	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.001	0.007	0.029	0.084	0.185	0.333	0.506	0.676	0.816	0.912	0.965	0.989	0.998	1−	1−	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.002	0.009	0.033	0.087	0.186	0.329	0.500	0.671	0.814	0.913	0.967	0.991	0.998	1−	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.011	0.035	0.088	0.184	0.324	0.494	0.667	0.815	0.916	0.971	0.993	1−	1−	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.003	0.011	0.035	0.087	0.180	0.317	0.488	0.666	0.818	0.923	0.977	0.996	1−	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.034	0.084	0.173	0.308	0.481	0.666	0.825	0.932	0.984	0.998	1−	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.011	0.032	0.078	0.163	0.297	0.474	0.668	0.837	0.946	0.991	1−	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.028	0.070	0.150	0.282	0.465	0.673	0.856	0.965	0.998	1−	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.023	0.059	0.133	0.263	0.455	0.684	0.885	0.987	1−	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.005	0.017	0.046	0.111	0.237	0.441	0.705	0.933	1−	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.010	0.031	0.083	0.198	0.420	0.755	0.985	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.014	0.046	0.135	0.377	0.826	19
20	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.182	0.642	0.878	0.961	0.988	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.017	0.264	0.608	0.824	0.931	0.976	0.992	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.075	0.323	0.595	0.794	0.909	0.965	0.988	0.996	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.016	0.133	0.352	0.589	0.775	0.893	0.956	0.984	0.995	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	4

Tabla II (*Continuación*)
PROBABILIDADES BINOMIALES ACUMULADAS

$$\sum_{x=r}^{n} b(x; n, p)$$

n	r	p 0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	0.99	r
20	5	0.0+	0.003	0.043	0.170	0.370	0.585	0.762	0.882	0.949	0.981	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.011	0.067	0.196	0.383	0.584	0.755	0.874	0.945	0.979	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.002	0.022	0.087	0.214	0.392	0.583	0.750	0.870	0.942	0.979	0.994	0.998	1−	1−	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.0+	0.006	0.032	0.102	0.228	0.399	0.584	0.748	0.868	0.942	0.979	0.994	0.999	1−	1−	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.001	0.010	0.041	0.113	0.238	0.404	0.586	0.748	0.869	0.943	0.980	0.995	1−	1−	1−	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.003	0.014	0.048	0.122	0.245	0.409	0.588	0.751	0.872	0.947	0.983	0.996	1−	1−	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.017	0.053	0.128	0.249	0.412	0.591	0.755	0.878	0.952	0.986	0.997	1−	1−	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.020	0.057	0.131	0.252	0.414	0.596	0.762	0.887	0.959	0.990	0.999	1−	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.021	0.058	0.132	0.252	0.416	0.601	0.772	0.898	0.968	0.994	1−	1−	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.021	0.058	0.130	0.250	0.417	0.608	0.786	0.913	0.978	0.998	1−	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.021	0.055	0.126	0.245	0.416	0.617	0.804	0.933	0.989	1−	1−	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.006	0.019	0.051	0.118	0.238	0.415	0.630	0.830	0.957	0.997	1−	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.005	0.016	0.044	0.107	0.225	0.411	0.648	0.867	0.984	1−	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.012	0.035	0.091	0.206	0.405	0.677	0.925	0.999	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.008	0.024	0.069	0.176	0.392	0.736	0.983	19
	20	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.039	0.122	0.358	0.818	20
21	0	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0
	1	0.190	0.659	0.891	0.967	0.991	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1
	2	0.019	0.283	0.635	0.845	0.942	0.981	0.994	0.999	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	2
	3	0.001	0.085	0.352	0.630	0.821	0.925	0.973	0.991	0.998	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	3
	4	0.0+	0.019	0.152	0.389	0.630	0.808	0.914	0.967	0.989	0.997	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	4
	5	0.0+	0.003	0.052	0.197	0.414	0.633	0.802	0.908	0.963	0.987	0.996	1−	1−	1−	1−	1−	1−	1−	1−	1−	1−	5
	6	0.0+	0.0+	0.014	0.083	0.231	0.433	0.637	0.799	0.904	0.961	0.987	0.996	1−	1−	1−	1−	1−	1−	1−	1−	1−	6
	7	0.0+	0.0+	0.003	0.029	0.109	0.256	0.449	0.643	0.800	0.904	0.961	0.987	0.996	1−	1−	1−	1−	1−	1−	1−	1−	7
	8	0.0+	0.0+	0.001	0.008	0.043	0.130	0.277	0.464	0.650	0.803	0.905	0.962	0.988	0.997	1−	1−	1−	1−	1−	1−	1−	8
	9	0.0+	0.0+	0.0+	0.002	0.014	0.056	0.148	0.294	0.476	0.659	0.808	0.909	0.965	0.989	0.998	1−	1−	1−	1−	1−	1−	9
	10	0.0+	0.0+	0.0+	0.0+	0.004	0.021	0.068	0.162	0.309	0.488	0.668	0.816	0.915	0.969	0.991	0.998	1−	1−	1−	1−	1−	10
	11	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.026	0.077	0.174	0.321	0.500	0.679	0.826	0.923	0.974	0.994	1−	1−	1−	1−	1−	11
	12	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.009	0.031	0.085	0.184	0.332	0.512	0.691	0.838	0.932	0.979	0.996	1−	1−	1−	1−	12
	13	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.002	0.011	0.035	0.091	0.192	0.341	0.524	0.706	0.852	0.944	0.986	0.998	1−	1−	1−	13
	14	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.012	0.038	0.095	0.197	0.350	0.536	0.723	0.870	0.957	0.992	1−	1−	1−	14
	15	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.039	0.096	0.200	0.357	0.551	0.744	0.891	0.971	0.997	1−	1−	15
	16	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.039	0.096	0.201	0.363	0.567	0.769	0.917	0.986	1−	1−	16
	17	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.004	0.013	0.037	0.092	0.198	0.367	0.586	0.803	0.948	0.997	1−	17
	18	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.003	0.011	0.033	0.086	0.192	0.370	0.611	0.848	0.981	1−	18
	19	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.009	0.027	0.075	0.179	0.370	0.648	0.915	0.999	19
	20	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.006	0.019	0.058	0.155	0.365	0.717	0.981	20
	21	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.0+	0.001	0.002	0.009	0.033	0.109	0.341	0.810	21

Tabla III
PROBABILIDADES ACUMULADAS DE POISSON

$$P(x;\lambda) = \sum_{r=0}^{x} \frac{\lambda^r}{r!}\, e^{-\lambda}$$

λ	x 0	1	2	3	4	5	6
0.01	0.990	1.000	1.000	1.000	1.000	1.000	1.000
0.02	0.980	1.000	1.000	1.000	1.000	1.000	1.000
0.03	0.970	1.000	1.000	1.000	1.000	1.000	1.000
0.04	0.961	0.999	1.000	1.000	1.000	1.000	1.000
0.05	0.951	0.999	1.000	1.000	1.000	1.000	1.000
0.06	0.942	0.998	1.000	1.000	1.000	1.000	1.000
0.07	0.932	0.998	1.000	1.000	1.000	1.000	1.000
0.08	0.923	0.997	1.000	1.000	1.000	1.000	1.000
0.09	0.914	0.996	1.000	1.000	1.000	1.000	1.000
0.10	0.905	0.995	1.000	1.000	1.000	1.000	1.000
0.11	0.896	0.994	1.000	1.000	1.000	1.000	1.000
0.12	0.887	0.993	1.000	1.000	1.000	1.000	1.000
0.13	0.878	0.992	1.000	1.000	1.000	1.000	1.000
0.14	0.869	0.991	1.000	1.000	1.000	1.000	1.000
0.15	0.861	0.990	0.999	1.000	1.000	1.000	1.000
0.16	0.852	0.988	0.999	1.000	1.000	1.000	1.000
0.17	0.844	0.987	0.999	1.000	1.000	1.000	1.000
0.18	0.835	0.986	0.999	1.000	1.000	1.000	1.000
0.19	0.827	0.984	0.999	1.000	1.000	1.000	1.000
0.20	0.819	0.982	0.999	1.000	1.000	1.000	1.000
0.25	0.779	0.974	0.998	1.000	1.000	1.000	1.000
0.30	0.741	0.963	0.996	1.000	1.000	1.000	1.000
0.35	0.705	0.951	0.994	1.000	1.000	1.000	1.000
0.40	0.670	0.938	0.992	0.999	1.000	1.000	1.000
0.45	0.638	0.925	0.989	0.999	1.000	1.000	1.000
0.50	0.607	0.910	0.986	0.998	1.000	1.000	1.000
0.55	0.577	0.894	0.982	0.998	1.000	1.000	1.000
0.60	0.549	0.878	0.977	0.997	1.000	1.000	1.000
0.65	0.522	0.861	0.972	0.996	0.999	1.000	1.000
0.70	0.497	0.844	0.966	0.994	0.999	1.000	1.000
0.75	0.472	0.827	0.959	0.993	0.999	1.000	1.000
0.80	0.449	0.809	0.953	0.991	0.999	1.000	1.000
0.85	0.427	0.791	0.945	0.989	0.998	1.000	1.000
0.90	0.407	0.772	0.937	0.987	0.998	1.000	1.000
0.95	0.387	0.754	0.929	0.984	0.997	1.000	1.000
1.00	0.368	0.736	0.920	0.981	0.996	0.999	1.000

