Manual Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Sep 2017

Departamento de Matemática Aplicada y Estadística
CEU San Pablo

Términos de la licencia

Contenidos

Introducción a la Estadística

¿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.

El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.

Datos

¿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.

El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.

¿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.

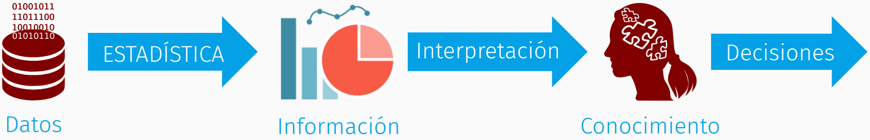
El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.

¿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.

El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.



¿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.

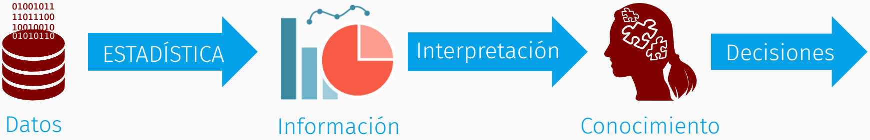
El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.

La estadística es imprescindible en cualquier disciplina científica o técnica donde se manejen datos, especialmente si son grandes volúmenes de datos, como por ejemplo en Física, Química, Medicina, Psicología, Economía o Ciencias Sociales.

Definición (Estadística)

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

El papel de la Estadística es extraer información de los datos para adquirir el conocimiento necesario para tomar decisiones.



La estadística es imprescindible en cualquier disciplina científica o técnica donde se manejen datos, especialmente si son grandes volúmenes de datos, como por ejemplo en Física, Química, Medicina, Psicología, Economía o Ciencias Sociales.

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.

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.

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

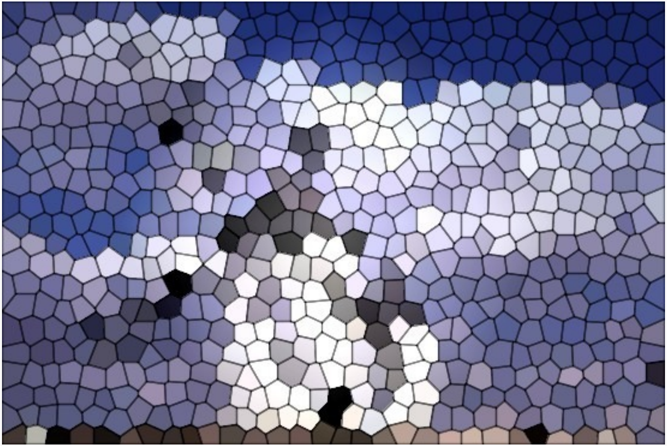
Una de las preguntas más interesantes que surge inmediatamente es:

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

La respuesta depende de varios factores, como la variabilidad de la población o la fiabilidad deseada para las extrapolaciones que se hagan hacia la población.

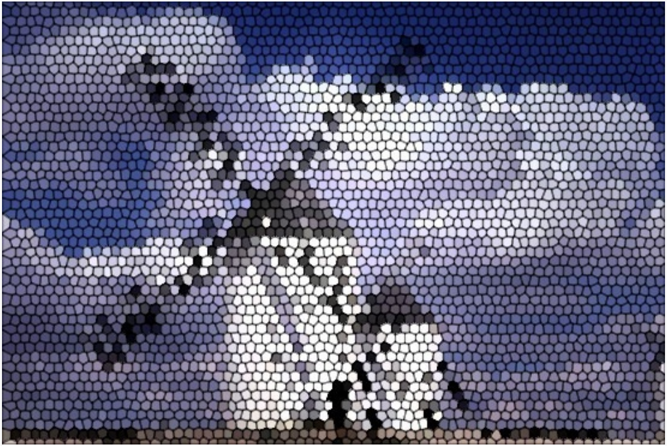
Por desgracia no se podrá responder hasta casi el final del curso, pero en general, cuantos más individuos haya en la muestra, más fiables serán las conclusiones sobre la población, pero también será más lento y costoso el estudio.

¿Puedes averiguar el motivo de la foto?



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

¿Eres capaz de adivinar el motivo de la foto ahora?



