

Curso Básico de Estadística

Santiago Angulo Díaz-Parreño (sangulo@ceu.es)
José Miguel Cárdenas Rebollo (cardenas@ceu.es)
José Rojo Montijano (jrojo.eps@ceu.es)
Anselmo Romero Limón (arlimon@ceu.es)
Alfredo Sánchez Alberca (asalber@ceu.es)



CEU
*Universidad
San Pablo*

Curso 2011-2012
©Copyleft

Curso básico de estadística

Alfredo Sánchez Alberca (asalber@gmail.com).

Esta obra está bajo una licencia Reconocimiento-No comercial-Compartir bajo la misma licencia 2.5 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-sa/2.5/es/> o envíe una carta a Creative Commons, 171 Second Street, Suite 300, San Francisco, California 94105, USA.

Con esta licencia eres libre de:

- Copiar, distribuir y mostrar este trabajo.
- Realizar modificaciones de este trabajo.

Bajo las siguientes condiciones:

- **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
- **No comercial.** No puede utilizar esta obra para fines comerciales.
- **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.
- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Contenidos

- 1 Introducción a la Estadística
- 2 Estadística Descriptiva
- 3 Regresión y Correlación
- 4 Teoría de la Probabilidad
- 5 Variables Aleatorias
- 6 Estimación de Parámetros
- 7 Contraste de hipótesis

- 1 Introducción a la Estadística
 - La estadística como herramienta científica
 - Población y muestra
 - Muestreo

¿Qué es la estadística?

Definición (Estadística)

La *estadística* es una rama de las matemáticas que se encarga de la recogida, análisis e interpretación de datos.

La estadística es imprescindible en cualquier disciplina científica o técnica donde se manejen datos, especialmente si son grandes volúmenes de datos, como por ejemplo, la física, la química, la medicina y las ciencias biosanitarias, pero también en la economía, la psicología o las ciencias sociales.

Pero,

¿Por qué es necesaria la estadística?

La variabilidad de nuestro mundo

El científico trata de estudiar el mundo que le rodea; un mundo que está lleno de variaciones que dificultan la determinación del comportamiento de las cosas.

¡La variabilidad del mundo real es el origen de la estadística!

La estadística actúa como disciplina puente entre la realidad del mundo y los modelos matemáticos que tratan de explicarla, proporcionando una metodología para evaluar las discrepancias entre la realidad y los modelos teóricos.

Esto la convierte en una herramienta indispensable en las ciencias aplicadas que requieran el análisis de datos y el diseño de experimentos.

Población estadística

Definición (Población)

Una *población* es un conjunto de elementos definido por una o más características que tienen todos los elementos, y sólo ellos. Cada elemento de la población se llama *individuo*.

Definición (Tamaño poblacional)

El número de individuos de una población se conoce como *tamaño poblacional* y se representa como N .

A veces, no todos los elementos de la población están accesibles para su estudio. Entonces se distingue entre:

Población Teórica: Conjunto de elementos a los que se quiere extrapolar los resultados del estudio.

Población Estudiada: Conjunto de elementos realmente accesibles en el estudio.

Inconvenientes en el estudio de la población

El científico estudia un determinado fenómeno en una población para comprenderlo, obtener conocimiento sobre el mismo, y así poder controlarlo.

Pero, para tener un conocimiento completo de la población *es necesario estudiar todos los individuos de la misma*.

Sin embargo, esto no siempre es posible por distintos motivos:

- El tamaño de la población es infinito, o bien es finito pero demasiado grande.
- Las pruebas a que se someten los individuos son destructivas.
- El coste, tanto de dinero como de tiempo, que supondría estudiar a todos los individuos es excesivo.

Muestra estadística

Cuando no es posible o conveniente estudiar todos los individuos de la población, se estudia sólo una parte de la misma.

Definición (Muestra)

Una *muestra* es un subconjunto de la población.

Definición (Tamaño muestral)

Al número de individuos que componen la muestra se le llama *tamaño muestral* y se representa por n .

Habitualmente, el estudio de una población se realiza a partir de muestras extraídas de dicha población.

Generalmente, el estudio de la muestra sólo aporta conocimiento aproximado de la población. Pero en muchos casos es *suficiente*.

Determinación del tamaño muestral

Una de las preguntas más interesantes que surge inmediatamente es:

¿cuántos individuos es necesario tomar en la muestra para tener un conocimiento aproximado pero suficiente de la población?

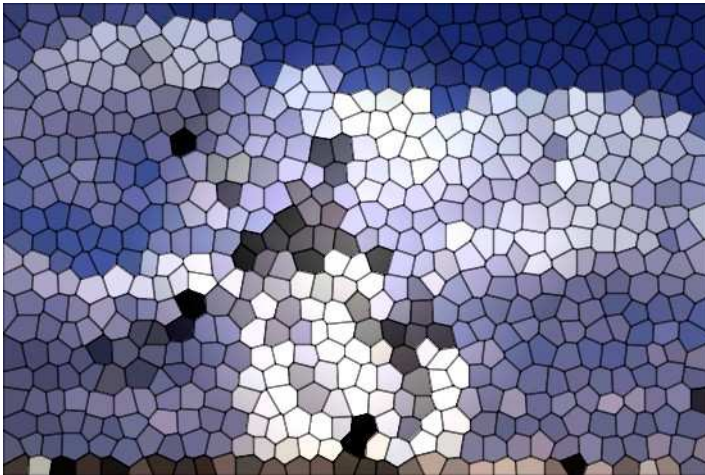
La respuesta depende de muchos factores, como la variabilidad de la población o la fiabilidad deseada para las extrapolaciones que se hagan hacia la población. Por desgracia no se podrá responder hasta casi el final del curso.

En general, cuantos más individuos haya en la muestra, más fiables serán las conclusiones sobre la población, pero también será más lento y costoso el estudio.

Determinación del tamaño muestral

Muestra pequeña de los píxeles de una imagen

¿De qué imagen se trata?



¡Con una muestra pequeña es difícil averiguar el contenido de la imagen!

Determinación del tamaño muestral

Muestra mayor de los píxeles de una imagen

¿De qué imagen se trata?



¡Con una muestra mayor es más fácil averiguar el contenido de la imagen!

Determinación del tamaño muestral

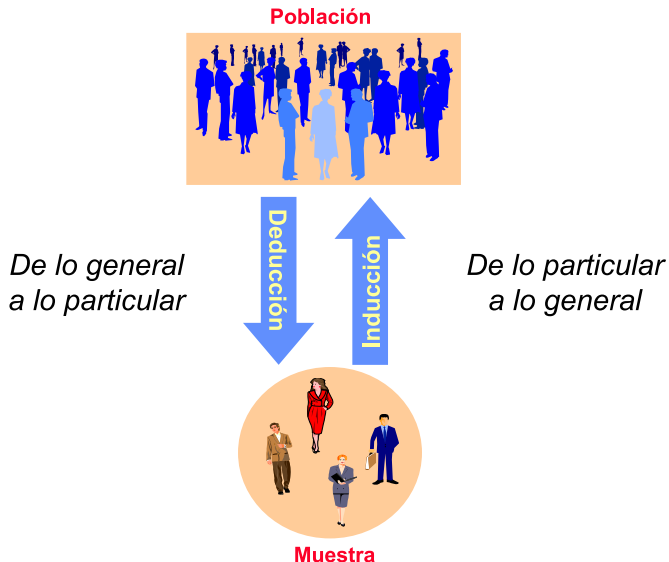
Población completa de los píxeles de una imagen

Y aquí está la población completa



¡No es necesario conocer todos los píxeles para averiguar la imagen!

Tipos de razonamiento



Tipos de razonamiento

Características de la deducción: Si las premisas son ciertas, garantiza la certeza de las conclusiones (es decir, si algo se cumple en la población, también se cumple en la muestra). Sin embargo, *¡no aporta conocimiento nuevo!*

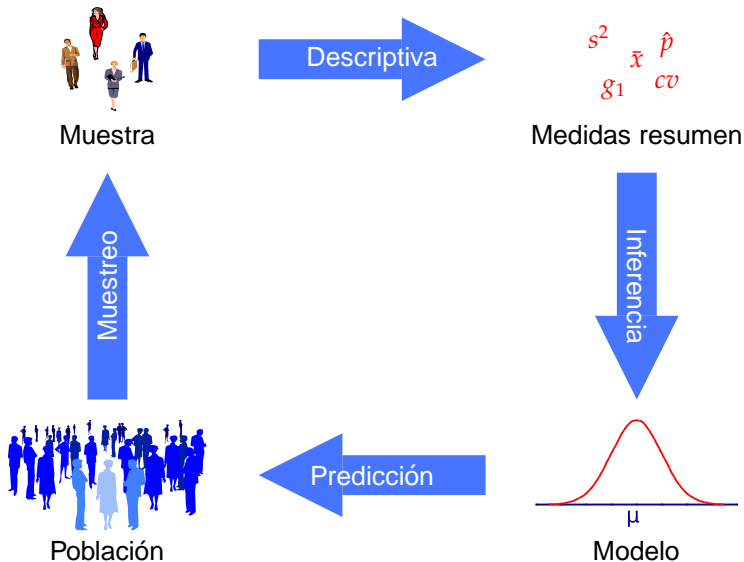
Características de la inducción: No garantiza la certeza de las conclusiones (si algo se cumple en la muestra, puede que no se cumpla en la población, así que ¡cuidado con las extrapolaciones!), pero *¡es la única forma de generar conocimiento nuevo!*

La estadística se apoya fundamentalmente en el razonamiento inductivo ya que utiliza la información obtenida a partir de muestras para sacar conclusiones sobre las poblaciones.

Normalmente un estudio estadístico pasa por 4 etapas:

- 1 El estudio de una población comienza por la selección de una muestra representativa de la misma. De esto se encarga el **muestreo**.
- 2 El siguiente paso consiste en estudiar las muestras extraídas y obtener resultados numéricos que resuman la información contenida en las mismas. De esto se encarga la **estadística descriptiva**.
- 3 La información obtenida es proyectada sobre un modelo matemático que intenta reflejar el comportamiento de la población. Tras construir el modelo, se realiza una crítica del mismo para validarlo. De todo esto se encarga la **inferencia estadística**.
- 4 Finalmente, el modelo validado nos permite hacer suposiciones y predicciones sobre la población de partida con cierta confianza.

El ciclo estadístico



Muestreo

Definición (Muestreo)

El proceso de selección de los elementos que compondrán una muestra se conoce como *muestreo*.



Población



Muestra

Para que una muestra refleje información fidedigna sobre la población global debe ser representativa de la misma.

El objetivo es obtener una muestra representativa de la población.

Modalidades de muestreo

Existen muchas técnicas de muestreo pero se pueden agrupar en dos categorías:

Muestreo Aleatorio Elección aleatoria de los individuos de la muestra. Todos tienen la misma probabilidad de ser elegidos (*equiprobabilidad*).

Muestreo No Aleatorio: Los individuos se eligen de forma no aleatoria.

Sólo las técnicas aleatorias evitan el sesgo de selección, y por tanto, garantizan la representatividad de la muestra extraída, y en consecuencia la validez de la inferencia.

Las técnicas no aleatorias no sirven para hacer generalizaciones, ya que no garantizan la representatividad de la muestra. Sin embargo, son menos costosas y pueden utilizarse en estudios exploratorios.

Muestreo aleatorio simple

Dentro de las modalidades de muestreo aleatorio, el tipo más conocido es el *muestreo aleatorio simple*, caracterizado por:

- Todos los individuos de la población tienen la misma probabilidad de ser elegidos para la muestra.
- La selección de individuos es con reemplazamiento (y por tanto no se altera la población de partida).
- Las sucesivas selecciones de un individuo son independientes.

La única forma de realizar un muestreo aleatorio es asignar un número a cada individuo de la población (*censo*) y realizar un sorteo aleatorio.

Estadística Descriptiva

- Variables estadísticas
- Distribución de frecuencias
- Representaciones gráficas
- Estadísticos muestrales
- Estadísticos de posición
- Estadísticos de dispersión
- Estadísticos de forma
- Transformaciones de variables

La estadística descriptiva es la parte de la estadística encargada de representar, analizar y resumir la información contenida en la muestra.

Tras el proceso de muestreo, es la siguiente etapa de todo estudio estadístico y suele consistir en:

- 1 Clasificar, agrupar y ordenar los datos de la muestra.
- 2 Representar dichos datos gráficamente y en forma de tablas.
- 3 Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

Su poder inferencial es mínimo, por lo que nunca deben sacarse conclusiones sobre la población a partir de las medidas resumen que aporta la estadística descriptiva.

Variables estadísticas y atributos

La característica objeto de estudio puede ser de dos tipos:

Atributos: De carácter cualitativo.

Variables estadísticas: De carácter cuantitativo.

A su vez, los atributos se dividen en:

Nominales: No existe un orden entre las modalidades.

Ejemplo: El color de ojos o de pelo.

Ordinales: Existe un orden entre las modalidades.

Ejemplo: El grado de gravedad de un paciente o la calificación de un curso.

Y las variables estadísticas en:

Discretas: Reciben valores aislados.

Ejemplo: El número de hijos o el número de coches.

Continuas: Pueden recibir cualquier valor de un intervalo.

Ejemplo: El peso o la estatura.

La matriz de datos

Las variables o atributos a estudiar se medirán en cada uno de los individuos de la muestra, obteniendo un conjunto de datos que suele organizarse en forma de matriz que se conoce como **matriz de datos**.

En esta matriz cada columna contiene la información de una variable y cada fila la información de un individuo.

Ejemplo

	Edad (años)	Sexo	Peso (Kg)	Altura (cm)
José Luis Martínez	18	H	85	179
Rosa Díaz	32	M	65	173
Javier García	24	H	71	181
Carmen López	35	M	65	170
Marisa López	46	M	51	158
Antonio Ruiz	68	H	66	174

Clasificación de la muestra

El estudio de una variable estadística comienza por medir la variable en los individuos de la muestra y clasificar los valores obtenidos.

Existen dos formas de clasificar estos valores:

- Sin agrupar:** Ordenar todos los valores obtenidos en la muestra de menor a mayor. Se utiliza con atributos y variables discretas con pocos valores diferentes.
- Agrupados:** Agrupar los valores en clases (intervalos) y ordenar dichas clases de menor a mayor. Se utiliza con variables discretas con muchos valores diferentes, y con variables continuas.

Clasificación de la muestra



$X = \text{Estatura}$

Clasificación



Recuento de frecuencias



$X = \text{Estatura}$

Frecuencias



Definición (Frecuencias muestrales)

Dada una muestra de tamaño n de una variable X , para cada valor de la variable x_i observado en la muestra, se define

- **Frecuencia absoluta n_i** : Es el número de individuos de la muestra que presentan el valor x_i .
- **Frecuencia relativa f_i** : Es la proporción de individuos de la muestra que presentan el valor x_i .

$$f_i = \frac{n_i}{n}$$

- **Frecuencia absoluta acumulada N_i** : Es el número de individuos de la muestra que presentan un valor menor o igual que x_i .

$$N_i = n_1 + \cdots + n_i$$

- **Frecuencia relativa acumulada F_i** : Es la proporción de individuos de la muestra que presentan un valor menor o igual que x_i .

$$F_i = \frac{N_i}{n}$$

Tabla de frecuencias

Al conjunto de valores observados en la muestra junto a sus respectivas frecuencias se le denomina **distribución muestral de frecuencias** y suele representarse mediante una **tabla de frecuencias**.

Valores de X	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
x_1	n_1	f_1	N_1	F_1
\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_i	f_i	N_i	F_i
\vdots	\vdots	\vdots	\vdots	\vdots
x_k	n_k	f_k	N_k	F_k

Tabla de frecuencias

Ejemplo de datos sin agrupar

En una encuesta a 25 matrimonios sobre el número de hijos que tenían se obtuvieron los siguientes datos:

1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2,
0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
2	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
Σ	25	1		

Tabla de frecuencias

Ejemplo de datos agrupados

Se ha medido la estatura (en cm) de 30 universitarios obteniendo:

179, 173, 181, 170, 158, 174, 172, 166, 194, 185,
162, 187, 198, 177, 178, 165, 154, 188, 166, 171,
175, 182, 167, 169, 172, 186, 172, 176, 168, 187.

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i	N_i	F_i
(150,160]	2	0,07	2	0,07
(160,170]	8	0,27	10	0,34
(170,180]	11	0,36	21	0,70
(180,190]	7	0,23	28	0,93
(190,200]	2	0,07	30	1
Σ	30	1		

Construcción de clases

Cada intervalo de agrupación de datos se denomina **clase** y el centro del intervalo se llama **marca de clase**.

A la hora de agrupar los datos en clases hay que tener en cuenta lo siguiente:

- El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo a la raíz cuadrada del tamaño muestral \sqrt{n} .
- Los intervalos no deben solaparse y deben cubrir todo el rango de valores. Es indiferente si se abren por la izquierda y se cierran por la derecha o al revés.
- El valor más pequeño debe caer dentro del primer intervalo y el más grande dentro del último.

Tabla de frecuencias

Ejemplo con un atributo

Los grupos sanguíneos de una muestra de 30 personas son:

A, B, B, A, AB, 0, 0, A, B, B, A, A, A, AB,
A, A, A, B, 0, B, B, B, A, A, A, 0, A, AB, 0.

La tabla de frecuencias asociada a esta muestra es

x_i	n_i	f_i
0	5	0,16
A	14	0,47
B	8	0,27
AB	3	0,10
Σ	30	1

¿Por qué en este caso no se construyen las columnas de frecuencias acumuladas?

Representaciones gráficas

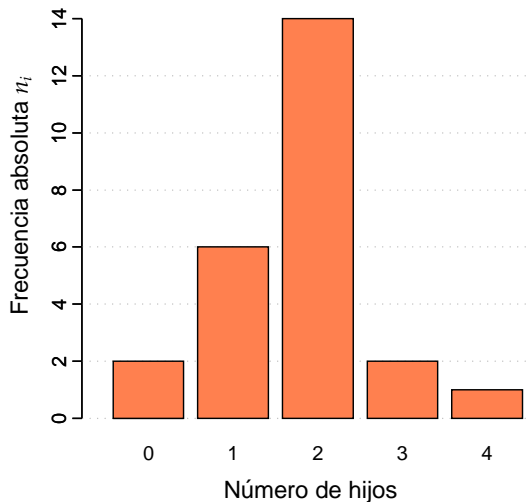
También es habitual representar la distribución muestral de frecuencias de forma gráfica. Dependiendo del tipo de variable y de si se han agrupado o no los datos, se utilizan distintos tipos de gráficos:

- **Diagrama de barras:** Consiste en un diagrama sobre el plano cartesiano en el que en el eje X se representan los valores de la variable y en el eje Y las frecuencias. Sobre cada valor de la variable se levanta una barra de altura la correspondiente frecuencia. Se utiliza con variables discretas no agrupadas.
- **Histograma:** Es similar a un diagrama de barras pero representando en el eje X las clases en que se agrupan los valores de la variable y levantando las barras sobre todo el intervalo de manera que las barras están pegadas unas a otras. Se utiliza con variables discretas agrupadas y con variables continuas.
- **Diagrama de sectores:** Consiste en un círculo dividido en sectores de área proporcional a la frecuencia de cada valor de la variable. Se utiliza sobre todo con atributos.

En cada uno de los diagramas pueden representarse los distintos tipos de frecuencias, siempre que estas existan.

Diagrama de barras de frecuencias absolutas

Datos sin agrupar



Polígono de frecuencias absolutas

Datos sin agrupar

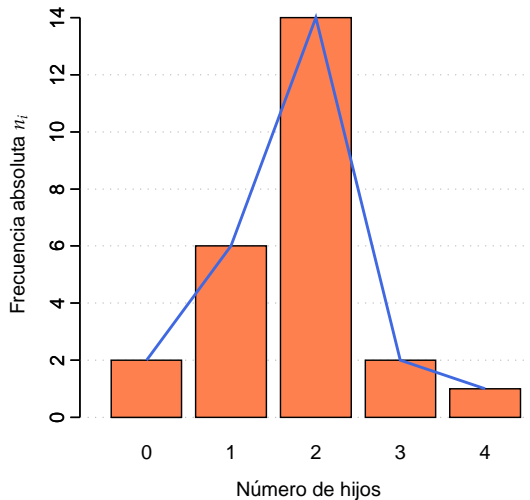
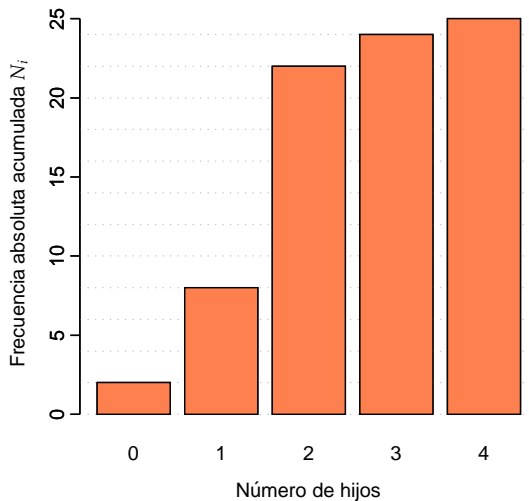


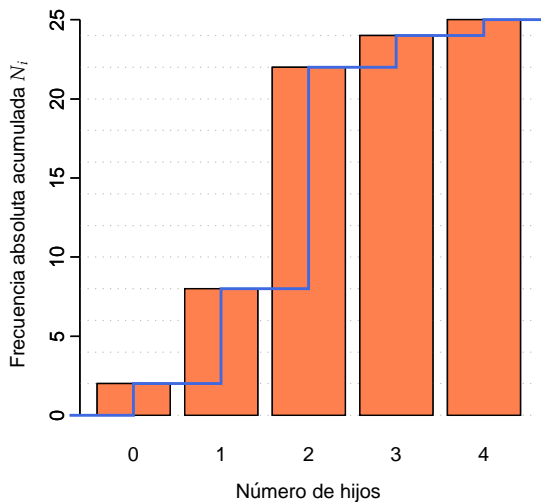
Diagrama de barras de frecuencias acumuladas

Datos sin agrupar



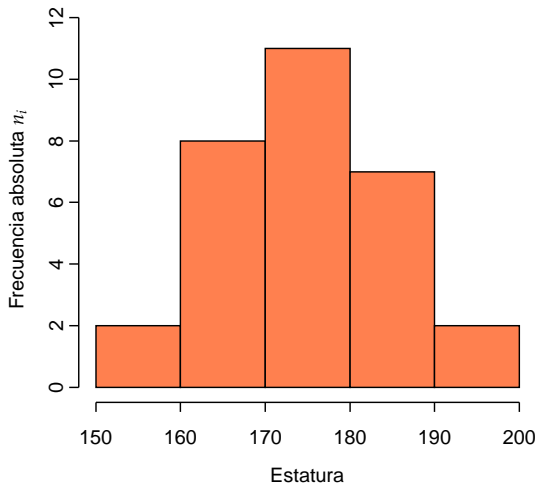
Polígono de frecuencias absolutas acumuladas

Datos sin agrupar



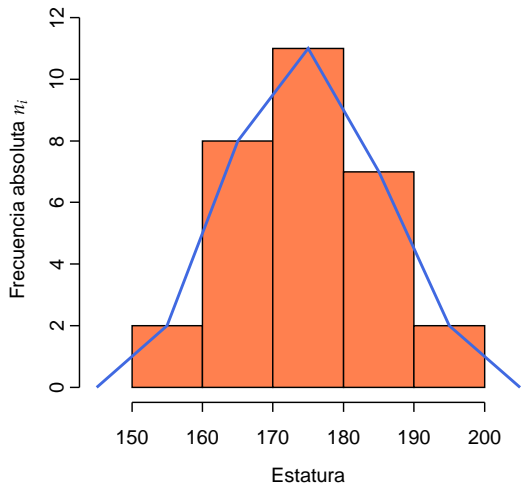
Histograma de frecuencias absolutas

Datos agrupados



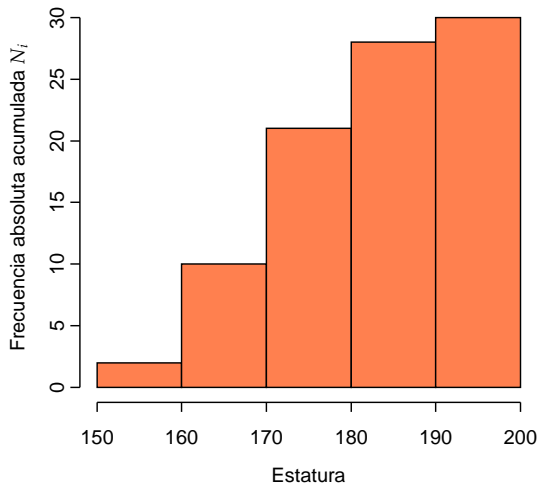
Polígono de frecuencias absolutas

Datos agrupados



Histograma de frecuencias absolutas acumuladas

Datos agrupados



Polígono de frecuencias absolutas acumuladas

Datos agrupados

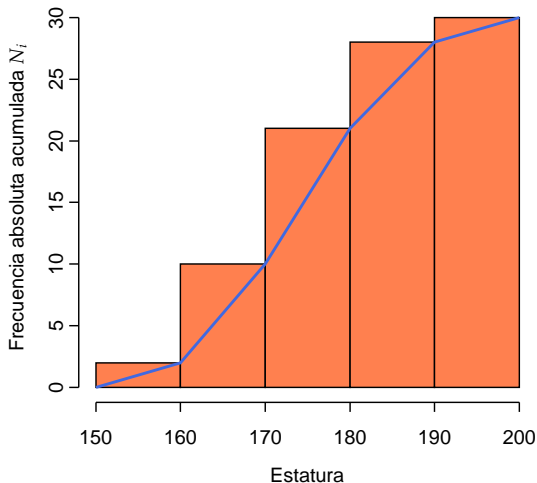
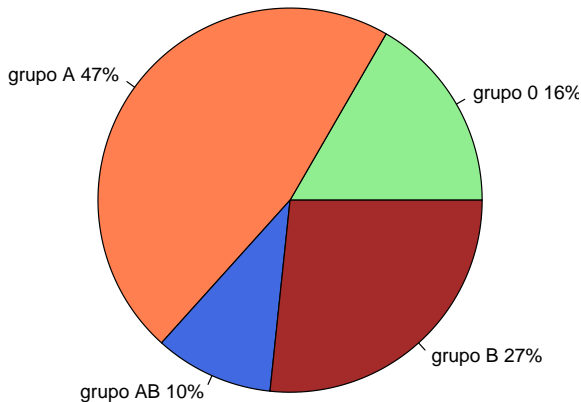


Diagrama de sectores

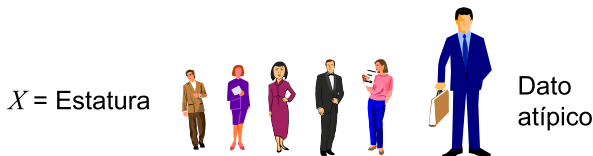
Atributos

Distribución del grupo sanguíneo



Datos atípicos

Uno de los principales problemas de las muestras son los datos atípicos. Los **datos atípicos** son valores de la variable que se diferencian mucho del resto de los valores.



Es muy importante detectar los datos atípicos antes de realizar cualquier análisis de los datos, pues *suelen distorsionar los resultados*.

Aparecen siempre en los extremos de la distribución, aunque más adelante veremos un diagrama para detectarlos.

Tratamiento de los datos atípicos

Cuando trabajemos con muestras grandes, los datos atípicos tienen menor influencia y pueden dejarse en la muestra.

Cuando trabajemos con muestras pequeñas tenemos varias opciones:

- Eliminarlo: Siempre que estemos seguros de que se trata de un error de medida.
- Sustituirlo: Si se trata de un individuo real pero que no concuerda con el modelo de distribución de la población. En tal caso se suele reemplazar por el mayor o menor dato no atípico.
- Dejarlo: Si se trata de un individuo real aunque no concuerde con el modelo de distribución. En tal caso se suele modificar el modelo de distribución supuesto.

Estadísticos muestrales

La tabla de frecuencias sintetiza la información de la variable estudiada en la muestra, pero en muchas ocasiones es insuficiente para describir determinados aspectos de la distribución.

Para describir adecuadamente el comportamiento de la variable se calculan unas medidas llamadas **estadísticos muestrales** que son indicadores de distintos aspectos de la distribución muestral.

Los estadísticos se clasifican en tres grupos:

Estadísticos de Posición: Miden en torno a qué valores se agrupan los datos y cómo se reparten en la distribución.

Estadísticos de Dispersión: Miden la heterogeneidad de los datos.

Estadísticos de Forma: Miden aspectos de la forma que tiene la distribución de los datos, como la simetría o el apuntamiento.

Estadísticos de posición

Pueden ser de dos tipos:

Estadísticos de Tendencia Central: Determinan valores alrededor de los cuales se agrupa la distribución. Estas medidas suelen utilizarse como valores representativos de la muestra. Las más importantes son:

- Media aritmética
- Mediana
- Moda

Otros estadísticos de Posición: Dividen la distribución en partes con el mismo número de observaciones. Las más importantes son:

- Cuantiles: Cuartiles, Deciles, Percentiles.

Definición (Media aritmética muestral \bar{x})

La *media aritmética muestral* de una variable X es la suma de los valores observados en la muestra dividida por el tamaño muestral

$$\bar{x} = \frac{\sum x_i}{n}$$

A partir de la tabla de frecuencias puede calcularse como:

$$\bar{x} = \frac{\sum x_i n_i}{n} = \sum x_i f_i$$

En la mayoría de los casos, la media aritmética es la medida que mejor representa a la muestra.

¡Ojo! No puede calcularse para atributos.

Cálculo de la media aritmética

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos tenemos

$$\bar{x} = \frac{1 + 2 + 4 + 2 + 2 + 2 + 3 + 2 + 1 + 1 + 0 + 2 + 2}{25} + \frac{0 + 2 + 2 + 1 + 2 + 2 + 3 + 1 + 2 + 2 + 1 + 2}{25} = \frac{44}{25} = 1,76 \text{ hijos.}$$

o bien, desde la tabla de frecuencias

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
0	2	0,08	0	0
1	6	0,24	6	0,24
2	14	0,56	28	1,12
3	2	0,08	6	0,24
4	1	0,04	4	0,16
Σ	25	1	44	1,76

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1,76 \quad \bar{x} = \sum x_i f_i = 1,76.$$

Es decir, el número de hijos que mejor representa a la muestra es 1,76

Cálculo de la media aritmética

Ejemplo con datos agrupados

En el ejemplo anterior de las estaturas se tiene

$$\bar{x} = \frac{179 + 173 + \dots + 187}{30} = 175,07 \text{ cm.}$$

o bien, desde la tabla de frecuencias utilizando las marcas de clase:

X	x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
(150, 160]	155	2	0,07	310	10,33
(160, 170]	165	8	0,27	1320	44,00
(170, 180]	175	11	0,36	1925	64,17
(180, 190]	185	7	0,23	1295	43,17
(190, 200]	195	2	0,07	390	13
Σ		30	1	5240	174,67

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174,67 \quad \bar{x} = \sum x_i f_i = 174,67.$$

Al agrupar datos el cálculo de estadísticos desde la tabla puede diferir ligeramente del valor real obtenido directamente desde la muestra, ya que no se trabaja con los datos reales sino con los representantes de las

Media ponderada

En algunos casos, los valores de la muestra no tienen la misma importancia. En este caso la media aritmética no es una buena medida de representatividad ya que en ella todos los valores de la muestra tienen el mismo peso. En este caso es mucho mejor utilizar otra medida de tendencia central conocida como media ponderada.

Definición (Media ponderada muestral \bar{x}_p)

Dada una muestra de n valores en la que cada valor x_i tiene asociado un peso p_i , la *media ponderada muestral* de la variable X es la suma de los productos de cada valor observado en la muestra por su peso, dividida por la suma de todos los pesos

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i}$$

A partir de la tabla de frecuencias puede calcularse como:

$$\bar{x}_p = \frac{\sum x_i p_i n_i}{\sum p_i}$$

Cálculo de la media ponderada

Supongase que un alumno quiere calcular la nota media de las asignaturas de un curso.

Asignatura	Créditos	Nota
Matemáticas	6	5
Lengua	4	3
Química	8	6

La media aritmética vale

$$\bar{x} = \frac{\sum x_i}{n} = \frac{5 + 3 + 6}{3} = 4,67 \text{ puntos,}$$

Sin embargo, esta nota no representa bien el rendimiento académico del alumno ya que en ella han tenido igual peso todas las asignaturas, cuando la química debería tener más peso que la lengua al tener más créditos.

Es más lógico calcular la media ponderada, tomando como pesos los créditos de cada asignatura:

$$\bar{x}_p = \frac{\sum x_i p_i}{\sum p_i} = \frac{5 \cdot 6 + 3 \cdot 4 + 6 \cdot 8}{6 + 4 + 8} = \frac{90}{18} = 5 \text{ puntos.}$$

Definición (Mediana muestral Me)

La *mediana muestral* de una variable X es el valor de la variable que, una vez ordenados los valores de la muestra de menor a mayor, deja el mismo número de valores por debajo y por encima de él.

La mediana cumple $N_{Me} = n/2$ y $F_{Me} = 0,5$.

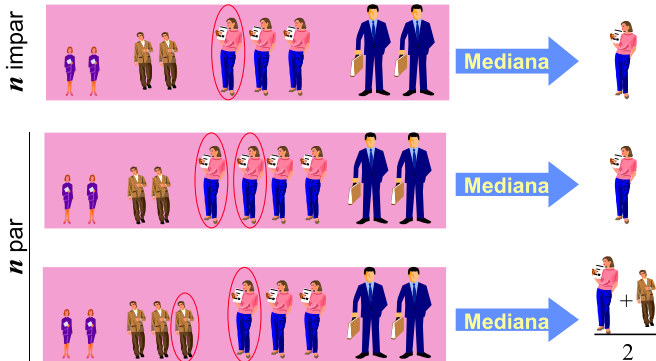
El cálculo de la mediana se realiza de forma distinta según se hayan agrupado los datos o no.

¡Ojo! No puede calcularse para atributos nominales.

Cálculo de la mediana con datos no agrupados

Con datos no agrupados pueden darse varios casos:

- Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.
- Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2} + 1$.



Cálculo de la mediana

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos, el tamaño muestral es 25, de manera que al ser impar se deben ordenar los datos de menor a mayor y buscar el que ocupa la posición $\frac{25+1}{2} = 13$.

0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, **2**, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4

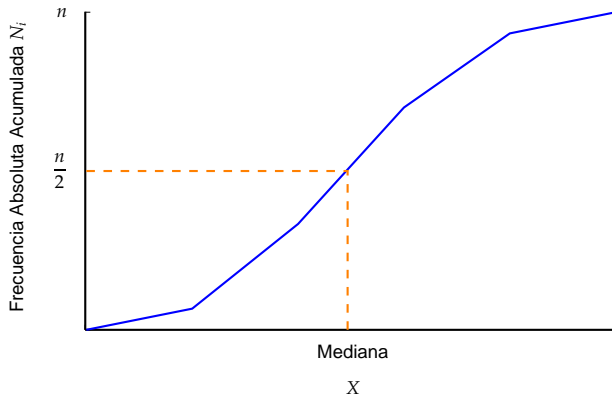
y la mediana es 2 hijos.

Si se trabaja con la tabla de frecuencias, se debe buscar el primer valor cuya frecuencia absoluta acumulada iguale o supere a 13, que es la posición que le corresponde a la mediana, o bien el primer valor cuya frecuencia relativa acumulada iguale o supere a 0,5:

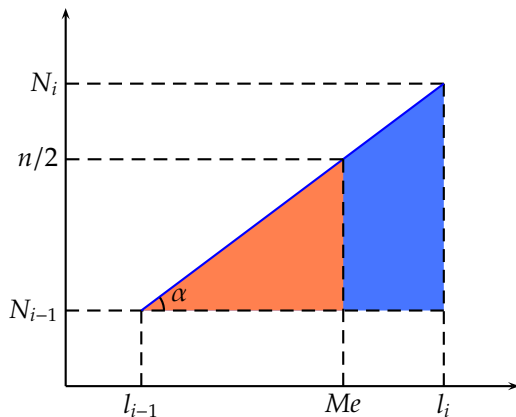
x_i	n_i	f_i	N_i	F_i
0	2	0,08	2	0,08
1	6	0,24	8	0,32
2	14	0,56	22	0,88
3	2	0,08	24	0,96
4	1	0,04	25	1
Σ	25	1		

Cálculo de la mediana con datos agrupados

Con datos agrupados la mediana se calcula interpolando en el polígono de frecuencias absolutas acumuladas para el valor $n/2$.



Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{N_i - N_{i-1}}{l_i - l_{i-1}}$$

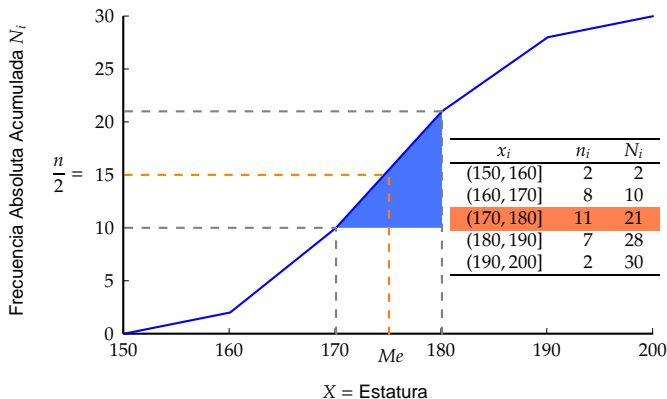
$$\operatorname{tg}(\alpha) = \frac{n/2 - N_{i-1}}{Me - l_{i-1}}$$

$$Me = l_{i-1} + \frac{n/2 - N_{i-1}}{N_i - N_{i-1}}(l_i - l_{i-1}) = l_{i-1} + \frac{n/2 - N_{i-1}}{n_i}a_i$$

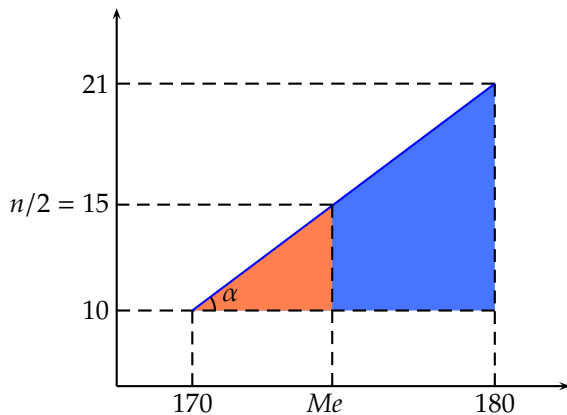
Cálculo de la mediana

Ejemplo con datos agrupados

En el ejemplo de las estaturas $n/2 = 30/2 = 15$. Si miramos en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo $(170, 180]$.