λ	x 0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1.10	0.333	0.699	0.900	0.974	0.995	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.20	0.301	0.663	0.879	0.966	0.992	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.30	0.273	0.627	0.857	0.957	0.989	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.40	0.247	0.592	0.833	0.946	0.986	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.50	0.223	0.558	0.809	0.934	0.981	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.60	0.202	0.525	0.783	0.921	0.976	0.994	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.70	0.183	0.493	0.757	0.907	0.970	0.992	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.80	0.165	0.463	0.731	0.891	0.964	0.990	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1.90	0.150	0.434	0.704	0.875	0.956	0.987	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.00	0.135	0.406	0.677	0.857	0.947	0.983	0.995	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.10	0.122	0.380	0.650	0.839	0.938	0.980	0.994	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.20	0.111	0.355	0.623	0.819	0.928	0.975	0.993	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.30	0.100	0.331	0.596	0.799	0.916	0.970	0.991	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.40	0.091	0.308	0.570	0.779	0.904	0.964	0.988	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.50	0.082	0.287	0.544	0.758	0.891	0.958	0.986	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.60	0.074	0.267	0.518	0.736	0.877	0.951	0.983	0.995	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.70	0.067	0.249	0.494	0.714	0.863	0.943	0.979	0.993	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.80	0.061	0.231	0.469	0.692	0.848	0.935	0.976	0.992	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
2.90	0.055	0.215	0.446	0.670	0.832	0.926	0.971	0.990	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3.00	0.050	0.199	0.423	0.647	0.815	0.916	0.966	0.988	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3.20	0.041	0.171	0.380	0.603	0.781	0.895	0.955	0.983	0.994	0.998	1.000	1.000	1.000	1.000	1.000	1.000	1.000
3.30	0.037	0.159	0.359	0.580	0.763	0.883	0.949	0.980	0.993	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
3.40	0.033	0.147	0.340	0.558	0.744	0.871	0.942	0.977	0.992	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000
3.50	0.030	0.136	0.321	0.537	0.725	0.858	0.935	0.973	0.990	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000
3.60	0.027	0.126	0.303	0.515	0.706	0.844	0.927	0.969	0.988	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000
3.70	0.025	0.116	0.285	0.494	0.687	0.830	0.918	0.965	0.986	0.995	0.998	1.000	1.000	1.000	1.000	1.000	1.000
3.80	0.022	0.107	0.269	0.473	0.668	0.816	0.909	0.960	0.984	0.994	0.998	0.999	1.000	1.000	1.000	1.000	1.000
3.90	0.020	0.099	0.253	0.453	0.648	0.801	0.899	0.955	0.981	0.993	0.998	0.999	1.000	1.000	1.000	1.000	1.000
4.00	0.018	0.092	0.238	0.433	0.629	0.785	0.889	0.949	0.979	0.992	0.997	0.999	1.000	1.000	1.000	1.000	1.000
4.10	0.017	0.085	0.224	0.414	0.609	0.769	0.879	0.943	0.976	0.990	0.997	0.999	1.000	1.000	1.000	1.000	1.000
4.20	0.015	0.078	0.210	0.395	0.590	0.753	0.867	0.936	0.972	0.989	0.996	0.999	1.000	1.000	1.000	1.000	1.000
4.30	0.014	0.072	0.197	0.377	0.570	0.737	0.856	0.929	0.968	0.987	0.995	0.998	0.999	1.000	1.000	1.000	1.000
4.40	0.012	0.066	0.185	0.359	0.551	0.720	0.844	0.921	0.964	0.985	0.994	0.998	0.999	1.000	1.000	1.000	1.000
4.50	0.011	0.061	0.174	0.342	0.532	0.703	0.831	0.913	0.960	0.983	0.993	0.998	0.999	1.000	1.000	1.000	1.000
4.60	0.010	0.056	0.163	0.326	0.513	0.686	0.818	0.905	0.955	0.980	0.992	0.997	0.999	1.000	1.000	1.000	1.000
4.70	0.009	0.052	0.152	0.310	0.495	0.668	0.805	0.896	0.950	0.978	0.991	0.997	0.999	1.000	1.000	1.000	1.000

Tabla III (*Continuación*)
PROBABILIDADES ACUMULADAS DE POISSON

$$P(x;\lambda) = \sum_{r=0}^{x} \frac{\lambda^r}{r!}\, e^{-\lambda}$$

λ	x																									
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
4.80	0.008	0.048	0.143	0.294	0.476	0.651	0.791	0.887	0.944	0.975	0.990	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.00	0.007	0.040	0.125	0.265	0.440	0.616	0.762	0.867	0.932	0.968	0.986	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.20	0.006	0.034	0.109	0.238	0.406	0.581	0.732	0.845	0.918	0.960	0.982	0.993	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.40	0.005	0.029	0.095	0.213	0.373	0.546	0.702	0.822	0.903	0.951	0.977	0.990	0.996	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.60	0.004	0.024	0.082	0.191	0.342	0.512	0.670	0.797	0.886	0.941	0.972	0.988	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
5.80	0.003	0.021	0.072	0.170	0.313	0.478	0.638	0.771	0.867	0.929	0.965	0.984	0.993	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.00	0.002	0.017	0.062	0.151	0.285	0.446	0.606	0.744	0.847	0.916	0.957	0.980	0.991	0.996	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.20	0.002	0.015	0.054	0.134	0.259	0.414	0.574	0.716	0.826	0.902	0.949	0.975	0.989	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.40	0.002	0.012	0.046	0.119	0.235	0.384	0.542	0.687	0.803	0.886	0.939	0.969	0.986	0.994	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.60	0.001	0.010	0.040	0.105	0.213	0.355	0.511	0.658	0.780	0.869	0.927	0.963	0.982	0.992	0.997	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
6.80	0.001	0.009	0.034	0.093	0.192	0.327	0.480	0.628	0.755	0.850	0.915	0.955	0.978	0.990	0.996	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7.00	0.001	0.007	0.030	0.082	0.173	0.301	0.450	0.599	0.729	0.830	0.901	0.947	0.973	0.987	0.994	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7.20	0.001	0.006	0.025	0.072	0.156	0.276	0.420	0.569	0.703	0.810	0.887	0.937	0.967	0.984	0.993	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7.40	0.001	0.005	0.022	0.063	0.140	0.253	0.392	0.539	0.676	0.788	0.871	0.926	0.961	0.980	0.991	0.996	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7.60	0.001	0.004	0.019	0.055	0.125	0.231	0.365	0.510	0.648	0.765	0.854	0.915	0.954	0.976	0.989	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
7.80	0.000	0.004	0.016	0.048	0.112	0.210	0.338	0.481	0.620	0.741	0.835	0.902	0.945	0.971	0.986	0.993	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8.00	0.000	0.003	0.014	0.042	0.100	0.191	0.313	0.453	0.593	0.717	0.816	0.888	0.936	0.966	0.983	0.992	0.996	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8.20	0.000	0.003	0.012	0.037	0.089	0.174	0.290	0.425	0.565	0.692	0.796	0.873	0.926	0.960	0.979	0.990	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8.40	0.000	0.002	0.010	0.032	0.079	0.157	0.267	0.399	0.537	0.666	0.774	0.857	0.915	0.952	0.975	0.987	0.994	0.997	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
8.60	0.000	0.002	0.009	0.028	0.070	0.142	0.246	0.373	0.509	0.640	0.752	0.840	0.903	0.945	0.970	0.985	0.993	0.997	0.999	0.999	1.000	1.000	1.000	1.000	1.000	1.000
8.80	0.000	0.001	0.007	0.024	0.062	0.128	0.226	0.348	0.482	0.614	0.729	0.822	0.890	0.936	0.965	0.982	0.991	0.996	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
9.00	0.000	0.001	0.006	0.021	0.055	0.116	0.207	0.324	0.456	0.587	0.706	0.803	0.876	0.926	0.959	0.978	0.989	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000
9.20	0.000	0.001	0.005	0.018	0.049	0.104	0.189	0.301	0.430	0.561	0.682	0.783	0.861	0.916	0.952	0.974	0.987	0.993	0.997	0.999	0.999	1.000	1.000	1.000	1.000	1.000
9.40	0.000	0.001	0.005	0.016	0.043	0.093	0.173	0.279	0.404	0.535	0.658	0.763	0.845	0.904	0.944	0.969	0.984	0.992	0.996	0.998	0.999	1.000	1.000	1.000	1.000	1.000
9.60	0.000	0.001	0.004	0.014	0.038	0.084	0.157	0.258	0.380	0.509	0.633	0.741	0.828	0.892	0.936	0.964	0.981	0.990	0.995	0.998	0.999	1.000	1.000	1.000	1.000	1.000
9.80	0.000	0.001	0.003	0.012	0.033	0.075	0.143	0.239	0.356	0.483	0.608	0.719	0.810	0.879	0.927	0.958	0.977	0.988	0.994	0.997	0.999	0.999	1.000	1.000	1.000	1.000
10.00	0.000	0.000	0.003	0.010	0.029	0.067	0.130	0.220	0.333	0.458	0.583	0.697	0.792	0.864	0.917	0.951	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000
10.20	0.000	0.000	0.002	0.009	0.026	0.060	0.118	0.203	0.311	0.433	0.558	0.674	0.772	0.849	0.906	0.944	0.968	0.983	0.991	0.996	0.998	0.999	1.000	1.000	1.000	1.000
10.40	0.000	0.000	0.002	0.008	0.023	0.053	0.107	0.186	0.290	0.409	0.533	0.650	0.752	0.834	0.894	0.936	0.963	0.980	0.989	0.995	0.997	0.999	0.999	1.000	1.000	1.000
10.60	0.000	0.000	0.002	0.007	0.020	0.048	0.097	0.171	0.269	0.385	0.508	0.627	0.732	0.817	0.882	0.927	0.957	0.976	0.987	0.994	0.997	0.999	0.999	1.000	1.000	1.000
10.80	0.000	0.000	0.001	0.006	0.017	0.042	0.087	0.157	0.250	0.363	0.484	0.603	0.710	0.799	0.868	0.918	0.951	0.972	0.985	0.992	0.996	0.998	0.999	1.000	1.000	1.000
11.00	0.000	0.000	0.001	0.005	0.015	0.038	0.079	0.143	0.232	0.341	0.460	0.579	0.689	0.781	0.854	0.907	0.944	0.968	0.982	0.991	0.995	0.998	0.999	1.000	1.000	1.000
11.20	0.000	0.000	0.001	0.004	0.013	0.033	0.071	0.131	0.215	0.319	0.436	0.555	0.667	0.762	0.839	0.896	0.936	0.963	0.979	0.989	0.994	0.997	0.999	0.999	1.000	1.000
11.40	0.000	0.000	0.001	0.004	0.012	0.029	0.064	0.119	0.198	0.299	0.413	0.532	0.644	0.743	0.823	0.885	0.928	0.957	0.976	0.987	0.993	0.997	0.998	0.999	1.000	1.000
11.60	0.000	0.000	0.001	0.003	0.010	0.026	0.057	0.108	0.183	0.279	0.391	0.508	0.622	0.723	0.807	0.872	0.919	0.951	0.972	0.984	0.992	0.996	0.998	0.999	1.000	1.000
11.80	0.000	0.000	0.001	0.003	0.009	0.023	0.051	0.099	0.169	0.260	0.369	0.485	0.599	0.702	0.790	0.859	0.909	0.944	0.967	0.982	0.990	0.995	0.998	0.999	0.999	1.000
12.00	0.000	0.000	0.001	0.002	0.008	0.020	0.046	0.090	0.155	0.242	0.347	0.462	0.576	0.682	0.772	0.844	0.899	0.937	0.963	0.979	0.988	0.994	0.997	0.999	0.999	1.000