¡Con una muestra mayor es más fácil averiguar el contenido de la 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!). Sin embargo, *¡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.

Definición (Muestreo)

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

Para que una muestra refleje información fidedigna sobre la población global debe ser representativa de la misma, lo que significa que debe reproducir a pequeña escala la variabilidad de la población.

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. Algunos individuos tienen más probabilidad de ser seleccionados que otros.

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 las conclusiones.

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.

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, es decir, cada individuo seleccionado es devuelto a la población antes de seleccionar al siguiente (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.

Todo estudio estadístico comienza por la identificación de las características que interesa estudiar en la población y que se medirán en los individuos de la muestra.

Definition (Variable estadística)

Una *variable estadística* es una propiedad o característica medida en los individuos de la población.

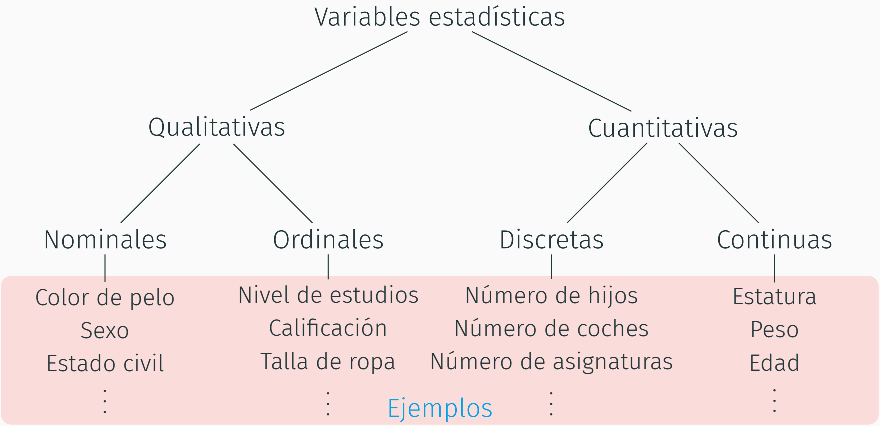
Los *datos* son los valores observados en las variables estadísticas.

Variables estadísticas y atributos

De acuerdo a la naturaleza de los valores y su escala, se tiene:

- Variables cualitativas o atributos: Miden cualidades no numéricas. Pueden ser:
  - Nominales: No existe un orden natural entre las categorías.
    Ejemplo: El color de pelo o el sexo.
  - Ordinales: Existe un orden natural entre las categorías.
    Ejemplo: El nivel de estudios o la gravedad de una enfermedad.
- Variables cuantitativas: Miden cantidades numéricas. Pueden ser:
  - Discretas: Toman valores numéricos aislados (habitualmente números enteros).
    Ejemplo: El número de hijos o de coches en una familia.
  - Continuas: Pueden tomar cualquier valor en un intervalo real.
    Ejemplo: La estatura, el peso, o la edad de una persona.

Tipos de variables estadísticas



Tipos de variables estadísticas

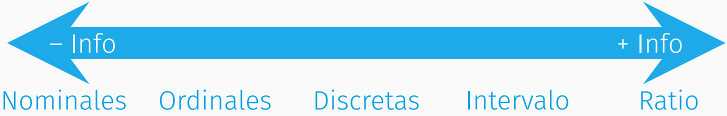
Eligiendo la variable adecuada

En ocasiones una característica puede medirse mediante variables de distinto tipo.

Ejemplo Si una persona fuma o no podría medirse de diferentes formas:

- Fuma: si/no. (Nominal)
- Nivel de fumador: No fuma / ocasional / moderado / bastante / empedernido. (Ordinal)
- Número de cigarros diarios: 0,1,2,...(Discreta)

En estos casos es preferible usar variables cuantitativas a cualitativas. Dentro de las cuantitativas es preferible usar las continuas a las discretas y dentro de las cualitativas es preferible usar ordinales a nominales pues aportan más información.



Tipos de variables estadísticas

De acuerdo al papel que juegan en el estudio:

- **Variables independientes:** Variables que supuestamente no dependen de otras variables en el estudio. Habitualmente son las variables manipuladas en el experimento para ver su efecto en las variables dependientes. Se conocen también como *variables predictivas*.
- **Variables dependientes:** Variables que supuestamente dependen de otras variables en el estudio. No son manipuladas en el experimento y también se conocen como *variables respuesta*.