Interpolación en el polígono de frecuencias absolutas acumuladas



$$\operatorname{tg}(\alpha) = \frac{21 - 10}{180 - 170}$$

$$\operatorname{tg}(\alpha) = \frac{15 - 10}{Me - 170}$$

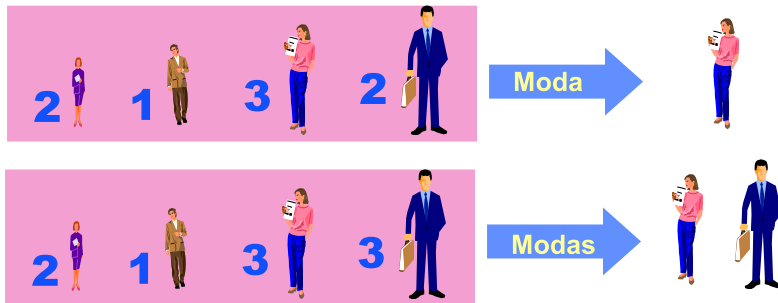
$$Med = 170 + \frac{15 - 10}{21 - 10}(180 - 170) = 170 + \frac{5}{11}10 = 174,54$$

Definición (Moda muestral M_o)

La *moda muestral* de una variable X es el valor de la variable más frecuente en la muestra.

Con datos agrupados se toma como clase modal la clase con mayor frecuencia en la muestra.

En ocasiones puede haber más de una moda.



Cálculo de la moda

En el ejemplo del número de hijos puede verse fácilmente en la tabla de frecuencias que la moda es $Mo = 2$ hijos.

x_i	n_i
0	2
1	6
2	14
3	2
4	1

Y en el ejemplo de las estaturas también puede verse en la tabla de frecuencias que la clase modal es $Mo = (170, 180]$.

x_i	n_i
(150, 160]	2
(160, 170]	8
(170, 180]	11
(180, 190]	7
(190, 200]	2

¿Qué estadístico de tendencia central usar?

En general, siempre que puedan calcularse conviene tomarlas en el siguiente orden:

- 1 Media. La media utiliza más información que el resto ya que para calcularla se tiene en cuenta la magnitud de los datos.
- 2 Mediana. La mediana utiliza menos información que la media, pero más que la moda, ya que para calcularla se tiene en cuenta el orden de los datos.
- 3 Moda. La moda es la que menos información utiliza ya que para calcularla sólo se tienen en cuenta las frecuencias absolutas.

Pero, ¡jojo! la media también es muy sensible a los datos atípicos, así que, tampoco debemos perder de vista la mediana.

Por ejemplo, consideremos la siguiente muestra del número de hijos de 7 matrimonios:

0, 0, 1, 1, 2, 2, 15

$\bar{x} = 3$ hijos y $Me = 1$ hijos

¿Qué representante de la muestra tomarías?

Cuantiles

Son valores de la variable que dividen la distribución, supuesta ordenada de menor a mayor, en partes que contienen el mismo número de datos.

Los más utilizados son:

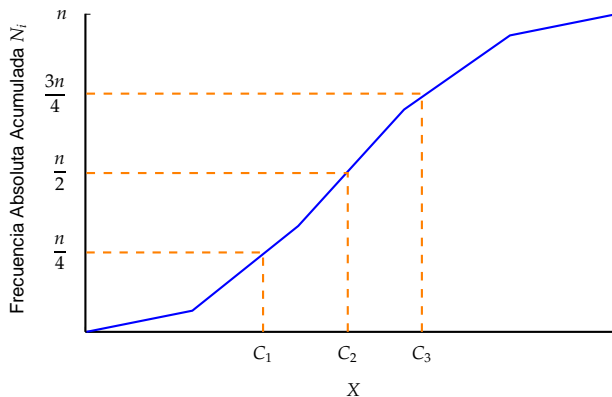
Cuartiles: Dividen la distribución en 4 partes iguales.
Hay tres cuartiles: C_1 (25 % acumulado) , C_2 (50 % acumulado), C_3 (75 % acumulado).

Deciles: Dividen la distribución en 10 partes iguales.
Hay 9 deciles: D_1 (10 % acumulado) , \dots , D_9 (90 % acumulado).

Percentiles: Dividen la distribución en 100 partes iguales.
Hay 99 percentiles: P_1 (1 % acumulado), \dots , P_{99} (99 % acumulado).

Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. Por ejemplo, en el caso de los cuantiles se buscan los valores que tienen frecuencias absolutas acumuladas $n/4$ (primer cuartil), $n/2$ (segundo cuartil) y $3n/4$ (tercer cuartil) y si se trata de datos agrupados se interpola sobre el polígono de frecuencias acumuladas.



Cálculo de los cuantiles

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos se tenían la siguientes frecuencias relativas acumuladas

x_i	F_i
0	0,08
1	0,32
2	0,88
3	0,96
4	1

$$F_{C_1} = 0,25 \Rightarrow C_1 = 1 \text{ hijos,}$$

$$F_{C_2} = 0,5 \Rightarrow C_2 = 2 \text{ hijos,}$$

$$F_{C_3} = 0,75 \Rightarrow C_3 = 2 \text{ hijos,}$$

$$F_{D_3} = 0,3 \Rightarrow D_3 = 1 \text{ hijos,}$$

$$F_{P_{92}} = 0,92 \Rightarrow P_{92} = 3 \text{ hijos.}$$

Recogen información respecto a la heterogeneidad de la variable y a la concentración de sus valores en torno a algún valor central.

Para las variables cuantitativas, las más empleadas son:

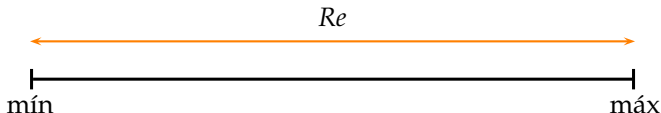
- Recorrido.
- Rango Intercuartílico.
- Varianza.
- Desviación Típica.
- Coeficiente de Variación.

Definición (Recorrido muestral Re)

El *recorrido muestral* de una variable X se define como la diferencia entre el máximo y el mínimo de los valores en la muestra.

$$Re = \max_{x_i} - \min_{x_i}$$

El recorrido da una idea de la máxima variación que hay entre los datos muestrales. No obstante, es muy sensible a datos atípicos ya que suelen aparecer justo en los extremos de la distribución, por lo que no se suele utilizar mucho.



Rango intercuartílico

Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

Definición (Rango intercuartílico muestral RI)

El *rango intercuartílico muestral* de una variable X se define como la diferencia entre el tercer y el primer cuartil de la muestra.

$$RI = C_3 - C_1$$

El rango intercuartílico da una idea de la variación que hay en el 50 % de los datos centrales.

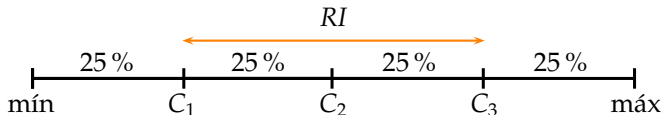


Diagrama de caja y bigotes

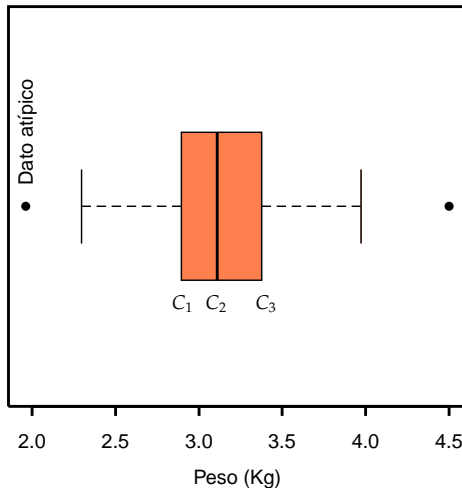
La dispersión de una variable suele representarse gráficamente mediante un **diagrama de caja y bigotes**, que consiste en una caja sobre un eje X donde el borde inferior de la caja es el primer cuartil, y el borde superior el tercer cuartil, y por tanto, la anchura de la caja es el rango intercuartílico. En ocasiones también se representa el segundo cuartil con una línea que divide la caja.

También se utiliza para detectar los valores atípicos mediante unos segmentos (bigotes) que salen de los extremos de la caja y que marcan el intervalo de normalidad de los datos.

Diagrama de caja y bigotes

Ejemplo con pesos de recién nacidos

Diagrama de caja y bigotes del peso de recién nacidos



Construcción del diagrama de caja y bigotes

- 1 Calcular los cuartiles.
- 2 Dibujar una caja de manera que el extremo inferior caiga sobre el primer cuartil y el extremo superior sobre el tercer cuartil.
- 3 Dividir la caja con una línea que caiga sobre el segundo cuartil.
- 4 Para los bigotes inicialmente se determina la posición de los puntos denominados *vallas* v_1 y v_2 restando y sumando respectivamente a primer y tercer cuartil 1,5 veces el rango intercuartílico RI :

$$v_1 = C_1 - 1,5RI$$

$$v_2 = C_3 + 1,5RI$$

A partir de las vallas se buscan los valores b_1 , que es el mínimo valor de la muestra mayor o igual que v_1 , y b_2 , que es máximo valor de la muestra menor o igual que v_2 . Para el bigote inferior se dibuja un segmento desde el borde inferior de la caja hasta b_1 y para el superior se dibuja un segmento desde el borde superior de la caja hasta b_2 .

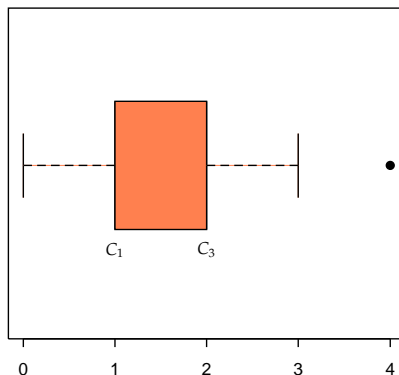
- 5 Finalmente, si en la muestra hay algún dato por debajo de v_1 o por encima de v_2 se dibuja un punto sobre dicho valor.

Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

- 1 Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
- 2 Dibujar la caja.
- 3 Calcular las vallas: $v_1 = 1 - 1,5 * 1 = -0,5$ y $v_2 = 2 + 1,5 * 1 = 3,5$.
- 4 Dibujar los bigotes: $b_1 = 0$ hijos y $b_2 = 3$ hijos.
- 5 Dibujar los datos atípicos: 4 hijos.

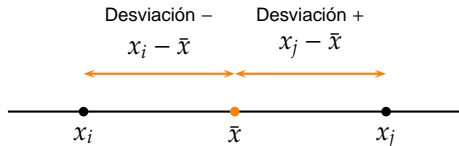
Diagrama de caja y bigotes del número de hijos



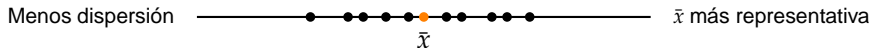
Desviaciones respecto de la media

Otra forma de medir la variabilidad de una variable es estudiar la concentración de los valores en torno a algún estadístico de tendencia central como por ejemplo la media.

Para ello se suele medir la distancia de cada valor a la media. A ese valor se le llama **desviación respecto de la media**.



Si las desviaciones son grandes la media no será tan representativa como cuando la desviaciones sean pequeñas.



Varianza y desviación típica

Definición (Varianza s^2)

La *varianza muestral* de una variable X se define como el promedio del cuadrado de las desviaciones de los valores de la muestra respecto de la media muestral.

$$s^2 = \frac{\sum (x_i - \bar{x})^2 n_i}{n} = \sum (x_i - \bar{x})^2 f_i$$

También puede calcularse de manera más sencilla mediante la fórmula

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \sum x_i^2 f_i - \bar{x}^2$$

La varianza tiene las unidades de la variable al cuadrado, por lo que para facilitar su interpretación se suele utilizar su raíz cuadrada:

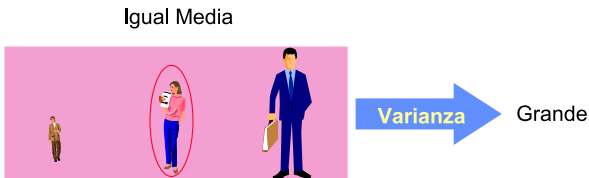
Definición (Desviación típica s)

La *desviación típica muestral* de una variable X se define como la raíz cuadrada positiva de su varianza muestral.

$$s = +\sqrt{s^2}$$

Interpretación de la varianza y la desviación típica

Tanto la varianza como la desviación típica sirven para cuantificar la dispersión de los datos en torno a la media.



Cálculo de la varianza y la desviación típica

Ejemplo con datos no agrupados

Para el número de hijos se puede calcular la varianza a partir de la tabla de frecuencias añadiendo una columna con los cuadrados de los valores:

x_i	n_i	$x_i^2 n_i$
0	2	0
1	6	6
2	14	56
3	2	18
4	1	16
Σ	25	96

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{96}{25} - 1,76^2 = 0,7424 \text{ hijos}^2.$$

Y la desviación típica es $s = \sqrt{0,7424} = 0,8616$ hijos.

Comparado este valor con el recorrido, que va de 0 a 4 hijos se observa que no es demasiado grande por lo que se puede concluir que no hay mucha dispersión y en consecuencia la media de 1,76 hijos representa bien a los matrimonios de la muestra.

Cálculo de la varianza y la desviación típica

Ejemplo con datos agrupados

En el ejemplo de las estaturas, al ser datos agrupados, el cálculo se realiza igual que antes pero tomando como valores de la variable las marcas de clase.

X	x_i	n_i	$x_i^2 n_i$
(150, 160]	155	2	48050
(160, 170]	165	8	217800
(170, 180]	175	11	336875
(180, 190]	185	7	239575
(190, 200]	195	2	76050
Σ		30	918350

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174,67^2 = 102,06 \text{ cm}^2.$$

Y la desviación típica es $s = \sqrt{102,06} = 10,1 \text{ cm}$.

Este valor es bastante pequeño, comparado con el recorrido de la variable, que va de 150 a 200 cm, por lo que la variable tiene poca dispersión y en consecuencia su media es muy representativa.

Coeficiente de variación

Tanto la varianza como la desviación típica tienen unidades y eso dificulta a veces su interpretación y su comparación.

Afortunadamente es fácil definir a partir de ellas una medida de dispersión adimensional que es más fácil de interpretar.

Definición (Coeficiente de variación muestral cv)

El *coeficiente de variación muestral* de una variable X se define como el cociente entre su desviación típica muestral y el valor absoluto de su media muestral.

$$cv = \frac{s}{|\bar{x}|}$$

El coeficiente de variación muestral mide la dispersión relativa de los valores de la muestra en torno a la media muestral.

Como no tiene unidades, es muy sencillo de interpretar: Cuanto mayor sea, mayor será la dispersión y menos representativa será la media.

También se utiliza para comparar la dispersión entre muestras distintas incluso si las variables tienen unidades diferentes.

Coeficiente de variación

Ejemplo

En el caso del número de hijos, como $\bar{x} = 1,76$ hijos y $s = 0,8616$ hijos, se tiene que el coeficiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{0,8616}{|1,76|} = 0,49.$$

En el caso de las estaturas, como $\bar{x} = 174,67$ cm y $s = 10,1$ cm, se tiene que el coeficiente de variación vale

$$cv = \frac{s}{|\bar{x}|} = \frac{10,1}{|174,67|} = 0,06.$$

Como se puede observar la dispersión relativa en la muestra de estaturas es mucho menor que en la del número de hijos, por lo que la media de las estaturas será más representativa que la media del número de hijos.

Estadísticos de forma

Son medidas que tratan de caracterizar aspectos de la forma de la distribución de una muestra.

Los aspectos más relevantes son:

Simetría: Miden la simetría de la distribución de frecuencias en torno a la media.

El estadístico más utilizado es el *Coeficiente de Asimetría de Fisher*.

Apuntamiento: Miden el apuntamiento de la distribución de frecuencias.

El estadístico más utilizado es el *Coeficiente de Apuntamiento o Curtosis*.

Definición (Coeficiente de asimetría muestral g_1)

El *coeficiente de asimetría muestral* de una variable X se define como el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas al cubo, dividido por la desviación típica al cubo.

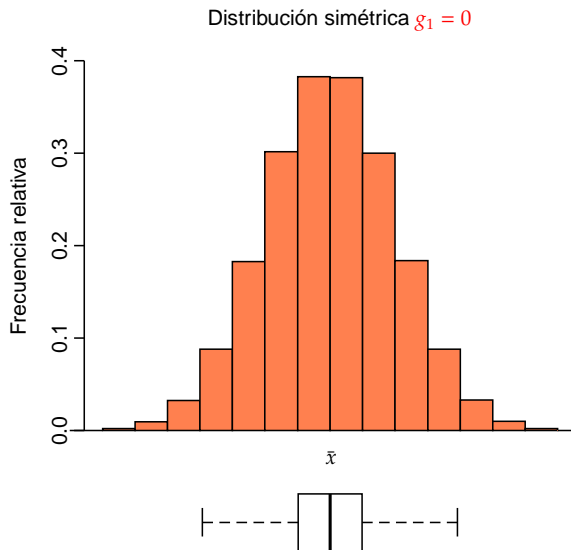
$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{\sum (x_i - \bar{x})^3 f_i}{s^3}$$

El coeficiente de asimetría muestral mide el grado de simetría de los valores de la muestra con respecto a la media muestral, de manera que:

- $g_1 = 0$ indica que hay el mismo número de valores a la derecha y a la izquierda de la media (simétrica).
- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media (asimétrica a la izquierda).
- $g_1 > 0$ indica que la mayoría de los valores son menores que la media (asimétrica a la derecha).

Coeficiente de asimetría

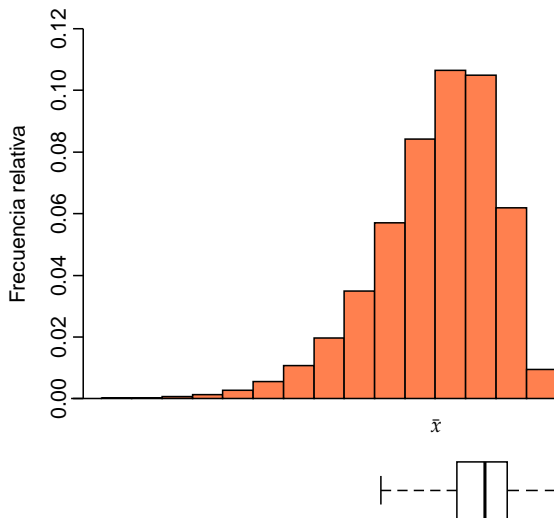
Ejemplo de distribución simétrica



Coeficiente de asimetría

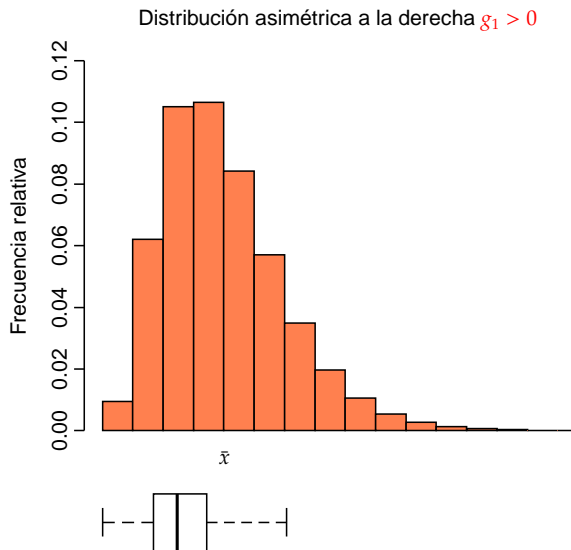
Ejemplo de distribución asimétrica hacia la izquierda

Distribución asimétrica a la izquierda $g_1 < 0$



Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la derecha



Cálculo del coeficiente de asimetría

Ejemplo con datos agrupados

Siguiendo con el ejemplo de las estaturas, podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con los cubos de las desviaciones a la media $\bar{x} = 174,67$ cm:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
(150, 160]	155	2	-19,67	-15221,00
(160, 170]	165	8	-9,67	-7233,85
(170, 180]	175	11	0,33	0,40
(180, 190]	185	7	10,33	7716,12
(190, 200]	195	2	20,33	16805,14
Σ		30		2066,81

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{2066,81/30}{10,1^3} = 0,07.$$

Al estar tan próximo a 0, este valor indica que la distribución es prácticamente simétrica con respecto a la media.

Coeficiente de apuntamiento o curtosis

Definición (Coeficiente de apuntamiento muestral g_2)

El *coeficiente de apuntamiento muestral* de una variable X se define como el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas a la cuarta, dividido por la desviación típica a la cuarta y al resultado se le resta 3.

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

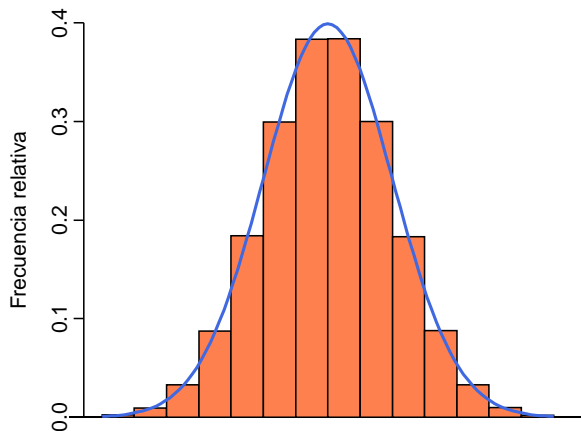
El coeficiente de apuntamiento muestral mide el grado de apuntamiento de los valores de la muestra con respecto a una distribución normal de referencia, de manera que:

- $g_2 = 0$ indica que la distribución tienen un apuntamiento normal (*mesocúrtica*).
- $g_2 < 0$ indica que la distribución tiene menos apuntamiento de lo normal (*platicúrtica*).
- $g_2 > 0$ indica que la distribución tiene más apuntamiento de lo normal (*leptocúrtica*).

Coeficiente de apuntamiento o curtosis

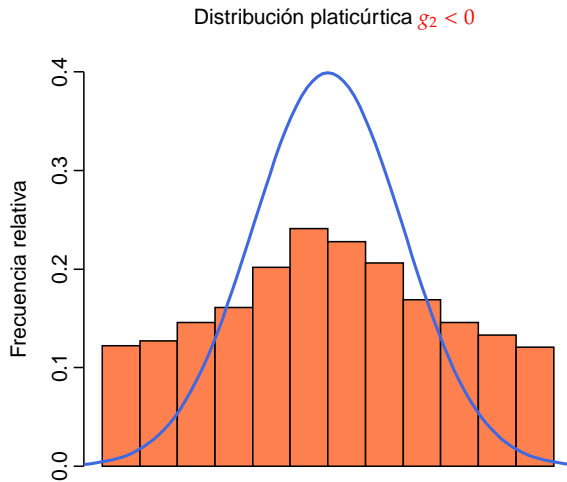
Ejemplo de distribución mesocúrtica

Distribución mesocúrtica $g_2 = 0$



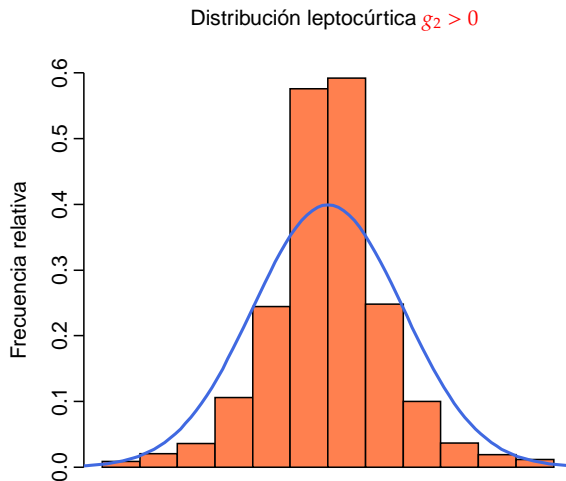
Coeficiente de apuntamiento o curtosis

Ejemplo de distribución platicúrtica



Coeficiente de apuntamiento o curtosis

Ejemplo de distribución leptocúrtica



Cálculo del coeficiente de apuntamiento

Ejemplo con datos agrupados

De nuevo para el ejemplo de las estaturas podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con las desviaciones a la media $\bar{x} = 174,67$ cm elevadas a la cuarta:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
(150, 160]	155	2	-19,67	299396,99
(160, 170]	165	8	-9,67	69951,31
(170, 180]	175	11	0,33	0,13
(180, 190]	185	7	10,33	79707,53
(190, 200]	195	2	20,33	341648,49
Σ		30		790704,45

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{790704,45/30}{10,1^4} - 3 = -0,47.$$

Como se trata de un valor negativo, aunque pequeño, podemos decir que la distribución es ligeramente platicúrtica.

Interpretación de los coeficientes de asimetría y apuntamiento

Como se verá más adelante en la parte de inferencia, muchas de las pruebas estadísticas solo pueden aplicarse a poblaciones normales.

Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

En general, se suele rechazar la hipótesis de normalidad de la población cuando g_1 o g_2 estén fuera del intervalo $[-2, 2]$.

En tal caso, lo habitual es aplicar alguna transformación a la variable para corregir la anormalidad.

Transformaciones de variables

En muchas ocasiones se suelen transformar los datos brutos para trabajar con unas unidades más cómodas, o bien para corregir alguna anomalía de la distribución.

Por ejemplo, si estamos trabajando con estaturas medidas en metros y tenemos los siguientes valores:

1,75m, 1,65m, 1,80m,

podemos evitar los decimales multiplicando por 100, es decir, pasando de metros a centímetros:

175cm, 165cm, 180cm,

Y si queremos reducir la magnitud de los datos podemos restarles a todos el menor de ellos, en este caso, 165cm:

10cm, 0cm, 15cm,

Está claro que este conjunto de datos es mucho más sencillo que el original. En el fondo lo que se ha hecho es aplicar a los datos la transformación:

$$Y = 100X - 165$$

Una de las transformaciones más habituales es la *transformación lineal*:

$$Y = a + bX.$$

Se puede comprobar fácilmente que la media y la desviación típica de la variable resultante cumplen:

$$\begin{aligned}\bar{y} &= a + b\bar{x}, \\ s_y &= |b|s_x\end{aligned}$$

Además, el coeficiente de curtosis no se altera y el de asimetría sólo cambia de signo si b es negativo.

Transformación de tipificación y puntuaciones típicas

Una de las transformaciones lineales más habituales es la *tipificación*:

Definición (Variable tipificada)

La *variable tipificada* de una variable estadística X es la variable que resulta de restarle su media y dividir por su desviación típica.

$$Z = \frac{X - \bar{x}}{s_x}$$

La tipificación es muy útil para eliminar la dependencia de una variable respecto de las unidades de medida empleadas.

Los valores tipificados se conocen como **puntuaciones típicas** y miden el número de desviaciones típicas que dista de la media cada observación, lo cual es útil para comparar variables con distintas unidades.

Otra propiedad de la variable tipificada es que tiene media 0 y desviación típica 1:

$$\bar{z} = 0 \quad s_z = 1$$

Transformación de tipificación y puntuaciones típicas

Ejemplo

Las notas de 5 alumnos en dos asignaturas X e Y son:

Alumno:	1	2	3	4	5		
X:	2	5	4	8	6	$\bar{x} = 5$	$s_x = 2$
Y:	1	9	8	5	2	$\bar{y} = 5$	$s_y = 3,16$

¿Han tenido el mismo rendimiento los alumnos que han sacado un 8?

Podría parecer que ambos alumnos han tenido el mismo rendimiento puesto que tienen la misma nota, pero si queremos ver el rendimiento relativo al resto del grupo, tendríamos que tener en cuenta la dispersión de cada muestra y medir sus puntuaciones típicas:

X:	-1,5	0	-0,5	1,5	0,5
Y:	-1,26	1,26	0,95	0	-0,95

Es decir, el alumno que tiene un 8 en X está 1,5 veces la desviación típica por encima de la media de su grupo, mientras que el alumno que tiene un 8 en Y sólo está 0,95 desviaciones típicas por encima de su media. Así pues, el primer alumno tuvo un rendimiento superior al segundo.

Transformación de tipificación y puntuaciones típicas

Ejemplo

Siguiendo con el ejemplo anterior

¿Cuál es el mejor alumno?

Si simplemente se suman las puntuaciones de cada asignatura se tiene:

Alumno:	1	2	3	4	5
X:	2	5	4	8	6
Y:	1	9	8	5	2
Σ	3	14	12	13	8

El mejor alumno sería el segundo.

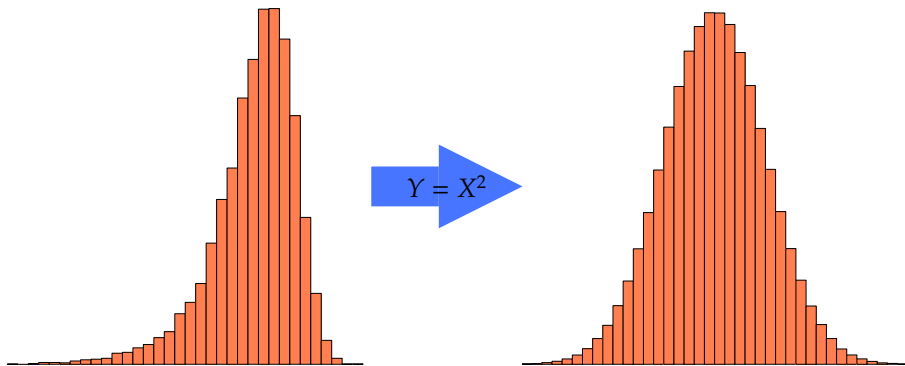
Pero si se considera el rendimiento relativo tomando las puntuaciones típicas se tiene:

Alumno:	1	2	3	4	5
X:	-1,5	0	-0,5	1,5	0,5
Y:	-1,26	1,26	0,95	0	-0,95
Σ	-2,76	1,26	0,45	1,5	-0,45

Y el mejor alumno sería el cuarto.

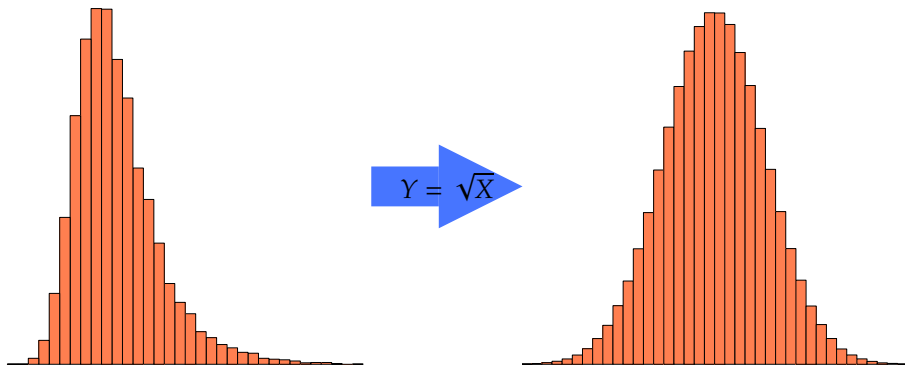
Transformaciones no lineales

La transformación $Y = X^2$ comprime la escala para valores pequeños y la expande para valores altos, de manera que es muy útil para corregir asimetrías hacia la izquierda.



Transformaciones no lineales

Las transformaciones $Y = \sqrt{x}$, $Y = \log X$ y $Y = 1/X$ comprimen la escala para valores altos y la expanden para valores pequeños, de manera que son útiles para corregir asimetrías hacia la derecha.



Variables clasificadoras o factores

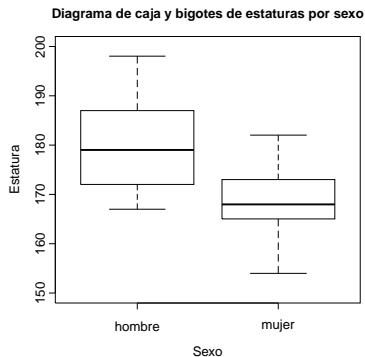
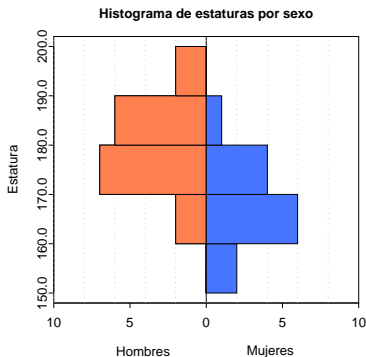
En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos, como por ejemplo, estudiar las estaturas en hombres y mujeres por separado.

En tal caso se utiliza una nueva variable, llamada **variable clasificadora** o **factor discriminante**, para dividir la muestra en grupos y posteriormente se realiza el estudio descriptivo de la variable principal en cada grupo.

Variables clasificadoras

Usando la misma muestra de estaturas, pero teniendo en cuenta el sexo, tenemos:

Mujeres	173, 158, 174, 166, 162, 177, 165, 154, 166, 182, 169, 172, 170, 168.
Hombres	179, 181, 172, 194, 185, 187, 198, 178, 188, 171, 175, 167, 186, 172, 176, 187.



- 3 Regresión y Correlación
 - Distribución de frecuencias bidimensional
 - Covarianza
 - Regresión
 - Recta de regresión
 - Correlación
 - Coeficientes de determinación y correlación
 - Regresión no lineal
 - Medidas de relación entre atributos

Relaciones entre variables

Hasta ahora se ha visto como describir el comportamiento de una variable, pero en los fenómenos naturales normalmente aparecen más de una variable que suelen estar relacionadas. Por ejemplo, en un estudio sobre el peso de las personas, deberíamos incluir todas las variables con las que podría tener relación: altura, edad, sexo, dieta, tabaco, ejercicio físico, etc.

Para comprender el fenómeno no basta con estudiar cada variable por separado y es preciso un estudio conjunto de todas las variables para ver cómo interactúan y qué relaciones se dan entre ellas. El objetivo de la estadística en este caso es dar medidas del grado y del tipo de relación entre dichas variables.

Generalmente, se considera una *variable dependiente* Y que se supone relacionada con otras variables X_1, \dots, X_n llamadas *variables independientes*.

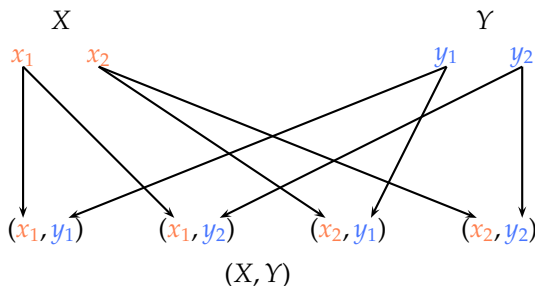
El caso más simple es el de una sola variable independiente, y en tal caso se habla de *estudio de dependencia simple*. Para más de una variable independiente se habla de *estudio de dependencia múltiple*.

En este tema se verán los estudios de dependencia simple que son más sencillos

Variables bidimensionales

Al estudiar la dependencia simple entre dos variables X e Y , no se pueden estudiar sus distribuciones por separado, sino que hay que estudiarlas en conjunto.

Para ello, conviene definir una **variable estadística bidimensional** (X, Y) , cuyos valores serán todos los pares formados por los valores de las variables X e Y .



Frecuencias de una variable bidimensional

Definición (Frecuencias muestrales de una variable bidimensional)

Dada una muestra de tamaño n de una variable bidimensional (X, Y) , para cada valor de la variable (x_i, y_j) observado en la muestra se define:

- **Frecuencia absoluta n_{ij}** : Es el número de individuos de la muestra que presentan simultáneamente el valor x_i de la variable X y el valor y_j de la variable Y .
- **Frecuencia relativa f_{ij}** : Es la proporción de individuos de la muestra que presentan simultáneamente el valor x_i de la variable X y el valor y_j de la variable Y .

$$f_{ij} = \frac{n_{ij}}{n}$$

¡Ojo! Para las variables bidimensionales no tienen sentido las frecuencias acumuladas.

Distribución de frecuencias bidimensional

Al conjunto de valores de la variable bidimensional y sus respectivas frecuencias muestrales se le denomina **distribución conjunta**.

La distribución conjunta de una variable bidimensional se suele representar mediante una **tabla de frecuencias bidimensional**.

$X \backslash Y$	y_1	\cdots	y_j	\cdots	y_q
x_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1q}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_{i1}	\cdots	n_{ij}	\cdots	n_{iq}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_p	n_{p1}	\cdots	n_{pj}	\cdots	n_{pq}

Distribución de frecuencias bidimensional

Ejemplo con estaturas y pesos

Se ha medido la estatura (en cm) y el peso (en Kg) de 30 universitarios obteniendo:

(179,85), (173,65), (181,71), (170,65), (158,51), (174,66), (172,62),
(166,60), (194,90), (185,75), (162,55), (187,78), (198,109), (177,61),
(178,70), (165,58), (154,50), (183,93), (166,51), (171,65), (175,70),
(182,60), (167,59), (169,62), (172,70), (186,71), (172,54), (176,68),
(168,67), (187,80).

X/Y	[50,60)	[60,70)	[70,80)	[80,90)	[90,100)	[100,110)
(150,160]	2	0	0	0	0	0
(160,170]	4	4	0	0	0	0
(170,180]	1	6	3	1	0	0
(180,190]	0	1	4	1	1	0
(190,200]	0	0	0	0	1	1

Diagrama de dispersión

A menudo, la información de la tabla de frecuencias bidimensional se representa también gráficamente.

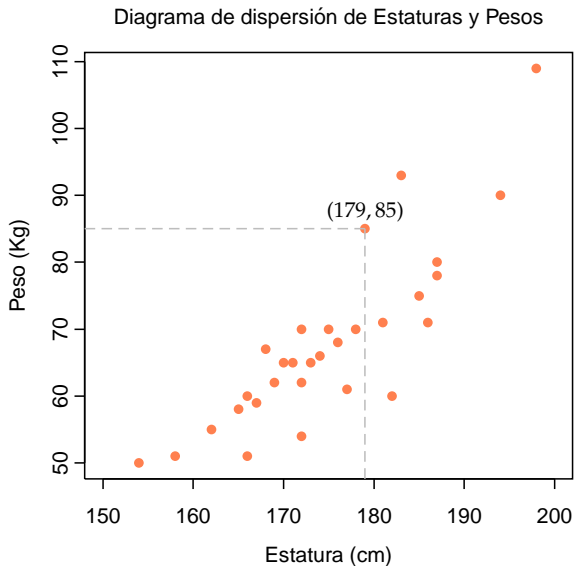
La representación gráfica que más se utiliza en el estudio de la dependencia de dos variables es el **diagrama de dispersión**, que consiste en representar sobre un plano cartesiano los puntos que se corresponden con los valores (x_i, y_j) de la variable bidimensional.