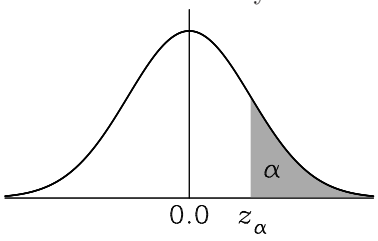
Tabla III (*Continuación*)
PROBABILIDADES ACUMULADAS DE POISSON

$$P(x;\lambda) = \sum_{r=0}^{x} \frac{\lambda^r}{r!}\, e^{-\lambda}$$

	x																									
λ	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
12.50	0.000	0.000	0.000	0.002	0.005	0.015	0.035	0.070	0.125	0.201	0.297	0.406	0.519	0.628	0.725	0.806	0.869	0.916	0.948	0.969	0.983	0.991	0.995	0.998	0.999	0.999
13.00	0.000	0.000	0.000	0.001	0.004	0.011	0.026	0.054	0.100	0.166	0.252	0.353	0.463	0.573	0.675	0.764	0.835	0.890	0.930	0.957	0.975	0.986	0.992	0.996	0.998	0.999
13.50	0.000	0.000	0.000	0.001	0.003	0.008	0.019	0.041	0.079	0.135	0.211	0.304	0.409	0.518	0.623	0.718	0.798	0.861	0.908	0.942	0.965	0.980	0.989	0.994	0.997	0.998
14.00	0.000	0.000	0.000	0.000	0.002	0.006	0.014	0.032	0.062	0.109	0.176	0.260	0.358	0.464	0.570	0.669	0.756	0.827	0.883	0.923	0.952	0.971	0.983	0.991	0.995	0.997
14.50	0.000	0.000	0.000	0.000	0.001	0.004	0.010	0.024	0.048	0.088	0.145	0.220	0.311	0.413	0.518	0.619	0.711	0.790	0.853	0.901	0.936	0.960	0.976	0.986	0.992	0.996
15.00	0.000	0.000	0.000	0.000	0.001	0.003	0.008	0.018	0.037	0.070	0.118	0.185	0.268	0.363	0.466	0.568	0.664	0.749	0.819	0.875	0.917	0.947	0.967	0.981	0.989	0.994
15.50	0.000	0.000	0.000	0.000	0.001	0.002	0.006	0.013	0.029	0.055	0.096	0.154	0.228	0.317	0.415	0.517	0.615	0.705	0.782	0.846	0.894	0.930	0.956	0.973	0.984	0.991

	x				
λ	26	27	28	29	30
12.50	1.000	1.000	1.000	1.000	1.000
13.00	1.000	1.000	1.000	1.000	1.000
13.50	0.999	1.000	1.000	1.000	1.000
14.00	0.999	0.999	1.000	1.000	1.000
14.50	0.998	0.999	0.999	1.000	1.000
15.00	0.997	0.998	0.999	1.000	1.000
15.50	0.995	0.997	0.999	0.999	1.000

Tabla IV
DISTRIBUCIÓN NORMAL TIPIFICADA
Tabla de áreas de las colas derechas, para valores de z_α
de centésima en centésima (tabla superior)
y de décima en décima (tabla inferior)

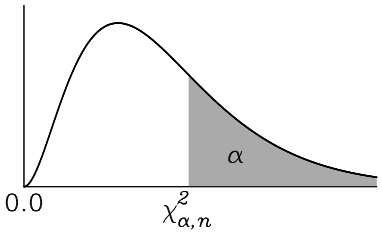


$$\alpha = \int_{z_\alpha}^{\infty} \frac{1}{\sqrt{2\pi}}\, e^{-z^2/2}\, \mathrm{d}z$$

z_α	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
2.0	0.02275	0.02222	0.02169	0.02118	0.02068	0.02018	0.01970	0.01923	0.01876	0.01831
2.1	0.01786	0.01743	0.01700	0.01659	0.01618	0.01578	0.01539	0.01500	0.01463	0.01426
2.2	0.01390	0.01355	0.01321	0.01287	0.01255	0.01222	0.01191	0.01160	0.01130	0.01101
2.3	0.01072	0.01044	0.01017	0.00990	0.00964	0.00939	0.00914	0.00889	0.00866	0.00842
2.4	0.00820	0.00798	0.00776	0.00755	0.00734	0.00714	0.00695	0.00676	0.00657	0.00639
2.5	0.00621	0.00604	0.00587	0.00570	0.00554	0.00539	0.00523	0.00508	0.00494	0.00480
2.6	0.00466	0.00453	0.00440	0.00427	0.00415	0.00402	0.00391	0.00379	0.00368	0.00357
2.7	0.00347	0.00336	0.00326	0.00317	0.00307	0.00298	0.00289	0.00280	0.00272	0.00264
2.8	0.00256	0.00248	0.00240	0.00233	0.00226	0.00219	0.00212	0.00205	0.00199	0.00193
2.9	0.00187	0.00181	0.00175	0.00169	0.00164	0.00159	0.00154	0.00149	0.00144	0.00139
3.0	0.001350	0.001306	0.001264	0.001223	0.001183	0.001144	0.001107	0.001070	0.001035	0.001001
3.1	0.000968	0.000935	0.000904	0.000874	0.000845	0.000816	0.000789	0.000762	0.000736	0.000711
3.2	0.000687	0.000664	0.000641	0.000619	0.000598	0.000577	0.000557	0.000538	0.000519	0.000501
3.3	0.000483	0.000466	0.000450	0.000434	0.000419	0.000404	0.000390	0.000376	0.000362	0.000349
3.4	0.000337	0.000325	0.000313	0.000302	0.000291	0.000280	0.000270	0.000260	0.000251	0.000242
3.5	0.000233	0.000224	0.000216	0.000208	0.000200	0.000193	0.000185	0.000178	0.000172	0.000165
3.6	0.000159	0.000153	0.000147	0.000142	0.000136	0.000131	0.000126	0.000121	0.000117	0.000112
3.7	0.000108	0.000104	0.000100	0.000096	0.000092	0.000088	0.000085	0.000082	0.000078	0.000075
3.8	0.000072	0.000069	0.000067	0.000064	0.000062	0.000059	0.000057	0.000054	0.000052	0.000050
3.9	0.000048	0.000046	0.000044	0.000042	0.000041	0.000039	0.000037	0.000036	0.000034	0.000033