Ejemplo En un estudio sobre el rendimiento de los alumnos de un curso, la inteligencia de los alumnos y el número de horas de estudio diarias serían variables independientes y la nota del curso sería una variable dependiente.

Tipos de estudios estadísticos

- **Experimentales:** Cuando las variables independientes son manipuladas para ver el efecto que producen en las variables dependientes.
  Ejemplo En un estudio sobre el rendimiento de los estudiantes en un test, el profesor manipula la metodología de estudio para crear dos o más grupos con metodologías de estudio distintas.
- **No experimentales:** Cuando las variables independientes no son manipuladas. Esto no significa que sea imposible hacerlo, sino que es difícil o poco ético hacerlo.
  Ejemplo En un estudio un investigador puede estar interesado en el efecto de fumar sobre el cáncer de pulmón. Aunque es posible, no sería ético pedirle a los pacientes que fumasen para ver el efecto que tiene sobre sus pulmones. En este caso, el investigador podría estudiar dos grupos de pacientes, uno con cáncer de pulmón y otro sin cáncer, y observar en cada grupo cuántos fuman o no.

Los estudios experimentales permiten identificar causas y efectos entre las variables del estudio, mientras que los no experimentales sólo permiten

Las variables 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 **tabla de datos**.

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

Ejemplo

Nombre	Edad	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

Normalmente un estudio estadístico pasa por las siguientes etapas:

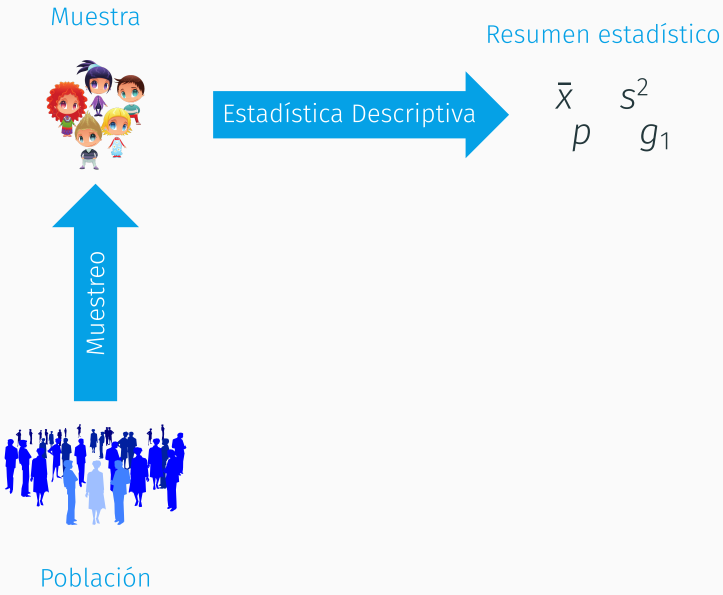
1. El estudio comienza por el diseño previo del mismo en el que se establezcan los objetivos del mismo, la población, las variables que se medirán y el tamaño muestral requerido.
2. A continuación se seleccionará una muestra representativa del tamaño establecido y se medirán las variables en los individuos de la muestra obteniendo la tabla de datos. De esto se encarga el **Muestreo**.
3. El siguiente paso consiste en describir y resumir la información que contiene la muestra. De esto se encarga la **Estadística Descriptiva**.
4. La información obtenida es proyectada sobre un modelo matemático que intenta explicar el comportamiento de la población y el modelo se valida. De todo esto se encarga la **Estadística Inferencial**.
5. Finalmente, el modelo validado nos permite hacer predicciones y sacar conclusiones sobre la población de partida con cierta confianza.