El conjunto de todos estos puntos recibe el nombre de *nube de puntos*.

En un diagrama de dispersión sólo se recogen los valores observados en la muestra, no las frecuencias de los mismos. Para reflejar las frecuencias tendríamos que recurrir a otro tipo de representación como un *diagrama de burbujas* o *histograma tridimensional*.

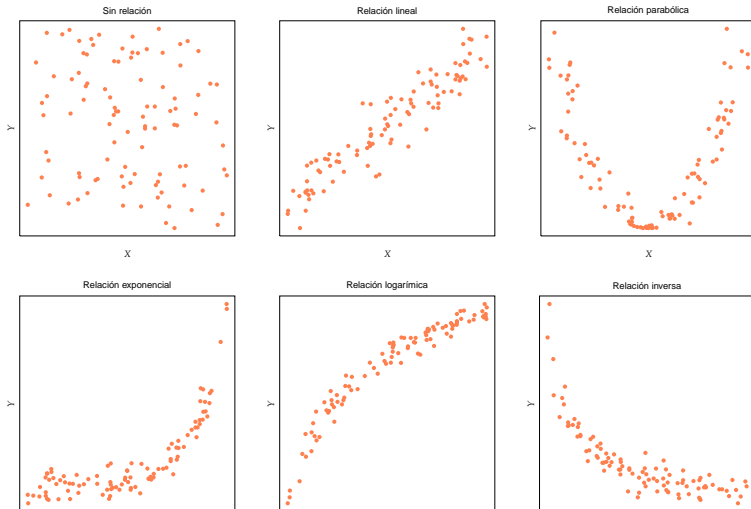
¡Ojo! No tiene sentido cuando alguna de las variables es un atributo.

Diagrama de dispersión



Interpretación del diagrama de dispersión

El diagrama de dispersión da información visual sobre el tipo de relación entre las variables.



Distribuciones marginales

A cada una de las distribuciones de las variables que conforman la variable bidimensional se les llama **distribuciones marginales**.

Las distribuciones marginales se pueden obtener a partir de la tabla de frecuencias bidimensional, sumando las frecuencias por filas y columnas.

$X \backslash Y$	y_1	\cdots	y_j	\cdots	y_q	n_x
x_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1q}	n_{x1}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_i	n_{i1}	$\xrightarrow{+}$	n_{ij}	$\xrightarrow{+}$	n_{iq}	n_{xi}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_p	n_{p1}	\cdots	n_{pj}	\cdots	n_{pq}	n_{xp}
n_y	n_{y1}	\cdots	n_{yj}	\cdots	n_{yq}	n

Distribuciones marginales

Ejemplo con estaturas y pesos

En el ejemplo anterior de las estaturas y los pesos, las distribuciones marginales son

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)	n_x
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

y los estadísticos asociados:

$$\bar{x} = 174,67 \text{ cm}$$

$$s_x^2 = 102,06 \text{ cm}^2$$

$$s_x = 10,1 \text{ cm}$$

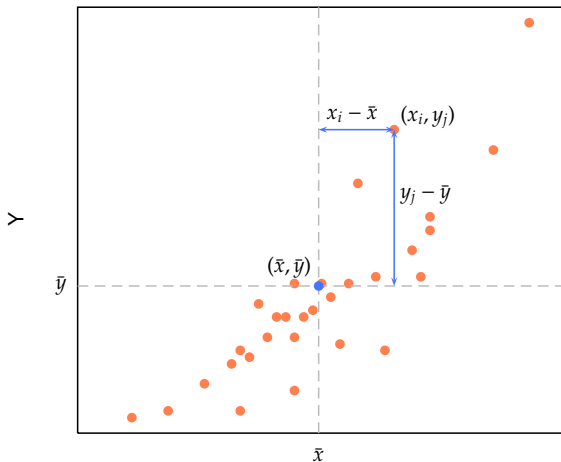
$$\bar{y} = 69,67 \text{ Kg}$$

$$s_y^2 = 164,42 \text{ Kg}^2$$

$$s_y = 12,82 \text{ Kg}$$

Desviaciones respecto de las medias

Para analizar la relación entre dos variables cuantitativas es importante hacer un estudio conjunto de las desviaciones respecto de la media de cada variable.

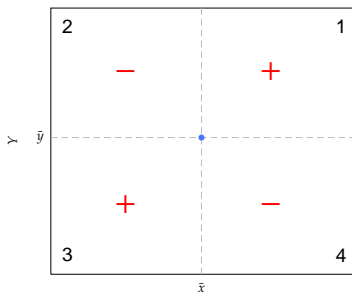


Estudio de las desviaciones respecto de las medias

Si dividimos la nube de puntos del diagrama de dispersión en 4 cuadrantes centrados en el punto de medias (\bar{x}, \bar{y}) , el signo de las desviaciones será:

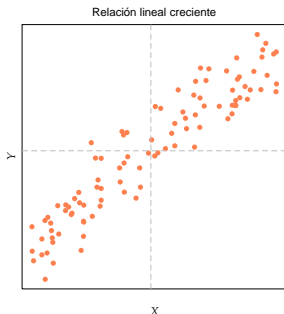
Cuadrante	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	-	+	-
3	-	-	+
4	+	-	-

Signo del producto de desviaciones



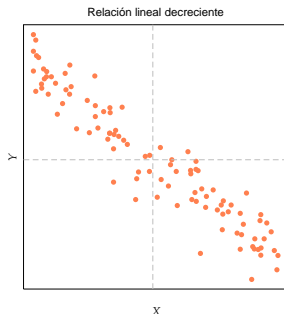
Estudio de las desviaciones respecto de las medias

Si la relación entre las variables es *lineal y creciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 1 y 3 y la suma de los productos de desviaciones será positiva.



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = +$$

Si la relación entre las variables es *lineal y decreciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 2 y 4 y la suma de los productos de desviaciones será negativa.



$$\sum (x_i - \bar{x})(y_j - \bar{y}) = -$$

Covarianza

Del estudio conjunto de las desviaciones respecto de la media surge el siguiente estadístico de relación lineal:

Definición (Covarianza muestral)

La *covarianza muestral* de una variable aleatoria bidimensional (X, Y) se define como el promedio de los productos de las respectivas desviaciones respecto de las medias de X e Y .

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}$$

También puede calcularse de manera más sencilla mediante la fórmula

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

La covarianza sirve para estudiar la relación lineal entre dos variables:

- Si $s_{xy} > 0$ existe una relación lineal creciente entre las variables.
- Si $s_{xy} < 0$ existe una relación lineal decreciente entre las variables.
- Si $s_{xy} = 0$ no existe relación lineal entre las variables.

Cálculo de la covarianza

Ejemplo con estaturas y pesos

En el ejemplo de las estaturas y pesos, teniendo en cuenta que

X/Y	[50, 60)	[60, 70)	[70, 80)	[80, 90)	[90, 100)	[100, 110)	n_x
(150, 160]	2	0	0	0	0	0	2
(160, 170]	4	4	0	0	0	0	8
(170, 180]	1	6	3	1	0	0	11
(180, 190]	0	1	4	1	1	0	7
(190, 200]	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

$$\bar{x} = 174,67 \text{ cm} \quad \bar{y} = 69,67 \text{ Kg}$$

la covarianza vale

$$\begin{aligned}s_{xy} &= \frac{\sum x_i y_j n_{ij}}{n} - \bar{x} \bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \dots + 195 \cdot 105 \cdot 1}{30} - 174,67 \cdot 69,67 \\&= \frac{368200}{30} - 12169,26 = 104,07 \text{ cm} \cdot \text{Kg},\end{aligned}$$

lo que indica que existe una relación lineal creciente entre la estatura y el peso.

Regresión

En muchos casos el objetivo de un estudio no es solo detectar una relación entre variables, sino explicarla mediante alguna función matemática.

La **regresión** es la parte de la estadística que trata de determinar la posible relación entre una variable numérica dependiente Y , y otro conjunto de variables numéricas independientes, X_1, X_2, \dots, X_n , de una misma población. Dicha relación se refleja mediante un modelo funcional

$$y = f(x_1, \dots, x_n).$$

El objetivo es determinar una ecuación mediante la que pueda estimarse el valor de la variable dependiente en función de los valores de las independientes.

El caso más sencillo se da cuando sólo hay una variable independiente X , entonces se habla de *regresión simple*. En este caso el modelo que explica la relación de Y como función de X es una función de una variable $y = f(x)$ que se conoce como **función de regresión**.

Modelos de regresión simple

Dependiendo de la forma de función de regresión, existen muchos tipos de regresión simple. Los más habituales son los que aparecen en la siguiente tabla:

Familia de curvas	Ecuación genérica
Lineal	$y = a + bx$
Parabólica	$y = a + bx + cx^2$
Polinómica de grado n	$y = a_0 + a_1x + \dots + a_nx^n$
Potencial	$y = a \cdot x^b$
Exponencial	$y = a \cdot e^{bx}$
Logarítmica	$y = a + b \log x$
Inverso	$y = a + \frac{b}{x}$
Curva S	$y = e^{a + \frac{b}{x}}$

La elección de un tipo u otro depende de la forma que tenga la nube de puntos del diagrama de dispersión.

Residuos o errores predictivos

Una vez elegida la familia de curvas que mejor se adapta a la nube de puntos, se determina, dentro de dicha familia, la curva que mejor se ajusta a la distribución.

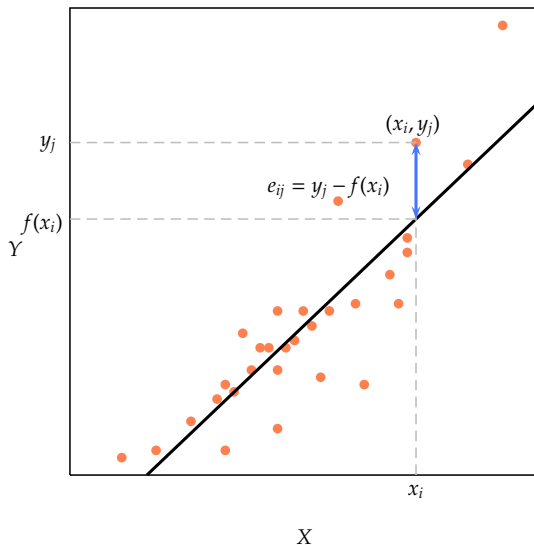
El objetivo es encontrar la función de regresión que haga mínimas las distancias entre los valores de la variable dependiente observados en la muestra, y los predichos por la función de regresión. Estas distancias se conocen como *residuos* o *errores predictivos*.

Definición (Residuos o Errores predictivos)

Dado el modelo de regresión $y = f(x)$ para una variable bidimensional (X, Y) , el *residuo* o *error predictivo* de un valor (x_i, y_j) observado en la muestra, es la diferencia entre el valor observado de la variable dependiente y_j y el predicho por la función de regresión para x_i :

$$e_{ij} = y_j - f(x_i).$$

Resíduos o errores predictivos en Y



Método de mínimos cuadrados

Una forma posible de obtener la función de regresión es mediante el método de *mínimos cuadrados* que consiste en calcular la función que haga mínima la suma de los cuadrados de los residuos

$$\sum e_{ij}^2.$$

En el caso de un modelo de regresión lineal $f(x) = a + bx$, como la recta depende de dos parámetros (el término independiente a y la pendiente b), la suma también dependerá de estos parámetros

$$\theta(a, b) = \sum e_{ij}^2 = \sum (y_j - f(x_i))^2 = \sum (y_j - a - bx_i)^2.$$

Así pues, todo se reduce a buscar los valores a y b que hacen mínima esta suma.

Cálculo de la recta de regresión

Método de mínimos cuadrados

Considerando la suma de los cuadrados de los residuos como una función de dos variables $\theta(a, b)$, se pueden calcular los valores de los parámetros del modelo que hacen mínima esta suma derivando e igualando a 0 las derivadas:

$$\begin{aligned}\frac{\partial \theta(a, b)}{\partial a} &= \frac{\partial \sum (y_j - a - bx_i)^2}{\partial a} = 0 \\ \frac{\partial \theta(a, b)}{\partial b} &= \frac{\partial \sum (y_j - a - bx_i)^2}{\partial b} = 0\end{aligned}$$

Tras resolver el sistema se obtienen los valores

$$a = \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x} \quad b = \frac{s_{xy}}{s_x^2}$$

Estos valores hacen mínimos los residuos en Y y por tanto dan la recta de regresión.

Definición (Recta de regresión)

Dada una variable bidimensional (X, Y) , la *recta de regresión* de Y sobre X es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}).$$

La recta de regresión de Y sobre X es la recta que hace mínimos los errores predictivos en Y , y por tanto es la recta que hará mejores predicciones de Y para cualquier valor de X .

Cálculo de la recta de regresión

Ejemplo con estaturas y pesos

Siguiendo con el ejemplo de las estaturas (X) y los pesos (Y) con los siguientes estadísticos:

$$\begin{array}{lll}\bar{x} = 174,67 \text{ cm} & s_x^2 = 102,06 \text{ cm}^2 & s_x = 10,1 \text{ cm} \\ \bar{y} = 69,67 \text{ Kg} & s_y^2 = 164,42 \text{ Kg}^2 & s_y = 12,82 \text{ Kg} \\ & s_{xy} = 104,07 \text{ cm} \cdot \text{Kg} & \end{array}$$

Entonces, la recta de regresión del peso sobre la estatura es:

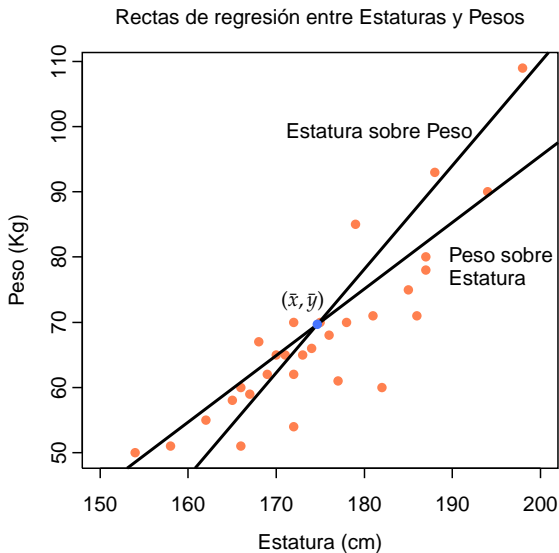
$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 69,67 + \frac{104,07}{102,06}(x - 174,67) = 1,02x - 108,49.$$

De igual modo, si en lugar de considerar el peso como variable dependiente, tomamos la estatura, entonces la recta de regresión de la estatura sobre el peso es:

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 174,67 + \frac{104,07}{164,42}(y - 69,67) = 0,63y + 130,78.$$

Rectas de regresión

Ejemplo de estaturas y pesos



Posición relativa de las rectas de regresión

Las rectas de regresión siempre se cortan en el punto de medias (\bar{x}, \bar{y}) .

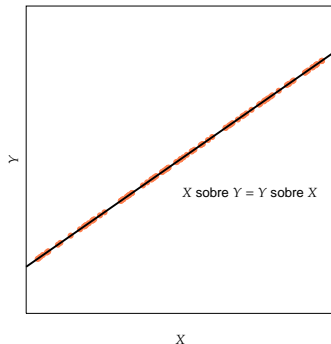
Si entre las variables la relación lineal es perfecta, entonces ambas rectas coinciden ya que sus residuos son nulos.

Si no hay relación lineal, entonces las ecuaciones de las rectas son

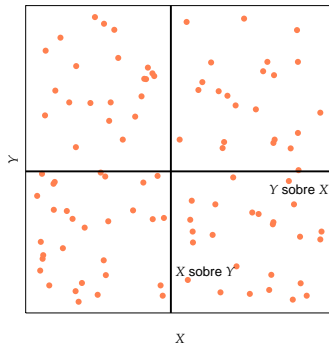
$$y = \bar{y}, \quad x = \bar{x},$$

y se cortan perpendicularmente

Relación lineal perfecta



Sin relación lineal



Definición (Coeficiente de regresión b_{yx})

Dada una variable bidimensional (X, Y) , el *coeficiente de regresión* de la recta de regresión de Y sobre X es su pendiente,

$$b_{yx} = \frac{s_{xy}}{s_x^2}$$

El coeficiente de regresión siempre tiene el mismo signo que la covarianza y refleja el crecimiento de la recta de regresión, ya que da el número de unidades que aumenta o disminuye la variable dependiente por cada unidad que aumenta la variable independiente, según la recta de regresión.

En el ejemplo de las estaturas y los pesos, el coeficiente de regresión del peso sobre la estatura es $b_{yx} = 1,02 \text{ Kg/cm}$, lo que indica que, según la recta de regresión del peso sobre la estatura, por cada cm más de estatura, la persona pesará 1,02 Kg más.

Predicciones con las rectas de regresión

Ejemplo con estaturas y pesos

Las rectas de regresión, y en general cualquier modelo de regresión, suele utilizarse con fines predictivos.

¡Ojo! Para predecir una variable, esta siempre debe considerarse como dependiente en el modelo de regresión que se utilice.

Así, en el ejemplo de las estaturas y los pesos, si se quiere predecir el peso de una persona que mide 180 cm, se debe utilizar la recta de regresión del peso sobre la estatura:

$$y = 1,02 \cdot 180 - 108,49 = 75,11 \text{ Kg.}$$

Y si se quiere predecir la estatura de una persona que pesa 79 Kg, se debe utilizar la recta de regresión de la estatura sobre el peso:

$$x = 0,63 \cdot 79 + 130,78 = 180,55 \text{ cm.}$$

Ahora bien, ¿qué fiabilidad tienen estas predicciones?

Una vez construido un modelo de regresión, para saber si se trata de un buen modelo predictivo, se tiene que analizar el grado de dependencia entre las variables según el tipo de dependencia planteada en el modelo. De ello se encarga la parte de la estadística conocida como **correlación**.

Para cada tipo de modelo existe el correspondiente tipo de correlación.

La correlación se basa en el estudio de los residuos. Cuanto menores sean éstos, más se ajustará la curva de regresión a los puntos, y más intensa será la correlación.

Varianza residual muestral

Una medida de la bondad del ajuste del modelo de regresión es la *varianza residual*.

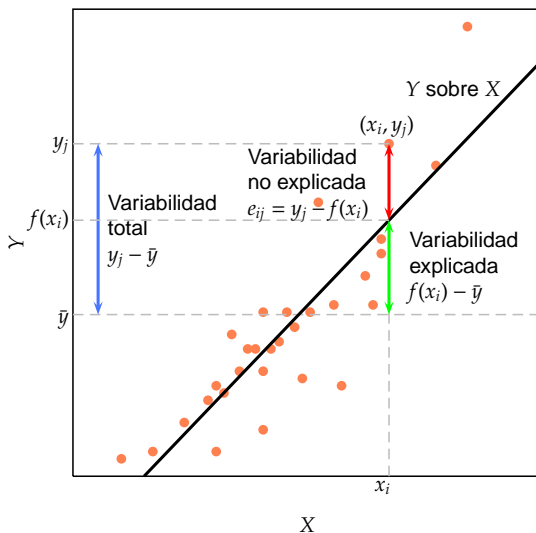
Definición (Varianza residual s_{ry}^2)

Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y) , su *varianza residual muestral* es el promedio de los cuadrados de los residuos para los valores de la muestra,

$$s_{ry}^2 = \frac{\sum e_{ij}^2 n_{ij}}{n} = \frac{\sum (y_j - f(x_i))^2 n_{ij}}{n}.$$

Cuanto más alejados estén los puntos de la curva de regresión, mayor será la varianza residual y menor la dependencia.

Descomposición de la variabilidad total: Variabilidad explicada y no explicada



Coeficiente de determinación

A partir de la varianza residual se puede definir otro estadístico más sencillo de interpretar.

Definición (Coeficiente de determinación muestral)

Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y) , su *coeficiente de determinación muestral* es

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2}$$

El coeficiente de determinación mide la proporción de variabilidad de la variable dependiente explicada por el modelo de regresión, y por tanto,

$$0 \leq r^2 \leq 1$$

Cuanto mayor sea r^2 , mejor explicará el modelo de regresión la relación entre las variables, en particular:

- Si $r^2 = 0$ entonces no existe relación del tipo planteado por el modelo.
- Si $r^2 = 1$ entonces la relación que plantea el modelo es perfecta.

Coefficiente de determinación lineal

En el caso de las rectas de regresión, la varianza residual vale

$$\begin{aligned}s_{ry}^2 &= \sum e_{ij}^2 f_{ij} = \sum (y_j - f(x_i))^2 f_{ij} = \sum \left(y_j - \bar{y} - \frac{s_{xy}}{s_x^2} (x_i - \bar{x}) \right)^2 f_{ij} = \\&= \sum \left((y_j - \bar{y})^2 + \frac{s_{xy}^2}{s_x^4} (x_i - \bar{x})^2 - 2 \frac{s_{xy}}{s_x^2} (x_i - \bar{x})(y_j - \bar{y}) \right) f_{ij} = \\&= \sum (y_j - \bar{y})^2 f_{ij} + \frac{s_{xy}^2}{s_x^4} \sum (x_i - \bar{x})^2 f_{ij} - 2 \frac{s_{xy}}{s_x^2} \sum (x_i - \bar{x})(y_j - \bar{y}) f_{ij} = \\&= s_y^2 + \frac{s_{xy}^2}{s_x^4} s_x^2 - 2 \frac{s_{xy}}{s_x^2} s_{xy} = s_y^2 - \frac{s_{xy}^2}{s_x^2}.\end{aligned}$$

y, por tanto, el coeficiente de determinación lineal vale

$$r^2 = 1 - \frac{s_{ry}^2}{s_y^2} = 1 - \frac{s_y^2 - \frac{s_{xy}^2}{s_x^2}}{s_y^2} = 1 - 1 + \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{s_{xy}^2}{s_x^2 s_y^2}.$$

Cálculo del coeficiente de determinación lineal

Ejemplo de estaturas y pesos

En el ejemplo de las estaturas y pesos se tenía

$$\begin{aligned}\bar{x} &= 174,67 \text{ cm} & s_x^2 &= 102,06 \text{ cm}^2 \\ \bar{y} &= 69,67 \text{ Kg} & s_y^2 &= 164,42 \text{ Kg}^2 \\ s_{xy} &= 104,07 \text{ cm} \cdot \text{Kg}\end{aligned}$$

De modo que el coeficiente de determinación lineal vale

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(104,07 \text{ cm} \cdot \text{Kg})^2}{102,06 \text{ cm}^2 \cdot 164,42 \text{ Kg}^2} = 0,65.$$

Esto indica que la recta de regresión del peso sobre la estatura explica el 65 % de la variabilidad del peso, y de igual modo, la recta de regresión de la estatura sobre el peso explica el 65 % de la variabilidad de la estatura.

Coefficiente de correlación lineal

Definición (Coeficiente de correlación lineal)

Dada una variable bidimensional (X, Y) , el *coeficiente de correlación lineal muestral* es la raíz cuadrada de su coeficiente de determinación lineal, con signo el de la covarianza

$$r = \sqrt{r^2} = \frac{s_{xy}}{s_x s_y}$$

Como r^2 toma valores entre 0 y 1, el coeficiente de correlación lineal tomará valores entre -1 y 1:

$$-1 \leq r \leq 1$$

El coeficiente de correlación lineal también mide el grado de dependencia lineal:

- Si $r = 0$ entonces no existe relación lineal.
- Si $r = 1$ entonces existe una relación lineal creciente perfecta.
- Si $r = -1$ entonces existe una relación lineal decreciente perfecta.

Coeficiente de correlación lineal

Ejemplo

En el ejemplo de las estaturas y los pesos, el coeficiente de correlación lineal vale

$$r = \frac{s_{xy}}{s_x s_y} = \frac{104,07 \text{ cm} \cdot \text{Kg}}{10,1 \text{ cm} \cdot 12,82 \text{ Kg}} = +0,8.$$

lo que indica que la relación lineal entre el peso y la estatura es fuerte, y además creciente.

Fiabilidad de las predicciones de un modelo de regresión

Aunque el coeficiente de determinación o el de correlación hablan de la bondad de un modelo de regresión, no es lo único que hay que tener en cuenta a la hora de hacer predicciones.

La fiabilidad de las predicciones que hagamos con un modelo de regresión depende de varias cosas:

- El coeficiente de determinación: Cuanto mayor sea, menores serán los errores predictivos y mayor la fiabilidad de las predicciones.
- La variabilidad de la población: Cuanto más variable es una población, más difícil es predecir y por tanto menos fiables serán las predicciones.
- El tamaño muestral: Cuanto mayor sea, más información tendremos y, en consecuencia, más fiables serán las predicciones.

Además, hay que tener en cuenta que un modelo de regresión es válido únicamente para el rango de valores observados en la muestra. Fuera de ese rango no hay información del tipo de relación entre las variables, por lo que no deben hacerse predicciones para valores lejos de los observados en la muestra.

El ajuste de un modelo de regresión no lineal es similar al del modelo lineal y también puede realizarse mediante la técnica de mínimos cuadrados.

No obstante, en determinados casos un ajuste no lineal puede convertirse en un ajuste lineal mediante una sencilla transformación de alguna de las variables del modelo.

Transformación de modelos de regresión no lineales

- **Modelo logarítmico:** Un modelo logarítmico $y = a + b \log x$ se convierte en un modelo lineal haciendo el cambio $t = \log x$:

$$y = a + b \log x = a + bt.$$

- **Modelo exponencial:** Un modelo exponencial $y = ae^{bx}$ se convierte en un modelo lineal haciendo el cambio $z = \log y$:

$$z = \log y = \log(ae^{bx}) = \log a + \log e^{bx} = a' + bx.$$

- **Modelo potencial:** Un modelo potencial $y = ax^b$ se convierte en un modelo lineal haciendo los cambios $t = \log x$ y $z = \log y$:

$$z = \log y = \log(ax^b) = \log a + b \log x = a' + bt.$$

- **Modelo inverso:** Un modelo inverso $y = a + b/x$ se convierte en un modelo lineal haciendo el cambio $t = 1/x$:

$$y = a + b(1/x) = a + bt.$$

- **Modelo curva S:** Un modelo curva S $y = e^{a+b/x}$ se convierte en un modelo lineal haciendo los cambios $t = 1/x$ y $z = \log y$:

$$z = \log y = \log(e^{a+b/x}) = a + b(1/x) = a + bt.$$

Ejemplo de ajuste de un modelo exponencial

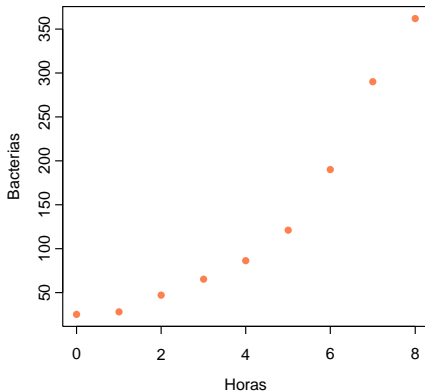
Evolución del número de bacterias de un cultivo

El número de bacterias de un cultivo evoluciona con el tiempo según la siguiente tabla:

El diagrama de dispersión asociado es

Horas	Bacterias
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

Diagrama de dispersión de Horas y Bacterias



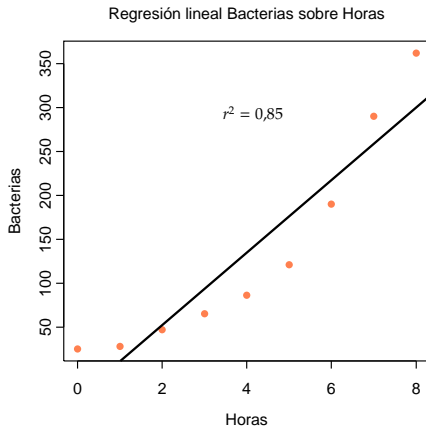
Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

Si realizamos un ajuste lineal, obtenemos la siguiente recta de regresión

Horas	Bacterias
0	25
1	28
2	47
3	65
4	86
5	121
6	190
7	290
8	362

$$\text{Bacterias} = -30,18 + 41,27 \text{ Horas}$$



¿Es un buen modelo?

Ejemplo de ajuste de un modelo exponencial

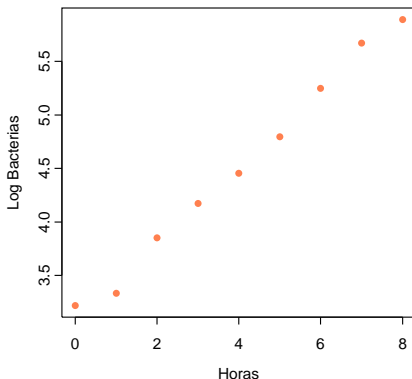
Evolución del número de bacterias de un cultivo

Aunque el modelo lineal no es malo, de acuerdo al diagrama de dispersión es más lógico construir un modelo exponencial o cuadrático.

Para construir el modelo exponencial $y = ae^{bx}$ hay que realizar la transformación $z = \log y$, es decir, aplicar el logaritmo a la variable dependiente.

Horas	Bacterias	Log Bacterias
0	25	3,22
1	28	3,33
2	47	3,85
3	65	4,17
4	86	4,45
5	121	4,80
6	190	5,25
7	290	5,67
8	362	5,89

Diagrama de dispersión de Horas y Bacterias



Ejemplo de ajuste de un modelo exponencial

Evolución del número de bacterias de un cultivo

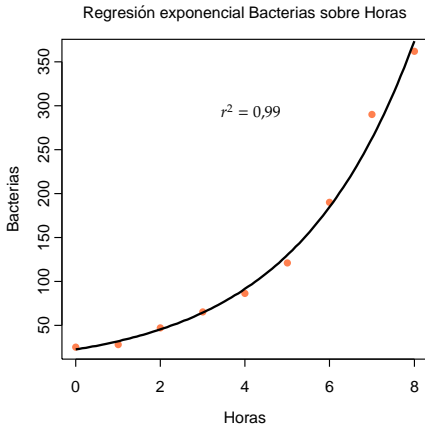
Ahora sólo queda calcular la recta de regresión del logaritmo de Bacterias sobre Horas

$\text{Log Bacterias} = 3,107 + 0,352 \text{ Horas}$.

Y deshaciendo el cambio de variable, se obtiene el modelo exponencial

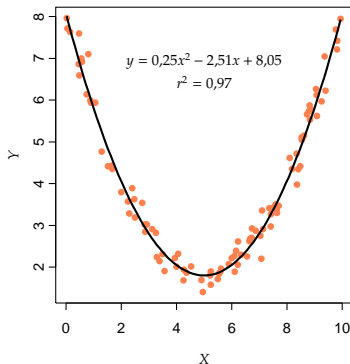
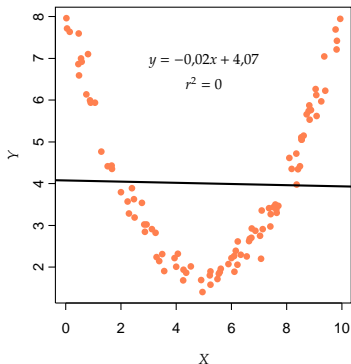
$$\text{Bacterias} = e^{3,107 + 0,352 \text{ Horas}},$$

que, a la vista del coeficiente de determinación, es mucho mejor modelo que el lineal.



Interpretación de un coeficiente de determinación pequeño

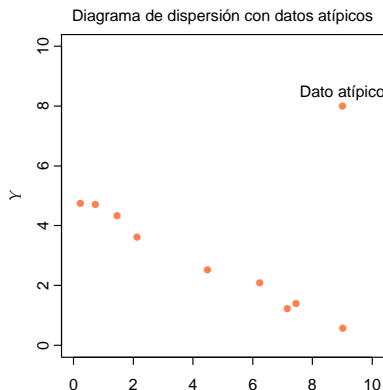
Tanto el coeficiente de determinación como el de correlación hacen referencia a un modelo concreto, de manera que un coeficiente $r^2 = 0$ significa que no existe relación entre las variables del tipo planteado por el modelo, pero *eso no quiere decir que las variables sean independientes*, ya que puede existir relación de otro tipo.



Datos atípicos en regresión

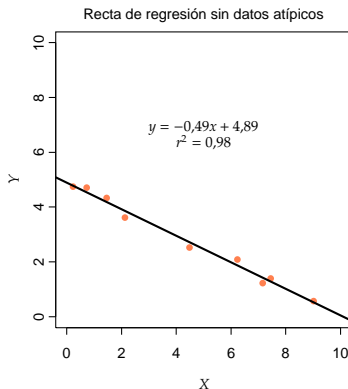
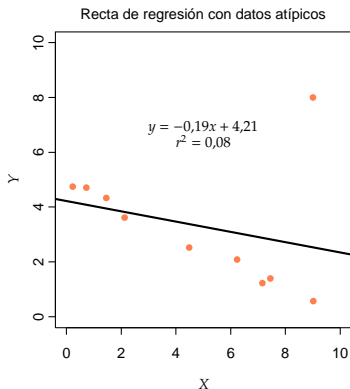
En un estudio de regresión es posible que aparezca algún individuo que se aleja notablemente de la tendencia del resto de individuos en la nube de puntos.

Aunque el individuo podría no ser un *dato atípico* al considerar las variables de manera separada, sí lo sería al considerarlas de manera conjunta.



Influencia de los datos atípicos en los modelos de regresión

Los datos atípicos en regresión suelen provocar cambios drásticos en el ajuste de los modelos de regresión, y por tanto, habrá que tener mucho cuidado con ellos.



Los modelos de regresión vistos sólo pueden aplicarse cuando las variables estudiadas son cuantitativas.

Cuando se desea estudiar la relación entre atributos, tanto ordinales como nominales, es necesario recurrir a otro tipo de medidas de relación o de asociación. En este tema veremos tres de ellas:

- Coeficiente de correlación de Spearman.
- Coeficiente chi-cuadrado.
- Coeficiente de contingencia.

Coefficiente de correlación de Spearman

Cuando se tengan atributos ordinales es posible ordenar sus categorías y asignarles valores ordinales, de manera que se puede calcular el coeficiente de correlación lineal entre estos valores ordinales.

Esta medida de relación entre el orden que ocupan las categorías de dos atributos ordinales se conoce como coeficiente de correlación de Spearman, y puede demostrarse fácilmente que puede calcularse a partir de la siguiente fórmula

Definición (Coeficiente de correlación de Spearman)

Dada una muestra de n individuos en los que se han medido dos atributos ordinales X e Y , el coeficiente de correlación de Spearman se define como:

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

donde d_i es la diferencia entre el valor ordinal de X y el valor ordinal de Y del individuo i .

Interpretación del coeficiente de correlación de Spearman

Como el coeficiente de correlación de Spearman es en el fondo el coeficiente de correlación lineal aplicado a los órdenes, se tiene:

$$-1 \leq r_s \leq 1,$$

de manera que:

- Si $r_s = 0$ entonces no existe relación entre los atributos ordinales.
- Si $r_s = 1$ entonces los órdenes de los atributos coinciden y existe una relación directa perfecta.
- Si $r_s = -1$ entonces los órdenes de los atributos están invertidos y existe una relación inversa perfecta.

En general, cuanto más cerca de 1 o -1 esté r_s , mayor será la relación entre los atributos, y cuanto más cerca de 0, menor será la relación.

Cálculo del coeficiente de correlación de Spearman

Ejemplo

Una muestra de 5 alumnos realizaron dos tareas diferentes X e Y , y se ordenaron de acuerdo a la destreza que manifestaron en cada tarea:

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	2	-1	1
Alumno 4	3	1	2	4
Alumno 5	4	5	-1	1
Σ			0	8

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8}{5(5^2 - 1)} = 0,6,$$

lo que indica que existe bastante relación directa entre las destrezas manifestadas en ambas tareas.

Cálculo del coeficiente de correlación de Spearman

Ejemplo con empates

Cuando hay empates en el orden de las categorías se atribuye a cada valor empatado la media aritmética de los valores ordinales que hubieran ocupado esos individuos en caso de no haber estado empatados.

Si en el ejemplo anterior los alumnos 4 y 5 se hubiesen comportado igual en la primera tarea y los alumnos 3 y 4 se hubiesen comportado igual en la segunda tarea, entonces se tendría

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	1,5	-0,5	0,25
Alumno 4	3,5	1,5	2	4
Alumno 5	3,5	5	-1,5	2,25
Σ			0	8,5

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8,5}{5(5^2 - 1)} = 0,58.$$

Relación entre atributos nominales

Cuando se quiere estudiar la relación entre atributos nominales no tiene sentido calcular el coeficiente de correlación de Spearman ya que las categorías no pueden ordenarse.

Para estudiar la relación entre atributos nominales se utilizan medidas basadas en las frecuencias de la tabla de frecuencias bidimensional, que para atributos se suele llamar *tabla de contingencia*.

Ejemplo En un estudio para ver si existe relación entre el sexo y el hábito de fumar se ha tomado una muestra de 100 personas. La tabla de contingencia resultante es

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

Si el hábito de fumar fuese independiente del sexo, la proporción de fumadores en mujeres y hombres sería la misma.

Frecuencias teóricas o esperadas

En general, dada una tabla de contingencia para dos atributos X e Y ,

$X \backslash Y$	y_1	\cdots	y_j	\cdots	y_q	n_{x_i}
x_1	n_{11}	\cdots	n_{1j}	\cdots	n_{1q}	n_{x_1}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_i	n_{i1}	\cdots	n_{ij}	\cdots	n_{iq}	n_{x_i}
\vdots	\vdots	\ddots	\vdots	\ddots	\vdots	\vdots
x_p	n_{p1}	\cdots	n_{pj}	\cdots	n_{pq}	n_{x_p}
n_{y_j}	n_{y_1}	\cdots	n_{y_j}	\cdots	n_{y_q}	n

si X e Y fuesen independientes, para cualquier valor y_j se tendría

$$\frac{n_{1j}}{n_{x_1}} = \frac{n_{2j}}{n_{x_2}} = \cdots = \frac{n_{pj}}{n_{x_p}} = \frac{n_{1j} + \cdots + n_{pj}}{n_{x_1} + \cdots + n_{x_p}} = \frac{n_{y_j}}{n},$$

de donde se deduce que

$$n_{ij} = \frac{n_{x_i} n_{y_j}}{n}.$$

A esta última expresión se le llama *frecuencia teórica* o *frecuencia esperada* del par (x_i, y_j) .