z_α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.0	0.500	0.460	0.421	0.382	0.345	0.309	0.274	0.242	0.212	0.184
1.0	0.159	0.136	0.115	0.968E-01	0.808E-01	0.668E-01	0.548E-01	0.446E-01	0.359E-01	0.287E-01
2.0	0.228E-01	0.179E-01	0.139E-01	0.107E-01	0.820E-02	0.621E-02	0.466E-02	0.347E-02	0.256E-02	0.187E-02
3.0	0.135E-02	0.968E-03	0.687E-03	0.483E-03	0.337E-03	0.233E-03	0.159E-03	0.108E-03	0.723E-04	0.481E-04
4.0	0.317E-04	0.207E-04	0.133E-04	0.854E-05	0.541E-05	0.340E-05	0.211E-05	0.130E-05	0.793E-06	0.479E-06
5.0	0.287E-06	0.170E-06	0.996E-07	0.579E-07	0.333E-07	0.190E-07	0.107E-07	0.599E-08	0.332E-08	0.182E-08
6.0	0.987E-09	0.530E-09	0.282E-09	0.149E-09	0.777E-10	0.402E-10	0.206E-10	0.104E-10	0.523E-11	0.260E-11

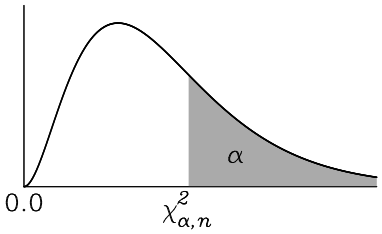
TABLA V
DISTRIBUCIÓN χ^2 DE PEARSON
Abcisas $\chi^2_{\alpha,n}$ que dejan a su derecha un área α bajo
la función con n grados de libertad



$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

n	α 0.995	0.990	0.980	0.975	0.950	0.900	0.800	0.750	0.700
1	.3928E-04	.1571E-03	.6284E-03	.9820E-03	.3932E-02	.1579E-01	.6419E-01	.1015	.1485
2	.1002E-01	.2010E-01	.4041E-01	.5064E-01	.1026	.2107	.4463	.5754	.7134
3	.7172E-01	.1148	.1848	.2158	.3518	.5844	1.005	1.213	1.424
4	.2070	.2971	.4294	.4844	.7107	1.064	1.649	1.923	2.195
5	.4118	.5543	.7519	.8312	1.145	1.610	2.343	2.675	3.000
6	.6757	.8721	1.134	1.237	1.635	2.204	3.070	3.455	3.828
7	.9892	1.239	1.564	1.690	2.167	2.833	3.822	4.255	4.671
8	1.344	1.647	2.032	2.180	2.733	3.490	4.594	5.071	5.527
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	5.899	6.393
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	6.737	7.267
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	7.584	8.148
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	8.438	9.034
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	9.299	9.926
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	10.165	10.821
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	11.037	11.721
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	11.912	12.624
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	12.792	13.531
18	6.265	7.015	7.906	8.231	9.391	10.865	12.857	13.675	14.440
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	14.562	15.352
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	15.452	16.266
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	16.344	17.182
22	8.643	9.543	10.600	10.982	12.338	14.041	16.314	17.240	18.101
23	9.260	10.196	11.293	11.689	13.090	14.848	17.187	18.137	19.021
24	9.887	10.856	11.992	12.401	13.848	15.659	18.062	19.037	19.943
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	19.939	20.867
26	11.160	12.198	13.409	13.844	15.379	17.292	19.820	20.843	21.792
27	11.808	12.879	14.125	14.573	16.151	18.114	20.703	21.749	22.719
28	12.461	13.565	14.847	15.308	16.928	18.939	21.588	22.657	23.647
29	13.121	14.262	15.574	16.047	17.708	19.768	22.475	23.567	24.577
30	13.787	14.953	16.306	16.790	18.493	20.599	23.364	24.478	25.508

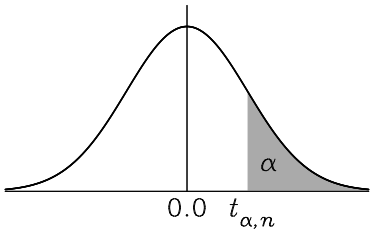
TABLA V (*Continuación*)
DISTRIBUCIÓN χ^2 DE PEARSON
Abcisas $\chi^2_{\alpha,n}$ que dejan a su derecha un área α bajo
la función con n grados de libertad



$$f(x) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} x^{(n/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$

	α										
n	0.500	0.300	0.250	0.200	0.100	0.050	0.025	0.020	0.010	0.005	0.001
1	.4549	1.074	1.323	1.642	2.706	3.841	5.024	5.412	6.635	7.880	10.827
2	1.386	2.408	2.773	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.816
3	2.366	3.665	4.108	4.642	6.251	7.815	9.348	9.838	11.345	12.838	16.266
4	3.357	4.878	5.385	5.989	7.779	9.488	11.143	11.668	13.277	14.861	18.464
5	4.351	6.064	6.626	7.289	9.236	11.071	12.832	13.388	15.086	16.749	20.514
6	5.348	7.231	7.841	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.460
7	6.346	8.383	9.037	9.803	12.017	14.067	16.013	16.623	18.486	20.278	24.321
8	7.344	9.524	10.219	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.124
9	8.343	10.656	11.389	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	9.342	11.781	12.549	13.442	15.987	18.307	20.483	21.161	23.209	25.189	29.589
11	10.341	12.899	13.701	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.281
12	11.340	14.011	14.845	15.812	18.549	21.026	23.337	24.054	26.217	28.299	32.910
13	12.340	15.119	15.984	16.985	19.812	22.362	24.736	25.471	27.688	29.820	34.529
14	13.339	16.222	17.117	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.124
15	14.339	17.322	18.245	19.311	22.307	24.996	27.488	28.260	30.578	32.801	37.697
16	15.339	18.418	19.369	20.465	23.542	26.296	28.845	29.633	32.000	34.266	39.253
17	16.338	19.511	20.489	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.793
18	17.338	20.601	21.605	22.760	25.989	28.869	31.526	32.346	34.805	37.157	42.314
19	18.338	21.689	22.718	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.821
20	19.337	22.775	23.828	25.037	28.412	31.410	34.170	35.020	37.566	39.997	45.314
21	20.337	23.858	24.935	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	21.337	24.939	26.039	27.301	30.813	33.924	36.850	37.660	40.289	42.796	48.269
23	22.337	26.018	27.141	28.429	32.007	35.172	38.076	38.968	41.638	44.182	49.728
24	23.337	27.096	28.241	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.178
25	24.337	28.172	29.339	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.622
26	25.336	29.246	30.435	31.795	35.563	38.885	41.923	42.856	45.642	48.290	54.052
27	26.336	30.319	31.528	32.912	36.741	40.113	43.194	44.139	46.963	49.645	55.477
28	27.336	31.391	32.620	34.027	37.916	41.337	44.461	45.419	48.278	50.996	56.893
29	28.336	32.461	33.711	35.139	39.087	42.557	45.722	46.693	49.588	52.336	58.301
30	29.336	33.530	34.800	36.250	40.256	43.773	46.979	47.962	50.892	53.672	59.703

TABLA VI
DISTRIBUCIÓN t DE STUDENT
Abcisas $t_{\alpha,n}$ que dejan a su derecha un área α bajo
la función con n grados de libertad



$$f(t) = \frac{1}{\sqrt{n}\beta\left(\frac{1}{2},\frac{n}{2}\right)}\left(1+\frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$

Para valores de $\alpha > 0.5$ se puede utilizar la relación
$t_{\alpha,n} = -t_{1-\alpha,n}$