El ciclo estadístico

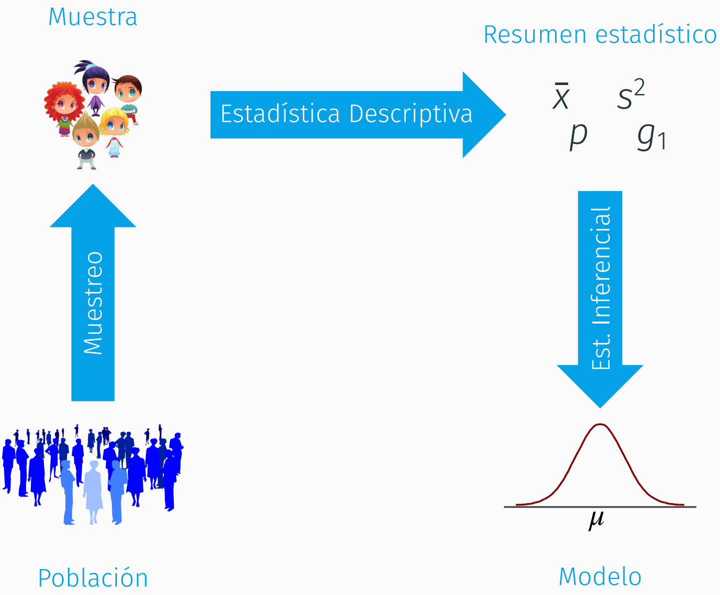
Población

El ciclo estadístico

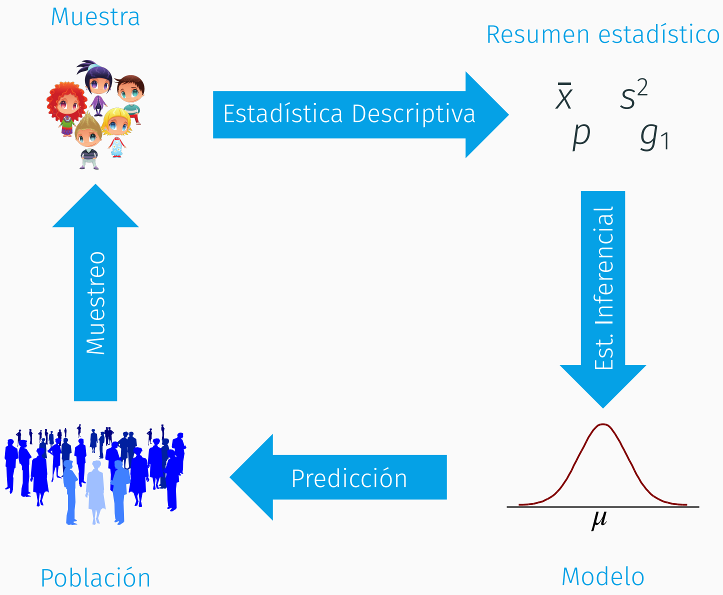
El ciclo estadístico



El ciclo estadístico



El ciclo estadístico



Distribución de frecuencias: Tabulación y gráficos

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. Tabular y representar gráficamente los datos de acuerdo a sus frecuencias.
3. Calcular medidas que resuman la información que contiene la muestra (*estadísticos muestrales*).

No tiene poder inferencial $\Rightarrow$ *No utilizar para sacar conclusiones sobre la población!*

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:

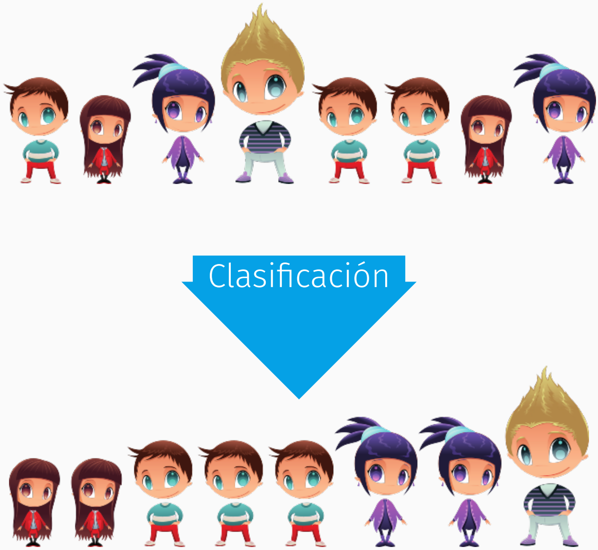
Sin agrupar : Ordenar todos los valores obtenidos en la muestra de menor a mayor (si existe orden). Se utiliza con atributos y variables discretas con pocos valores diferentes.