Coeficiente chi-cuadrado χ^2

Es posible estudiar la relación entre dos atributos X e Y comparando las frecuencias reales con las esperadas:

Definición (Coeficiente chi-cuadrado χ^2)

Dada una muestra de tamaño n en la que se han medido dos atributos X e Y , se define el coeficiente χ^2 como

$$\chi^2 = \sum_{i=1}^p \sum_{j=1}^q \frac{\left(n_{ij} - \frac{n_{x_i} n_{y_j}}{n}\right)^2}{\frac{n_{x_i} n_{y_j}}{n}},$$

donde p es el número de categorías de X y q el número de categorías de Y .

Por ser suma de cuadrados, se cumple que

$$\chi^2 \geq 0,$$

de manera que $\chi^2 = 0$ cuando los atributos son independientes, y crece a medida que aumenta la dependencia entre las variables.

Cálculo del coeficiente chi-cuadrado χ^2

Ejemplo

Siguiendo con el ejemplo anterior, a partir de la tabla de contingencia

Sexo\Fuma	Si	No	n_i
Mujer	12	28	40
Hombre	26	34	60
n_j	38	62	100

se obtienen las siguientes frecuencias esperadas:

Sexo	Si	No	n_i
Mujer	$\frac{40 \cdot 38}{100} = 15,2$	$\frac{40 \cdot 62}{100} = 24,8$	40
Hombre	$\frac{60 \cdot 38}{100} = 22,8$	$\frac{60 \cdot 62}{100} = 37,2$	60
n_j	38	62	100

y el coeficiente χ^2 vale

$$\chi^2 = \frac{(12 - 15,2)^2}{15,2} + \frac{(28 - 24,8)^2}{24,8} + \frac{(26 - 22,8)^2}{22,8} + \frac{(34 - 37,2)^2}{37,2} = 1,81,$$

lo que indica que no existe gran relación entre el sexo y el hábito de fumar.

Coeficiente de contingencia

El coeficiente χ^2 depende del tamaño muestral, ya que al multiplicar por una constante las frecuencias de todas las casillas, su valor queda multiplicado por dicha constante, lo que podría llevarnos al equívoco de pensar que ha aumentado la relación, incluso cuando las proporciones se mantienen. En consecuencia el valor de χ^2 no está acotado superiormente y resulta difícil de interpretar.

Para evitar estos problemas se suele utilizar el siguiente estadístico:

Definición (Coeficiente de contingencia)

Dada una muestra de tamaño n en la que se han medido dos atributos X e Y , se define el *coeficiente de contingencia* como

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

Interpretación del coeficiente de contingencia

De la definición anterior se deduce que

$$0 \leq C \leq 1,$$

de manera que cuando $C = 0$ las variables son independientes, y crece a medida que aumenta la relación.

Aunque C nunca puede llegar a valer 1, se puede demostrar que para tablas de contingencia con k filas y k columnas, el valor máximo que puede alcanzar C es $\sqrt{(k-1)/k}$.

Ejemplo En el ejemplo anterior el coeficiente de contingencia vale

$$C = \sqrt{\frac{1,81}{1,81 + 100}} = 0,13.$$

Como se trata de una tabla de contingencia de 2×2 , el valor máximo que podría tomar el coeficiente de contingencia es $\sqrt{(2-1)/2} = \sqrt{1/2} = 0,707$, y como 0,13 está bastante lejos de este valor, se puede concluir que no existe demasiada relación entre el hábito de fumar y el sexo.

Teoría de la Probabilidad

- Experimentos y sucesos aleatorios
- Teoría de conjuntos
- Definición de probabilidad
- Probabilidad condicionada
- Dependencia e independencia de sucesos
- Teorema de la probabilidad total
- Teorema de Bayes
- Tests diagnósticos

La estadística descriptiva permite describir el comportamiento y las relaciones entre las variables en la muestra, pero no permite sacar conclusiones sobre el resto de la población.

Ha llegado el momento de dar el salto de la muestra a la población y pasar de la estadística descriptiva a la inferencia estadística, y el puente que lo permite es la **teoría de la probabilidad**.

Hay que tener en cuenta que el conocimiento que se puede obtener de la población a partir de la muestra es limitado, pero resulta evidente que la aproximación a la realidad de la población será mejor cuanto más representativa sea la muestra de ésta. Y recordemos que para que la muestra sea representativa de la población deben utilizarse técnicas de muestreo aleatorio, es decir, en la que los individuos se seleccionen al *azar*.

La teoría de la probabilidad precisamente se encarga de controlar ese azar para saber hasta qué punto son fiables las conclusiones obtenidas a partir de una muestra.

Experimentos y sucesos aleatorios

El estudio de una característica en una población se realiza a través de experimentos aleatorios.

Definición (Experimento aleatorio)

Un *experimento aleatorio* es aquel en el que se conoce cuál es el conjunto de resultados posibles antes de su realización pero se desconoce cuál será el resultado concreto del mismo.

Un ejemplo sencillo de experimentos aleatorios son los juegos de azar. Por ejemplo, el lanzamiento de un dado es un experimento aleatorio ya que:

- Se conoce el conjunto posibles de resultados $\{1, 2, 3, 4, 5, 6\}$.
- Antes de lanzar el dado, es imposible predecir con absoluta certeza el valor que saldrá.

Otro ejemplo de experimento aleatorio sería la selección de un individuo de una población al azar y la determinación de su grupo sanguíneo.

En general, la obtención de cualquier muestra mediante procedimientos aleatorios será un experimento aleatorio.

Definición (Espacio muestral)

Al conjunto E de todos los posibles resultados de un experimento aleatorio se le llama *espacio muestral*.

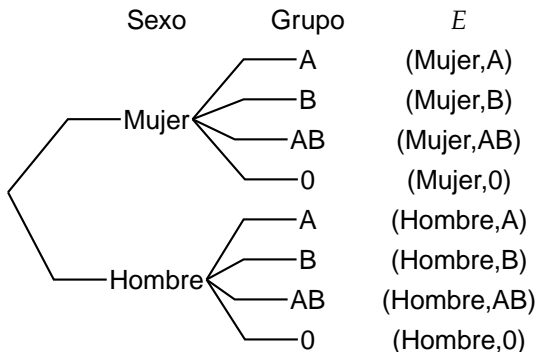
Algunos ejemplos de espacios muestrales son:

- Lanzamiento de una moneda: $E = \{c, x\}$.
- Lanzamiento de un dado: $E = \{1, 2, 3, 4, 5, 6\}$.
- Grupo sanguíneo de un individuo seleccionado al azar:
 $E = \{A, B, AB, 0\}$.
- Estatura de un individuo seleccionado al azar: \mathbb{R}^+ .

Construcción del espacio muestral

En los experimentos donde se miden más de una variable, la construcción del espacio muestral puede complicarse. En tales casos, es recomendable utilizar un **diagrama de árbol** de manera que cada nivel del árbol es una variable observada y cada rama un posible valor.

Por ejemplo, si el experimento consiste en observar el sexo y el grupo sanguíneo de una persona, el espacio muestral podría construirse mediante el siguiente árbol:



Definición (Suceso aleatorio)

Un *suceso aleatorio* es cualquier subconjunto del espacio muestral E de un experimento aleatorio.

Existen distintos tipos de sucesos:

Suceso imposible: Es el subconjunto vacío \emptyset . El suceso nunca ocurre.

Sucesos elementales: Son los subconjuntos formados por un solo elemento.

Sucesos compuestos: Son los subconjuntos formados por dos o más elementos.

Suceso seguro: Es el propio espacio muestral. El suceso seguro siempre ocurre.

Definición (Espacio de sucesos)

Dado un espacio muestral E de un experimento aleatorio, el conjunto formado por todos los posibles sucesos de E se llama *espacio de sucesos de E* y se denota $\mathcal{P}(E)$.

Ejemplo. Dado el espacio muestral $E = \{a, b, c\}$, se tiene

$$\mathcal{P}(E) = \{\emptyset, \{a\}, \{b\}, \{c\}, \{a, b\}, \{a, c\}, \{b, c\}, \{a, b, c\}\}$$

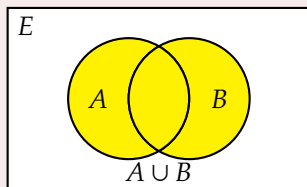
Puesto que los sucesos son conjuntos, por medio de la teoría de conjuntos se pueden definir las siguientes operaciones entre sucesos:

- Unión.
- Intersección.
- Complementario.
- Diferencia.

Definición (Suceso unión)

Dados dos sucesos $A, B \in \mathcal{P}(E)$, se llama *suceso unión* de A y B , y se denota $A \cup B$, al suceso formado por los elementos de A junto a los elementos de B , es decir,

$$A \cup B = \{x \mid x \in A \text{ o } x \in B\}.$$



El suceso unión $A \cup B$ ocurre siempre que ocurre $A \cup B$.

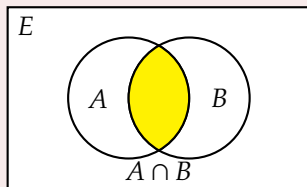
Ejemplo. Sea $E = \{1, 2, 3, 4, 5, 6\}$, el conjunto de los números de un dado, y $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$. Entonces $A \cup B = \{1, 2, 3, 4, 6\}$.

Intersección de sucesos

Definición (Suceso intersección)

Dados dos sucesos $A, B \in \mathcal{P}(E)$, se llama *suceso intersección* de A y B , y se denota $A \cap B$, al suceso formado por los elementos comunes de A y B , es decir,

$$A \cap B = \{x \mid x \in A \text{ y } x \in B\}.$$



El suceso intersección $A \cap B$ ocurre siempre que ocurren A y B .

Ejemplo. Sea $E = \{1, 2, 3, 4, 5, 6\}$, el conjunto de los números de un dado, y $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$. Entonces $A \cap B = \{2, 4\}$.

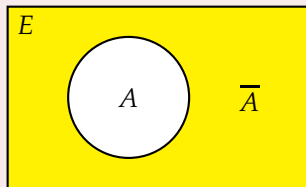
Diremos que dos sucesos son **incompatibles** si su intersección es vacía. Por ejemplo $A = \{2, 4, 6\}$ y $C = \{1, 3\}$ son incompatibles.

Contrario de un suceso

Definición (Suceso contrario)

Dado un conjunto $A \in \mathcal{P}(E)$, se llama *suceso contrario o complementario* de A , y se denota \bar{A} , al suceso formado por los elementos de E que no pertenecen a A , es decir,

$$\bar{A} = \{x \mid x \notin A\}.$$



El suceso contrario \bar{A} ocurre siempre que **no** ocurre A .

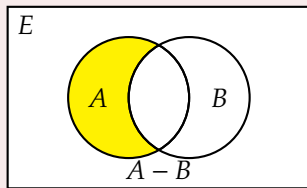
Ejemplo. Sea $E = \{1, 2, 3, 4, 5, 6\}$, el conjunto de los números de un dado, y $A = \{2, 4, 6\}$. Entonces $\bar{A} = \{1, 3, 5\}$.

Diferencia de sucesos

Definición (Suceso diferencia)

Dados dos sucesos $A, B \in \mathcal{P}(E)$, se llama *suceso diferencia* de A y B , y se denota $A - B$, al suceso formado por los elementos de A que no pertenecen a B , es decir,

$$A - B = \{x \mid x \in A \text{ y } x \notin B\}.$$



El suceso diferencia $A - B$ ocurre siempre que ocurre A pero no ocurre B , y también puede expresarse como $A \cap \bar{B}$.

Ejemplo. Sea $E = \{1, 2, 3, 4, 5, 6\}$, el conjunto de los números de un dado, y $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$. Entonces $A - B = \{6\}$.

Dados los sucesos $A, B, C \in \mathcal{P}(E)$, se cumplen las siguientes propiedades:

- 1 $A \cup A = A, A \cap A = A$ (idempotencia).
- 2 $A \cup B = B \cup A, A \cap B = B \cap A$ (conmutativa).
- 3 $(A \cup B) \cup C = A \cup (B \cup C), (A \cap B) \cap C = A \cap (B \cap C)$ (asociativa).
- 4 $(A \cup B) \cap C = (A \cap C) \cup (B \cap C), (A \cap B) \cup C = (A \cup C) \cap (B \cup C)$ (distributiva).
- 5 $A \cup \emptyset = A, A \cap E = A$ (elemento neutro).
- 6 $A \cup E = E, A \cap \emptyset = \emptyset$ (elemento absorbente).
- 7 $A \cup \overline{A} = E, A \cap \overline{A} = \emptyset$ (elemento simétrico complementario).
- 8 $\overline{\overline{A}} = A$ (doble contrario).
- 9 $\overline{A \cup B} = \overline{A} \cap \overline{B}, \overline{A \cap B} = \overline{A} \cup \overline{B}$ (leyes de Morgan).
- 10 $A \cap B \subseteq A \cup B$.

Definición clásica de probabilidad

Definición (Probabilidad Clásica de Laplace)

Para un experimento aleatorio donde todos los elementos del espacio muestral E son equiprobables, se define la *probabilidad* de un suceso $A \subseteq E$ como el cociente entre el número de elementos de A y el número de elementos de E :

$$P(A) = \frac{|A|}{|E|} = \frac{\text{nº casos favorables a } A}{\text{nº casos posibles}}$$

Esta definición es ampliamente utilizada, aunque tiene importantes restricciones:

- No puede utilizarse con espacios muestrales infinitos, o de los que no se conoce el número de casos posibles.
- Es necesario que todos los elementos del espacio muestral tengan la misma probabilidad de ocurrir (*equiprobabilidad*).

¡Ojo! Esto no se cumple en muchos experimentos aleatorios reales.

Definición frecuentista de probabilidad

Teorema (Ley de los grandes números)

Cuando un experimento aleatorio se repite un gran número de veces, las frecuencias relativas de los sucesos del experimento tienden a estabilizarse en torno a cierto número, que es precisamente su probabilidad.

De acuerdo al teorema anterior, podemos dar la siguiente definición

Definición (Probabilidad frecuentista)

Para un experimento aleatorio reproducible, se define la *probabilidad* de un suceso $A \subseteq E$ como la frecuencia relativa del suceso A en infinitas repeticiones del experimento:

$$P(A) = \lim_{n \rightarrow \infty} \frac{n_A}{n}$$

Aunque esta definición es muy útil en experimentos científicos reproducibles, también tiene serios inconvenientes, ya que

- Sólo se calcula una aproximación de la probabilidad real.
- La repetición del experimento debe ser en las mismas condiciones

Definición (Kolmogórov)

Se llama *probabilidad* a toda aplicación que asocia a cada suceso A del espacio de sucesos de un experimento aleatorio, un número real $P(A)$, que cumple los siguientes axiomas:

- 1 La probabilidad de un suceso cualquiera es positiva o nula:

$$P(A) \geq 0.$$

- 2 La probabilidad de la unión de dos sucesos incompatibles es igual a la suma de las probabilidades de cada uno de ellos:

$$P(A \cup B) = P(A) + P(B).$$

- 3 La probabilidad del suceso seguro es igual a la unidad:

$$P(E) = 1.$$

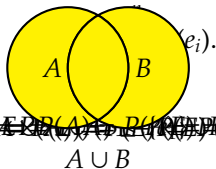
Consecuencias de los axiomas de probabilidad

A partir de los axiomas de la definición de probabilidad se pueden deducir los siguientes resultados:

- 1 $P(\bar{A}) = 1 - P(A)$.
- 2 $P(\emptyset) = 0$.
- 3 Si $A \subseteq B$ entonces $P(A) \leq P(B)$.
- 4 $P(A) \leq 1$.
- 5 Si A y B son sucesos compatibles, es decir, su intersección no es vacía, entonces

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

- 6 Si el suceso A está compuesto por los sucesos elementales e_1, e_2, \dots, e_n , entonces



$$A = \{e_1, e_2, \dots, e_n\} \Rightarrow P(A) = P(e_1 \cup e_2 \cup \dots \cup e_n) = P(e_1) + P(e_2) + \dots + P(e_n) = \sum_{i=1}^n P(e_i) = P(A) \cdot P(\{e_i\}).$$

Experimentos condicionados

En algunas ocasiones puede que haya que calcular la probabilidad de algún suceso A sabiendo que ha ocurrido otro B . En tal caso se dice que el suceso B es un *condicionante*, y la probabilidad del suceso condicionado suele escribirse como

$$P(A/B)$$

Los condicionantes, en el fondo, cambian el espacio muestral del experimento y por tanto las probabilidades de sus sucesos.

Ejemplo. Supongamos que hemos observado las siguientes frecuencias de aprobados en un grupo de 100 hombres y 100 mujeres:

	Aprobados	Suspensos
Mujeres	80	20
Hombres	60	40

Entonces, la probabilidad de que una persona elegida al azar haya aprobado es $P(\text{Aprobado}) = 140/200 = 0,7$.

Sin embargo, si se sabe que la persona elegida es mujer, entonces se tiene $P(\text{Aprobado}/\text{Mujer}) = 80/100 = 0,8$.

Probabilidad condicionada

Definición (Probabilidad condicionada)

Dados dos sucesos A y B de un mismo espacio de sucesos de un experimento aleatorio, la probabilidad de A condicionada por B es

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

siempre y cuando, $P(B) \neq 0$.

Esta definición permite calcular probabilidades sin tener que alterar el espacio muestral original del experimento.

Ejemplo. En el ejemplo anterior se tiene que la probabilidad del suceso Aprobado condicionada por el suceso Mujer es

$$P(\text{Aprobado}/\text{Mujer}) = \frac{P(\text{Aprobado} \cap \text{Mujer})}{P(\text{Mujer})} = \frac{80/200}{100/200} = \frac{80}{100} = 0,8.$$

De esta definición se deduce que la probabilidad de la intersección es

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

Definición (Sucesos independientes)

Dados dos sucesos A y B de un mismo espacio de sucesos de un experimento aleatorio, se dice que A es *independiente* de B , si la probabilidad de A no se ve alterada al condicionar por B , es decir,

$$P(A/B) = P(A).$$

Si A es independiente de B , también se cumple que B es independiente de A , y en general simplemente se dice que A y B son independientes.

También se cumple que si A y B son independientes, entonces

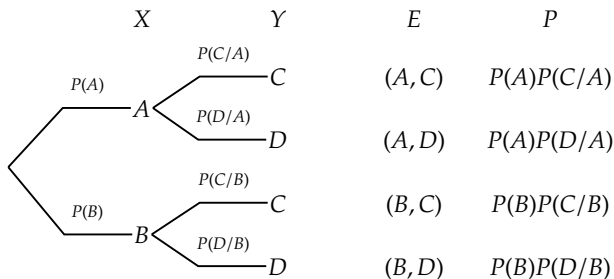
$$P(A \cap B) = P(A)P(B).$$

Árboles de probabilidad

Ya se vio que en experimentos donde se medía más de una variable, era conveniente construir el espacio muestral mediante un diagrama de árbol.

Dicho diagrama también es útil para calcular las probabilidades de cada uno de los elementos del espacio muestral del siguiente modo:

- 1 Para cada nodo del árbol, etiquetar su rama con la probabilidad de que la variable correspondiente tome el valor del nodo, condicionada por la ocurrencia de todos los nodos que conducen hasta el actual.
- 2 La probabilidad de cada suceso elemental se calcula multiplicando las probabilidades que etiquetan las ramas que conducen hasta él.

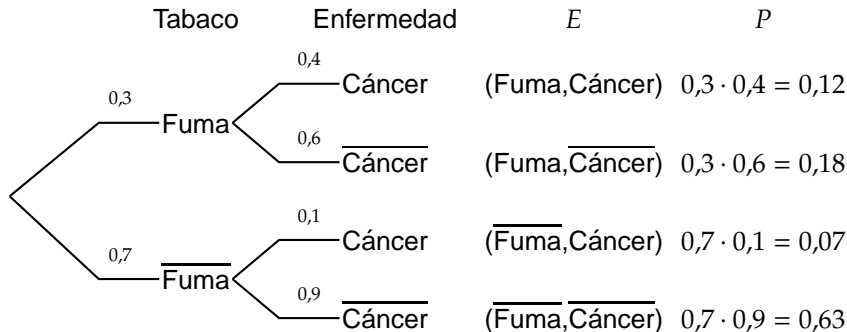


Árboles de probabilidad con variables dependientes

Ejemplo de dependencia del cáncer con respecto al tabaco

Sea una población en la que el 30 % de las personas fuman, y que la incidencia del cáncer de pulmón en fumadores es del 40 % mientras que en los no fumadores es del 10 %.

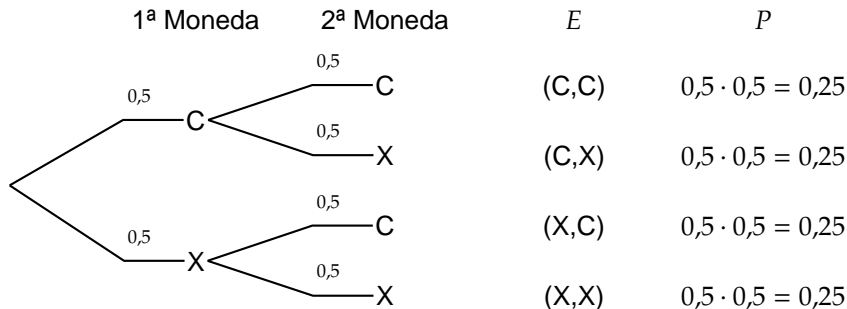
El árbol de probabilidad que expresa este experimento es:



Árboles de probabilidad con variables independientes

Ejemplo de independencia en el lanzamiento de dos monedas

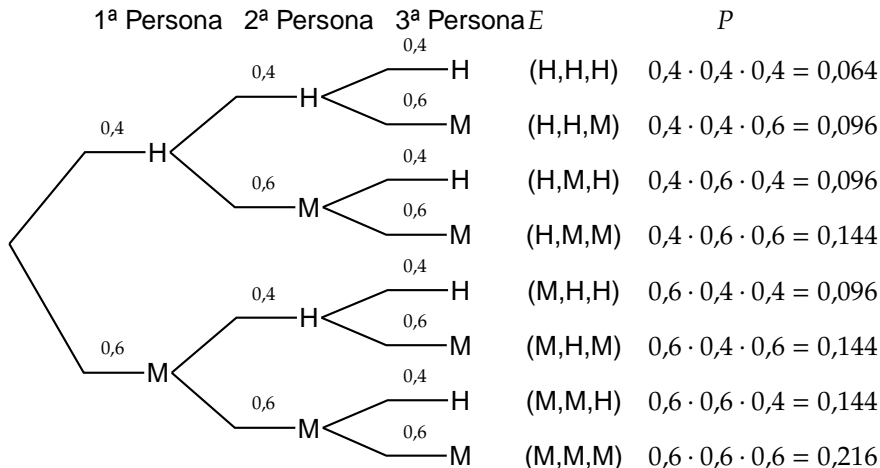
El árbol de probabilidad asociado al experimento aleatorio que consiste en el lanzamiento de dos monedas es:



Árboles de probabilidad con variables independientes

Ejemplo de independencia en la elección de una muestra aleatoria de tamaño 3

Dada una población en la que hay un 40 % de hombres y un 60 % de mujeres, el experimento aleatorio que consiste en tomar una muestra aleatoria de tres personas tiene el siguiente árbol de probabilidad:

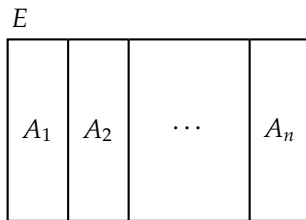


Sistema completo de sucesos

Definición (Sistema completo de sucesos)

Una colección de sucesos A_1, A_2, \dots, A_n de un mismo espacio de sucesos es un *sistema completo* si cumple las siguientes condiciones:

- 1 La unión de todos es el espacio muestral: $A_1 \cup \dots \cup A_n = E$.
- 2 Son incompatibles dos a dos: $A_i \cap A_j = \emptyset \ \forall i \neq j$.



En realidad un sistema completo de sucesos es una partición del espacio muestral de acuerdo a algún atributo, como por ejemplo el sexo o el grupo sanguíneo.

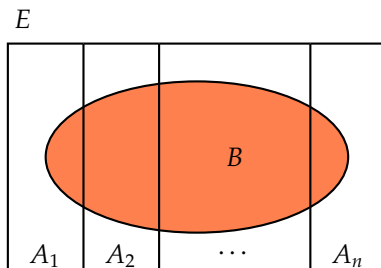
Teorema de la probabilidad total

Conocer las probabilidades de un determinado suceso en cada una de las partes de un sistema completo puede ser útil para calcular su probabilidad.

Teorema (Probabilidad total)

Dado un sistema completo de sucesos A_1, \dots, A_n y un suceso B de un mismo espacio de sucesos, se cumple

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



Teorema de la probabilidad total

Demostración

La demostración del teorema es sencilla, ya que al ser A_1, \dots, A_n un sistema completo tenemos

$$B = B \cap E = B \cap (A_1 \cup \dots \cup A_n) = (B \cap A_1) \cup \dots \cup (B \cap A_n)$$

y como estos sucesos son incompatibles entre sí, se tiene

$$\begin{aligned} P(B) &= P((B \cap A_1) \cup \dots \cup (B \cap A_n)) = P(B \cap A_1) + \dots + P(B \cap A_n) = \\ &= P(A_1)P(B/A_1) + \dots + P(A_n)P(B/A_n) = \sum_{i=1}^n P(A_i)P(B/A_i). \end{aligned}$$

Teorema de la probabilidad total

Un ejemplo de diagnóstico

Un determinado síntoma B puede ser originado por una enfermedad A pero también lo pueden presentar las personas sin la enfermedad. Sabemos que en la población la tasa de personas con la enfermedad A es $0,2$. Además, de las personas que presentan la enfermedad, el 90% presentan el síntoma, mientras que de las personas sin la enfermedad sólo lo presentan el 40% .

Si se toma una persona al azar de la población, *¿qué probabilidad hay de que tenga el síntoma?*

Para responder a la pregunta hay que fijarse en que el conjunto de sucesos $\{A, \bar{A}\}$ es un sistema completo, ya que $A \cup \bar{A} = E$ y $A \cap \bar{A} = \emptyset$, de modo que se puede aplicar el teorema de la probabilidad total:

$$P(B) = P(A)P(B/A) + P(\bar{A})P(B/\bar{A}) = 0,2 \cdot 0,9 + 0,8 \cdot 0,4 = 0,5.$$

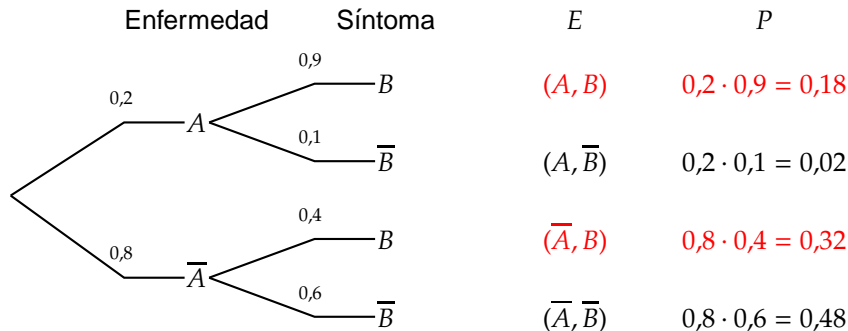
Es decir, la mitad de la población tendrá el síntoma.

¡En el fondo se trata de una media ponderada de probabilidades!

Teorema de la probabilidad total

Cálculo con el árbol de probabilidad

La respuesta a la pregunta anterior es evidente a la luz del árbol de probabilidad del experimento.



$$\begin{aligned}P(B) &= P(A, B) + P(\bar{A}, B) = \\&= P(A)P(B/A) + P(\bar{A})P(B/\bar{A}) = 0,2 \cdot 0,9 + 0,8 \cdot 0,4 = 0,18 + 0,32 = 0,5.\end{aligned}$$

Teorema de Bayes

Los sucesos de un sistema completo de sucesos A_1, \dots, A_n también pueden verse como las distintas hipótesis ante un determinado hecho B .

En estas condiciones resulta útil poder calcular las probabilidades a posteriori $P(A_i/B)$ de cada una de las hipótesis.

Teorema (Bayes)

Dado un sistema completo de sucesos A_1, \dots, A_n y un suceso B de un mismo espacio de sucesos, se cumple

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

Teorema de Bayes

Un ejemplo de diagnóstico

En el ejemplo anterior se ha visto cómo calcular la probabilidad de que una persona elegida al azar presente el síntoma, pero desde un punto de vista de diagnóstico clínico, una pregunta más interesante es:

Si llega a la consulta una persona que presenta el síntoma, *¿qué se debe diagnosticar?*

En este caso, las hipótesis ante las que hay que decidir son A y \bar{A} y sus probabilidades “a priori” son $P(A) = 0,2$ y $P(\bar{A}) = 0,8$.

Esto quiere decir que si no hubiese ninguna información sobre la persona, el diagnóstico sería que no tiene la enfermedad pues es mucho más probable que que la tenga.

Sin embargo, si al reconocer a la persona se observa que presenta el síntoma, dicha información condiciona a las hipótesis, y para decidir entre ellas es necesario calcular sus probabilidades “a posteriori”, es decir

$$P(A/B) \text{ y } P(\bar{A}/B)$$

Teorema de Bayes

Un ejemplo de diagnóstico

Para calcular las probabilidades “a posteriori” se puede utilizar el teorema de Bayes:

$$P(A/B) = \frac{P(A)P(B/A)}{P(A)P(B/A) + P(\bar{A})P(B/\bar{A})} = \frac{0,2 \cdot 0,9}{0,2 \cdot 0,9 + 0,8 \cdot 0,4} = \frac{0,18}{0,5} = 0,36,$$

$$P(\bar{A}/B) = \frac{P(\bar{A})P(B/\bar{A})}{P(A)P(B/A) + P(\bar{A})P(B/\bar{A})} = \frac{0,8 \cdot 0,4}{0,2 \cdot 0,9 + 0,8 \cdot 0,4} = \frac{0,32}{0,5} = 0,64.$$

Según esto, a pesar de que la probabilidad de estar enfermo ha aumentado, seguiríamos diagnosticando que no lo está, puesto que es más probable.

En este caso se dice que el síntoma B *no es determinante* a la hora de diagnosticar la enfermedad, pues la información que aporta no sirve para cambiar el diagnóstico en ningún caso.

Tests diagnósticos

En epidemiología es común el uso de tests para diagnosticar enfermedades.

Generalmente estos tests no son totalmente fiables, sino que hay cierta probabilidad de acierto o fallo en el diagnóstico, que suele representarse en la siguiente tabla:

	Presencia de la enfermedad (E)	Ausencia de la enfermedad (\bar{E})
Test positivo (+)	Diagnóstico acertado $P(+/E)$ Sensibilidad	Diagnóstico erróneo $P(+/\bar{E})$
Test negativo (-)	Diagnóstico erróneo $P(-/E)$	Diagnóstico acertado $P(-/\bar{E})$ Especificidad

Tests diagnósticos

La validez de una prueba diagnóstica depende de estas dos probabilidades:

Sensibilidad Es el porcentaje de positivos entre las personas enfermas:
 $P(+/E)$.

Especificidad Es el porcentaje de negativos entre las personas sanas:
 $P(-/\bar{E})$.

Pero lo realmente interesante de un test diagnóstico es su capacidad predictiva para diagnosticar, lo cual se mide mediante las siguientes probabilidades a posteriori:

Valor predictivo positivo Es el porcentaje de enfermos entre los positivos:
 $P(E/+)$.

Valor predictivo negativo Es el porcentaje de sanos entre los negativos:
 $P(\bar{E}/-)$.

Sin embargo, estos últimos valores dependen del porcentaje de enfermos en la población $P(E)$, lo que se conoce como, **tasa o prevalencia** de la enfermedad.

Ejemplo

Un test para diagnosticar la gripe tiene una sensibilidad del 95 % y una especificidad del 90 %. Según esto, las probabilidades de acierto y fallo del test son:

	Gripe	No gripe
Test +	0,95	0,10
Test -	0,05	0,90

Si la prevalencia de la gripe en la población es del 10 % y al aplicar el test a un individuo da positivo, *¿cuál es la probabilidad de que tenga gripe?*

Aplicando el teorema de Bayes, se tiene que el valor predictivo positivo del test vale

$$\begin{aligned}P(\text{Gripe}/+) &= \frac{P(\text{Gripe})P(+/\text{Gripe})}{P(\text{Gripe})P(+/\text{Gripe}) + P(\overline{\text{Gripe}})P(+/\overline{\text{Gripe}})} = \\&= \frac{0,1 \cdot 0,95}{0,1 \cdot 0,95 + 0,9 \cdot 0,1} = 0,5135.\end{aligned}$$

Aunque con esta probabilidad se diagnosticaría la enfermedad en caso de que el test diese positivo, se trata de un valor predictivo positivo muy bajo.

Ejemplo

	Gripe	No gripe
Test +	0,95	0,10
Test -	0,05	0,90

Y si el test da negativo, *¿cuál es la probabilidad de que no tenga gripe?*

De nuevo, aplicando el teorema de Bayes, se tiene que el valor predictivo negativo del test vale

$$\begin{aligned}P(\overline{\text{Gripe}}/-) &= \frac{P(\overline{\text{Gripe}})P(-/\overline{\text{Gripe}})}{P(\text{Gripe})P(-/\text{Gripe}) + P(\overline{\text{Gripe}})P(-/\overline{\text{Gripe}})} = \\&= \frac{0,9 \cdot 0,9}{0,1 \cdot 0,05 + 0,9 \cdot 0,9} = 0,9939.\end{aligned}$$

De manera que el valor predictivo negativo de este test es mucho más alto que el valor predictivo positivo.

Variables Aleatorias

- Variables Aleatorias Discretas
 - Distribución Uniforme
 - Distribución Binomial
 - Distribución de Poisson
- Variables aleatorias continuas
 - Distribución Uniforme continua
 - Distribución Normal
 - Distribución Chi-cuadrado
 - Distribución T de Student
 - Distribución F de Fisher-Snedecor

Variable aleatoria

Cuando seleccionamos una muestra al azar de una población estamos realizando un experimento aleatorio y cualquier variable estadística medida a partir de la muestra será una variable aleatoria porque sus valores dependerán del azar.

Definición (Variable Aleatoria)

Una *variable aleatoria* X es una función que asocia un número real a cada elemento del espacio muestral de un experimento aleatorio.

$$X : E \rightarrow \mathbb{R}$$

Al conjunto de posibles valores que puede tomar la variable aleatoria se le llama *rango* o *recorrido* de la variable.

En el fondo, una variable aleatoria es una variable cuyos valores provienen de la realización de un experimento aleatorio, y por tanto, tendrá asociada una determinada distribución de probabilidad.

Un ejemplo de variable aleatoria es la que mide el resultado del lanzamiento de un dado.

Las variables aleatorias se clasifican en dos tipos:

Discretas (VAD): Toman valores aislados (recorrido finito o infinito numerable).

Ejemplo. Número de hijos, número de accidentes, número de cigarrillos, etc.

Continuas (VAC): Toman valores en un intervalo real.

Ejemplo. Peso, estatura, nivel de colesterol, tiempo de respuesta a un fármaco, etc.

Los modelos probabilísticos de cada tipo de variables tienen características diferenciadas y por eso se estudiarán por separado.

Distribución de probabilidad de una variable discreta

Como los valores de una variable aleatoria están asociados a los sucesos elementales del correspondiente experimento aleatorio, cada valor tendrá asociada una probabilidad.

Definición (Función de probabilidad)

La *función de probabilidad* de una variable aleatoria discreta X es una función $f(x)$ que asocia a cada valor su probabilidad

$$f(x_i) = P(X = x_i).$$

Las probabilidades también pueden acumularse, al igual que se acumulaban las frecuencias en las muestras.

Definición (Función de distribución)

La *función de distribución* de una variable aleatoria discreta X es una función $F(x)$ que asocia a cada valor x_i la probabilidad de que la variable tome un valor menor o igual que dicho valor.

$$F(x_i) = P(X \leq x_i) = f(x_1) + \cdots + f(x_i).$$

Distribución de probabilidad de una variable discreta

Al recorrido de la variable, junto a su función de probabilidad o de distribución, se le llama **Distribución de probabilidad** de la variable.

Tanto la función de probabilidad como la de distribución suelen representarse en forma de tabla

X	x_1	x_2	\cdots	x_n	Σ
$f(x)$	$f(x_1)$	$f(x_2)$	\cdots	$f(x_n)$	1
$F(x)$	$F(x_1)$	$F(x_2)$	\cdots	$F(x_n) = 1$	

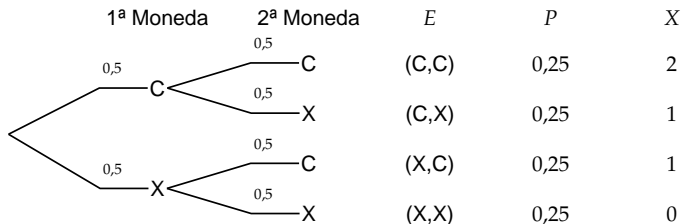
Al igual que la distribución de frecuencias de una variable reflejaba cómo se distribuían los valores de la variable en una muestra, la distribución de probabilidad de una variable aleatoria sirve para reflejar cómo se distribuyen los valores de dicha variable en toda la población.

Distribución de probabilidad de una variable discreta

Ejemplo del lanzamiento de dos monedas

Sea X la variable aleatoria que mide el número de caras en el lanzamiento de dos monedas.

El árbol de probabilidad asociado al experimento es



y según esto, su distribución de probabilidad es

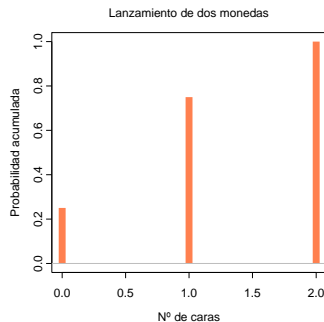
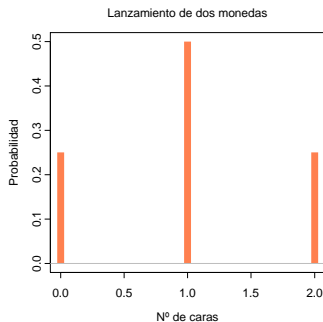
X	0	1	2
$f(x)$	0,25	0,5	0,25
$F(x)$	0,25	0,75	1

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 0,25 & \text{si } 0 \leq x < 1 \\ 0,75 & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

Gráfico de la función de probabilidad

Ejemplo del lanzamiento de dos monedas

Tanto la función de probabilidad como la de distribución también suelen representarse gráficamente:



Estadísticos poblacionales

Al igual que para describir las variables medidas en las muestras se utilizan estadísticos descriptivos, para describir determinadas características de las variables aleatorias se utilizan también estadísticos poblacionales.