	α										
n	0.50	0.40	0.30	0.20	0.10	0.050	0.025	0.010	0.005	0.001	0.0005
1	0.000	0.325	0.727	1.376	3.078	6.320	12.706	31.820	63.656	318.390	636.791
2	0.000	0.289	0.617	1.061	1.886	2.920	4.303	6.964	9.925	22.315	31.604
3	0.000	0.277	0.584	0.978	1.638	2.353	3.182	4.541	5.841	10.214	12.925
4	0.000	0.271	0.569	0.941	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.267	0.559	0.920	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.265	0.553	0.906	1.440	1.943	2.447	3.143	3.707	5.208	5.958
7	0.000	0.263	0.549	0.896	1.415	1.895	2.365	2.998	3.499	4.784	5.408
8	0.000	0.262	0.546	0.889	1.397	1.860	2.306	2.897	3.355	4.501	5.041
9	0.000	0.261	0.543	0.883	1.383	1.833	2.262	2.821	3.250	4.297	4.782
10	0.000	0.260	0.542	0.879	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.260	0.540	0.876	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.259	0.539	0.873	1.356	1.782	2.179	2.681	3.055	3.929	4.318
13	0.000	0.259	0.538	0.870	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.258	0.537	0.868	1.345	1.761	2.145	2.624	2.977	3.787	4.141
15	0.000	0.258	0.536	0.866	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.258	0.535	0.865	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.257	0.534	0.863	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.257	0.534	0.862	1.330	1.734	2.101	2.552	2.878	3.610	3.921
19	0.000	0.257	0.533	0.861	1.328	1.729	2.093	2.539	2.861	3.579	3.884
20	0.000	0.257	0.533	0.860	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.257	0.532	0.859	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.256	0.532	0.858	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.256	0.532	0.858	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.256	0.531	0.857	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.256	0.531	0.856	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.256	0.531	0.856	1.315	1.706	2.056	2.479	2.779	3.435	3.704
27	0.000	0.256	0.531	0.855	1.314	1.703	2.052	2.473	2.771	3.421	3.689
28	0.000	0.256	0.530	0.855	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.256	0.530	0.854	1.311	1.699	2.045	2.462	2.756	3.396	3.660
30	0.000	0.256	0.530	0.854	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.255	0.529	0.851	1.303	1.684	2.021	2.423	2.704	3.307	3.551
50	0.000	0.255	0.528	0.849	1.299	1.676	2.009	2.403	2.678	3.261	3.496
60	0.000	0.254	0.527	0.848	1.296	1.671	2.000	2.390	2.660	3.232	3.460
70	0.000	0.254	0.527	0.847	1.294	1.667	1.994	2.381	2.648	3.211	3.435
80	0.000	0.254	0.527	0.846	1.292	1.664	1.990	2.374	2.639	3.195	3.416
90	0.000	0.254	0.526	0.846	1.291	1.662	1.987	2.368	2.632	3.183	3.404
100	0.000	0.254	0.526	0.845	1.290	1.661	1.984	2.364	2.626	3.174	3.390
200	0.000	0.254	0.525	0.843	1.286	1.653	1.972	2.345	2.601	3.132	3.340
300	0.000	0.254	0.525	0.843	1.284	1.650	1.968	2.339	2.592	3.118	3.323
400	0.000	0.254	0.525	0.843	1.284	1.649	1.966	2.336	2.588	3.111	3.341
500	0.000	0.253	0.525	0.842	1.283	1.648	1.965	2.334	2.586	3.107	3.310
∞	0.000	0.253	0.524	0.842	1.282	1.645	1.960	2.326	2.576	3.090	3.291

Tabla VII
DISTRIBUCIÓN F DE FISHER

Abcisas $F_{\alpha;n_1,n_2}$ que dejan a su derecha un área α bajo la función con n_1 y n_2 grados de libertad.

Para valores de α próximos a uno se puede utilizar la relación $F_{1-\alpha;n_2,n_1} = \dfrac{1}{F_{\alpha;n_1,n_2}}$.

$\alpha = 0.10$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.863	49.500	53.593	55.833	57.240	58.204	58.906	59.438	59.857	60.195	60.705	61.222	61.741	62.002	62.265	62.529	62.794	63.061	63.325
2	8.5263	9.0000	9.1618	9.2434	9.2926	9.3255	9.3491	9.3667	9.3806	9.3916	9.4082	9.4248	9.4414	9.4500	9.4579	9.4662	9.4746	9.4829	9.4912
3	5.5383	5.4624	5.3908	5.3426	5.3092	5.2847	5.2662	5.2517	5.2400	5.2304	5.2156	5.2003	5.1845	5.1762	5.1681	5.1598	5.1512	5.1425	5.1337
4	4.5448	4.3246	4.1909	4.1072	4.0506	4.0097	3.9790	3.9549	3.9357	3.9199	3.8955	3.8703	3.8443	3.8310	3.8174	3.8037	3.7896	3.7753	3.7607
5	4.0604	3.7797	3.6195	3.5202	3.4530	3.4045	3.3679	3.3393	3.3163	3.2974	3.2682	3.2380	3.2067	3.1905	3.1741	3.1572	3.1402	3.1228	3.1050
6	3.7760	3.4633	3.2888	3.1809	3.1075	3.0546	3.0145	2.9830	2.9577	2.9369	2.9047	2.8712	2.8363	2.8183	2.8000	2.7812	2.7620	2.7423	2.7222
7	3.5894	3.2574	3.0740	2.9605	2.8833	2.8273	2.7849	2.7516	2.7247	2.7025	2.6681	2.6322	2.5947	2.5753	2.5555	2.5351	2.5142	2.4928	2.4708
8	3.4579	3.1131	2.9238	2.8064	2.7265	2.6683	2.6241	2.5893	2.5612	2.5380	2.5020	2.4642	2.4246	2.4041	2.3830	2.3614	2.3391	2.3162	2.2926
9	3.3604	3.0065	2.8129	2.6927	2.6106	2.5509	2.5053	2.4694	2.4403	2.4163	2.3789	2.3396	2.2983	2.2768	2.2547	2.2320	2.2085	2.1843	2.1592
10	3.2850	2.9245	2.7277	2.6053	2.5216	2.4606	2.4141	2.3772	2.3473	2.3226	2.2840	2.2435	2.2007	2.1784	2.1554	2.1317	2.1072	2.0818	2.0554
12	3.1765	2.8068	2.6055	2.4801	2.3940	2.3310	2.2828	2.2446	2.2135	2.1878	2.1474	2.1049	2.0597	2.0360	2.0115	1.9861	1.9597	1.9323	1.9036
15	3.0732	2.6952	2.4898	2.3614	2.2730	2.2081	2.1582	2.1185	2.0862	2.0593	2.0171	1.9722	1.9243	1.8990	1.8728	1.8454	1.8168	1.7867	1.7551
20	2.9747	2.5893	2.3801	2.2489	2.1582	2.0913	2.0397	1.9985	1.9649	1.9367	1.8924	1.8449	1.7938	1.7667	1.7382	1.7083	1.6768	1.6432	1.6074
24	2.9271	2.5383	2.3274	2.1949	2.1030	2.0351	1.9826	1.9407	1.9063	1.8775	1.8319	1.7831	1.7302	1.7019	1.6721	1.6407	1.6073	1.5715	1.5327
30	2.8807	2.4887	2.2761	2.1422	2.0492	1.9803	1.9269	1.8841	1.8490	1.8195	1.7727	1.7223	1.6673	1.6377	1.6065	1.5732	1.5376	1.4989	1.4564
40	2.8354	2.4404	2.2261	2.0909	1.9968	1.9269	1.8725	1.8289	1.7929	1.7627	1.7146	1.6624	1.6052	1.5741	1.5411	1.5056	1.4672	1.4248	1.3769
60	2.7911	2.3932	2.1774	2.0410	1.9457	1.8747	1.8194	1.7748	1.7380	1.7070	1.6574	1.6034	1.5435	1.5107	1.4755	1.4373	1.3952	1.3476	1.2915
120	2.7478	2.3473	2.1300	1.9923	1.8959	1.8238	1.7675	1.7220	1.6842	1.6524	1.6012	1.5450	1.4821	1.4472	1.4094	1.3676	1.3203	1.2646	1.1926
∞	2.7055	2.3026	2.0838	1.9448	1.8473	1.7741	1.7167	1.6702	1.6315	1.5987	1.5458	1.4871	1.4206	1.3832	1.3419	1.2951	1.2400	1.1686	1.1000

TABLA VII (*Continuación*)
DISTRIBUCIÓN F DE FISHER

Abcisas $F_{\alpha;n_1,n_2}$ que dejan a su derecha un área α bajo la función con n_1 y n_2 grados de libertad.

Para valores de α próximos a uno se puede utilizar la relación $F_{1-\alpha;n_2,n_1} = \dfrac{1}{F_{\alpha;n_1,n_2}}$.