Agrupados : Agrupar los valores en clases (intervalos) y ordenar dichas clases de menor a mayor. Se utiliza con variables continuas y con variables discretas con muchos valores diferentes.

X =Estatura

Clasificación de la muestra

X =Estatura



Recuento de frecuencias

$X =$Estatura

Recuento de frecuencias

X =Estatura

Definición (Frecuencias muestrales)

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

- **Frecuencia absoluta n_i**: Es el número de veces que el valor x_i aparece en la muestra.
- **Frecuencia relativa f_i**: Es la proporción de veces que el valor x_i aparece en la muestra.
$$f_i = \frac{n_i}{n}$$
- **Frecuencia absoluta acumulada N_i**: Es el número de valores en la muestra menores o iguales que x_i.
$$N_i = n_1 + \cdots + n_i$$
- **Frecuencia relativa acumulada F_i**: Es la proporción de valores en la muestra menores o iguales que x_i.
$$F_i = \frac{N_i}{}$$

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

El número de hijos en 25 familias es

$$1, 2, 4, 2, 2, 2, 3, 2, 1, 1, 0, 2, 2,$$
$$0, 2, 2, 1, 2, 2, 3, 1, 2, 2, 1, 2.$$

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
$\sum$	25	1		

Las estaturas (en cm) de 30 estudiantes es

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

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
$\sum$	30	1		

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 $\sqrt{n}$ o $\log_2(n)$.
- 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.

Los grupos sanguíneos de 30 personas son

A, B, B, A, AB, 0, 0, A, B, B, A, 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
$\sum$	30	1

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

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
- Histograma
- Diagrama de líneas
- Digrama de sectores

Diagrama de barras

Un diagrama de barras consiste en un conjunto de barras, una para cada valor o categoría de la variabe, dibujadas en unos ejes cartesianos.

Habitualmente os valores o categorías de la variable se representan en el eje X, y las frecuencias en el eje Y. Para cada valor o categoría de la variabe se dibuja una barra de altura la correspondiente frecuencia. La anchura de la barra es indiferente pero debe haber una separación clara entre las barras.

Dependiendo de la frecuencia representada en el eje Y se tienen distintos tipos de diagramas de barras.

A veces se dibuja un polígono, conocido como polígono de frecuencias, uniendo los puntos más altos de cada barra con segmentos.