La definición de los estadísticos poblacionales es análoga a la de los muestrales, pero utilizando probabilidades en lugar de frecuencias relativas.

Los más importantes son¹:

- **Media o esperanza matemática:**

$$\mu = E(X) = \sum_{i=1}^n x_i f(x_i)$$

- **Varianza:**

$$\sigma^2 = Var(X) = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2$$

- **Desviación típica:**

$$\sigma = +\sqrt{\sigma^2}$$

¹Para distinguirlos de los muestrales se suelen representar con letras griegas

Estadísticos poblacionales

Ejemplo de cálculo en el caso del lanzamiento de dos monedas

En el ejemplo del lanzamiento de dos monedas, a partir de la distribución de probabilidad

X	0	1	2
f(x)	0,25	0,5	0,25
F(x)	0,25	0,75	1

se pueden calcular fácilmente los estadísticos poblacionales:

$$\mu = \sum_{i=1}^n x_i f(x_i) = 0 \cdot 0,25 + 1 \cdot 0,5 + 2 \cdot 0,25 = 1 \text{ cara},$$

$$\sigma^2 = \sum_{i=1}^n x_i^2 f(x_i) - \mu^2 = (0^0 \cdot 0,25 + 1^2 \cdot 0,5 + 2^2 \cdot 0,25) - 1^2 = 0,5 \text{ caras}^2,$$

$$\sigma = +\sqrt{0,5} = 0,71 \text{ caras}.$$

En teoría, para obtener la distribución de probabilidad de una variable aleatoria en una población es necesario conocer el valor de la variable en todos los individuos de la población, lo cual muchas veces es imposible.

Sin embargo, dependiendo de la naturaleza del experimento, a veces es posible obtener la distribución de probabilidad de una variable aleatoria sin medirla en toda la población.

Dependiendo del tipo de experimento, existen diferentes modelos de distribución de probabilidad discretos. Los más habituales son:

- Distribución Uniforme.
- Distribución Binomial.
- Distribución de Poisson.

Distribución Uniforme $U_d(a, b)$

Cuando por la simetría del experimento, todos los valores $a = x_1, \dots, x_k = b$ de una variable discreta X son igualmente probables, se dice que la variable sigue un *modelo de distribución uniforme*.

Definición (Distribución uniforme $U_d(a, b)$)

Se dice que una variable aleatoria X sigue un *modelo de distribución uniforme* de parámetros a, b , y se nota, $X \sim U_d(a, b)$, si su recorrido es $Re(X) = \{a = x_1, \dots, x_k = b\}$, y su función de probabilidad vale

$$f(x_i) = \frac{1}{k}.$$

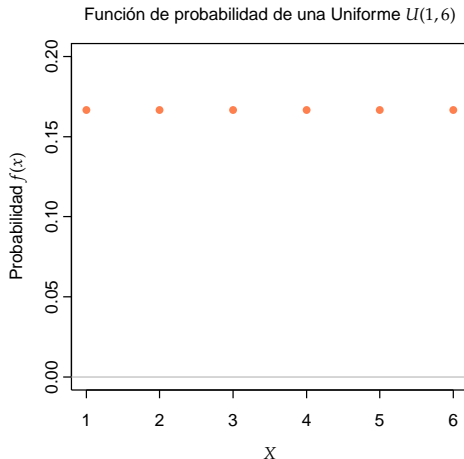
Su media y varianza valen

$$\mu = \sum_{i=1}^k x_i \frac{1}{k} \quad \sigma^2 = \sum_{i=1}^k (x_i - \mu)^2 \frac{1}{k}.$$

Distribución Uniforme $U_d(a, b)$

Ejemplo del lanzamiento de un dado

En el lanzamiento de un dado la variable que mide el número obtenido sigue un modelo de distribución uniforme $U_d(1, 6)$.



Distribución Binomial

Sea un experimento aleatorio con las siguientes características:

- El experimento consiste en una secuencia de n repeticiones de un mismo ensayo aleatorio.
- Los ensayos se realizan bajo idénticas condiciones, y cada uno de ellos tiene únicamente dos posibles resultados, que habitualmente se denotan por éxito (A) o fracaso (\bar{A}).
- Los ensayos son independientes, por lo que el resultado de cualquier ensayo en particular no influye sobre el resultado de cualquier otro.
- La probabilidad de éxito es idéntica para todos los ensayos y vale $P(A) = p$.

En estas condiciones, la variable aleatoria X que mide le número de éxitos obtenidos en los n ensayos sigue un *modelo de distribución binomial* de parámetros n y p .

Distribución Binomial $B(n, p)$

Definición (Distribución Binomial $B(n, p)$)

Se dice que una variable aleatoria X sigue un *modelo de distribución binomial* de parámetros n y p , si su recorrido es $Re(X) = \{0, 1, \dots, n\}$, y su función de probabilidad vale

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

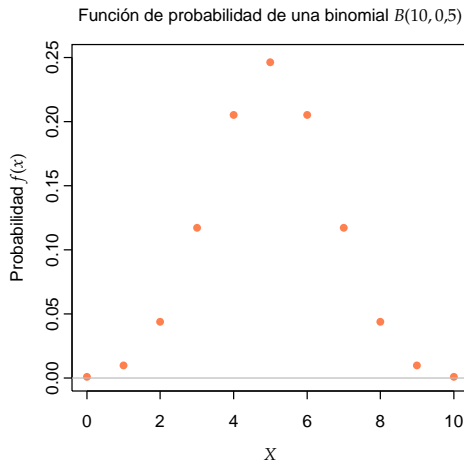
Su media y varianza valen

$$\mu = n \cdot p \quad \sigma^2 = n \cdot p \cdot (1-p).$$

Distribución Binomial $B(n, p)$

Ejemplo de 10 lanzamientos de una moneda

La variable que mide el número de caras obtenidos al lanzar 10 veces una moneda sigue un modelo de distribución binomial $B(10, 0,5)$.



Distribución Binomial $B(n, p)$

Ejemplo de 10 lanzamientos de una monedas

Sea $X \sim B(10, 0,5)$ la variable que mide el número de caras en 10 lanzamientos de una moneda. Entonces:

- La probabilidad de sacar 4 caras es

$$f(4) = \binom{10}{4} 0,5^4 (1 - 0,5)^{10-4} = \frac{10!}{4!6!} 0,5^4 0,5^6 = 210 \cdot 0,5^{10} = 0,2051.$$

- La probabilidad de sacar dos o menos caras es

$$\begin{aligned} F(2) &= f(0) + f(1) + f(2) = \\ &= \binom{10}{0} 0,5^0 (1 - 0,5)^{10-0} + \binom{10}{1} 0,5^1 (1 - 0,5)^{10-1} + \binom{10}{2} 0,5^2 (1 - 0,5)^{10-2} = \\ &= 0,0547. \end{aligned}$$

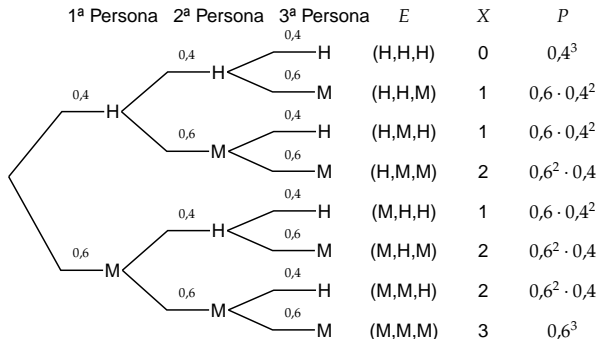
- Y el número esperado de caras es

$$\mu = 10 \cdot 0,5 = 5 \text{ caras.}$$

Distribución Binomial $B(n, p)$

Ejemplo de una muestra aleatoria con reemplazamiento

Dada una población con un 40 % de hombres y un 60 % de mujeres, la variable que mide el número de mujeres en una muestra aleatoria de tamaño 3, sigue una distribución binomial $X \sim B(3, 0,6)$.



$$f(0) = \binom{3}{0} 0,6^0 (1 - 0,6)^{3-0} = 0,4^3,$$

$$f(1) = \binom{3}{1} 0,6^1 (1 - 0,6)^{3-1} = 3 \cdot 0,6 \cdot 0,4^2,$$

$$f(2) = \binom{3}{2} 0,6^2 (1 - 0,6)^{3-2} = 3 \cdot 0,6^2 \cdot 0,4,$$

$$f(3) = \binom{3}{3} 0,6^3 (1 - 0,6)^{3-3} = 0,6^3.$$

Distribución de Poisson

Sea un experimento aleatorio con las siguientes características:

- El experimento consiste en observar la aparición de fenómenos puntuales sobre un soporte continuo, ya sea espacial o temporal. Por ejemplo: averías de máquinas en un espacio de tiempo, recepción de llamadas en una centralita, nº de linfocitos en un volumen de sangre, etc.
- El experimento produce, a largo plazo, un número medio constante de fenómenos puntuales por unidad de soporte continuo que llamaremos λ .

En estas circunstancias, la variable aleatoria X que mide el número de ocurrencias del fenómeno por unidad de soporte continuo sigue un *modelo de distribución de Poisson* de parámetro λ .

Distribución de Poisson $P(\lambda)$

Definición (Distribución de Poisson $P(\lambda)$)

Se dice que una variable aleatoria X sigue un *modelo de distribución de Poisson* de parámetro λ si su recorrido es $Re(X) = \{0, 1, \dots, \infty\}$, y su función de probabilidad vale

$$f(x) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

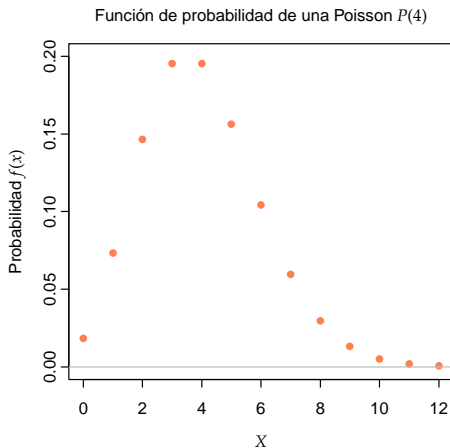
Su media y varianza valen

$$\mu = \lambda \quad \sigma^2 = \lambda.$$

Distribución de Poisson $P(\lambda)$

Ejemplo del número de ingresos en un hospital

Sea un hospital en el que se producen por término medio 4 ingresos diarios. Entonces la variable aleatoria X que mide el número de ingresos en un día sigue un modelo de distribución de Poisson $X \sim P(4)$.



Distribución de Poisson $P(\lambda)$

Ejemplo del número de ingresos en un hospital

Sea $X \sim P(4)$ la variable que mide el número de ingresos diarios en un hospital. Entonces:

- La probabilidad de que un día cualquiera se produzcan 5 ingresos es

$$f(5) = e^{-4} \frac{4^5}{5!} = 0,1563.$$

- La probabilidad de que un día se produzcan menos de 2 ingresos es

$$F(1) = f(0) + f(1) = e^{-4} \frac{4^0}{0!} + e^{-4} \frac{4^1}{1!} = 5e^{-4} = 0,0916.$$

- La probabilidad de que un día se produzcan más de un 1 ingresos es

$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - 0,0916 = 0,9084.$$

Aproximación del modelo Binomial mediante el Poisson

La ley de los casos raros

En realidad, el modelo de distribución de Poisson surge a partir del modelo de distribución Binomial, cuando el número de ensayos es muy grande $n \rightarrow \infty$ y la probabilidad de “éxito” es muy pequeña $p \rightarrow 0$.

En tales circunstancias, la variable $X \sim B(n, p)$ puede aproximarse mediante el modelo de distribución de Poisson $P(n \cdot p)$.

$$\lim_{n \rightarrow \infty, p \rightarrow 0} \binom{n}{x} p^x (1-p)^{n-x} = e^{-\lambda} \frac{\lambda^x}{x!}.$$

En la práctica, esta aproximación suele utilizarse para $n \geq 30$ y $p \leq 0,1$.

Aproximación del modelo Binomial mediante el Poisson

Ejemplo

Se sabe que una vacuna produce una reacción adversa en el 4 % de los casos. Si se vacunan 50 personas, ¿cuál es la probabilidad de que haya más de 2 personas con reacción adversa?

Está claro que la variable que mide el número de personas con reacción adversa entre las 50 personas vacunadas sigue un modelo de distribución binomial $X \sim B(50, 0,04)$, pero como $n = 50 > 30$ y $p = 0,04 < 0,1$, se cumplen las condiciones de la ley de los casos raros y se puede aproximar mediante una distribución de Poisson $P(50 \cdot 0,04) = P(2)$.

Así pues, utilizando la fórmula de la función de probabilidad de la distribución de Poisson, se tiene

$$\begin{aligned} P(X > 2) &= 1 - P(X \leq 2) = 1 - f(0) - f(1) - f(2) = 1 - e^{-2} \frac{2^0}{0!} - e^{-2} \frac{2^1}{1!} - e^{-2} \frac{2^2}{2!} = \\ &= 1 - 5e^{-2} = 0,3233. \end{aligned}$$

Variables aleatorias continuas

Las variables aleatorias continuas, a diferencia de las discretas, se caracterizan porque pueden tomar cualquier valor en un intervalo real. Es decir el conjunto de valores que pueden tomar no sólo es infinito, sino que además es no numerable.

Tal densidad de valores hace imposible el cálculo de las probabilidades de cada uno de ellos, y por tanto no podemos definir los modelos de distribución de probabilidad por medio de una función de probabilidad como en el caso discreto.

Por otro lado, la medida de una variable aleatoria continua suele estar limitada por las imprecisiones del proceso o instrumento de medida. Por ejemplo, cuando se dice que una estatura es 1,68 m, no se está diciendo que es exactamente 1,68 m, sino que la estatura está entre 1,675 y 1,685 m, ya que el instrumento de medida sólo es capaz de precisar hasta cm.

Así pues, en el caso de variables continuas, *no tiene sentido medir probabilidades de valores aislados, sino que se medirán probabilidades de intervalos.*

Función de densidad

Para conocer cómo se distribuye la probabilidad entre los valores de una variable aleatoria continua se utiliza la función de densidad.

Definición (Función de densidad)

La *función de densidad* de una variable aleatoria continua X es una función $f(x)$ que cumple las siguientes propiedades:

- Es no negativa: $f(x) \geq 0 \ \forall x \in \mathbb{R}$,
- El área acumulada entre la función y el eje de abscisas es 1, es decir,

$$\int_{-\infty}^{\infty} f(x) \, dx = 1.$$

La probabilidad de que la variable tome un valor dentro un intervalo cualquiera $[a, b]$ es

$$P(a \leq X \leq b) = \int_a^b f(x) \, dx$$

¡Ojo! $f(x)$ no es la probabilidad de que la variable tome el valor x .

Función de distribución

Al igual que para las variables discretas, también tiene sentido medir probabilidades acumuladas por debajo de un determinado valor.

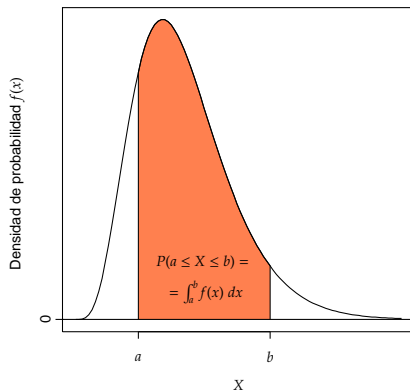
Definición (Función de distribución)

La *función de distribución* de una variable aleatoria continua X es una función $F(x)$ que asocia a cada valor a la probabilidad de que la variable tome un valor menor o igual que dicho valor.

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

Cálculo de probabilidades como áreas

La función de densidad nos permite calcular la probabilidad un intervalo $[a, b]$ como el área acumulada por debajo de la función en dicho intervalo.



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

Cálculo de probabilidades como áreas

Ejemplo

Dada la siguiente función:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0, \end{cases}$$

veamos que se trata de una función de densidad. Para ello hay que comprobar que es no negativa, lo cual es evidente al tratarse de una función exponencial, y que el área por debajo de ella es 1:

$$\begin{aligned} \int_{-\infty}^{\infty} f(x) dx &= \int_{-\infty}^0 f(x) dx + \int_0^{\infty} f(x) dx = \int_{-\infty}^0 0 dx + \int_0^{\infty} e^{-x} dx = \\ &= [-e^{-x}]_0^{\infty} = -e^{-\infty} + e^0 = 1. \end{aligned}$$

Ahora, a partir de ella, se puede calcular por ejemplo la probabilidad de que la variable tome un valor entre 0 y 2.

$$P(0 \leq X \leq 2) = \int_0^2 f(x) dx = \int_0^2 e^{-x} dx = [-e^{-x}]_0^2 = -e^{-2} + e^0 = 0,8646.$$

Estadísticos poblacionales

El cálculo de los estadísticos poblacionales es similar al caso discreto, pero utilizando la función de densidad, en lugar de la función de probabilidad, y extendiendo la suma discreta a la integral en todo el recorrido de la variable.

Los más importantes son:

- **Media o esperanza matemática:**

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x) dx$$

- **Varianza:**

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

- **Desviación típica:**

$$\sigma = + \sqrt{\sigma^2}$$

Cálculo de los estadísticos poblacionales

Ejemplo

Sea la función de densidad del ejemplo anterior:

$$f(x) = \begin{cases} 0 & \text{si } x < 0 \\ e^{-x} & \text{si } x \geq 0 \end{cases}$$

Su media es

$$\begin{aligned} \mu &= \int_{-\infty}^{\infty} xf(x) dx = \int_{-\infty}^0 xf(x) dx + \int_0^{\infty} xf(x) dx = \int_{-\infty}^0 0 dx + \int_0^{\infty} xe^{-x} dx = \\ &= [-e^{-x}(1+x)]_0^{\infty} = 1. \end{aligned}$$

y su varianza vale

$$\begin{aligned} \sigma^2 &= \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \int_{-\infty}^0 x^2 f(x) dx + \int_0^{\infty} x^2 f(x) dx - \mu^2 = \\ &= \int_{-\infty}^0 0 dx + \int_0^{\infty} x^2 e^{-x} dx - \mu^2 = [-e^{-x}(x^2 + 2x + 2)]_0^{\infty} - 1^2 = 2e^0 - 1 = 2. \end{aligned}$$

Existen varios modelos de distribución de probabilidad que aparecen bastante a menudo en la naturaleza y también como consecuencia de los procesos de muestreo aleatorio simple.

A continuación veremos los más importantes:

- Distribución Uniforme continua.
- Distribución Normal.
- Distribución T de Student.
- Distribución Chi-cuadrado.
- Distribución F de Fisher-Snedecor.

Distribución Uniforme continua $U(a, b)$

Cuando todos los valores de una variable continua son equiprobables, se dice que la variable sigue un *modelo de distribución uniforme continuo*.

Definición (Distribución Uniforme continua)

Una variable aleatoria continua X , cuyo recorrido es el intervalo $[a, b]$, sigue un modelo de distribución *uniforme* $U(a, b)$, si todos los valores de la variable son equiprobables, y por tanto, su función de densidad es constante en todo el intervalo:

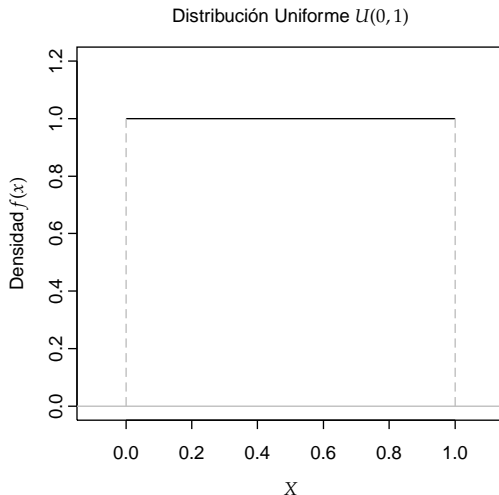
$$f(x) = \frac{1}{b-a} \quad \forall x \in [a, b]$$

Su media y varianza valen

$$\mu = \frac{a+b}{2} \quad \sigma^2 = \frac{(b-a)^2}{12}.$$

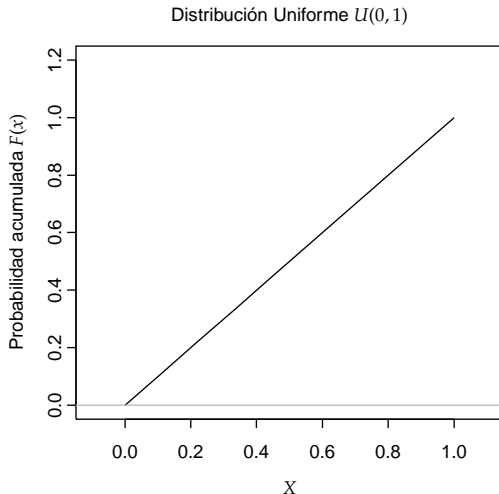
Función de densidad de la Uniforme continua $U(a, b)$

La generación aleatoria de un número real entre 0 y 1 sigue un modelo de distribución uniforme continuo $U(0, 1)$.



Función de distribución de la Uniforme continua $U(a, b)$

Como la función de densidad es constante, la función de distribución presenta un crecimiento lineal.



Cálculo de probabilidades con una Uniforme continua

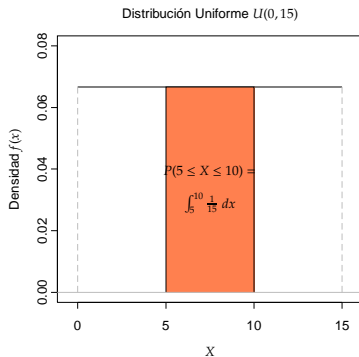
Ejemplo de espera de un autobús

Supóngase que un autobús pasa por una parada cada 15 minutos. Si una persona puede llegar a la parada en cualquier instante, *¿cuál es la probabilidad de que espere entre 5 y 10 minutos?*

En este caso, la variable X que mide el tiempo de espera sigue un modelo de distribución uniforme continua $U(0, 15)$ ya que cualquier valor entre los 0 y los 15 minutos es equiprobable.

Así pues, la probabilidad que nos piden es

$$\begin{aligned} P(5 \leq X \leq 10) &= \int_5^{10} \frac{1}{15} dx = \left[\frac{x}{15} \right]_5^{10} = \\ &= \frac{10}{15} - \frac{5}{15} = \frac{1}{3}. \end{aligned}$$



Además, el tiempo medio de espera será $\mu = \frac{0+15}{2} = 7,5$ minutos.

Distribución Normal $N(\mu, \sigma)$

El modelo de distribución normal es, sin duda, el modelo de distribución continuo más importante, ya que es el que más a menudo se presenta en la naturaleza.

Definición (Distribución Normal)

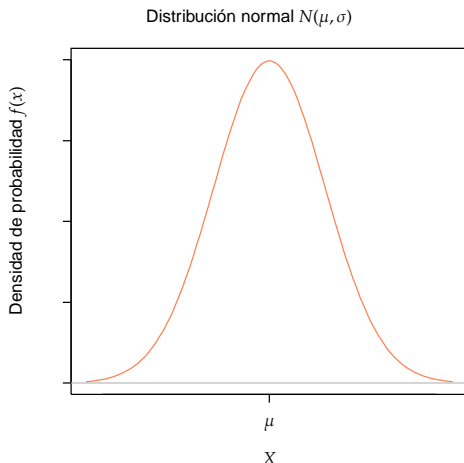
Una variable aleatoria continua X sigue un modelo de distribución *normal* $N(\mu, \sigma)$ si su recorrido es \mathbb{R} y su función de densidad vale

$$f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

La distribución normal depende de dos parámetros μ y σ que son, precisamente, su media y desviación típica.

Función de densidad de la Normal $N(\mu, \sigma)$

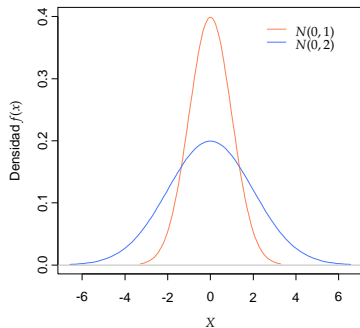
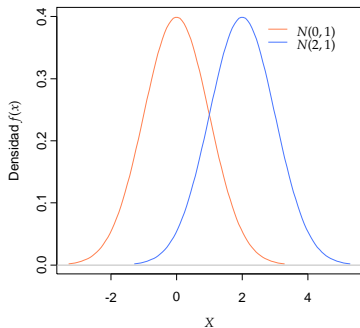
La gráfica de la función de densidad de la distribución normal tiene forma de una especie de campana, conocida como *campana de Gauss* (en honor a su descubridor), y está centrada en la media μ .



Función de densidad de la Normal $N(\mu, \sigma)$

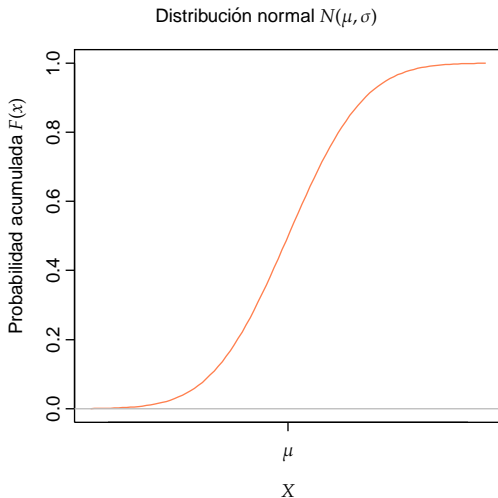
La forma de la campana de Gauss depende de sus dos parámetros:

- La media μ determina dónde está centrada.
- La desviación típica σ determina su anchura.



Función de distribución de la Normal $N(\mu, \sigma)$

Por su parte, la gráfica de la función de distribución tiene forma de S.



Propiedades de la distribución Normal

- La función de densidad es simétrica respecto a la media y por tanto, su coeficiente de asimetría es $g_1 = 0$.
- También es mesocúrtica, y por tanto, su coeficiente de apuntamiento vale $g_2 = 0$.
- La media, la mediana y la moda coinciden

$$\mu = Me = Mo.$$

- Tiende asintóticamente a 0 cuando x tiende a $\pm\infty$.

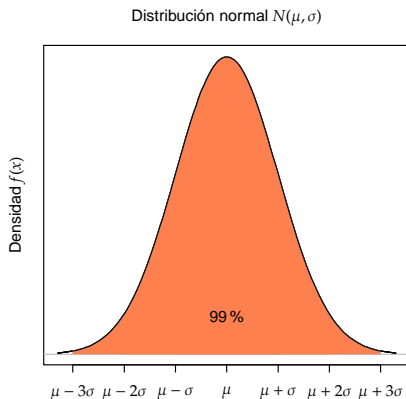
Propiedades de la distribución Normal

- Se cumple que

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0,68,$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0,95,$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0,99.$$



Propiedades de la distribución Normal

Ejemplo

En un estudio se ha comprobado que el nivel de colesterol total en mujeres sanas de entre 40 y 50 años sigue una distribución normal de media de 210 mg/dl y desviación típica 20 mg/dl. *¿Qué quiere decir esto?*

Atendiendo a las propiedades de la campana de Gauss, se tiene que

- El 68 % de las mujeres sanas tendrán el colesterol entre 210 ± 20 mg/dl, es decir, entre 190 y 230 mg/dl.
- El 95 % de las mujeres sanas tendrán el colesterol entre $210 \pm 2 \cdot 20$ mg/dl, es decir, entre 170 y 250 mg/dl.
- El 99 % de las mujeres sanas tendrán el colesterol entre $210 \pm 3 \cdot 20$ mg/dl, es decir, entre 150 y 270 mg/dl.

En la analítica sanguínea suele utilizarse el intervalo $\mu \pm 2\sigma$ para detectar posibles patologías. En el caso del colesterol, dicho intervalo es [170 mg/dl, 250 mg/dl]. Cuando una persona tiene el colesterol fuera de estos límites, se tiende a pensar que tiene alguna patología, aunque ciertamente podría estar sana, pero la probabilidad de que eso ocurra es sólo de un 5 %.

El teorema central del límite

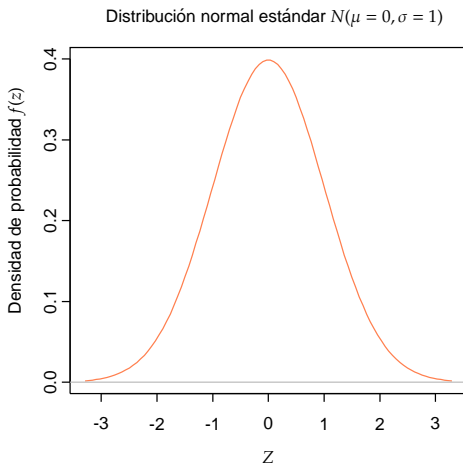
El comportamiento anterior lo presentan muchas variables continuas físicas y biológicas.

Si se piensa por ejemplo en la distribución de las estaturas, se verá que la mayor parte de los individuos presentan estaturas en torno a la media, tanto por arriba, como por debajo, pero que a medida que van alejándose de la media, cada vez hay menos individuos con dichas estaturas.

La justificación de que la distribución normal aparezca de manera tan frecuente en la naturaleza la encontramos en el **teorema central del límite**, que veremos más adelante, y que establece que si una variable aleatoria X proviene de un experimento aleatorio cuyos resultados son debidos a un conjunto muy grande de causas independientes que actúan sumando sus efectos, entonces X sigue una distribución aproximadamente normal.

La distribución Normal estándar $N(0, 1)$

De todas las distribuciones normales, la más importante es la que tiene media $\mu = 0$ y desviación típica $\sigma = 1$, que se conoce como **normal estándar** y se designa por Z .



Cálculo de probabilidades con la Normal estándar

Manejo de la tabla de la función de distribución

Para evitar tener que calcular probabilidades integrando la función de densidad de la normal estándar se suele utilizar su función de distribución.

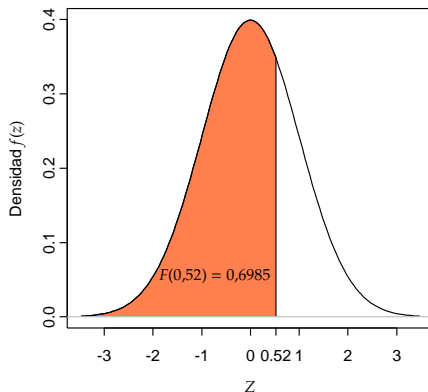
Habitualmente se suele manejar una tabla con los valores de la función de distribución tabulados cada centésima.

Ejemplo $P(Z \leq 0,52)$

z	0,00	0,01	0,02	...
0,0	0,5000	0,5040	0,5080	...
0,1	0,5398	0,5438	0,5478	...
0,2	0,5793	0,5832	0,5871	...
0,3	0,6179	0,6217	0,6255	...
0,4	0,6554	0,6591	0,6628	...
0,5	0,6915	0,6950	0,6985	...
⋮	⋮	⋮	⋮	⋮

0,52 → fila 0,5 + columna 0,02

Distribución normal estándar $N(\mu = 0, \sigma = 1)$



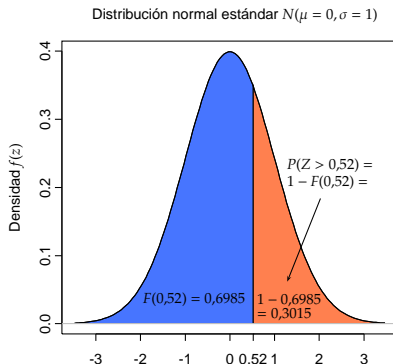
Cálculo de probabilidades con la Normal estándar

Probabilidades acumuladas por encima de un valor

Cuando tengamos que calcular probabilidades acumuladas por encima de un determinado valor, podemos hacerlo por medio de la probabilidad del suceso contrario.

Por ejemplo

$$P(Z > 0,52) = 1 - P(Z \leq 0,52) = 1 - F(0,52) = 1 - 0,6985 = 0,3015.$$



Tipificación

Ya se ha visto cómo calcular probabilidades con una distribución normal estándar, pero *¿qué hacer cuando la distribución normal no es la estándar?*

Afortunadamente, siempre se puede transformar una variable normal para convertirla en una normal estándar.

Teorema (Tipificación)

Si X es una variable normal de media μ y desviación típica σ , entonces la variable resultante de restarle a X su media y dividir por su desviación típica, sigue un modelo de distribución normal estándar:

$$X \sim N(\mu, \sigma) \Rightarrow Z = \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Esta transformación lineal se conoce como *transformación de tipificación* y la variable resultante Z se conoce como *normal tipificada*.

Así pues, para calcular probabilidades de una variable normal que no sea la normal estándar, se aplica primero la transformación de tipificación y después se puede utilizar la función de distribución de la normal estándar.

Cálculo de probabilidades tipificando

Ejemplo

Supóngase que la nota de un examen sigue un modelo de distribución de probabilidad normal $N(\mu = 6, \sigma = 1,5)$. *¿Qué porcentaje de suspensos habrá en la población?*

Para responder a esta pregunta necesitamos calcular la probabilidad $P(X < 5)$. Como X no es la normal estándar, se le aplica la transformación de tipificación $Z = \frac{X - \mu}{\sigma} = \frac{X - 6}{1,5}$:

$$P(X < 5) = P\left(\frac{X - 6}{1,5} < \frac{5 - 6}{1,5}\right) = P(Z < -0,67).$$

Después se mira en la tabla de la función de distribución de la normal estándar:

$$P(Z < -0,67) = F(-0,67) = 0,2514.$$

Así pues, habrán suspendido el 25,14 % de los alumnos.

Distribución chi-cuadrado $\chi^2(n)$

Definición (Distribución chi-cuadrado $\chi^2(n)$)

Si Z_1, \dots, Z_n son n variables aleatorias normales estándar independientes, entonces la suma de sus cuadrados sigue un modelo de distribución *chi-cuadrado de n grados de libertad*:

$$\chi^2(n) = Z_1^2 + \dots + Z_n^2.$$

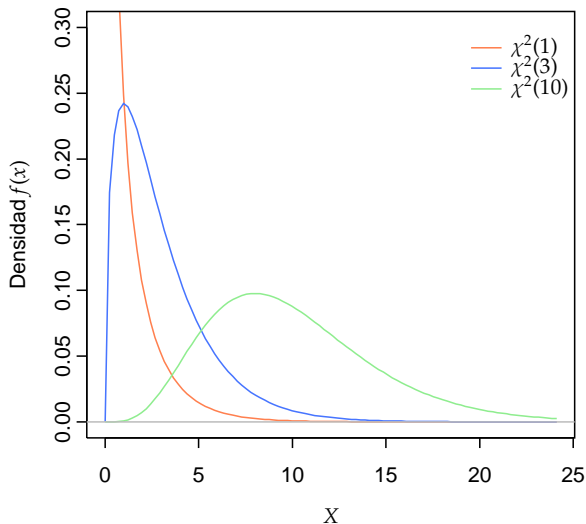
Su recorrido es \mathbb{R}^+ y su media y varianza valen

$$\mu = n, \quad \sigma^2 = 2n.$$

Como se verá más adelante, la distribución chi-cuadrado juega un papel importante en la estimación de la varianza poblacional y en el estudio de la relación entre variables cualitativas.

Función de densidad de la distribución chi-cuadrado

Distintas distribuciones chi-cuadrado



Propiedades de la distribución chi-cuadrado $\chi^2(n)$

- No toma valores negativos.
- Si $X \sim \chi^2(n)$ e $Y \sim \chi^2(m)$, entonces

$$X + Y \sim \chi^2(n + m).$$

- Al aumentar el número de grados de libertad, se aproxima asintóticamente a una normal.

Distribución T de Student $T(n)$

Definición (Distribución T de Student $T(n)$)

Si $Z \sim N(0, 1)$ es una variable aleatoria normal estándar y $X \sim \chi^2(n)$ es una variable aleatoria chi-cuadrado de n grados de libertad, ambas independientes, entonces la variable

$$T = \frac{Z}{\sqrt{X/n}},$$

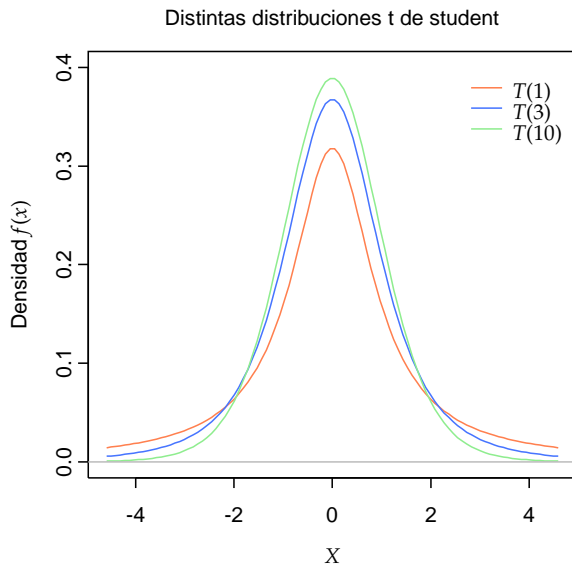
sigue un modelo de distribución *T de Student de n grados de libertad*.

Su recorrido es \mathbb{R} y su media y varianza valen

$$\mu = 0, \quad \sigma^2 = \frac{n}{n-2} \text{ si } n > 2.$$

Como se verá más adelante, la distribución T de Student juega un papel importante en la estimación la media poblacional.

Función de densidad de la distribución T de Student



Propiedades de la distribución T de Student $T(n)$

- Es simétrica con respecto a su media $\mu = 0$.
- Es muy similar a la normal estándar, pero algo más platicúrtica. Además, a medida que aumentan los grados de libertad, la gráfica de la distribución tiende hacia la de la normal estándar, hasta llegar a ser prácticamente iguales para $n \geq 30$.

$$T(n) \stackrel{n \rightarrow \infty}{\approx} N(0, 1).$$

Distribución F de Fisher-Snedecor $F(m, n)$

Definición (Distribución F de Fisher-Snedecor $F(m, n)$)

Si $X \sim \chi^2(m)$ es una variable aleatoria chi-cuadrado de m grados de libertad e $Y \sim \chi^2(n)$ es otra variable aleatoria chi-cuadrado de n grados de libertad, ambas independientes, entonces la variable

$$F = \frac{X/m}{Y/n},$$

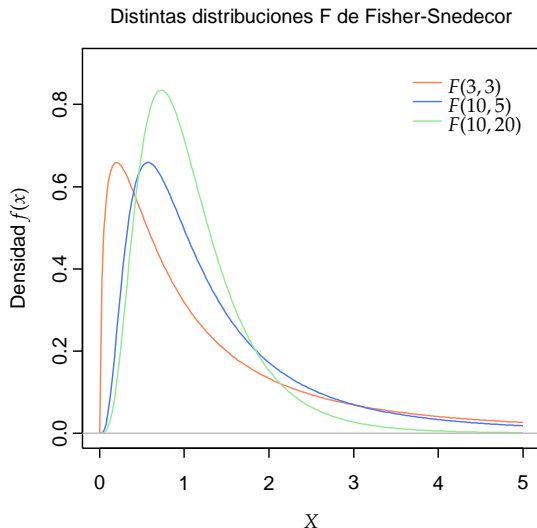
sigue un modelo de distribución *F de Fisher-Snedecor de m y n grados de libertad*.