$\alpha = 0.05$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.45	199.70	215.71	224.58	230.15	233.99	236.76	238.88	240.54	241.89	243.90	245.90	248.03	249.05	250.09	251.14	252.20	253.25	254.32
2	18.513	19.000	19.164	19.247	19.296	19.329	19.353	19.371	19.385	19.396	19.425	19.429	19.446	19.454	19.463	19.471	19.479	19.487	19.496
3	10.128	9.5521	9.2766	9.1156	9.0135	8.9406	8.8867	8.8452	8.8121	8.7855	8.7446	8.7029	8.6602	8.6385	8.6166	8.5944	8.5720	8.5493	8.5264
4	7.7087	6.9443	6.5914	6.3883	6.2563	6.1631	6.0942	6.0411	5.9987	5.9644	5.9117	5.8578	5.8027	5.7744	5.7459	5.7170	5.6877	5.6580	5.6280
5	6.6079	5.7863	5.4095	5.1922	5.0503	4.9503	4.8759	4.8183	4.7725	4.7351	4.6777	4.6188	4.5582	4.5271	4.4957	4.4638	4.4314	4.3984	4.3650
6	5.9874	5.1433	4.7571	4.5337	4.3874	4.2839	4.2067	4.1468	4.0990	4.0602	3.9999	3.9381	3.8742	3.8415	3.8082	3.7743	3.7398	3.7047	3.6689
7	5.5914	4.7374	4.3468	4.1219	3.9715	3.8660	3.7870	3.7257	3.6767	3.6363	3.5747	3.5107	3.4445	3.4105	3.3758	3.3402	3.3043	3.2675	3.2297
8	5.3177	4.4590	4.0662	3.8378	3.6875	3.5806	3.5004	3.4381	3.3881	3.3472	3.2839	3.2184	3.1503	3.1152	3.0794	3.0428	3.0053	2.9669	2.9276
9	5.1173	4.2565	3.8625	3.6331	3.4817	3.3737	3.2927	3.2296	3.1789	3.1373	3.0729	3.0061	2.9365	2.9005	2.8636	2.8259	2.7872	2.7475	2.7067
10	4.9646	4.1028	3.7083	3.4781	3.3258	3.2172	3.1355	3.0717	3.0204	2.9782	2.9130	2.8450	2.7740	2.7372	2.6995	2.6609	2.6211	2.5801	2.5379
12	4.7472	3.8853	3.4903	3.2592	3.1059	2.9961	2.9134	2.8486	2.7964	2.7534	2.6866	2.6168	2.5436	2.5055	2.4663	2.4259	2.3842	2.3410	2.2962
15	4.5431	3.6823	3.2874	3.0556	2.9013	2.7905	2.7066	2.6408	2.5876	2.5437	2.4753	2.4034	2.3275	2.2878	2.2468	2.2043	2.1601	2.1141	2.0658
20	4.3512	3.4928	3.0984	2.8661	2.7109	2.5990	2.5140	2.4471	2.3928	2.3479	2.2776	2.2033	2.1242	2.0825	2.0391	1.9938	1.9464	1.8963	1.8432
24	4.2597	3.4028	3.0088	2.7763	2.6206	2.5082	2.4226	2.3551	2.3002	2.2547	2.1834	2.1077	2.0267	1.9838	1.9390	1.8920	1.8424	1.7896	1.7330
30	4.1709	3.3158	2.9223	2.6896	2.5336	2.4205	2.3343	2.2662	2.2107	2.1646	2.0921	2.0148	1.9317	1.8874	1.8409	1.7918	1.7396	1.6835	1.6223
40	4.0847	3.2317	2.8388	2.6060	2.4495	2.3359	2.2490	2.1802	2.1240	2.0772	2.0035	1.9244	1.8389	1.7929	1.7444	1.6928	1.6373	1.5766	1.5089
60	4.0012	3.1504	2.7581	2.5252	2.3683	2.2541	2.1666	2.0970	2.0401	1.9926	1.9174	1.8364	1.7480	1.7001	1.6491	1.5943	1.5343	1.4673	1.3893
120	3.9201	3.0718	2.6802	2.4472	2.2898	2.1750	2.0868	2.0164	1.9588	1.9104	1.8337	1.7505	1.6587	1.6084	1.5543	1.4952	1.4290	1.3519	1.2539
∞	3.8415	2.9957	2.6049	2.3719	2.2141	2.0986	2.0096	1.9384	1.8799	1.8307	1.7522	1.6664	1.5705	1.5173	1.4591	1.3940	1.3180	1.2214	1.1000

Tabla VII (*Continuación*)
DISTRIBUCIÓN F DE FISHER

Abcisas $F_{\alpha;n_1,n_2}$ que dejan a su derecha un área α bajo la función con n_1 y n_2 grados de libertad.

Para valores de α próximos a uno se puede utilizar la relación $F_{1-\alpha;n_2,n_1} = \dfrac{1}{F_{\alpha;n_1,n_2}}$.

$\alpha = 0.025$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	647.80	799.70	864.18	899.58	921.80	937.10	948.23	956.65	963.28	968.65	976.70	984.88	993.30	997.20	1001.4	1005.5	1009.9	1014.0	1018.3
2	38.513	39.000	39.166	39.247	39.298	39.332	39.355	39.373	39.387	39.398	39.414	39.438	39.448	39.450	39.465	39.473	39.475	39.490	39.498
3	17.443	16.044	15.439	15.101	14.885	14.735	14.624	14.540	14.473	14.419	14.337	14.252	14.167	14.124	14.081	14.036	13.992	13.948	13.902
4	12.218	10.649	9.9791	9.6045	9.3645	9.1973	9.0741	8.9795	8.9031	8.8439	8.7508	8.6564	8.5600	8.5109	8.4612	8.4109	8.3604	8.3090	8.2572
5	10.007	8.4336	7.7636	7.3875	7.1463	6.9777	6.8530	6.7571	6.6809	6.6192	6.5246	6.4273	6.3286	6.2781	6.2269	6.1751	6.1225	6.0693	6.0153
6	8.8131	7.2598	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4609	5.3662	5.2687	5.1684	5.1188	5.0652	5.0125	4.9590	4.9045	4.8491
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.4150	4.3624	4.3089	4.2545	4.1989	4.1423
8	7.5709	6.0594	5.4159	5.0525	4.8173	4.6517	4.5285	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9473	3.8940	3.8398	3.7844	3.7279	3.6702
9	7.2094	5.7147	5.0750	4.7181	4.4844	4.3197	4.1971	4.1023	4.0260	3.9637	3.8682	3.7693	3.6669	3.6142	3.5604	3.5055	3.4493	3.3922	3.3328
10	6.9367	5.4563	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6209	3.5217	3.4186	3.3654	3.3110	3.2554	3.1984	3.1399	3.0798
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3735	3.2773	3.1772	3.0728	3.0187	2.9633	2.9063	2.8478	2.7874	2.7250
15	6.1995	4.7650	4.1528	3.8042	3.5764	3.4147	3.2938	3.1987	3.1227	3.0602	2.9641	2.8621	2.7559	2.7006	2.6437	2.5850	2.5242	2.4611	2.3953
20	5.8715	4.4613	3.8587	3.5146	3.2891	3.1283	3.0074	2.9128	2.8365	2.7737	2.6759	2.5731	2.4645	2.4076	2.3486	2.2873	2.2234	2.1562	2.0853
24	5.7167	4.3188	3.7211	3.3794	3.1548	2.9946	2.8738	2.7791	2.7027	2.6396	2.5411	2.4374	2.3273	2.2693	2.2090	2.1460	2.0799	2.0099	1.9353
30	5.5676	4.1821	3.5894	3.2499	3.0266	2.8667	2.7460	2.6512	2.5750	2.5112	2.4120	2.3072	2.1952	2.1359	2.0739	2.0089	1.9400	1.8664	1.7867
40	5.4239	4.0510	3.4633	3.1261	2.9037	2.7444	2.6238	2.5289	2.4519	2.3882	2.2882	2.1819	2.0677	2.0069	1.9429	1.8752	1.8028	1.7242	1.6371
60	5.2856	3.9252	3.3425	3.0077	2.7863	2.6274	2.5068	2.4117	2.3344	2.2702	2.1692	2.0613	1.9445	1.8817	1.8152	1.7440	1.6668	1.5810	1.4822
120	5.1523	3.8046	3.2269	2.8943	2.6740	2.5154	2.3948	2.2994	2.2217	2.1570	2.0548	1.9450	1.8249	1.7597	1.6899	1.6141	1.5299	1.4327	1.3104
∞	5.0239	3.6889	3.1161	2.7858	2.5665	2.4082	2.2875	2.1918	2.1136	2.0483	1.9447	1.8326	1.7085	1.6402	1.5660	1.4835	1.3883	1.2684	1.1000

TABLA VII (*Continuación*)
DISTRIBUCIÓN F DE FISHER

Abcisas $F_{\alpha;n_1,n_2}$ que dejan a su derecha un área α bajo la función con n_1 y n_2 grados de libertad.

Para valores de α próximos a uno se puede utilizar la relación $F_{1-\alpha;n_2,n_1} = \dfrac{1}{F_{\alpha;n_1,n_2}}$.