Su recorrido es \mathbb{R}^+ y su media y varianza valen

$$\mu = \frac{n}{n-2}, \quad \sigma^2 = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \text{ si } n > 4.$$

Como se verá más adelante, la distribución F de Fisher-Snedecor juega un papel importante en la comparación de varianzas poblacionales y en el análisis de la varianza.

Función de densidad de la distribución F de Fisher-Snedecor $F(m, n)$



Propiedades de la distribución F de Fisher-Snedecor

$F(m, n)$

- No está definida para valores negativos.
- De la definición se deduce que

$$F(m, n) = \frac{1}{F(n, m)}$$

de manera que si llamamos $f(m, n)_p$ al valor que cumple que $P(F(m, n) \leq f(m, n)_p) = p$, entonces se cumple

$$f(m, n)_p = \frac{1}{f(n, m)_{1-p}}$$

Esto resulta muy útil para utilizar las tablas de su función de distribución.



Estimación de Parámetros

- Distribuciones muestrales
- Estimadores
- Estimación puntual
- Estimación por intervalos
- Intervalos de confianza para una población
- Intervalos de confianza para la comparación dos poblaciones

Los modelos de distribución de probabilidad vistos en el tema anterior explican el comportamiento de las variables aleatorias, pero para ello debemos saber qué modelo de distribución sigue una determinada variable. Este es el primer paso de la etapa de *Inferencia Estadística*.

Para determinar con exactitud el modelo de distribución hay que conocer la característica estudiada en todos los individuos de la población, lo cual no es posible en la mayoría de los casos (inviabilidad económica, física, temporal, etc.).

Para evitar estos inconvenientes se recurre al estudio de una muestra, a partir de la cual se trata de averiguar, de manera *aproximada*, el modelo de distribución de la variable aleatoria.

Ventajas e inconvenientes del muestreo

Estudiar un número reducido de individuos de una muestra en lugar de toda la población tiene indudables ventajas:

- Menor coste.
- Mayor rapidez.
- Mayor facilidad.

Pero también presenta algunos inconvenientes:

- Necesidad de conseguir una muestra representativa.
- Posibilidad de cometer errores (sesgos).

Afortunadamente, estos errores pueden ser superados: La representatividad de la muestra se consigue eligiendo la modalidad de muestreo más apropiada para el tipo de estudio; en el caso de los errores, aunque no se pueden evitar, se tratará de reducirlos al máximo y acotarlos.

Los valores de una variable X en una muestra de tamaño n de una población puede entenderse como un valor de una variable aleatoria n -dimensional.

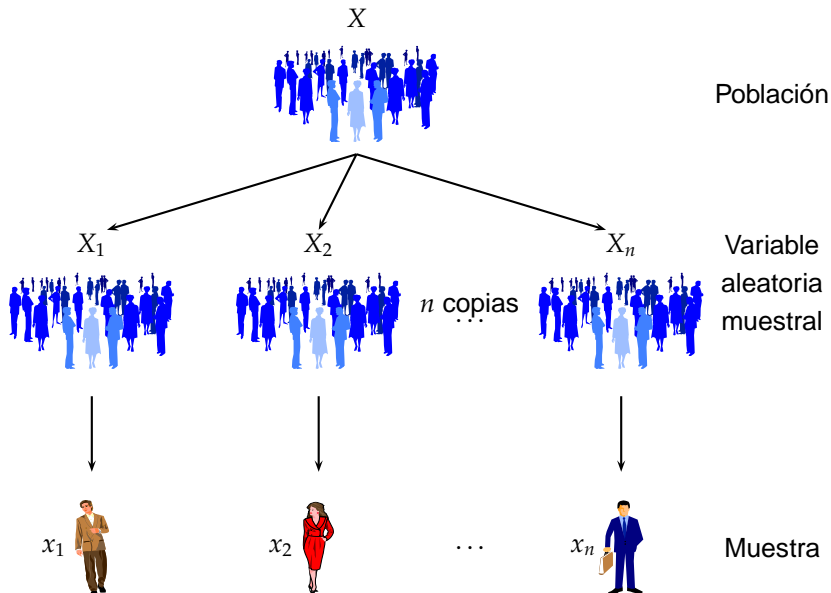
Definición (Variable aleatoria muestral)

Una *variable aleatoria muestral* de una variable X estudiada en una población es una colección de n variables aleatorias X_1, \dots, X_n tales que:

- Cada una de las variables X_i sigue la misma distribución de probabilidad que la variable X en la población.
- Todas las variables X_i son mutuamente independientes.

Los valores que puede tomar esta variable n dimensional, serán todas las posibles muestras de tamaño n que pueden extraerse de la población.

Obtención de una muestra



Estimación de parámetros

Las tres cuestiones fundamentales respecto a la variable aleatoria muestral son:

Homogeneidad : Las n variables que componen la variable aleatoria muestral siguen la misma distribución.

Independencia : Las variables son independientes entre sí.

Modelo de distribución : El modelo de distribución que siguen las n variables.

Las dos primeras cuestiones pueden resolverse si se utiliza muestreo aleatorio simple para obtener la muestra. En cuanto a la última, hay que responder, a su vez, a dos cuestiones:

- ¿Qué modelo de distribución se ajusta mejor a nuestro conjunto de datos? Esto se resolverá, en parte, mediante la utilización de técnicas no paramétricas.
- Una vez seleccionado el modelo de distribución más apropiado, ¿qué estadístico del modelo nos interesa y cómo determinar su valor? De esto último se encarga la parte de la inferencia estadística conocida como **Estimación de Parámetros**.

Parámetros a estimar

En este tema se abordará la segunda cuestión, es decir, suponiendo que se conoce el modelo de distribución de una población, se intentará estimar los principales parámetros que la definen. Por ejemplo, los principales parámetros que definen las distribuciones vistas en el tema anterior son:

Distribución	Parámetro
Binomial	n, p
Poisson	λ
Uniforme	a, b
Normal	μ, σ
Chi-cuadrado	n
T-Student	n
F-Fisher	m, n

Distribución de la variable aleatoria muestral

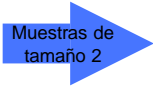
La distribución de probabilidad de los valores de la variable muestral depende claramente de la distribución de probabilidad de los valores de la población.

Ejemplo: Sea una población en la que la cuarta parte de las familias no tienen hijos, la mitad de las familias tiene 1 hijo, y el resto tiene 2 hijos.

Distribución Poblacional

X	$P(x)$
0	0,25
1	0,50
2	0,25

Muestras de tamaño 2



Distribución muestral

(X_1, X_2)	$P(x_1, x_2)$
(0, 0)	0,0625
(0, 1)	0,1250
(0, 2)	0,0625
(1, 0)	0,1250
(1, 1)	0,2500
(1, 2)	0,1250
(2, 0)	0,0625
(2, 1)	0,1250
(2, 2)	0,0625

Distribución de un estadístico muestral

Por ser función de una variable aleatoria, un estadístico en el muestreo es también una variable aleatoria.

Por tanto, su distribución de probabilidad también depende de la distribución de la población y de los parámetros que la determinan (μ, σ, p, \dots) .

Ejemplo: Si se toma la media muestral \bar{X} de las muestras de tamaño 2 del ejemplo anterior, su distribución de probabilidad es

Distribución muestral

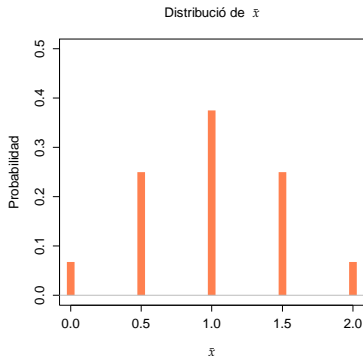
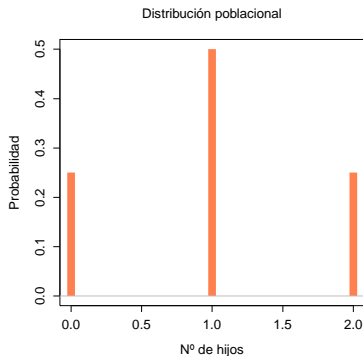
(X_1, X_2)	$P(x_1, x_2)$
(0, 0)	0,0625
(0, 1)	0,1250
(0, 2)	0,0625
(1, 0)	0,1250
(1, 1)	0,2500
(1, 2)	0,1250
(2, 0)	0,0625
(2, 1)	0,1250
(2, 2)	0,0625

Muestras de
tamaño 2

Distribución
de \bar{x}

\bar{X}	$P(x)$
0	0,0625
0,5	0,2500
1	0,3750
1,5	0,2500
2	0,0625

Distribución de un estadístico muestral



¿Cuál es la probabilidad de obtener una media muestral que aproxime la media poblacional con un error máximo de 0.5?

Teorema central del límite

Como hemos visto, para conocer la distribución de un estadístico muestral, es necesario conocer la distribución de la población, lo cual no siempre es posible. Afortunadamente, para muestras grandes es posible aproximar la distribución de algunos estadísticos como la media, gracias al siguiente teorema:

Teorema (Teorema central del límite)

Si X_1, \dots, X_n son variables aleatorias independientes ($n \geq 30$) con medias y varianzas $\mu_i = E(X_i)$, $\sigma_i^2 = Var(X_i)$, $i = 1, \dots, n$ respectivamente, entonces la variable aleatoria $X = X_1 + \dots + X_n$ sigue una distribución aproximadamente normal de media la suma de las medias y varianza la suma de las varianzas

$$X = X_1 + \dots + X_n \stackrel{n \geq 30}{\sim} N \left(\sum_{i=1}^n \mu_i, \sqrt{\sum_{i=1}^n \sigma_i^2} \right)$$

Este teorema además es la explicación de que la mayoría de las variables biológicas presenten una distribución normal, ya que suelen ser causa de múltiples factores que suman sus efectos de manera independiente.

Distribución de la media muestral

Muestras grandes ($n \geq 30$)

La media muestral de una muestra aleatoria de tamaño n es la suma de n variables aleatorias independientes, idénticamente distribuidas:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{X_1}{n} + \cdots + \frac{X_n}{n}$$

De acuerdo a las propiedades de las transformaciones lineales, la media y la varianza de cada una de estas variables son

$$E\left(\frac{X_i}{n}\right) = \frac{\mu}{n} \quad \text{y} \quad \text{Var}\left(\frac{X_i}{n}\right) = \frac{\sigma^2}{n^2}$$

con μ y σ^2 la media y la varianza de la población de partida.

Entonces, si el tamaño de la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la media muestral será normal:

$$\bar{X} \sim N\left(\sum_{i=1}^n \frac{\mu}{n}, \sqrt{\sum_{i=1}^n \frac{\sigma^2}{n^2}}\right) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

Distribución de la media muestral

Ejemplo para muestras grandes ($n \geq 30$)

Supongase que se desea estimar el número medio de hijos de una población con media $\mu = 2$ hijos y desviación típica $\sigma = 1$ hijo.

¿Qué probabilidad hay de estimar μ a partir de \bar{x} con un error menor de 0,2?

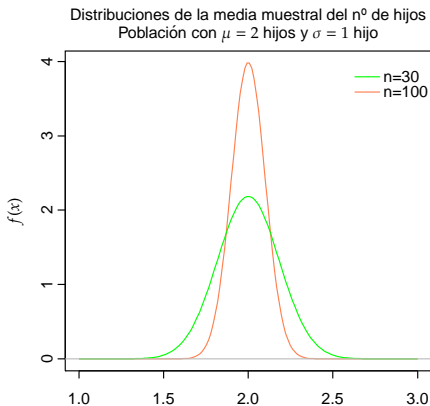
De acuerdo al teorema central del límite se tiene:

- Para $n = 30$,
 $\bar{x} \sim N(2, 1/\sqrt{30})$ y

$$P(1,8 < \bar{x} < 2,2) = 0,7267.$$

- Para $n = 100$,
 $\bar{x} \sim N(2, 1/\sqrt{100})$ y

$$P(1,8 < \bar{x} < 2,2) = 0,9545.$$



Distribución de una proporción muestral

Muestras grandes ($n \geq 30$)

Una proporción p poblacional puede calcularse como la media de una variable dicotómica (0,1). Esta variable se conoce como *variable de Bernoulli* $B(p)$, que es un caso particular de la binomial para $n = 1$. Por tanto, para una muestra aleatoria de tamaño n , una proporción muestral \hat{p} también puede expresarse como la suma de n variables aleatorias independientes, idénticamente distribuidas:

$$\hat{p} = \bar{X} = \frac{X_1 + \cdots + X_n}{n} = \frac{X_1}{n} + \cdots + \frac{X_n}{n}, \text{ con } X_i \sim B(p)$$

y con media y varianza

$$E\left(\frac{X_i}{n}\right) = \frac{p}{n} \quad \text{y} \quad \text{Var}\left(\frac{X_i}{n}\right) = \frac{p(1-p)}{n^2}$$

Entonces, si el tamaño de la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la proporción muestral también será normal:

$$\hat{p} \sim N\left(\sum_{i=1}^n \frac{p}{n}, \sqrt{\sum_{i=1}^n \frac{p(1-p)}{n^2}}\right) = N\left(p, \sqrt{\frac{p(1-p)}{n}}\right).$$

Estimador y estimación

Los estadísticos muestrales pueden utilizarse para aproximar los parámetros de la población, y cuando un estadístico se utiliza con este fin se le llama *estimador del parámetro*.

Definición (Estimador y estimación)

Un *estimador* es una función de la variable aleatoria muestral

$$\hat{\theta} = F(X_1, \dots, X_n).$$

Dada una muestra concreta (x_1, \dots, x_n) , el valor del estimador aplicado a ella se conoce como *estimación*

$$\hat{\theta}_0 = F(x_1, \dots, x_n).$$

Por ser una función de la variable aleatoria muestral, un estimador es, a su vez, una variable aleatoria cuya distribución depende de la población de partida.

Mientras que el estimador es una función que es única, la estimación no es única, sino que depende de la muestra tomada.

Estimador y estimación

Distribución de la población

X

Parámetro poblacional

θ



Variable aleatoria muestral

(X_1, \dots, X_n)

Estimador

$\hat{\theta} = F(X_1, \dots, X_n)$



Muestra de tamaño n

(x_1, \dots, x_n)

Estimación

$\hat{\theta}_0 = F(x_1, \dots, x_n)$



Estimador y estimación

Ejemplo

Supóngase que se quiere saber la proporción p de fumadores en una ciudad. En ese caso, la variable dicotómica que mide si una persona fuma (1) o no (0), sigue una distribución de Bernouilli $B(p)$.

Si se toma una muestra aleatoria de tamaño 5, $(X_1, X_2, X_3, X_4, X_5)$, de esta población, se puede utilizar la proporción de fumadores en la muestra como estimador para la proporción de fumadores en la población:

$$\hat{p} = \frac{\sum_{i=1}^5 X_i}{5}$$

Este estimador es una variable que se distribuye $\hat{p} \sim \frac{1}{n}B\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$.

Si se toman distintas muestras, se obtienen diferentes estimaciones:

Muestra	Estimación
(1, 0, 0, 1, 1)	3/5
(1, 0, 0, 0, 0)	1/5
(0, 1, 0, 0, 1)	2/5
...	...

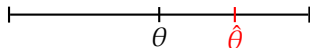
Tipos de estimación

La estimación de parámetros puede realizarse de dos formas:

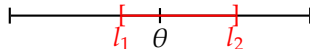
Estimación puntual : Se utiliza un único estimador que proporciona un valor o estimación aproximada del parámetro. El principal inconveniente de este tipo de estimación es que no se especifica la bondad de la estimación.

Estimación por intervalos : Se utilizan dos estimadores que proporcionan los extremos de un intervalo dentro del cual se cree que está el verdadero valor del parámetro con un cierto grado de seguridad. Esta forma de estimar sí permite controlar el error cometido en la estimación.

Estimación puntual



Estimación por intervalos



Estimación puntual

La estimación puntual utiliza un único estimador para estimar el valor del parámetro desconocido de la población.

En teoría pueden utilizarse distintos estimadores para estimar un mismo parámetro. Por ejemplo, en el caso de estimar la proporción de fumadores en una ciudad, podrían haberse utilizado otros posibles estimadores además de la proporción muestral, como pueden ser:

$$\hat{\theta}_1 = \sqrt[5]{X_1 X_2 X_3 X_4 X_5}$$

$$\hat{\theta}_2 = \frac{X_1 + X_5}{2}$$

$$\hat{\theta}_3 = X_1 \cdots$$

¿Cuál es el mejor estimador?

La respuesta a esta cuestión depende de las propiedades de cada estimador.

Aunque la estimación puntual no proporciona ninguna medida del grado de bondad de la estimación, existen varias propiedades que garantizan dicha bondad.

Las propiedades más deseables en un estimador son:

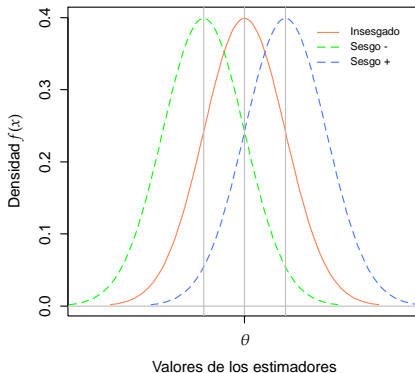
- Insesgadez
- Eficiencia
- Consistencia
- Normalidad asintótica
- Suficiencia

Definición (Estimador insesgado)

Un estimador $\hat{\theta}$ es *insesgado* para un parámetro θ si su esperanza es precisamente θ , es decir,

$$E(\hat{\theta}) = \theta.$$

Distribuciones de estimadores sesgados e insesgados



Sesgo de un estimador

Cuando un estimador no es insesgado, a la diferencia entre su esperanza y el valor del parámetro θ se le llama *sesgo*:

$$\text{Sesgo}(\hat{\theta}) = E(\hat{\theta}) - \theta.$$

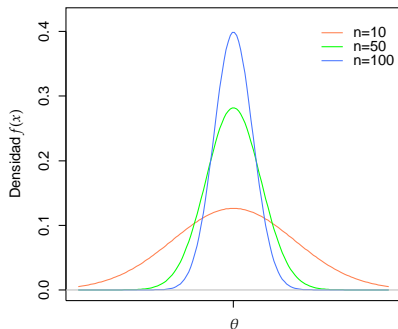
Cuanto menor sea el sesgo de un estimador, mejor se aproximarán sus estimaciones al verdadero valor del parámetro.

Definición (Estimador consistente)

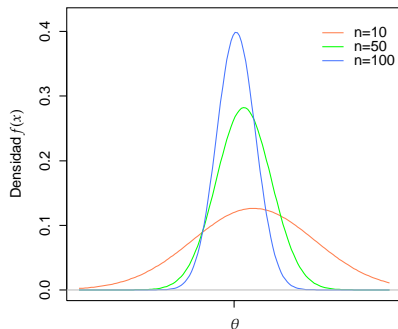
Un estimador $\hat{\theta}_n$ para muestras de tamaño n es *consistente* para un parámetro θ si para cualquier valor $\epsilon > 0$ se cumple

$$\lim_{n \rightarrow \infty} P(|\hat{\theta}_n - \theta| < \epsilon) = 1.$$

Distribuciones de estimadores consistentes



Distribuciones de estimadores consistentes sesgados



Valores de los estimadores

Valores de los estimadores

Condiciones para la consistencia

Las condiciones suficientes para que un estimador sea consistente son:

- $Sesgo(\hat{\theta}_n) = 0$ o $\lim_{n \rightarrow \infty} Sesgo(\hat{\theta}_n) = 0$.
- $\lim_{n \rightarrow \infty} Var(\hat{\theta}_n) = 0$.

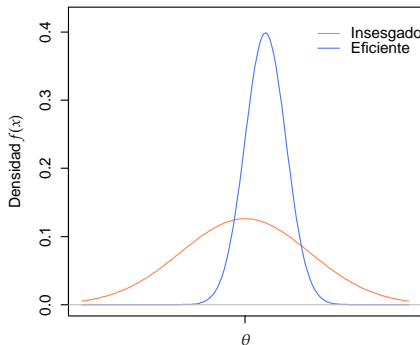
Así pues, si la varianza y el sesgo disminuyen a medida que aumenta el tamaño de la muestra, el estimador será consistente.

Definición (Estimador eficiente)

Un estimador $\hat{\theta}$ de un parámetro θ es *eficiente* si tiene el menor error cuadrático medio

$$ECM(\hat{\theta}) = \text{Sesgo}(\hat{\theta})^2 + \text{Var}(\theta).$$

Distribuciones de estimadores insesgado y eficiente sesgado

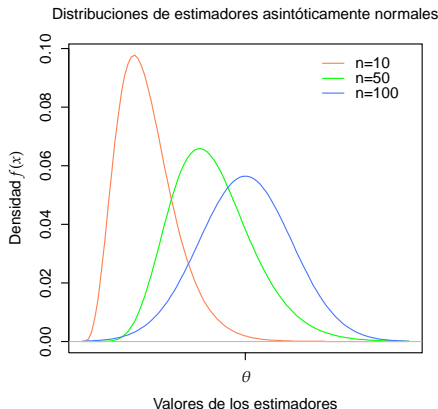


Valores de los estimadores

Normalidad asintótica

Definición (Estimador asintóticamente normal)

Un estimador $\hat{\theta}$ es *asintóticamente normal* si, independientemente de la distribución de la variable aleatoria muestral, su distribución es normal si el tamaño de la muestra es suficientemente grande.



Definición (Estimador suficiente)

Un estimador $\hat{\theta}$ es *suficiente* para un parámetro θ , si la distribución condicional de la variable aleatoria muestral, una vez dada la estimación $\hat{\theta} = \hat{\theta}_0$, no depende de θ .

Esto significa que cuando se obtiene una estimación, cualquier otra información es irrelevante para θ .

Estimador de la media poblacional

El estimador que se suele utilizar para estimar la media poblacional es la media muestral.

Para muestras de tamaño n resulta la siguiente variable aleatoria:

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

Si la población de partida tiene media μ y varianza σ^2 se cumple

$$E(\bar{X}) = \mu \quad \text{y} \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

Así pues, la media muestral es un estimador insesgado, y como su varianza disminuye a medida que aumenta el tamaño muestral, también es consistente y eficiente.

Estimador para la varianza poblacional: La cuasivarianza

Sin embargo, la varianza muestral

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

es un estimador sesgado para la varianza poblacional, ya que

$$E(S^2) = \frac{n-1}{n} \sigma^2.$$

No obstante, resulta sencillo corregir este sesgo para llegar a un estimador insesgado:

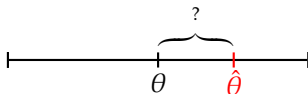
Definición (Cuasivarianza muestral)

Dada una muestra de tamaño n de una variable aleatoria X , se define la *cuasivarianza muestral* como

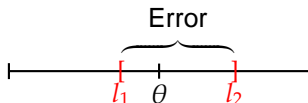
$$\hat{S}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{n}{n-1} S^2.$$

Estimación por intervalos

El principal problema de la estimación puntual es que, una vez seleccionada la muestra y hecha la estimación, resulta imposible saber el error cometido.



Para controlar el error de la estimación es mejor utilizar la estimación por intervalos



La estimación por intervalos trata de construir a partir de la muestra un intervalo dentro del cual se supone que se encuentra el parámetro a estimar con un cierto grado de confianza. Para ello se utilizan dos estimadores, uno para el límite inferior del intervalo y otro para el superior.

Definición (Intervalo de confianza)

Dados dos estimadores $\hat{l}_i(X_1, \dots, X_n)$ y $\hat{l}_s(X_1, \dots, X_n)$, y sus respectivas estimaciones l_1 y l_2 para una muestra concreta, se dice que el intervalo $I = [l_1, l_2]$ es un intervalo de confianza para un parámetro poblacional θ , con un nivel de confianza $1 - \alpha$ (o nivel de significación α), si se cumple

$$P(\hat{l}_i(X_1, \dots, X_n) \leq \theta \leq \hat{l}_s(X_1, \dots, X_n)) = 1 - \alpha.$$

Nivel de confianza

Un intervalo de confianza nunca garantiza con absoluta certeza que el parámetro se encuentra dentro él.

Tampoco se puede decir que la probabilidad de que el parámetro esté dentro del intervalo es $1 - \alpha$, ya que una vez calculado el intervalo, las variables aleatorias que determinan sus extremos han tomado un valor concreto y ya no tiene sentido hablar de probabilidad, es decir, o el parámetro está dentro, o está fuera, pero con absoluta certeza.

Lo que si se deduce de la definición es que el $(1 - \alpha) \%$ de los intervalos correspondientes a las todas las posibles muestras aleatorias, contendrán al parámetro. Es por eso que se habla de **confianza** y no de probabilidad.

Para que un intervalo sea útil su nivel de confianza debe ser alto:

$$1 - \alpha = 0,90 \text{ o } \alpha = 0,10$$

$$1 - \alpha = 0,95 \text{ o } \alpha = 0,05$$

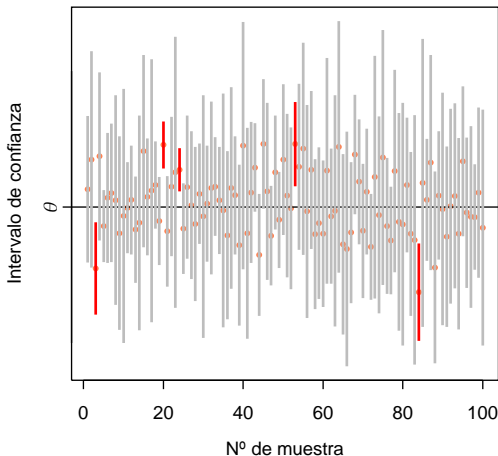
$$1 - \alpha = 0,99 \text{ o } \alpha = 0,01$$

siendo 0,95 el nivel de confianza más habitual y 0,99 en casos críticos.

Nivel de confianza

Teóricamente, de cada 100 intervalos para estimar un parámetro θ con nivel de confianza $1 - \alpha = 0,95$, 95 contendrían a θ y sólo 5 lo dejarían fuera.

100 intervalos de confianza del 95 % para estimar θ



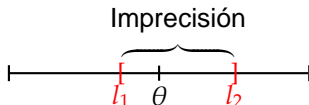
Error o imprecisión de un intervalo

Otro de los aspectos más importantes de un intervalo de confianza es su *error o imprecisión*.

Definición (Error o imprecisión de un intervalo)

El *error* o la *imprecisión* de un intervalo de confianza $[l_i, l_s]$ es su amplitud

$$A = l_s - l_i.$$



Para que un intervalo sea útil no debe ser demasiado impreciso.

¿De qué depende la imprecisión de un intervalo?

En general, la precisión de un intervalo depende de tres factores:

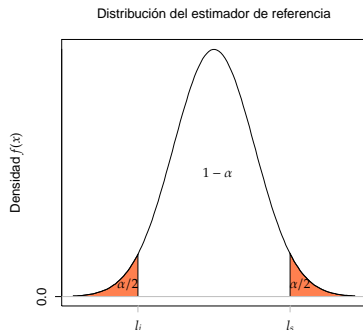
- La dispersión de la población. Cuanto más dispersa sea, menos preciso será el intervalo.
- El nivel de confianza. Cuanto mayor sea el nivel de confianza, menos preciso será el intervalo.
- El tamaño muestral. Cuanto mayor sea el tamaño muestral, más preciso será el intervalo.

Si la confianza y la precisión están reñidas, ¿cómo se puede ganar precisión sin perder confianza?

Cálculo de los intervalos de confianza

Habitualmente, para calcular un intervalo de confianza se suele partir de un estimador puntual del que se conoce su distribución muestral.

A partir de este estimador se calculan los extremos del intervalo sobre su distribución, buscando los valores que dejan encerrada una probabilidad $1 - \alpha$. Estos valores suelen tomarse de manera simétrica, de manera que el extremo inferior deje una probabilidad acumulada inferior $\alpha/2$ y el extremo superior deje una probabilidad acumulada superior también de $\alpha/2$.



Intervalos de confianza más importantes

Intervalos para una población:

- Intervalo para la media de una población normal con varianza conocida.
- Intervalo para la media de una población normal con varianza desconocida.
- Intervalo para la media de una población con varianza desconocida a partir de muestras grandes.
- Intervalo para la varianza de una población normal.
- Intervalo para una proporción de una población.

Intervalos para la comparación de dos poblaciones:

- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas conocidas.
- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas desconocidas pero iguales.
- Intervalo para la diferencia de medias de dos poblaciones normales con varianzas desconocidas y diferentes.
- Intervalo para el cociente de varianzas de dos poblaciones normales.
- Intervalo para la diferencia de proporciones de dos poblaciones.

Intervalo de confianza para la media de una población normal con varianza conocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- La media μ es desconocida, pero su varianza σ^2 es conocida.

Bajo estas hipótesis, la media muestral, para muestras de tamaño n , sigue también una distribución normal

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Tipificando la variable se tiene

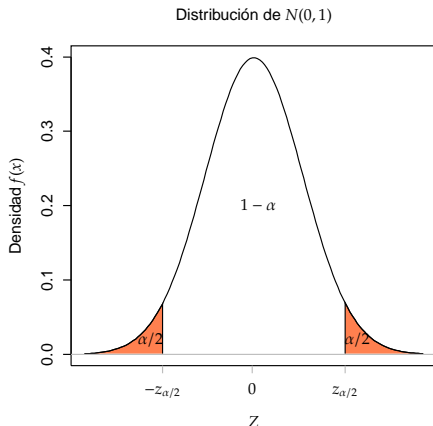
$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

Sobre esta distribución resulta sencillo calcular los valores z_i y z_s de manera que

$$P(z_i \leq Z \leq z_s) = 1 - \alpha.$$

Intervalo de confianza para la media de una población normal con varianza conocida

Como la distribución normal estándar es simétrica respecto al 0, lo mejor es tomar valores opuestos $-z_{\alpha/2}$ y $z_{\alpha/2}$ que dejen sendas colas de probabilidad acumulada $\alpha/2$.



Intervalo de confianza para la media de una población normal con varianza conocida

A partir de aquí, deshaciendo la tipificación, resulta sencillo llegar a los estimadores que darán los extremos del intervalo de confianza:

$$\begin{aligned}1 - \alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = P\left(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \leq z_{\alpha/2}\right) = \\&= P\left(-z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = \\&= P\left(-\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = \\&= P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right).\end{aligned}$$

Así pues, el intervalo de confianza para la media de una población normal con varianza conocida es:

$$\left[\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right] \text{ o bien } \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Características del intervalo

De la fórmula del intervalo de confianza

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

se deducen varias características:

- El intervalo está centrado en la media muestral \bar{X} que era el mejor estimador de la media poblacional.
- La amplitud o imprecisión del intervalo es

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

de manera que depende de:

- σ : cuanto mayor sea la varianza poblacional, mayor será la imprecisión.
- $z_{\alpha/2}$: que a su vez depende del nivel de confianza, y cuanto mayor sea $1 - \alpha$, mayor será la imprecisión.
- n : cuanto mayor sea el tamaño de la muestra, menor será la imprecisión.

Por tanto, la única forma de reducir la imprecisión del intervalo, manteniendo la confianza, es aumentando el tamaño muestral.

Control de la imprecisión mediante el tamaño muestral

Teniendo en cuenta que la amplitud o imprecisión del intervalo para la media de una población normal con varianza conocida es

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Leftrightarrow \sqrt{n} = 2z_{\alpha/2} \frac{\sigma}{A},$$

de donde se deduce

$$n = 4z_{\alpha/2}^2 \frac{\sigma^2}{A^2}$$

Intervalo de confianza para la media de una población normal con varianza conocida

Ejemplo

Sea una población de estudiantes en la que la puntuación obtenida en un examen sigue una distribución normal $X \sim N(\mu, \sigma = 1,5)$.

Para estimar la nota media μ , se toma una muestra de 10 estudiantes:

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

A partir de esta muestra, podemos calcular el intervalo de confianza para μ con un nivel de confianza $1 - \alpha = 0,95$ (nivel de significación $\alpha = 0,05$):

- $\bar{X} = \frac{4+\dots+3}{10} = \frac{53}{10} = 5,3$ puntos.
- $z_{\alpha/2} = z_{0,025}$ es el valor de la normal estándar que deja una probabilidad acumulada superior de 0,025, que vale aproximadamente 1,96.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5,3 \pm 1,96 \frac{1,5}{\sqrt{10}} = 5,3 \pm 0,93 = [4,37, 6,23].$$

Es decir, μ estaría entre 4,37 y 6,23 puntos con un 95 % de confianza.

Control de la imprecisión mediante el tamaño muestral

Ejemplo

La imprecisión del intervalo anterior es de $\pm 0,93$ puntos.

Si se desea reducir esta imprecisión a $\pm 0,5$ puntos, ¿qué tamaño muestral sería necesario?

$$n = 4z_{\alpha/2}^2 \frac{\sigma^2}{A^2} = 4 \cdot 1,96^2 \frac{1,5^2}{(2 \cdot 0,5)^2} = 34,57.$$

Por tanto, se necesitaría una muestra de al menos 35 estudiantes para conseguir un intervalo del 95 % de confianza y una precisión de $\pm 0,5$ puntos.

Intervalo de confianza para la media de una población normal con varianza desconocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Cuando se desconoce la varianza poblacional se suele estimar mediante la cuasivarianza \hat{S}^2 . Como consecuencia, el estimador de referencia ya no sigue una distribución normal como en el caso de conocer la varianza, sino un T de Student de $n - 1$ grados de libertad:

$$\left. \begin{array}{l} \bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \\ \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1) \end{array} \right\} \Rightarrow \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \sim T(n-1),$$

Intervalo de confianza para la media de una población normal con varianza desconocida

Como la distribución T de Student, al igual que la normal, también es simétrica respecto al 0, se pueden tomar dos valores opuestos $-t_{\alpha/2}^{n-1}$ y $t_{\alpha/2}^{n-1}$ de manera que

$$P\left(-t_{\alpha/2}^{n-1} \leq \frac{\bar{X} - \mu}{\hat{S}/\sqrt{n}} \leq t_{\alpha/2}^{n-1}\right) = 1 - \alpha.$$

y a partir de aquí se llega, razonando como antes, al intervalo

$$\left[\bar{X} - t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}, \bar{X} + t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \right] \text{ o bien } \bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

Control de la imprecisión mediante el tamaño muestral

Al igual que antes, teniendo en cuenta que la amplitud o imprecisión del intervalo para la media de una población con varianza desconocida es

$$A = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} \Leftrightarrow \sqrt{n} = 2t_{\alpha/2}^{n-1} \frac{\hat{S}}{A},$$

de donde se deduce

$$n = 4(t_{\alpha/2}^{n-1})^2 \frac{\hat{S}^2}{A^2}$$

El único problema, a diferencia del caso anterior en que σ era conocida, es que se necesita \hat{S} , por lo que se suele tomar una muestra pequeña previa para calcularla. Por otro lado, el valor de la T de student suele aproximarse asintóticamente por el de la normal estándar $t_{\alpha/2}^{n-1} \approx z_{\alpha/2}$.

Intervalo de confianza para la media de una población normal con varianza desconocida

Ejemplo

Supóngase que en el ejemplo anterior no se conoce la varianza poblacional de las puntuaciones.

Trabajando con la misma muestra de las puntuaciones de 10 estudiantes

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

se puede calcular el intervalo de confianza para μ con un nivel de confianza $1 - \alpha = 0,95$ (nivel de significación $\alpha = 0,05$):

- $\bar{X} = \frac{4+\dots+3}{10} = \frac{53}{10} = 5,3$ puntos.
- $\hat{S}^2 = \frac{(4-5,3)^2+\dots+(3-5,3)^2}{9} = 3,5667$ y $\hat{S} = \sqrt{3,5667} = 1,8886$ puntos.
- $t_{\alpha/2}^{n-1} = t_{0,025}^9$ es el valor de la T de Student de 9 grados de libertad, que deja una probabilidad acumulada superior de 0,025, que vale 2,2622.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}} = 5,3 \pm 2,2622 \frac{1,8886}{\sqrt{10}} = 5,3 \pm 1,351 = [3,949, 6,651].$$

Control de la imprecisión mediante el tamaño muestral

Ejemplo

Como se puede apreciar, la imprecisión del intervalo anterior es de $\pm 1,8886$ puntos, que es significativamente mayor que en el caso de conocer la varianza de la población. Esto es lógico pues al tener que estimar la varianza de la población, el error de la estimación se agrega al error del intervalo.

Ahora, el tamaño muestral necesario para reducir la imprecisión a $\pm 0,5$ puntos es

$$n = 4(z_{\alpha/2})^2 \frac{\hat{S}^2}{A^2} = 4 \cdot 1,96^2 \frac{3,5667}{(2 \cdot 0,5)^2} = 54,81.$$

Por tanto, si se desconoce la varianza de la población se necesita una muestra de al menos 55 estudiantes para conseguir un intervalo del 95 % de confianza y una precisión de $\pm 0,5$ puntos.

Intervalo de confianza para la media de una población no normal con varianza desconocida y muestras grandes

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución no es normal.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Si la población no es normal las distribuciones de los estimadores de referencia cambian, de manera que los intervalos anteriores no son válidos.

No obstante, si la muestra es grande ($n \geq 30$), de acuerdo al teorema central del límite, la distribución de la media muestral se aproximará a una normal, de modo que sigue siendo cierto

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

y en consecuencia, sigue siendo válido el intervalo

$$\bar{X} \pm t_{\alpha/2}^{n-1} \frac{\hat{S}}{\sqrt{n}}$$

Intervalo de confianza para la varianza de una población normal

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Para estimar la varianza de una población normal, se parte del estimador de referencia

$$\frac{nS^2}{\sigma^2} = \frac{(n-1)\hat{S}^2}{\sigma^2} \sim \chi^2(n-1),$$

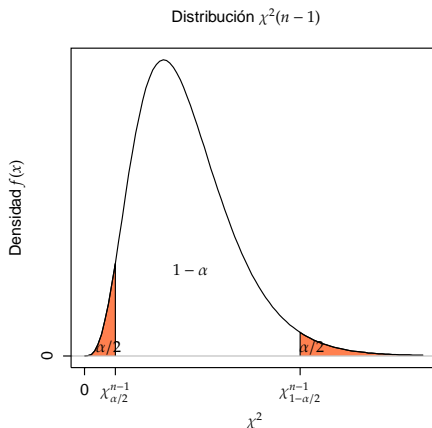
que sigue una distribución chi-cuadrado de $n - 1$ grados de libertad.