$\alpha = 0.01$

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	4052.1	4999.7	5404.1	5624.5	5763.3	5858.9	5928.5	5980.9	6021.7	6055.7	6106.5	6156.9	6208.9	6234.5	6260.5	6286.9	6312.9	6339.3	6365.7
2	98.500	99.100	99.169	99.200	99.300	99.331	99.363	99.373	99.400	99.300	99.419	99.431	99.448	99.456	99.469	99.473	99.481	99.494	99.300
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.491	27.344	27.229	27.052	26.872	26.689	26.598	26.505	26.409	26.316	26.222	26.125
4	21.198	18.000	16.695	15.977	15.519	15.207	14.975	14.799	14.659	14.546	14.373	14.198	14.020	13.929	13.838	13.745	13.652	13.558	13.463
5	16.258	13.274	12.060	11.392	10.967	10.672	10.455	10.289	10.158	10.051	9.8875	9.7223	9.5527	9.4665	9.3793	9.2910	9.2021	9.1118	9.0205
6	13.745	10.925	9.7795	9.1483	8.7457	8.4662	8.2600	8.1016	7.9761	7.8740	7.7183	7.5594	7.3958	7.3127	7.2289	7.1433	7.0566	6.9690	6.8800
7	12.246	9.5465	8.4514	7.8467	7.4604	7.1906	6.9929	6.8402	6.7250	6.6200	6.4690	6.3143	6.1554	6.0744	5.9920	5.9085	5.8236	5.7373	5.6495
8	11.259	8.6490	7.5910	7.0061	6.6316	6.3707	6.1775	6.0289	5.9106	5.8143	5.6667	5.5150	5.3591	5.2792	5.1981	5.1125	5.0316	4.9461	4.8588
9	10.562	8.0215	6.9919	6.4221	6.0570	5.8020	5.6128	5.4671	5.3512	5.2564	5.1115	4.9621	4.8080	4.7289	4.6485	4.5666	4.4831	4.3978	4.3109
10	10.044	7.5595	6.5523	5.9945	5.6359	5.3858	5.2001	5.0567	4.9424	4.8492	4.7059	4.5581	4.4054	4.3270	4.2469	4.1653	4.0818	3.9961	3.9086
12	9.3302	6.9266	5.9527	5.4120	5.0643	4.8206	4.6396	4.4994	4.3875	4.2960	4.1552	4.0097	3.8584	3.7805	3.7008	3.6192	3.5354	3.4495	3.3608
15	8.6832	6.3589	5.4169	4.8932	4.5556	4.3183	4.1415	4.0044	3.8948	3.8049	3.6663	3.5222	3.3719	3.2940	3.2141	3.1319	3.0471	2.9594	2.8684
20	8.0960	5.8490	4.9382	4.4307	4.1026	3.8714	3.6987	3.5644	3.4567	3.3682	3.2311	3.0881	2.9377	2.8563	2.7785	2.6947	2.6077	2.5168	2.4213
24	7.8229	5.6136	4.7180	4.2185	3.8951	3.6667	3.4959	3.3629	3.2560	3.1682	3.0316	2.8887	2.7380	2.6591	2.5773	2.4923	2.4035	2.3100	2.2107
30	7.5750	5.3904	4.5097	4.0180	3.6988	3.4735	3.3045	3.1726	3.0665	2.9791	2.8431	2.7002	2.5487	2.4689	2.3860	2.2992	2.2078	2.1108	2.0063
40	7.3141	5.1781	4.3125	3.8283	3.5138	3.2906	3.1238	2.9930	2.8875	2.8005	2.6648	2.5216	2.3689	2.2880	2.2034	2.1142	2.0194	1.9172	1.8047
60	7.0771	4.9774	4.1259	3.6490	3.3389	3.1187	2.9530	2.8233	2.7184	2.6318	2.4961	2.3523	2.1978	2.1154	2.0285	1.9360	1.8363	1.7263	1.6006
120	6.8509	4.7865	3.9491	3.4795	3.1735	2.9559	2.7918	2.6629	2.5586	2.4721	2.3363	2.1916	2.0346	1.9500	1.8600	1.7629	1.6557	1.5330	1.3805
∞	6.6349	4.6051	3.7816	3.3192	3.0173	2.8020	2.6394	2.5113	2.4073	2.3209	2.1848	2.0385	1.8783	1.7908	1.6964	1.5923	1.4730	1.3246	1.1000

Capítulo 20

Apéndice B: Tablas con Intervalos de Confianza

En este apéndice aparecen tabulados los intervalos de confianza más habituales.

Parámetro a estimar	**Estimador**	**Distribución**	**Intervalo**
Media de una $N(\mu,\sigma)$ σ^2 conocida	$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	Normal: $N\left(\mu,\frac{\sigma}{\sqrt{n}}\right)$	$I=\left[\overline{X}\pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$
Media de una $N(\mu,\sigma)$ σ^2 desconocida $n>30$	$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	Normal: $N\left(\mu,\frac{S}{\sqrt{n}}\right)$	$I=\left[\overline{X}\pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$
Media de una $N(\mu,\sigma)$ σ^2 desconocida $n\leq 30$	$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	$T=\frac{\overline{X}-\mu}{S/\sqrt{n}}$ sigue una t de Student con $(n-1)$ g.l.	$I=\left[\overline{X}\pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right]$
Media de cualquier población muestras grandes	$\overline{X}=\frac{\sum_{i=1}^{n}X_i}{n}$	Normal: $N\left(\mu,\frac{S}{\sqrt{n}}\right)$	$I=\left[\overline{X}\pm z_{\alpha/2}\frac{S}{\sqrt{n}}\right]$
p de Binomial	$\overline{P}=\frac{\text{número de éxitos}}{\text{número de ensayos}}$	Normal: $N\left(\overline{P},\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right)$	$I=\left[\overline{P}\pm z_{\alpha/2}\sqrt{\frac{\overline{P}(1-\overline{P})}{n}}\right]$
λ de Poisson	$\overline{\lambda}=\frac{\sum_{i=1}^{n}X_i}{n}$	Normal: $N\left(\overline{\lambda},\sqrt{\frac{\overline{\lambda}}{n}}\right)$	$I=\left[\overline{\lambda}\pm z_{\alpha/2}\sqrt{\frac{\overline{\lambda}}{n}}\right]$
Diferencia de medias poblaciones normales σ_1^2 y σ_2^2 conocidas	$\overline{X_1}-\overline{X_2}$	$N\left(\mu_1-\mu_2,\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right)$	$I=\left[(\overline{X_1}-\overline{X_2})\pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]$
Diferencia de medias poblaciones normales σ_1^2 y σ_2^2 desconocidas $n_1+n_2>30$ $(n_1\simeq n_2)$	$\overline{X_1}-\overline{X_2}$	$N\left(\mu_1-\mu_2,\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right)$	$I=\left[(\overline{X_1}-\overline{X_2})\pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$
Diferencia de medias poblaciones normales σ_1^2 y σ_2^2 desconocidas $\sigma_1=\sigma_2$ (muestras pequeñas)	$\overline{X_1}-\overline{X_2}$	$T=\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$ sigue una t de Student con (n_1+n_2-2) g.l. donde $S_p^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}$	$I=\left[(\overline{X_1}-\overline{X_2})\pm t_{\alpha/2,n_1+n_2-2}S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}\right]$
Diferencia de medias poblaciones normales σ_1^2 y σ_2^2 desconocidas $\sigma_1\neq\sigma_2$ (muestras pequeñas)	$\overline{X_1}-\overline{X_2}$	$T=\frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$ sigue una t de Student con f g.l. donde $f=\frac{\left(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1+1}+\frac{(S_2^2/n_2)^2}{n_2+1}}-2$	$I=\left[(\overline{X_1}-\overline{X_2})\pm t_{\alpha/2,f}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}\right]$

Parámetro a estimar	Estimador	Distribución	Intervalo
Diferencia de medias poblaciones no normales muestras grandes	$\overline{X_1} - \overline{X_2}$	$N\left(\mu_1 - \mu_2, \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right)$	$I = \left[(\overline{X_1} - \overline{X_2}) \pm z_{\alpha/2}\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}\right]$
Diferencia de proporciones muestras grandes	$\overline{P_1} - \overline{P_2}$	$N\left(p_1 - p_2, \sqrt{\frac{\overline{P_1}(1-\overline{P_1})}{n_1} + \frac{\overline{P_2}(1-\overline{P_2})}{n_2}}\right)$	$I = \left[(\overline{P_1} - \overline{P_2}) \pm z_{\alpha/2}\sqrt{\frac{\overline{P_1}(1-\overline{P_1})}{n_1} + \frac{\overline{P_2}(1-\overline{P_2})}{n_2}}\right]$

Parámetro a estimar	Estimador	Distribución	Intervalo
Varianza de una $N(\mu, \sigma)$	S^2	$\chi^2_{n-1} = \frac{(n-1)S^2}{\sigma^2}$	$I = \left[\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}\right]$
Razón de varianzas dos poblaciones normales	S_1^2/S_2^2	$F_{n_1-1,n_2-1} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$	$I = \left[\frac{S_1^2}{S_2^2}\frac{1}{F_{\alpha/2;n_1-1,n_2-1}}, \frac{S_1^2}{S_2^2}F_{\alpha/2;n_2-1,n_1-1}\right]$

Capítulo 21

Apéndice C: Tablas con Contrastes de Hipótesis

En este apéndice aparecen tabulados los contrastes de hipótesis más habituales.

CONTRASTE PARA LA MEDIA DE UNA POBLACIÓN						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL σ^2 conocida	$\mu = \mu_0$	$\mu \neq \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$	Normal	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{\sigma/\sqrt{n}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{\sigma/\sqrt{n}} > z_{\alpha/2}$
UNILATERAL σ^2 conocida	$\mu \leq \mu_0$	$\mu > \mu_0$			$\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} \leq z_{\alpha}$	$\dfrac{\overline{x} - \mu_0}{\sigma/\sqrt{n}} > z_{\alpha}$
BILATERAL σ^2 desconocida $n > 30$	$\mu = \mu_0$	$\mu \neq \mu_0$	$z = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$	Normal	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{s/\sqrt{n}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{s/\sqrt{n}} > z_{\alpha/2}$
UNILATERAL σ^2 desconocida $n > 30$	$\mu \leq \mu_0$	$\mu > \mu_0$			$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq z_{\alpha}$	$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} > z_{\alpha}$
BILATERAL σ^2 desconocida $n \leq 30$	$\mu = \mu_0$	$\mu \neq \mu_0$	$t = \dfrac{\overline{x} - \mu_0}{s/\sqrt{n}}$	t de Student	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{s/\sqrt{n}} \leq t_{\alpha/2,n-1}$	$\dfrac{\lvert\overline{x} - \mu_0\rvert}{s/\sqrt{n}} > t_{\alpha/2,n-1}$
UNILATERAL σ^2 desconocida $n \leq 30$	$\mu \leq \mu_0$	$\mu > \mu_0$			$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} \leq t_{\alpha,n-1}$	$\dfrac{\overline{x} - \mu_0}{s/\sqrt{n}} > t_{\alpha,n-1}$

CONTRASTE DE UNA PROPORCIÓN						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL	$p = p_0$	$p \neq p_0$	$z = \dfrac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}}$	Normal	$\dfrac{\lvert\overline{p} - p_0\rvert}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{p} - p_0\rvert}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} > z_{\alpha/2}$
UNILATERAL	$p \leq p_0$	$p > p_0$			$\dfrac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} \leq z_{\alpha}$	$\dfrac{\overline{p} - p_0}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}} > z_{\alpha}$

CONTRASTE DE LA VARIANZA DE UNA POBLACIÓN NORMAL						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL	$\sigma^2 = \sigma_0^2$	$\sigma^2 \neq \sigma_0^2$	$\chi^2 = \dfrac{(n-1)s^2}{\sigma_0^2}$	χ^2	$\dfrac{(n-1)s^2}{\sigma_0^2} \in [\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}]$	$\dfrac{(n-1)s^2}{\sigma_0^2} \notin [\chi^2_{1-\alpha/2,n-1}, \chi^2_{\alpha/2,n-1}]$
UNILATERAL	$\sigma^2 \leq \sigma_0^2$	$\sigma^2 > \sigma_0^2$			$\dfrac{(n-1)s^2}{\sigma_0^2} \leq \chi^2_{\alpha,n-1}$	$\dfrac{(n-1)s^2}{\sigma_0^2} > \chi^2_{\alpha,n-1}$

CONTRASTE PARA LA IGUALDAD DE MEDIAS DE DOS POBLACIONES NORMALES						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL σ^2 conocida	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Normal	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha/2}$
UNILATERAL σ^2 conocida	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$			$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha}$	$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} > z_{\alpha}$
BILATERAL σ^2 desconocida $n_1 + n_2 > 30,\ (n_1 \simeq n_2)$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$z = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	Normal	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_{\alpha/2}$
UNILATERAL σ^2 desconocida $n_1 + n_2 > 30,\ (n_1 \simeq n_2)$	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$			$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq z_{\alpha}$	$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > z_{\alpha}$
BILATERAL σ^2 desconocida, $\sigma_1 = \sigma_2$ $n_1 + n_2 \leq 30$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$t = \dfrac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$	t de Student	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha/2,n_1+n_2-2}$	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha/2,n_1+n_2-2}$
UNILATERAL σ^2 desconocida, $\sigma_1 = \sigma_2$ $n_1 + n_2 \leq 30$	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$s_p^2 = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$		$\dfrac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \leq t_{\alpha,n_1+n_2-2}$	$\dfrac{\overline{x_1} - \overline{x_2}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} > t_{\alpha,n_1+n_2-2}$
BILATERAL σ^2 desconocida, $\sigma_1 \neq \sigma_2$ $n_1 + n_2 \leq 30$	$\mu_1 = \mu_2$	$\mu_1 \neq \mu_2$	$t = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	t de Student	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha/2,f}$	$\dfrac{\lvert\overline{x_1} - \overline{x_2}\rvert}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\alpha/2,f}$
UNILATERAL σ^2 desconocida, $\sigma_1 \neq \sigma_2$ $n_1 + n_2 \leq 30$	$\mu_1 \leq \mu_2$	$\mu_1 > \mu_2$	$f = \dfrac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{(s_1^2/n_1)^2}{n_1+1}+\frac{(s_2^2/n_2)^2}{n_2+1}} - 2$		$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \leq t_{\alpha,f}$	$\dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} > t_{\alpha,f}$

CONTRASTE DE LA IGUALDAD ENTRE DOS PROPORCIONES						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL	$p_1 = p_2$	$p_1 \neq p_2$	$z = \dfrac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}}$	Normal	$\dfrac{\lvert\overline{p_1} - \overline{p_2}\rvert}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq z_{\alpha/2}$	$\dfrac{\lvert\overline{p_1} - \overline{p_2}\rvert}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} > z_{\alpha/2}$
UNILATERAL	$p_1 \leq p_2$	$p_1 > p_2$			$\dfrac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} \leq z_{\alpha}$	$\dfrac{\overline{p_1} - \overline{p_2}}{\sqrt{\frac{\overline{p_1}(1-\overline{p_1})}{n_1} + \frac{\overline{p_2}(1-\overline{p_2})}{n_2}}} > z_{\alpha}$

CONTRASTE DE LA IGUALDAD DE VARIANZAS DE DOS POBLACIONES NORMALES						
Tipo de contraste	H_0	H_1	**Estadístico**	**Distribución**	**Se acepta si**	**Se rechaza si**
BILATERAL	$\sigma_1^2 = \sigma_2^2$	$\sigma_1^2 \neq \sigma_2^2$	$F = \dfrac{s_1^2}{s_2^2}$	F de Fisher	$\dfrac{s_1^2}{s_2^2} \in [F_{1-\alpha/2,n_1-1,n_2-1}, F_{\alpha/2,n_1-1,n_2-1}]$	$\dfrac{s_1^2}{s_2^2} \notin [F_{1-\alpha/2,n_1-1,n_2-1}, F_{\alpha/2,n_1-1,n_2-1}]$
UNILATERAL	$\sigma_1^2 \leq \sigma_2^2$	$\sigma_1^2 > \sigma_2^2$			$\dfrac{s_1^2}{s_2^2} \leq F_{\alpha,n_1-1,n_2-1}$	$\dfrac{s_1^2}{s_2^2} > F_{\alpha,n_1-1,n_2-1}$

Este libro se ha escrito utilizado LaTeX, software libre disponible bajo licencia LPPL.

La Estadística ha demostrado aportar un conocimiento esencial para la formación de los estudiantes de la Licenciatura en Física. Estamos convencidos de que este tipo de conocimiento es básico para cualquier estudiante de ciencias.

Aunque la bibliografía en este campo es extensa, hemos considerado oportuno redactar un libro restringido a los contenidos específicos que se incluyen en un curso introductorio de Estadística. Pretendemos así delimitar, y en lo posible simplificar, el trabajo del estudiante, mostrándole de forma precisa los conceptos más fundamentales. Una vez consolidados estos conceptos, esperamos que los estudiantes de ciencias encuentren menos dificultades para aprender y profundizar en las técnicas estadísticas más avanzadas que son de uso común en el trabajo diario del científico.

Los autores son astrofísicos, Profesores Titulares del Departamento de Astrofísica y CC. de la Atmósfera de la Facultad de CC. Físicas de la Universidad Complutense de Madrid. Los tres han sido profesores de la asignatura "Estadística", introducida en el plan de estudios de la Licenciatura en Física en el año 1995.

ISBN 978-84-691-8981-8
9 788469 189818