Sobre esta distribución hay que calcular los valores χ_i y χ_s tales que

$$P(\chi_i \leq \chi^2(n-1) \leq \chi_s) = 1 - \alpha.$$

Intervalo de confianza para la varianza de una población normal

Como la distribución chi-cuadrado no es simétrica respecto al 0, se toman dos valores $\chi_{\alpha/2}^{n-1}$ y $\chi_{1-\alpha/2}^{n-1}$ que dejen sendas colas de probabilidad acumulada inferior de $\alpha/2$ y $1 - \alpha/2$ respectivamente.



Intervalo de confianza para la varianza de una población normal

Así pues, se tiene

$$\begin{aligned} 1 - \alpha &= P\left(\chi_{\alpha/2}^{n-1} \leq \frac{nS^2}{\sigma^2} \leq \chi_{1-\alpha/2}^{n-1}\right) = P\left(\frac{1}{\chi_{\alpha/2}^{n-1}} \geq \frac{\sigma^2}{nS^2} \geq \frac{1}{\chi_{1-\alpha/2}^{n-1}}\right) = \\ &= P\left(\frac{1}{\chi_{1-\alpha/2}^{n-1}} \leq \frac{\sigma^2}{nS^2} \leq \frac{1}{\chi_{\alpha/2}^{n-1}}\right) = P\left(\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}} \leq \sigma^2 \leq \frac{nS^2}{\chi_{\alpha/2}^{n-1}}\right), \end{aligned}$$

y el intervalo de confianza para la varianza de una población normal es:

$$\left[\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}}, \frac{nS^2}{\chi_{\alpha/2}^{n-1}} \right]$$

Intervalo de confianza para la varianza de una población normal

Ejemplo

Siguiendo con el ejemplo de las puntuaciones en un examen, si se quiere estimar la varianza a partir de la muestra:

$$4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

para el intervalo de confianza para σ^2 con un nivel de confianza $1 - \alpha = 0,95$ (nivel de significación $\alpha = 0,05$) se tiene:

- $S^2 = \frac{(4-5,3)^2 + \dots + (3-5,3)^2}{10} = 3,21$ puntos².
- $\chi_{\alpha/2}^{n-1} = \chi_{0,025}^9$ es el valor de la chi-cuadrado de 9 grados de libertad, que deja una probabilidad acumulada inferior de 0,025, y vale 2,7.
- $\chi_{1-\alpha/2}^{n-1} = \chi_{0,975}^9$ es el valor de la chi-cuadrado de 9 grados de libertad, que deja una probabilidad acumulada inferior de 0,975, y vale 19.

Sustituyendo estos valores en la fórmula del intervalo, se llega a

$$\left[\frac{nS^2}{\chi_{1-\alpha/2}^{n-1}}, \frac{nS^2}{\chi_{\alpha/2}^{n-1}} \right] = \left[\frac{10 \cdot 3,21}{19}, \frac{10 \cdot 3,21}{2,7} \right] = [1,69, 11,89] \text{ puntos}^2.$$

Intervalo de confianza para la proporción de una población y muestras grandes

Para estimar la proporción p de individuos de una población que presentan una determinada característica, se parte de la variable que mide el número de individuos que la presentan en una muestra de tamaño n . Dicha variable sigue una distribución binomial

$$X \sim B(n, p)$$

Como ya se vio, si el tamaño muestral es suficientemente grande (en realidad basta que se cumpla $np \geq 5$ y $n(1-p) \geq 5$), el teorema central de límite asegura que X tendrá una distribución aproximadamente normal

$$X \sim N(np, \sqrt{np(1-p)}).$$

En consecuencia, la proporción muestral \hat{p} también será normal

$$\hat{p} = \frac{X}{n} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right),$$

que es el estimador de referencia.

Intervalo de confianza para la proporción de una población y muestras grandes

Trabajando con la distribución del estimador de referencia

$$\hat{p} \sim N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

tras tipificar, se pueden encontrar fácilmente, al igual que hicimos antes, valores $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplan

$$P\left(-z_{\alpha/2} \leq \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right).$$

Finalmente, deshaciendo la tipificación y razonando como antes, se llega fácilmente a la fórmula del intervalo

$$\left[\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right] \text{ o bien } \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Control de la imprecisión mediante el tamaño muestral

La amplitud o imprecisión del intervalo para la proporción de una población es

$$A = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

así que se puede calcular fácilmente el tamaño muestral necesario para conseguir un intervalo de amplitud A con confianza $1 - \alpha$:

$$A = 2z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \Leftrightarrow A^2 = 4z_{\alpha/2}^2 \frac{\hat{p}(1 - \hat{p})}{n},$$

de donde se deduce

$$n = 4z_{\alpha/2}^2 \frac{\hat{p}(1 - \hat{p})}{A^2}$$

Para poder hacer el cálculo se necesita una estimación de la proporción \hat{p} , por lo que suele tomarse una muestra previa pequeña para calcularla. En el peor de los casos, si no se dispone de una muestra previa, puede tomarse $\hat{p} = 0,5$.

Intervalo de confianza para la proporción de una población y muestras grandes

Ejemplo

Supóngase que se quiere estimar la proporción de fumadores que hay en una determinada población. Para ello se toma una muestra de 20 personas y se observa si fuman (1) o no (0):

0 - 1 - 1 - 0 - 0 - 0 - 1 - 0 - 0 - 1 - 0 - 0 - 0 - 1 - 1 - 0 - 1 - 1 - 0 - 0

Entonces:

- $\hat{p} = \frac{8}{20} = 0,4$, por tanto, se cumple $np = 20 \cdot 0,4 = 8 \geq 5$ y $n(1 - p) = 20 \cdot 0,6 = 12 \geq 5$.
- $z_{\alpha/2} = z_{0,025}$ es el valor de la normal estándar que deja una probabilidad acumulada superior de 0,025, que vale aproximadamente 1,96.

Sustituyendo estos valores en la fórmula del intervalo, se tiene

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} = 0,4 \pm 1,96 \sqrt{\frac{0,4 \cdot 0,6}{10}} = 0,4 \pm 0,3 = [0,1, 0,7].$$

Es decir, p estaría entre 0,1 y 0,7 con un 95 % de confianza.

Control de la imprecisión mediante el tamaño muestral

Ejemplo

Como se puede apreciar la imprecisión del intervalo anterior es $\pm 0,3$, que es enorme teniendo en cuenta que se trata de un intervalo para una proporción.

Para conseguir intervalos precisos para estimar proporciones se necesitan tamaños muestrales bastante grandes. Si por ejemplo se quiere una precisión de $\pm 0,05$, el tamaño muestral necesario sería:

$$n = 4z_{\alpha/2}^2 \frac{\hat{p}(1 - \hat{p})}{A^2} = 4 \cdot 1,96^2 \frac{0,4 \cdot 0,6}{(2 \cdot 0,05)^2} = 368,79.$$

Es decir, se necesitarían al menos 369 individuos para conseguir un intervalo para la proporción con una confianza del 95 %.

Comparación de dos poblaciones

En muchos estudios el objetivo en sí no es averiguar el valor de un parámetro, sino compararlo con el de otra población. Por ejemplo, comparar si un determinado parámetro vale lo mismo en la población de hombres y en la de mujeres.

En estos casos no interesa realmente estimar los dos parámetros por separado, sino hacer una estimación que permita su comparación.

Se verán tres casos:

Comparación de medias : Se estima la diferencia de medias $\mu_1 - \mu_2$.

Comparación de varianzas : Se estima la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$.

Comparación de proporciones : Se estima la diferencia de proporciones $\hat{p}_1 - \hat{p}_2$.

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas conocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 y μ_2 son desconocidas, pero sus varianzas σ_1^2 y σ_2^2 son conocidas.

Bajo estas hipótesis, si se toman dos muestras independientes, una de cada población, de tamaños n_1 y n_2 respectivamente, la diferencia de las medias muestrales sigue una distribución normal

$$\left. \begin{array}{l} \bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \\ \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \end{array} \right\} \Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right).$$

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas conocidas

A partir de aquí, tipificando, se pueden buscar los valores de la normal estándar $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplen:

$$P\left(-z_{\alpha/2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

Y deshaciendo la tipificación, se llega fácilmente al intervalo

$$\left[\bar{X}_1 - \bar{X}_2 - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 y μ_2 son desconocidas y sus varianzas también, pero son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Cuando se desconoce la varianza poblacional se puede estimar a partir de las muestras de tamaños n_1 y n_2 de ambas poblaciones mediante la *cuasivarianza ponderada*:

$$\hat{S}_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2}{n_1 + n_2 - 2}.$$

El estimador de referencia en este caso sigue una distribución T de Student:

$$\left. \begin{aligned} \bar{X}_1 - \bar{X}_2 &\sim N\left(\mu_1 - \mu_2, \sigma \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\right) \\ \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} &\sim \chi^2(n_1 + n_2 - 2) \end{aligned} \right\} \Rightarrow \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \sim T(n_1 + n_2 - 2).$$

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales

A partir de aquí, se pueden buscar los valores de la T de Student $-t_{\alpha/2}^{n_1+n_2-2}$ y $t_{\alpha/2}^{n_1+n_2-2}$ que cumplen

$$P\left(-t_{\alpha/2}^{n_1+n_2-2} \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}} \leq t_{\alpha/2}^{n_1+n_2-2}\right) = 1 - \alpha,$$

de donde se llega al intervalo

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}^{n_1+n_2-2} \hat{S}_p \sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

Interpretación del intervalo de confianza para la diferencia de medias de dos poblaciones

Si $[l_i, l_s]$ es un intervalo de confianza de nivel $1 - \alpha$ para la diferencia de medias $\mu_1 - \mu_2$, entonces

$$\mu_1 - \mu_2 \in [l_i, l_s]$$

con una confianza del $1 - \alpha$ %.

Por consiguiente, según los valores del intervalo de confianza se tiene:

- Si todos los valores del intervalo son negativos ($l_s < 0$), entonces se puede concluir que $\mu_1 - \mu_2 < 0$ y por tanto $\mu_1 < \mu_2$.
- Si todos los valores del intervalo son positivos ($l_i > 0$), entonces se puede concluir que $\mu_1 - \mu_2 > 0$ y por tanto $\mu_1 > \mu_2$.
- Si el intervalo tiene tanto valores positivos como negativos, y por tanto contiene al 0 ($0 \in [l_i, l_s]$), entonces no se puede afirmar que una media sea mayor que la otra. En este caso se suele asumir la hipótesis de que las medias son iguales $\mu_1 = \mu_2$.

Tanto en el primer como en el segundo caso se dice que entre las medias hay diferencias *estadísticamente significativas*.

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales

Ejemplo

Supóngase que se quiere comparar el rendimiento académico de dos grupos de alumnos, uno con 10 alumnos y otro con 12, que han seguido metodologías diferentes. Para ello se les realiza un examen y se obtienen las siguientes puntuaciones:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

Si se supone que ambas variables tienen la misma varianza, se tiene

- $\bar{X}_1 = \frac{4+\dots+3}{10} = 5,3$ y $\bar{X}_2 = \frac{8+\dots+7}{12} = 6,75$ puntos.
- $S_1^2 = \frac{4^2+\dots+3^2}{10} - 5,3^2 = 3,21$ y $S_2^2 = \frac{8^2+\dots+7^2}{12} - 6,75^2 = 2,6875$ puntos².
- $\hat{S}_p^2 = \frac{10 \cdot 3,21 + 12 \cdot 2,6875}{10+12-2} = 3,2175$ puntos², y $\hat{S}_p = 1,7937$.
- $t_{\alpha/2}^{n_1+n_2-2} = t_{0,025}^{20}$ es el valor de la T de Student de 20 grados de libertad que deja una probabilidad acumulada superior de 0,025, y que vale aproximadamente 2,09.

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas e iguales

Ejemplo

Y sustituyendo en la fórmula del intervalo llegamos a

$$5,3 - 6,75 \pm 2,086 \cdot 1,7937 \sqrt{\frac{10 + 12}{10 \cdot 12}} = -1,45 \pm 1,6021 = [-3,0521, 0,1521] \text{ puntos.}$$

Es decir, la diferencia de puntuaciones medias $\mu_1 - \mu_2$ está entre $-3,0521$ y $0,1521$ puntos con una confianza del 95 %.

A la vista del intervalo se puede concluir que, puesto que el intervalo contiene tanto valores positivos como negativos, y por tanto contiene al 0, no puede afirmarse que una de las medias sea mayor que la otra, de modo que se supone que son iguales y no se puede decir que haya diferencias significativas entre los grupos.

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1, μ_2 y varianzas σ_1^2, σ_2^2 , son desconocidas, pero $\sigma_1^2 \neq \sigma_2^2$.

En este caso el estimador de referencia sigue una distribución T de Student

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim T(g),$$

donde el número de grados de libertad es

$$g = n_1 + n_2 - 2 - \Delta \quad \text{siendo } \Delta = \frac{(\frac{n_2-1}{n_1} \hat{S}_1^2 - \frac{n_1-1}{n_2} \hat{S}_2^2)^2}{\frac{n_2-1}{n_1^2} \hat{S}_1^4 + \frac{n_1-1}{n_2^2} \hat{S}_2^4}.$$

Intervalo de confianza para la diferencia de medias de dos poblaciones normales con varianzas desconocidas

A partir de aquí, una vez más, se pueden buscar los valores de la T de Student $-t_{\alpha/2}^g$ y $t_{\alpha/2}^g$ que cumplen

$$P\left(-t_{\alpha/2}^g \leq \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \leq t_{\alpha/2}^g\right) = 1 - \alpha,$$

de donde llegamos al intervalo

$$\left[\bar{X}_1 - \bar{X}_2 - t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}, \bar{X}_1 - \bar{X}_2 + t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}} \right]$$

o bien

$$\bar{X}_1 - \bar{X}_2 \pm t_{\alpha/2}^g \sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}$$

Elección del intervalo de confianza para la diferencia de medias en función de las varianzas

Como se acaba de ver, existen dos intervalos posibles para estimar la diferencia de medias: uno para cuando las varianzas poblacionales son iguales y otro para cuando no lo son.

Ahora bien, si las varianzas poblacionales son desconocidas,

¿cómo saber qué intervalo utilizar?

La respuesta está en el próximo intervalo que se verá, que permite estimar la razón de varianzas $\frac{\sigma_2^2}{\sigma_1^2}$ y por tanto, su comparación.

Así pues, antes de calcular el intervalo de confianza para la comparación de medias, cuando las varianzas poblacionales sean desconocidas, es necesario calcular el intervalo de confianza para la razón de varianzas y elegir el intervalo para la comparación de medias en función del valor de dicho intervalo.

Intervalo de confianza para el cociente de varianzas de dos poblaciones normales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

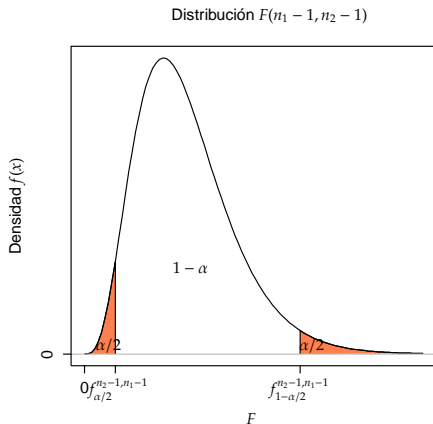
- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 , μ_2 y varianzas σ_1^2 , σ_2^2 son desconocidas.

En este caso, para muestras de ambas poblaciones de tamaños n_1 y n_2 respectivamente, el estimador de referencia sigue una distribución F de Fisher-Snedecor:

$$\left. \begin{array}{l} \frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \\ \frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1) \end{array} \right\} \Rightarrow \frac{\frac{\frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2}}{n_2 - 1}}{\frac{\frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2}}{n_1 - 1}} = \frac{\sigma_1^2}{\sigma_2^2} \frac{\hat{S}_2^2}{\hat{S}_1^2} \sim F(n_2 - 1, n_1 - 1).$$

Intervalo de confianza para el cociente de varianzas de dos poblaciones normales

Como la distribución F de Fisher-Snedecor no es simétrica respecto al 0, se toman dos valores $f_{\alpha/2}^{n_2-1, n_1-1}$ y $f_{1-\alpha/2}^{n_2-1, n_1-1}$ que dejen sendas colas de probabilidad acumulada inferior de $\alpha/2$ y $1 - \alpha/2$ respectivamente.



Intervalo de confianza para el cociente de varianzas de dos poblaciones normales

Así pues, se tiene

$$\begin{aligned} 1 - \alpha &= P \left(f_{\alpha/2}^{n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \frac{\hat{S}_2^2}{\hat{S}_1^2} \leq f_{1-\alpha/2}^{n_2-1, n_1-1} \right) = \\ &= P \left(f_{\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq f_{1-\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \right) \end{aligned}$$

y el intervalo de confianza para la comparación de varianzas de dos poblaciones normales es:

$$\left[f_{\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2}, f_{1-\alpha/2}^{n_2-1, n_1-1} \frac{\hat{S}_1^2}{\hat{S}_2^2} \right]$$

Interpretación del intervalo de confianza para el cociente de varianzas de dos poblaciones

Si $[l_i, l_s]$ es un intervalo de confianza de nivel $1 - \alpha$ para la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$, entonces

$$\frac{\sigma_1^2}{\sigma_2^2} \in [l_i, l_s]$$

con una confianza del $1 - \alpha$ %.

Por consiguiente, según los valores del intervalo de confianza se tiene:

- Si todos los valores del intervalo son menores que 1 ($l_s < 1$), entonces se puede concluir que $\frac{\sigma_1^2}{\sigma_2^2} < 1$ y por tanto $\sigma_1^2 < \sigma_2^2$.
- Si todos los valores del intervalo son mayores que 1 ($l_i > 1$), entonces se puede concluir que $\frac{\sigma_1^2}{\sigma_2^2} > 1$ y por tanto $\sigma_1^2 > \sigma_2^2$.
- Si el intervalo tiene tanto valores mayores como menores que 1, y por tanto contiene al 1 ($1 \in [l_i, l_s]$), entonces no se puede afirmar que una varianza sea mayor que la otra. En este caso se suele asumir la hipótesis de que las varianzas son iguales $\sigma_1^2 = \sigma_2^2$.

Intervalo de confianza para el cociente de varianzas de dos poblaciones normales

Ejemplo

Siguiendo con el ejemplo de las puntuaciones en dos grupos:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

Para calcular el intervalo de confianza para la razón de varianzas con una confianza del 95 %, se tiene:

- $\bar{X}_1 = \frac{4+\dots+3}{10} = 5,3$ puntos y $\bar{X}_2 = \frac{8+\dots+7}{12} = 6,75$ puntos.
- $\hat{S}_1^2 = \frac{(4-5,3)^2+\dots+(3-5,3)^2}{9} = 3,5667$ puntos² y $\hat{S}_2^2 = \frac{(8-6,75)^2+\dots+(3-6,75)^2}{11} = 2,9318$ puntos².
- $f_{\alpha/2}^{n_2-1, n_1-1} = f_{0,025}^{11,9}$ es el valor de la F de Fisher de 11 y 9 grados de libertad que deja una probabilidad acumulada inferior de 0,025, y que vale aproximadamente 0,2787.
- $f_{1-\alpha/2}^{n_2-1, n_1-1} = f_{0,975}^{11,9}$ es el valor de la F de Fisher de 11 y 9 grados de libertad que deja una probabilidad acumulada inferior de 0,975, y que vale aproximadamente 3,9121.

Intervalo de confianza para la razón de de varianzas de dos poblaciones normales

Ejemplo

Sustituyendo en la fórmula del intervalo se llega a

$$\left[0,2787 \frac{3,5667}{2,9318}, 3,9121 \frac{3,5667}{2,9318} \right] = [0,3391, 4,7591] \text{ puntos}^2.$$

Es decir, la razón de varianzas $\frac{\sigma_1^2}{\sigma_2^2}$ está entre 0,3391 y 4,7591 con una confianza del 95 %.

Como el intervalo tiene tanto valores menores como mayores que 1, no se puede concluir que una varianza sea mayor que la otra, y por tanto se mantiene la hipótesis de que ambas varianzas son iguales.

Si ahora se quisiesen comparar las medias de ambas poblaciones, el intervalo de confianza para la diferencia de medias que habría que tomar es el que parte de la hipótesis de igualdad de varianzas, que precisamente es el que se ha utilizado antes.

Intervalo de confianza para la diferencia de proporciones de dos poblaciones y muestras grandes

Para comparar las proporciones p_1 y p_2 de individuos que presentan una determinada característica en dos poblaciones independientes, se estima su diferencia $p_1 - p_2$.

Si se toma una muestra de cada población, de tamaños n_1 y n_2 respectivamente, las variables que miden el número de individuos que presentan la característica en cada una de ellas siguen distribuciones

$$X_1 \sim B(n_1, p_1) \quad \text{y} \quad X_2 \sim B(n_2, p_2)$$

Cuando los tamaños muestrales son grandes (en realidad basta que se cumpla $n_1 p_1 \geq 5$, $n_1(1 - p_1) \geq 5$, $n_2 p_2 \geq 5$ y $n_2(1 - p_2) \geq 5$), el teorema central de límite asegura que X_1 y X_2 tendrán distribuciones normales

$$X_1 \sim N(n_1 p_1, \sqrt{n_1 p_1(1 - p_1)}) \quad \text{y} \quad X_2 \sim N(n_2 p_2, \sqrt{n_2 p_2(1 - p_2)}),$$

y las proporciones muestrales

$$\hat{p}_1 = \frac{X_1}{n_1} \sim N\left(p_1, \sqrt{\frac{p_1(1 - p_1)}{n_1}}\right) \quad \text{y} \quad \hat{p}_2 = \frac{X_2}{n_2} \sim N\left(p_2, \sqrt{\frac{p_2(1 - p_2)}{n_2}}\right)$$

Intervalo de confianza para la diferencia de proporciones de dos poblaciones y muestras grandes

A partir de las proporciones muestrales se construye el estimador de referencia

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}\right).$$

Tipificando, se buscan valores $-z_{\alpha/2}$ y $z_{\alpha/2}$ que cumplan

$$P\left(-z_{\alpha/2} \leq \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \leq z_{\alpha/2}\right).$$

Finalmente, deshaciendo la tipificación, se llega fácilmente a la fórmula del intervalo

$$\left[\hat{p}_1 - \hat{p}_2 - z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, \hat{p}_1 - \hat{p}_2 + z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right]$$

Intervalo de confianza para la diferencia de proporciones de dos poblaciones y muestras grandes

Ejemplo

Supóngase que se quieren comparar las proporciones o porcentajes de aprobados en dos grupos que han seguido metodologías distintas. En el primer grupo han aprobado 24 alumnos de un total de 40, mientras que en el segundo han aprobado 48 de 60.

Para calcular el intervalo de confianza para la diferencia de proporciones con un nivel de confianza del 95 %, se tiene:

- $\hat{p}_1 = 24/40 = 0,6$ y $\hat{p}_2 = 48/60 = 0,8$, de manera que se cumplen las hipótesis $n_1\hat{p}_1 = 40 \cdot 0,6 = 24 \geq 5$, $n_1(1 - \hat{p}_1) = 40(1 - 0,6) = 26 \geq 5$, $n_2\hat{p}_2 = 60 \cdot 0,8 = 48 \geq 5$ y $n_2(1 - \hat{p}_2) = 60(1 - 0,8) = 12 \geq 5$.
- $z_{\alpha/2} = z_{0,025} = 1,96$.

Sustituyendo en la fórmula del intervalo se tiene

$$0,6 - 0,8 \pm 1,96 \sqrt{\frac{0,6(1 - 0,6)}{40} + \frac{0,8(1 - 0,8)}{60}} = -0,2 \pm 0,17 = [-0,37, -0,03].$$

Como el intervalo es negativo se tiene $p_1 - p_2 < 0 \Rightarrow p_1 < p_2$, y se puede concluir que hay diferencias significativas en el porcentaje de aprobados.

7

Contraste de hipótesis

- Hipótesis estadísticas y tipos de contrastes de hipótesis
- Planteamiento de un contraste de hipótesis
- Estadístico del contraste
- Regiones de aceptación y de rechazo
- Errores en un contraste de hipótesis
- Potencia de un contraste
- p -valor de un contraste
- Pruebas de conformidad
- Pruebas de homogeneidad
- Realización de contrastes mediante intervalos de confianza

Hipótesis estadística

En muchos estudios estadísticos, el objetivo, más que estimar el valor de un parámetro desconocido en la población, es comprobar la veracidad de una hipótesis formulada sobre la población objeto de estudio.

El investigador, de acuerdo a su experiencia o a estudios previos, suele tener conjeturas sobre la población estudiada que expresa en forma de hipótesis.

Definición (Hipótesis estadística)

Una *hipótesis estadística* es cualquier afirmación o conjetura que determina, total o parcialmente, la distribución una o varias variables de la población.

Por ejemplo, si estamos interesados en el rendimiento académico de un grupo de alumnos en una determinada asignatura, podríamos plantear la hipótesis de si el porcentaje de aprobados es mayor del 50 %.

Contraste de hipótesis

En general nunca se sabrá con absoluta certeza si una hipótesis estadística es cierta o falsa, ya que para ello habría que estudiar a todos los individuos de la población.

Para comprobar la veracidad o falsedad de estas hipótesis hay que contrastarlas con los resultados empíricos obtenidos de las muestras. Si los resultados observados en las muestras coinciden, dentro de un margen de error admisible, con lo que cabría esperar en caso de que la hipótesis fuese cierta, la hipótesis se aceptará como verdadera, mientras que en caso contrario se rechazará como falsa y se buscarán nuevas hipótesis capaces de explicar los datos observados.

Como las muestras se obtienen aleatoriamente, *la decisión de aceptar o rechazar una hipótesis estadística se tomará sobre una base de probabilidad.*

La metodología que se encarga de contrastar la veracidad de las hipótesis estadísticas se conoce como *contraste de hipótesis*.

Tipos de contrastes de hipótesis

- **Pruebas de bondad de ajuste:** El objetivo es comprobar una hipótesis sobre la forma de la distribución de la población.
Por ejemplo, ver si las notas de un grupo de alumnos siguen una distribución normal.
- **Pruebas de conformidad:** El objetivo es comprobar una hipótesis sobre alguno de los parámetros de la población.
Por ejemplo, ver si la nota media en un grupo de alumnos es igual a 5.
- **Pruebas de homogeneidad:** El objetivo es comparar dos poblaciones con respecto a alguno de sus parámetros.
Por ejemplo, ver si el rendimiento de dos grupos de alumnos es el mismo comparando sus notas medias.
- **Pruebas de independencia:** El objetivo es comprobar si existe relación entre dos variables de la población.
Por ejemplo, ver si existe relación entre las notas de dos asignaturas diferentes.

Cuando las hipótesis se plantean sobre parámetros de la población, también se habla de **pruebas paramétricas**.

Hipótesis nula e hipótesis alternativa

En la mayoría de los casos un contraste supone tomar una decisión entre dos hipótesis antagonistas:

Hipótesis nula Es la hipótesis conservadora, ya que se mantendrá mientras que los datos de las muestras no reflejen claramente su falsedad. Se representa como H_0 .

Hipótesis alternativa Es la negación de la hipótesis nula y generalmente representa la afirmación que se pretende probar. Se representa como H_1 .

Ambas hipótesis se eligen de acuerdo con el principio de simplicidad científica:

“Solamente se debe abandonar un modelo simple por otro más complejo cuando la evidencia a favor del último sea fuerte.” (Navaja de Occam)

Elección de las hipótesis nula y alternativa

Analogía con un juicio

En el caso de un juicio, en el que el juez debe decidir si el acusado es culpable o inocente, la elección de hipótesis debería ser

H_0 : Inocente

H_1 : Culpable

ya que la inocencia se asume, mientras que la culpabilidad hay que demostrarla.

Según esto, el juez sólo aceptaría la hipótesis alternativa cuando hubiese pruebas significativas de la culpabilidad del acusado.

El investigador jugaría el papel del fiscal, ya que su objetivo consistiría en intentar rechazar la hipótesis nula, es decir, demostrar culpabilidad del acusado.

¡Esta metodología siempre favorece a la hipótesis nula!

Contrastes de hipótesis paramétricos

En muchos contrastes, sobre todo en las pruebas de conformidad y de homogeneidad, las hipótesis se formulan sobre parámetros desconocidos de la población como puede ser una media, una varianza o una proporción.

En tal caso, la hipótesis nula siempre asigna al parámetro un valor concreto, mientras que la alternativa suele ser una hipótesis abierta que, aunque opuesta a la hipótesis nula, no fija el valor del parámetro.

Esto da lugar a tres tipos de contrastes:

Bilateral	Unilateral de menor	Unilateral de mayor
$H_0: \theta = \theta_0$ $H_1: \theta \neq \theta_0$	$H_0: \theta = \theta_0$ $H_1: \theta < \theta_0$	$H_0: \theta = \theta_0$ $H_1: \theta > \theta_0$

Elección del tipo de contraste

Ejemplo

Supóngase que existen sospechas de que en una población hay menos hombres que mujeres.

¿Qué tipo de contraste debería plantearse para validar o refutar esta sospecha?

- 1 Las sospechas se refieren al porcentaje o la proporción p de hombres en la población, por lo que se trata de un *contraste paramétrico*.
- 2 El objetivo es averiguar el valor de p , por lo que se trata de una *prueba de conformidad*. En la hipótesis nula el valor de p se fijará a 0,5 ya que, de acuerdo a las leyes de la genética, en la población debería haber la misma proporción de hombres que de mujeres.
- 3 Finalmente, existen sospechas de que el porcentaje de mujeres sea mayor que el de hombres, por lo que la hipótesis alternativa será de menor $p < 0,5$.

Así pues, el contraste que debería plantearse es el siguiente:

$$H_0: p = 0,5,$$

$$H_1: p < 0,5.$$

Estadístico del contraste

La aceptación o rechazo de la hipótesis nula depende, en última instancia, de lo que se observe en la muestra.

La decisión se tomará según el valor que presente algún estadístico de la muestra relacionado con el parámetro o característica que se esté contrastando, y cuya distribución de probabilidad debe ser conocida suponiendo cierta la hipótesis nula y una vez fijado el tamaño de la muestra. Este estadístico recibe el nombre de **estadístico del contraste**.

Para cada muestra, el estadístico dará una estimación a partir de la cual se tomará la decisión: *Si la estimación difiere demasiado del valor esperado bajo la hipótesis H_0 , entonces se rechazará, y en caso contrario se aceptará.*

La lógica que guía la decisión es la de mantener la hipótesis nula a no ser que en la muestra haya pruebas contundentes de su falsedad. Siguiendo con el símil del juicio, se trataría de mantener la inocencia mientras no haya pruebas claras de culpabilidad.

Estadístico del contraste

Ejemplo

Volviendo al ejemplo del contraste sobre la proporción de hombres de una población

$$H_0: p = 0,5,$$

$$H_1: p < 0,5.$$

Si para resolver el contraste se toma una muestra aleatoria de 10 personas, podría tomarse como estadístico del contraste X el número de hombres en la muestra.

Suponiendo cierta la hipótesis nula, el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0,5)$, de manera que el número esperado de hombres en la muestra sería 5.

Así pues, es lógico aceptar la hipótesis nula si en la muestra se obtiene un número de hombres próximo a 5 y rechazarla cuando el número de hombres sea muy inferior a 5. Pero,

¿dónde poner el límite entre los valores X que lleven a la aceptación y los que lleven al rechazo?

Regiones de aceptación y de rechazo

Una vez elegido el estadístico del contraste, lo siguiente es decidir para qué valores de este estadístico se decidirá aceptar la hipótesis nula y para que valores se rechazará. Esto divide del conjunto de valores posibles del estadístico en dos regiones:

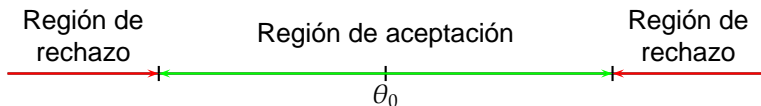
Región de aceptación : Es el conjunto de valores del estadístico del contraste a partir de los cuales se decidirá aceptar la hipótesis nula.

Región de rechazo : Es el conjunto de valores del estadístico del contraste a partir de los cuales se decidirá rechazar la hipótesis nula.

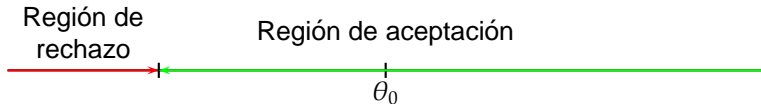
Ubicación de las regiones de aceptación y de rechazo

Dependiendo de la dirección del contraste, la región de rechazo quedará a un lado u otro del valor esperado del estadístico del contraste según la hipótesis nula:

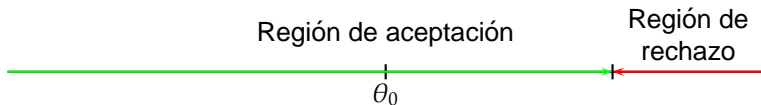
- Contraste bilateral $H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$.



- Contraste unilateral de menor $H_0 : \theta = \theta_0 \quad H_1 : \theta < \theta_0$.



- Contraste unilateral de mayor $H_0 : \theta = \theta_0 \quad H_1 : \theta > \theta_0$.



Regiones de aceptación y de rechazo

Ejemplo

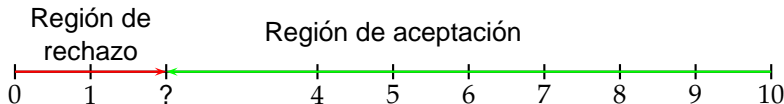
Siguiendo con el ejemplo del contraste sobre la proporción de hombres de una población

$$H_0: p = 0,5,$$

$$H_1: p < 0,5.$$

Como el estadístico del contraste tenía una distribución binomial $X \sim B(10, 0,5)$ suponiendo cierta la hipótesis nula, su recorrido será de 0 a 10 y su valor esperado 5, por lo que, al tratarse de un contraste unilateral de menor, la región de rechazo quedará por debajo del 5. Pero,

¿dónde poner el límite entre las regiones de aceptación y de rechazo?



¡Todo dependerá del riesgo de equivocarse!

Errores en un contraste de hipótesis

Hemos visto que un contraste de hipótesis se realiza mediante una regla de decisión que permite aceptar o rechazar la hipótesis nula dependiendo del valor que tome el estadístico del contraste.

Al final el contraste se resuelve tomando una decisión de acuerdo a esta regla. El problema es que nunca se conocerá con absoluta certeza la veracidad o falsedad de una hipótesis, de modo que al aceptarla o rechazarla es posible que se esté tomando una decisión equivocada.

Los errores que se pueden cometer en un contraste de hipótesis son de dos tipos:

- **Error de tipo I.** Se comete cuando se rechaza la hipótesis nula siendo esta verdadera.
- **Error de tipo II.** Se comete cuando se acepta la hipótesis nula siendo esta falsa.

Riesgos de los errores de un contraste de hipótesis

Los riesgos de cometer cada tipo de error se cuantifican mediante probabilidades:

Definición (Riesgos α y β)

En un contraste de hipótesis, se define el *riesgo* α como la probabilidad de cometer un error de tipo I, es decir,

$$\alpha = P(\text{Rechazar } H_0 / H_0)$$

y se define el *riesgo* β como la probabilidad de cometer un error de tipo II, es decir,

$$\beta = P(\text{Aceptar } H_0 / H_1)$$

Decisión	Hipótesis verdadera	
	H_0	H_1
Aceptar H_0	Decisión correcta $1 - \alpha$	Error de tipo II $\beta = P(\text{Aceptar } H_0 / H_1)$
Rechazar H_0	Error de tipo I $\alpha = P(\text{Rechazar } H_0 / H_0)$	Decisión correcta $1 - \beta$

Interpretación del riesgo α

En principio, puesto que esta metodología favorece a la hipótesis nula, el error del tipo I suele ser más grave que el error del tipo II, y por tanto, el riesgo α suele fijarse a niveles bajos de 0,1, 0,05 o 0,01, siendo 0,05 lo más habitual.

Debe tenerse cuidado al interpretar el riesgo α ya que se trata de una probabilidad condicionada a que la hipótesis nula sea cierta. Por tanto, cuando se rechaza la hipótesis nula con un riesgo $\alpha = 0,05$, es erróneo decir 5 de cada 100 veces nos equivocaremos, ya que esto sería cierto sólo si la hipótesis nula fuese siempre verdadera.

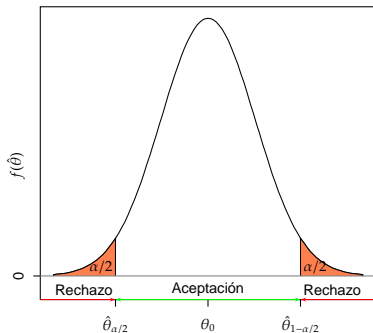
Tampoco tiene sentido hablar de la probabilidad de haberse equivocado una vez tomada una decisión a partir de una muestra concreta, pues en tal caso, si se ha tomado la decisión acertada, la probabilidad de error es 0 y si se ha tomado la decisión equivocada, la probabilidad de error es 1.

Determinación de las regiones de aceptación y de rechazo en función del riesgo α

Una vez fijado el riesgo α que se está dispuesto a tolerar, es posible delimitar las regiones de aceptación y de rechazo para el estadístico del contraste de manera que la probabilidad acumulada en la región de aceptación sea α , suponiendo cierta la hipótesis nula.

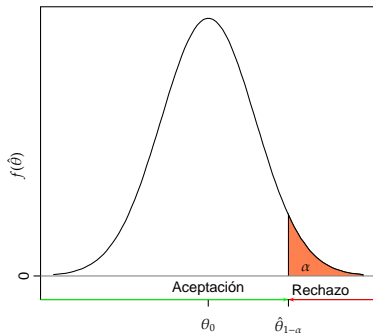
Contraste bilateral

Distribución del estadístico del contraste



Contraste unilateral

Distribución del estadístico del contraste



Determinación de las regiones de aceptación y de rechazo en función del riesgo α

Ejemplo

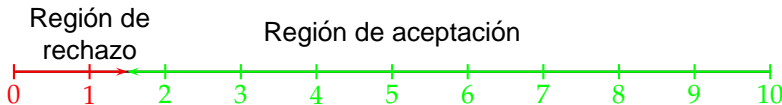
Siguiendo con el contraste sobre la proporción de hombres de una población, como el estadístico del contraste sigue una distribución binomial $X \sim B(10, 0,5)$, si se decide rechazar la hipótesis nula cuando en la muestra haya 2 o menos hombres, la probabilidad de cometer un error de tipo I será

$$P(X \leq 2) = f(0) + f(1) + f(2) = 0,0010 + 0,0098 + 0,0439 = 0,0547.$$

Si riesgo máximo de error de tipo I que se está dispuesto a tolerar es $\alpha = 0,05$, ¿qué valores del estadístico permitirán rechazar la hipótesis nula?

$$P(X \leq 1) = f(0) + f(1) = 0,0010 + 0,0098 = 0,0107.$$

Es decir, sólo se podría rechazar la hipótesis nula con 0 o 1 hombres en la muestra.



Aunque el error de tipo II pueda parecer menos grave, también interesa que el riesgo β sea bajo, ya que de lo contrario será difícil rechazar la hipótesis nula (que es lo que se persigue la mayoría de las veces), aunque haya pruebas muy claras de su falsedad.

El problema, en el caso de contrastes paramétricos, es que la hipótesis alternativa es una hipótesis abierta en la que no se fija el valor del parámetro a contrastar, de modo que, para poder calcular el riesgo β es necesario fijar dicho valor.

Lo normal es fijar el valor del parámetro del contraste a la mínima cantidad para admitir diferencias significativas desde un punto de vista práctico o clínico.

La mínima diferencia δ que se considera como clínicamente significativa no depende de la muestra y debe fijarla el investigador a priori.

Potencia de un contraste $1 - \beta$

Puesto que el objetivo del investigador suele ser rechazar la hipótesis nula, a menudo, lo más interesante de un contraste es su capacidad para detectar la falsedad de la hipótesis nula cuando realmente hay diferencias mayores que δ entre el verdadero valor del parámetro y el que establece la hipótesis nula.

Definición (Potencia de un contraste)

La *potencia* de un contraste de hipótesis se define como

$$\text{Potencia} = P(\text{Rechazar } H_0/H_1) = 1 - P(\text{Aceptar } H_0/H_1) = 1 - \beta.$$

Así pues, al reducir el riesgo β se aumentará la potencia del contraste.

Un contraste poco potente no suele ser interesante ya que no permitirá rechazar la hipótesis nula aunque haya evidencias en su contra.

Cálculo del riesgo β y de la potencia $1 - \beta$

Ejemplo

Supóngase que en el contraste sobre la proporción de hombres no se considera importante una diferencia de menos de un 10 % con respecto al valor que establece la hipótesis nula, es decir, $\delta = 0,1$.

Esto permite fijar la hipótesis alternativa

$$H_1 : p = 0,5 - 0,1 = 0,4.$$

Suponiendo cierta esta hipótesis el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0,4)$.

En tal caso, el riesgo β para las regiones de aceptación y rechazo fijadas antes será

$$\beta = P(\text{Aceptar } H_0/H_1) = P(X \geq 2) = 1 - P(X < 2) = 1 - 0,0464 = 0,9536.$$

Como puede apreciarse, se trata de un riesgo β muy alto, por lo que la potencia del contraste sería sólo de

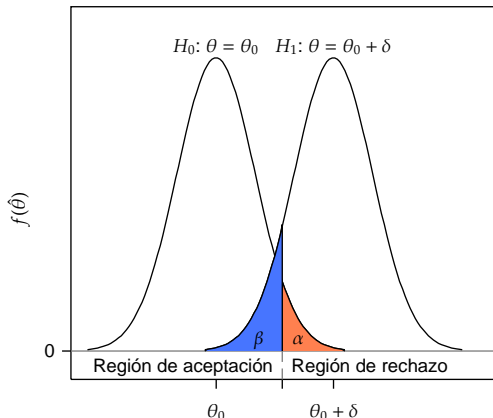
$$1 - \beta = 1 - 0,9536 = 0,0464,$$

lo que indica que no se trataría de un buen contraste para detectar diferencias de un 10 % en el valor del parámetro.

Relación del riesgo β y la mínima diferencia importante δ

El riesgo β depende directamente de la mínima diferencia δ que se desea detectar con respecto al valor del parámetro que establece la hipótesis nula.

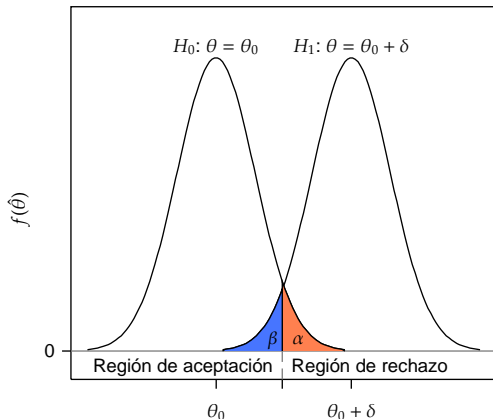
Relación entre el riesgo β y la mínima diferencia importante δ



Relación del riesgo β y la mínima diferencia importante δ

El riesgo β depende directamente de la mínima diferencia δ que se desea detectar con respecto al valor del parámetro que establece la hipótesis nula.

Relación entre el riesgo β y la mínima diferencia importante δ



Relación del riesgo β y la mínima diferencia importante δ

Si en el contraste sobre la proporción de hombres se deseara detectar una diferencia de al menos un 20 % con respecto al valor que establece la hipótesis nula, es decir, $\delta = 0,2$, entonces la hipótesis alternativa se fijaría a

$$H_1 : p = 0,5 - 0,2 = 0,3,$$

y bajo esta hipótesis el estadístico del contraste seguiría una distribución binomial $X \sim B(10, 0,3)$.

En tal caso, el riesgo β para las regiones de aceptación y rechazo fijadas antes sería

$$\beta = P(\text{Aceptar } H_0/H_1) = P(X \geq 2) = 1 - P(X < 2) = 1 - 0,1493 = 0,8507,$$

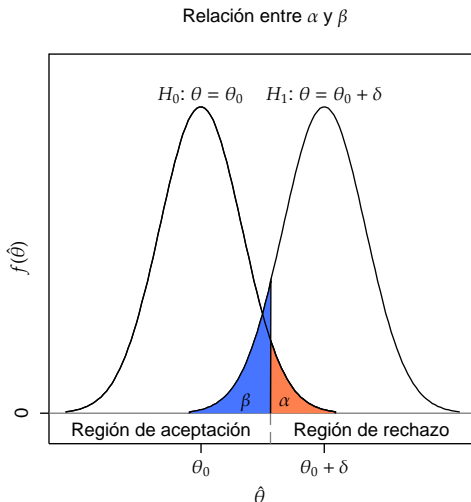
por lo que el riesgo β disminuiría y la potencia del contraste aumentaría

$$1 - \beta = 1 - 0,8507 = 0,1493,$$

aunque seguiría siendo un contraste poco potente.

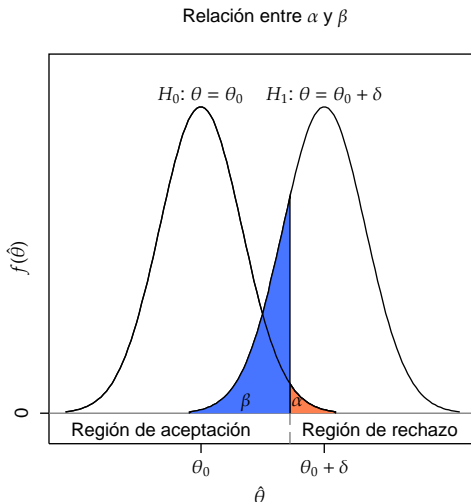
Relación entre los riesgos α y β

Los riesgos α y β están enfrentados, es decir, cuando uno aumenta el otro disminuye y viceversa.



Relación entre los riesgos α y β

Los riesgos α y β están enfrentados, es decir, cuando uno aumenta el otro disminuye y viceversa.



Relación entre los riesgos α y β

Ejemplo

Si en el contraste sobre la proporción de hombres toma como riesgo $\alpha = 0,1$, entonces la región de rechazo sería $X \leq 2$ ya que, suponiendo cierta la hipótesis nula, $X \sim B(10, 0,5)$, y

$$P(X \leq 2) = 0,0547 \leq 0,1 = \alpha.$$

Entonces, para una diferencia mínima $\delta = 0,1$ y suponiendo cierta la hipótesis alternativa, $X \sim B(10, 0,4)$, el riesgo β será

$$\beta = P(\text{Aceptar } H_0/H_1) = P(X \geq 3) = 1 - P(X < 3) = 1 - 0,1673 = 0,8327,$$

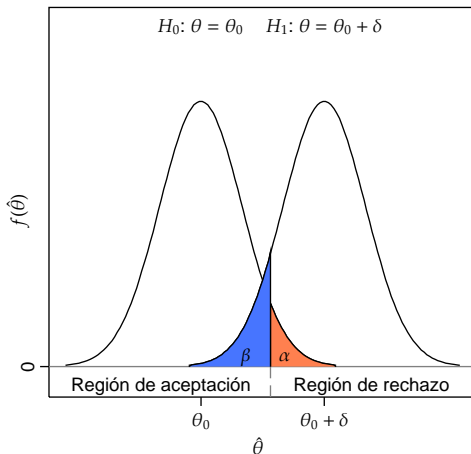
y ahora la potencia ha subido hasta

$$1 - \beta = 1 - 0,8327 = 0,1673.$$

Relación de los riesgos de error y el tamaño muestral

Los riesgos de error también dependen el tamaño de la muestra, ya que al aumentar el tamaño de la muestra, la dispersión del estadístico del contraste disminuye y con ello también lo hacen los riesgos de error.

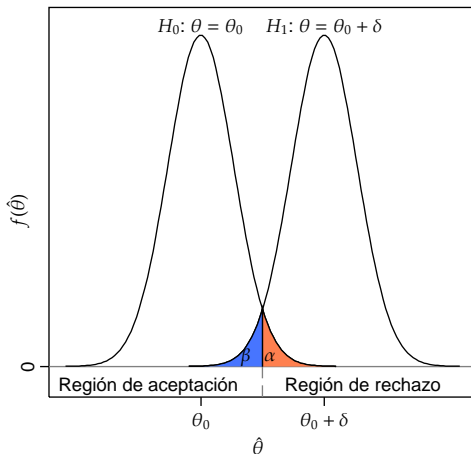
Riesgos de error para muestras pequeñas



Relación de los riesgos de error y el tamaño muestral

Los riesgos de error también dependen el tamaño de la muestra, ya que al aumentar el tamaño de la muestra, la dispersión del estadístico del contraste disminuye y con ello también lo hacen los riesgos de error.

Riesgos de error para muestras pequeñas



Relación de los riesgos de error y el tamaño muestral

Ejemplo

Si para realizar el contraste sobre la proporción de hombres se hubiese tomado una muestra de tamaño 100, en lugar de 10, entonces, bajo la suposición de certeza de la hipótesis nula, el estadístico del contraste seguiría una distribución binomial $B(100, 0,5)$, y ahora la región de rechazo sería $X \leq 41$, ya que

$$P(X \leq 41) = 0,0443 \leq 0,05 = \alpha.$$

Entonces, para $\delta = 0,1$ y suponiendo cierta la hipótesis alternativa, $X \sim B(100, 0,4)$, el riesgo β sería

$$\beta = P(\text{Aceptar } H_0/H_1) = P(X \geq 42) = 0,3775,$$

y ahora la potencia habría aumentado considerablemente

$$1 - \beta = 1 - 0,3775 = 0,6225.$$

Este contraste sería mucho más útil para detectar una diferencia de al menos un 10 % con respecto al valor del parámetro que establece la hipótesis nula.

Curva de potencia

La potencia de un contraste depende del valor del parámetro que establezca la hipótesis alternativa y, por tanto, es una función de este

$$\text{Potencia}(x) = P(\text{Rechazar } H_0 / \theta = x).$$

Esta función da la probabilidad de rechazar la hipótesis nula para cada valor del parámetro y se conoce como **curva de potencia**.

Cuando no se puede fijar el valor concreto del parámetro en la hipótesis alternativa, resulta útil representar esta curva para ver la bondad del contraste cuando no se rechaza la hipótesis nula. También es útil cuando sólo se dispone de un número determinado de individuos en la muestra, para ver si merece la pena hacer el estudio.

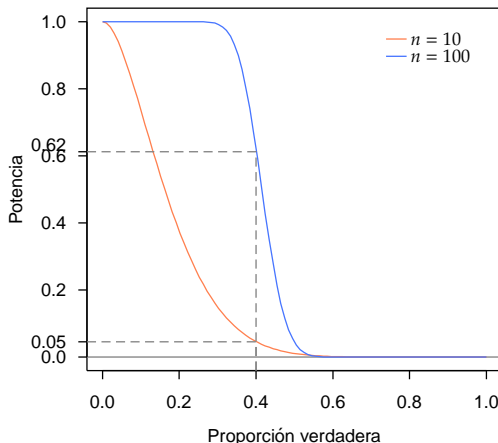
¡Un contraste será mejor cuanto mayor sea el área encerrada por debajo de la curva de potencia!

Curva de potencia

Ejemplo

La curva de potencia correspondiente al contraste sobre la proporción de hombres en la población es la siguiente

Curvas de potencia de un contraste unilateral de menor con $\alpha = 0,05$



p -valor de un contraste de hipótesis

En general, siempre que la estimación del estadístico caiga dentro de la región de rechazo, rechazaremos la hipótesis nula, pero evidentemente, si dicha estimación se aleja bastante de la región de aceptación tendremos más confianza en el rechazo que si la estimación está cerca del límite entre las regiones de aceptación y rechazo.

Por este motivo, al realizar un contraste, también se calcula la probabilidad de obtener una discrepancia mayor o igual a la observada entre la estimación del estadístico del contraste y su valor esperado según la hipótesis nula.

Definición (p -valor)

En un contraste de hipótesis, para cada estimación x_0 del estadístico del contraste X , dependiendo del tipo de contraste, se define el p -valor del contraste como

Contraste bilateral:	$2P(X \geq x_0/H_0)$
Contraste unilateral de menor:	$P(X \leq x_0/H_0)$
Contraste unilateral de mayor:	$P(X \geq x_0/H_0)$

Realización del contraste con el p -valor

En cierto modo, el p -valor expresa la confianza que se tiene al tomar la decisión de rechazar la hipótesis nula. Cuanto más próximo esté el p -valor a 1, mayor confianza existe al aceptar la hipótesis nula, y cuanto más próximo esté a 0, mayor confianza hay al rechazarla.

Una vez fijado el riesgo α , la regla de decisión para realizar un contraste también puede expresarse de la siguiente manera:

$$\begin{aligned}\text{Si } p\text{-valor} \leq \alpha &\rightarrow \text{Rechazar } H_0, \\ \text{Si } p\text{-valor} > \alpha &\rightarrow \text{Aceptar } H_0,\end{aligned}$$

De este modo, el p -valor nos da información de para qué niveles de significación puede rechazarse la hipótesis nula y para cuales no.

Cálculo del p -valor de un contraste de hipótesis

Ejemplo

Si el contraste sobre la proporción de hombres se toma una muestra de tamaño 10 y se observa 1 hombre, entonces el p -valor, bajo a supuesta certeza de la hipótesis nula, $X \sim B(10, 0,5)$, será

$$p = P(X \leq 1) = 0,0107,$$

mientras que si en la muestra se observan 0 hombres, entonces el p -valor será

$$p = P(X \leq 0) = 0,001.$$

En el primer caso se rechazaría la hipótesis nula para un riesgo $\alpha = 0,05$, pero no podría rechazarse par un riesgo $\alpha = 0,01$, mientras que en el segundo caso también se rechazaría para $\alpha = 0,01$. Es evidente que en el segundo la decisión de rechazar la hipótesis nula se tomaría con mayor confianza.

Pasos para la realización de un contraste de hipótesis

- 1 Formular la hipótesis nula H_0 y la alternativa H_1 .
- 2 Fijar los riesgos α y β deseados.
- 3 Seleccionar el estadístico del contraste.
- 4 Fijar la mínima diferencia clínicamente significativa δ .
- 5 Calcular el tamaño muestral necesario n .
- 6 Delimitar las regiones de aceptación y rechazo.
- 7 Tomar una muestra de tamaño n .
- 8 Calcular el estadístico del contraste en la muestra.
- 9 Rechazar la hipótesis nula si la estimación cae en la región de rechazo o bien si el p -valor es menor que el riesgo α y aceptarla en caso contrario.

Contrastes paramétricos más importantes

Pruebas de conformidad:

- Contraste para la media de una población normal con varianza conocida.
- Contraste para la media de una población normal con varianza desconocida.
- Contraste para la media de una población con varianza desconocida a partir de muestras grandes.
- Contraste para la varianza de una población normal.
- Contraste para una proporción de una población.

Pruebas de homogeneidad:

- Contraste de comparación de medias de dos poblaciones normales con varianzas conocidas.
- Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas pero iguales.
- Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas y diferentes.
- Contraste de comparación de varianzas de dos poblaciones normales.
- Contraste de comparación de proporciones de dos poblaciones.

Contraste para la media de una población normal con varianza conocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- La media μ es desconocida, pero su varianza σ^2 es conocida.

Contraste:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Estadístico del contraste:

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0, 1).$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $z \leq z_{\alpha/2}$ y $z \geq z_{1-\alpha/2}$.

Contraste para la media de una población normal con varianza desconocida

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Estadístico del contraste: Utilizando la cuasivarianza como estimador de la varianza poblacional se tiene

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow T = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \sim T(n-1).$$

Región de aceptación: $t_{\alpha/2}^{n-1} < T < t_{1-\alpha/2}^{n-1}$.

Región de rechazo: $T \leq t_{\alpha/2}^{n-1}$ y $T \geq t_{1-\alpha/2}^{n-1}$.

Contraste para la media de una población normal con varianza desconocida

Ejemplo

En un grupo de alumnos se quiere contrastar si la nota media de estadística es mayor que 5 puntos. Para ello se toma la siguiente muestra:

6,3, 5,4, 4,1, 5,0, 8,2, 7,6, 6,4, 5,6, 4,3, 5,2

El contraste que se plantea es

$$H_0 : \mu = 5 \quad H_1 : \mu > 5$$

Para realizar el contraste se tiene:

- $\bar{x} = \frac{6,3+\dots+5,2}{10} = \frac{58,1}{10} = 5,81$ puntos.
- $\hat{s}^2 = \frac{(6,3-5,56)^2+\dots+(5,2-5,56)^2}{9} = \frac{15,949}{9} = 1,7721$ puntos², y $\hat{s} = 1,3312$ puntos.

Y el estadístico del contraste vale

$$T = \frac{\bar{x} - \mu_0}{\hat{s} / \sqrt{n}} = \frac{5,81 - 5}{1,3312 / \sqrt{10}} = 1,9246.$$

El p -valor del contraste es $P(T(9) \geq 1,9246) = 0,04323$, lo que indica que se rechazaría la hipótesis nula para $\alpha = 0.05$.

Contraste para la media de una población normal con varianza desconocida

Ejemplo

La región de rechazo es

$$T = \frac{\bar{x} - 5}{1,3312 / \sqrt{10}} \geq t_{0,95}^9 = 1,8331 \Leftrightarrow \bar{x} \geq 5 + 1,8331 \frac{1,3312}{\sqrt{10}} = 5,7717,$$

de modo que se rechazará la hipótesis nula siempre que la media de la muestra sea mayor que 5,7717 y se aceptará en caso contrario.

Suponiendo que en la práctica la mínima diferencia importante en la nota media fuese de un punto $\delta = 1$, entonces bajo la hipótesis alternativa $H_1 : \mu = 6$, si se decidiese rechazar la hipótesis nula, el riesgo β sería

$$\beta = P\left(T(9) \leq \frac{5,7717 - 6}{1,3312 \sqrt{10}}\right) = P(T(9) \leq -0,5424) = 0,3004,$$

de manera que la potencia del contraste para detectar una diferencia de $\delta = 1$ punto sería $1 - \beta = 1 - 0,3004 = 0,6996$.

Determinación del tamaño muestral en un contraste para la media

Se ha visto que para un riesgo α la región de rechazo era

$$T = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \geq t_{1-\alpha}^{n-1} \approx z_{1-\alpha} \quad \text{para } n \geq 30.$$

o lo que es equivalente

$$\bar{x} \geq \mu_0 + z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}}.$$

Si la mínima diferencia clínicamente significativa es δ , para una hipótesis alternativa $H_1 : \mu = \mu_0 + \delta$, el riesgo β es

$$\beta = P\left(Z < \frac{\mu_0 + z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - (\mu_0 + \delta)}{\frac{\hat{s}}{\sqrt{n}}}\right) = P\left(Z < \frac{z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - \delta}{\frac{\hat{s}}{\sqrt{n}}}\right).$$

de modo que

$$z_\beta = \frac{z_{1-\alpha} \frac{\hat{s}}{\sqrt{n}} - \delta}{\frac{\hat{s}}{\sqrt{n}}} \Leftrightarrow \delta = (z_{1-\alpha} - z_\beta) \frac{\hat{s}}{\sqrt{n}} \Leftrightarrow n = (z_{1-\alpha} - z_\beta)^2 \frac{\hat{s}^2}{\delta^2} = (z_\alpha + z_\beta)^2 \frac{\hat{s}^2}{\delta^2}.$$

Determinación del tamaño muestral en un contraste para la media

Ejemplo

Se ha visto en el ejemplo anterior que la potencia del contraste para detectar una diferencia en la nota media de 1 punto era del 69,96 %. Para aumentar la potencia del test hasta un 90 %, ¿cuántos alumnos habría que tomar en la muestra?

Como se desea una potencia $1 - \beta = 0,9$, el riesgo $\beta = 0,1$ y mirando en la tabla de la normal estándar se puede comprobar que $z_\beta = z_{0,1} = 1,2816$.

Aplicando la fórmula anterior para determinar el tamaño muestral necesario, se tiene

$$n = (z_\alpha + z_\beta)^2 \frac{\hat{s}^2}{\delta^2} = (1,6449 + 1,2816)^2 \frac{1,7721}{1^2} = 15,18,$$

de manera que habría que haber tomado al menos 16 alumnos.

Contraste para la media de una población con varianza desconocida y muestras grandes $n \geq 30$

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución puede ser de cualquier tipo.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Estadístico del contraste: Utilizando la cuasivarianza como estimador de la varianza poblacional y gracias al teorema central del límite por tratarse de muestras grandes ($n \geq 30$) se tiene

$$\bar{x} \sim N\left(\mu_0, \frac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \frac{\bar{x} - \mu_0}{\hat{s}/\sqrt{n}} \sim N(0, 1).$$

Región de aceptación: $-z_{\alpha/2} < Z < z_{\alpha/2}$.

Región de rechazo: $z \leq -z_{\alpha/2}$ y $z \geq z_{\alpha/2}$.

Contraste para la varianza de una población normal

Sea X una variable aleatoria que cumple las siguientes hipótesis:

- Su distribución es normal $X \sim N(\mu, \sigma)$.
- Tanto su media μ como su varianza σ^2 son desconocidas.

Contraste:

$$H_0 : \sigma = \sigma_0$$

$$H_1 : \sigma \neq \sigma_0$$

Estadístico del contraste: Partiendo de la cuasivarianza muestral como estimador de la varianza poblacional, se tiene

$$J = \frac{nS^2}{\sigma_0^2} = \frac{(n-1)\hat{S}^2}{\sigma_0^2} \sim \chi^2(n-1),$$

que sigue una distribución chi-cuadrado de $n - 1$ grados de libertad.

Región de aceptación: $\chi_{\alpha/2}^{n-1} < J < \chi_{1-\alpha/2}^{n-1}$.

Región de rechazo: $J \leq \chi_{\alpha/2}^{n-1}$ y $J \geq \chi_{1-\alpha/2}^{n-1}$.

Contraste para la varianza de una población normal

Ejemplo

En un grupo de alumnos se quiere contrastar si la desviación típica de la nota es mayor de 1 punto. Para ello se toma la siguiente muestra:

6,3, 5,4, 4,1, 5,0, 8,2, 7,6, 6,4, 5,6, 4,3, 5,2

El contraste que se plantea es

$$H_0 : \sigma = 1 \quad H_1 : \sigma > 1$$

Para realizar el contraste se tiene:

$$\begin{aligned} - \bar{x} &= \frac{6,3 + \dots + 5,2}{10} = \frac{58,1}{10} = 5,81 \text{ puntos.} \\ - \hat{s}^2 &= \frac{(6,3-5,81)^2 + \dots + (5,2-5,81)^2}{9} = \frac{15,949}{9} = 1,7721 \text{ puntos}^2. \end{aligned}$$

El estadístico del contraste vale

$$J = \frac{(n-1)\hat{s}^2}{\sigma_0^2} = \frac{9 \cdot 1,7721}{1^2} = 15,949,$$

y el p -valor del contraste es $P(\chi(9) \geq 15,949) = 0,068$, por lo que no se puede rechazar la hipótesis nula para $\alpha = 0,05$.

Contraste para proporción de una población

Sea p la proporción de individuos de una población que tienen una determinada característica.

Contraste:

$$H_0 : p = p_0$$

$$H_1 : p \neq p_0$$

Estadístico del contraste: La variable que mide el número de individuos con la característica en una muestra aleatoria de tamaño n sigue una distribución binomial $X \sim B(n, p_0)$. De acuerdo al teorema central del límite, para muestras grandes ($np \geq 5$ y $n(1 - p) \geq 5$), $X \sim N(np_0, \sqrt{np_0(1 - p_0)})$, y se cumple

$$\hat{p} = \frac{X}{n} \sim N\left(p_0, \sqrt{\frac{p_0(1 - p_0)}{n}}\right) \Rightarrow Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} \sim N(0, 1).$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $z \leq z_{\alpha/2}$ y $z \geq z_{1-\alpha/2}$.

Contraste para proporción de una población

Ejemplo

En un grupo de alumnos se desea estimar si el porcentaje de aprobados es mayor del 50 %. Para ello se toma una muestra de 80 alumnos entre los que hay 50 aprobados.

El contraste que se plantea es

$$H_0 : p = 0,5$$

$$H_1 : p > 0,5$$

Para realizar el contraste se tiene que $\hat{p} = 50/80 = 0,625$ y como se cumple $n\hat{p} = 80 \cdot 0,625 = 50 \geq 5$ y $n(1 - \hat{p}) = 80(1 - 0,625) = 30 \geq 5$, el estadístico del contraste vale

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}} = \frac{0,625 - 0,5}{\sqrt{0,5(1 - 0,5)/80}} = 2,2361.$$

y el p -valor del contraste es $P(Z \geq 2,2361) = 0,0127$, por lo que se rechaza la hipótesis nula para $\alpha = 0,05$ y se concluye que el porcentaje de aprobados es mayor de la mitad.

Contraste de comparación de medias de dos poblaciones normales con varianzas conocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 y μ_2 son desconocidas, pero sus varianzas σ_1^2 y σ_2^2 son conocidas.

Contraste:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Estadístico del contraste:

$$\left. \begin{array}{l} \bar{X}_1 \sim N\left(\mu_1, \frac{\sigma_1^2}{n_1}\right) \\ \bar{X}_2 \sim N\left(\mu_2, \frac{\sigma_2^2}{n_2}\right) \end{array} \right\} \Rightarrow \bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right) \Rightarrow Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1).$$

Región de aceptación: $-z_{\alpha/2} < Z < z_{\alpha/2}$.

Región de rechazo: $z \leq -z_{\alpha/2}$ y $z \geq z_{\alpha/2}$.

Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 y μ_2 son desconocidas y sus varianzas también, pero son iguales $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Contraste:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Estadístico del contraste:

$$\left. \begin{aligned} \bar{X}_1 - \bar{X}_2 &\sim N\left(\mu_1 - \mu_2, \sigma \sqrt{\frac{n_1 + n_2}{n_1 n_2}}\right) \\ \frac{n_1 S_1^2 + n_2 S_2^2}{\sigma^2} &\sim \chi^2(n_1 + n_2 - 2) \end{aligned} \right\} \Rightarrow T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} \sim T(n_1 + n_2 - 2).$$

Región de aceptación: $-t_{\alpha/2}^{n_1+n_2-2} < T < t_{\alpha/2}^{n_1+n_2-2}$.

Región de rechazo: $T < -t_{\alpha/2}^{n_1+n_2-2} \vee T > t_{\alpha/2}^{n_1+n_2-2}$

Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales

Ejemplo

Se quiere comparar el rendimiento académico de dos grupos de alumnos, uno con 10 alumnos y otro con 12, que han seguido metodologías diferentes. Para ello se les realiza un examen y se obtienen las siguientes puntuaciones:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

El contraste que se plantea es

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Para realizar el contraste, se tiene

- $\bar{X}_1 = \frac{4+\dots+3}{10} = 5,3$ puntos y $\bar{X}_2 = \frac{8+\dots+7}{12} = 6,75$ puntos.
- $S_1^2 = \frac{(4^2+\dots+3^2)}{10} - 5,3^2 = 3,21$ puntos² y $S_2^2 = \frac{8^2+\dots+3^2}{12} - 6,75^2 = 2,69$ puntos².
- $\hat{S}_p^2 = \frac{10 \cdot 3,21 + 12 \cdot 2,6875}{10+12-2} = 3,2175$ puntos², y $\hat{S}_p = 1,7937$.

Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas e iguales

Ejemplo

Si se suponen varianzas iguales, el estadístico del contraste vale

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\hat{S}_p \sqrt{\frac{n_1 + n_2}{n_1 n_2}}} = \frac{5,3 - 6,75}{1,7937 \sqrt{\frac{10+12}{10 \cdot 12}}} = -1,8879,$$

y el p -valor del contraste es $2P(T(20) \leq -1,8879) = 0,0736$, de modo que no se puede rechazar la hipótesis nula y se concluye que no hay diferencias significativas entre las notas medias de los grupos.

Contraste de comparación de medias de dos poblaciones normales con varianzas desconocidas

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 , μ_2 y varianzas σ_1^2 , σ_2^2 , son desconocidas, pero $\sigma_1^2 \neq \sigma_2^2$.

Contraste:

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : \mu_1 \neq \mu_2$$

Estadístico del contraste:

$$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\hat{S}_1^2}{n_1} + \frac{\hat{S}_2^2}{n_2}}} \sim T(g), \quad \text{con } g = n_1 + n_2 - 2 - \Delta \quad \text{y } \Delta = \frac{\left(\frac{n_2-1}{n_1} \hat{S}_1^2 - \frac{n_1-1}{n_2} \hat{S}_2^2\right)^2}{\frac{n_2-1}{n_1^2} \hat{S}_1^4 + \frac{n_1-1}{n_2^2} \hat{S}_2^4}.$$

Región de aceptación: $-t_{\alpha/2}^g < T < t_{\alpha/2}^g$.

Región de rechazo: $T \leq -t_{\alpha/2}^g$ y $T \geq t_{\alpha/2}^g$.

Contraste de comparación de varianzas de dos poblaciones normales

Sean X_1 y X_2 dos variables aleatorias que cumplen las siguientes hipótesis:

- Su distribución es normal $X_1 \sim N(\mu_1, \sigma_1)$ y $X_2 \sim N(\mu_2, \sigma_2)$.
- Sus medias μ_1 , μ_2 y varianzas σ_1^2 , σ_2^2 son desconocidas.

Contraste:

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 \neq \sigma_2$$

Estadístico del contraste:

$$\left. \begin{array}{l} \frac{(n_1 - 1)\hat{S}_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \\ \frac{(n_2 - 1)\hat{S}_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1) \end{array} \right\} \Rightarrow F = \frac{\frac{\frac{(n_1-1)\hat{S}_1^2}{\sigma_1^2}}{n_1-1}}{\frac{\frac{(n_2-1)\hat{S}_2^2}{\sigma_2^2}}{n_2-1}} = \frac{\sigma_2^2}{\sigma_1^2} \frac{\hat{S}_1^2}{\hat{S}_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Región de aceptación: $F_{\alpha/2}^{n_1-1, n_2-1} < F < F_{1-\alpha/2}^{n_1-1, n_2-1}$.

Región de rechazo: $F \leq F_{\alpha/2}^{n_1-1, n_2-1}$ y $F \geq F_{1-\alpha/2}^{n_1-1, n_2-1}$.

Contraste de comparación de varianzas de dos poblaciones normales

Siguiendo con el ejemplo de las puntuaciones en dos grupos:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

si se desea comparar las varianzas, el contraste que se plantea es

$$H_0 : \sigma_1 = \sigma_2 \quad H_1 : \sigma_1 \neq \sigma_2$$

Para realizar el contraste, se tiene

- $\bar{X}_1 = \frac{4+\dots+3}{10} = 5,3$ puntos y $\bar{X}_2 = \frac{8+\dots+7}{12} = 6,75$ puntos.
- $\hat{S}_1^2 = \frac{(4-5,3)^2+\dots+(3-5,3)^2}{9} = 3,5667$ y $\hat{S}_2^2 = \frac{(8-6,75)^2+\dots+(3-6,75)^2}{11} = 2,9318$ puntos².

El estadístico del contraste vale

$$F = \frac{\hat{S}_1^2}{\hat{S}_2^2} = \frac{3,5667}{2,9318} = 1,2165,$$

y el p -valor del contraste es $2P(F(9, 11) \leq 1,2165) = 0,7468$, por lo que se mantiene la hipótesis de igualdad de varianzas.

Contraste de comparación de proporciones de dos poblaciones

Sean p_1 y p_2 las respectivas proporciones de individuos que presentan una determinada característica en dos poblaciones.

Contraste:

$$H_0 : p_1 = p_2 \quad H_1 : p_1 \neq p_2$$

Estadístico del contraste: Las variables que miden el número de individuos con la característica en dos muestras aleatorias de tamaños n_1 y n_2 respectivamente, siguen distribuciones binomiales $X_1 \sim B(n_1, p_1)$ y $X_2 \sim B(n_2, p_2)$. Si las muestras son grandes ($n_i p_i \geq 5$ y $n_i(1 - p_i) \geq 5$), de acuerdo al teorema central del límite, $X_1 \sim N(np_1, \sqrt{np_1(1 - p_1)})$ y $X_2 \sim N(np_2, \sqrt{np_2(1 - p_2)})$, y se cumple

$$\left. \begin{aligned} \hat{p}_1 = \frac{X_1}{n_1} &\sim N\left(p_1, \sqrt{\frac{p_1(1-p_1)}{n_1}}\right) \\ \hat{p}_2 = \frac{X_2}{n_2} &\sim N\left(p_2, \sqrt{\frac{p_2(1-p_2)}{n_2}}\right) \end{aligned} \right\} \Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$$

Región de aceptación: $z_{\alpha/2} < Z < z_{1-\alpha/2}$.

Región de rechazo: $z \leq z_{\alpha/2}$ y $z \geq z_{1-\alpha/2}$.

Contraste de comparación de proporciones de dos poblaciones

Se quiere comparar los porcentajes de aprobados en dos grupos que han seguido metodologías distintas. En el primer grupo han aprobado 24 alumnos de un total de 40, mientras que en el segundo han aprobado 48 de 60.

El contraste que se plantea es

$$H_0 : p_1 = p_2 \quad H_1 : p_1 \neq p_2$$

Para realizar el contraste, se tiene $\hat{p}_1 = 24/40 = 0,6$ y $\hat{p}_2 = 48/60 = 0,8$, de manera que se cumplen las condiciones $n_1\hat{p}_1 = 40 \cdot 0,6 = 24 \geq 5$, $n_1(1 - \hat{p}_1) = 40(1 - 0,6) = 26 \geq 5$, $n_2\hat{p}_2 = 60 \cdot 0,8 = 48 \geq 5$ y $n_2(1 - \hat{p}_2) = 60(1 - 0,8) = 12 \geq 5$, y el estadístico del contraste vale

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{0,6 - 0,8}{\sqrt{\frac{0,6(1-0,6)}{40} + \frac{0,8(1-0,8)}{60}}} = -2,1483,$$

y el p -valor del contraste es $2P(Z \leq -2,1483) = 0,0317$, de manera que se rechaza la hipótesis nula para $\alpha = 0,05$ y se concluye que hay diferencias.

Realización de un contraste mediante un intervalo de confianza

Una interesante alternativa a la realización de un contraste

$$H_0 : \theta = \theta_0 \quad H_1 : \theta \neq \theta_0$$

con un riesgo α , es calcular el intervalo de confianza para θ con un nivel de confianza $1 - \alpha$, ya que este intervalo se puede interpretar como el conjunto aceptable de hipótesis para θ , de manera que si θ_0 está fuera del intervalo, la hipótesis nula es poco creíble y puede rechazarse, mientras que si está dentro la hipótesis es creíble y se acepta.

Cuando el contraste sea unilateral de menor, el contraste se realizaría comparando θ_0 con el límite superior del intervalo de confianza para θ con un nivel de confianza $1 - 2\alpha$, mientras que si el contraste es unilateral de mayor, se comparará con el límite inferior del intervalo.

Contraste	Intervalo de confianza	Decisión
Bilateral	$[l_i, l_s]$ con nivel de confianza $1 - \alpha$	Rechazar H_0 si $\theta_0 \notin [l_i, l_s]$
Unilateral menor	$[l_i, l_s]$ con nivel de confianza $1 - 2\alpha$	Rechazar H_0 si $\theta_0 \geq l_s$
Unilateral mayor	$[l_i, l_s]$ con nivel de confianza $1 - 2\alpha$	Rechazar H_0 si $\theta_0 \leq l_i$

Realización de un contraste mediante un intervalo de confianza

Ejemplo

Volviendo al contraste para comparar el rendimiento académico de dos grupos de alumnos que han obtenido las siguientes puntuaciones:

$$X_1 : 4 - 6 - 8 - 7 - 7 - 6 - 5 - 2 - 5 - 3$$

$$X_2 : 8 - 9 - 5 - 3 - 8 - 7 - 8 - 6 - 8 - 7 - 5 - 7$$

El contraste que se planteaba era

$$H_0 : \mu_1 = \mu_2 \quad H_1 : \mu_1 \neq \mu_2$$

Como se trata de un contraste bilateral, el intervalo de confianza para la diferencia de medias $\mu_1 - \mu_2$ con nivel de confianza $1 - \alpha = 0,95$, suponiendo varianzas iguales, vale $[-3,0521, 0,1521]$ puntos. Y como según la hipótesis nula $\mu_1 - \mu_2 = 0$, y el 0 cae dentro del intervalo, se acepta la hipótesis nula.

La ventaja del intervalo es que, además de permitirnos realizar el contraste, nos da una idea de la magnitud de la diferencia entre las medias de los grupos.