Diagrama de barras de frecuencias absolutas
Datos sin agrupar

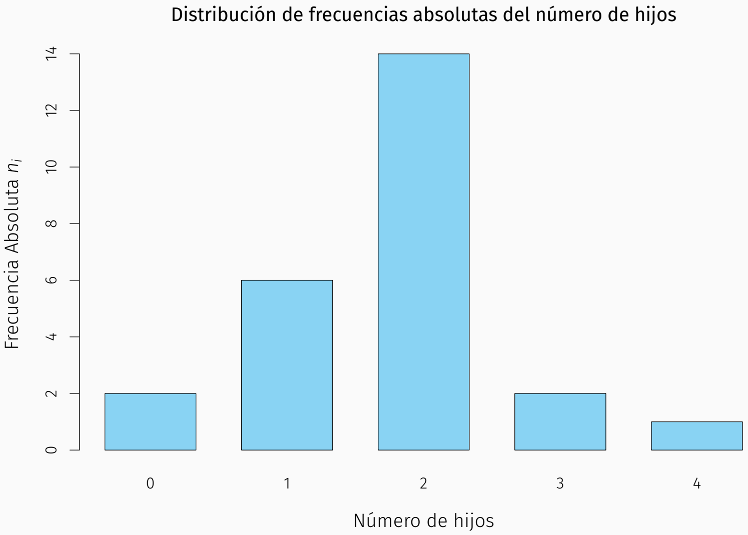


Diagrama de líneas o Polígono de frecuencias absolutas
Datos sin agrupar

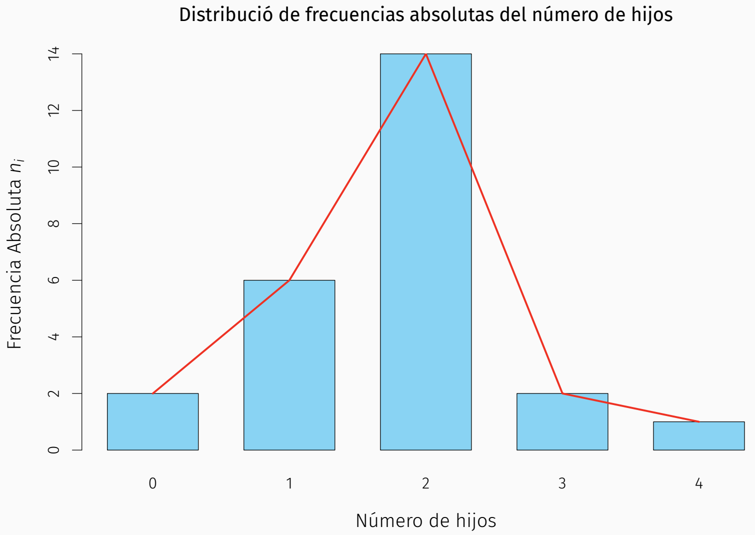


Diagrama de barras de frecuencias acumuladas
Datos sin agrupar

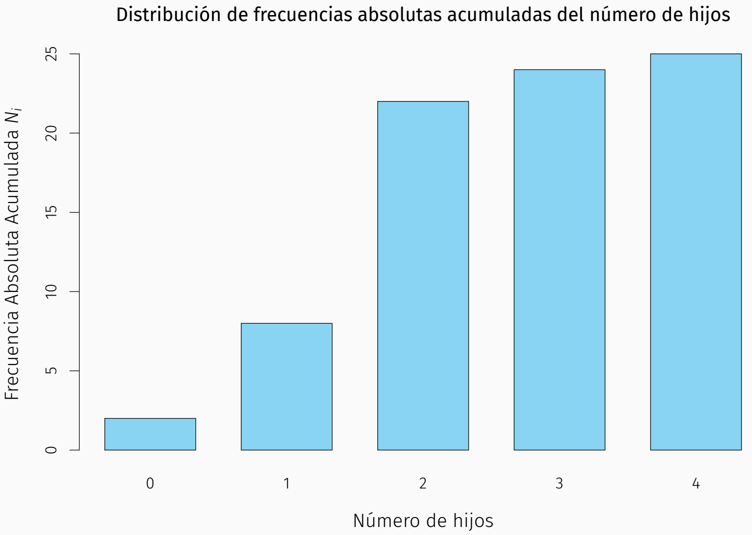
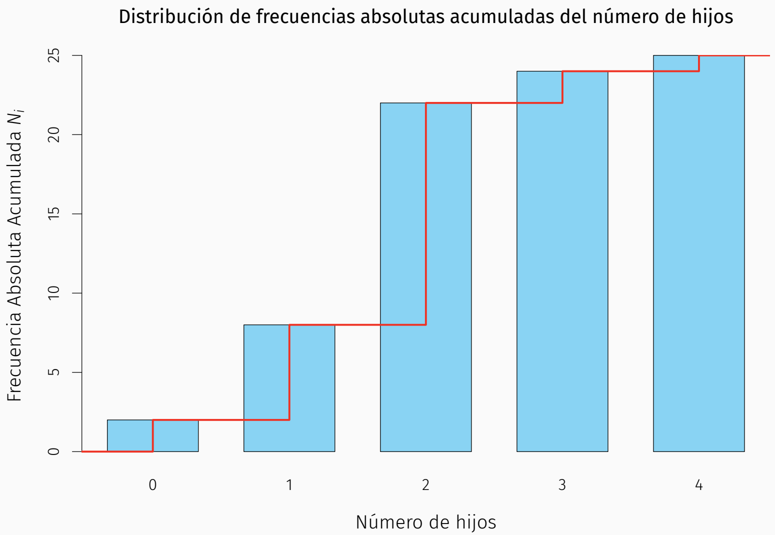


Diagrama de línea o polígono de frecuencias absolutas acumuladas

Datos sin agrupar



Un histograma es similar a un diagrama de barras pero para datos agrupados.

Habitualmente las clases o intervalos de agrupación se representan en el eje X, y las frecuencias en el eje Y.

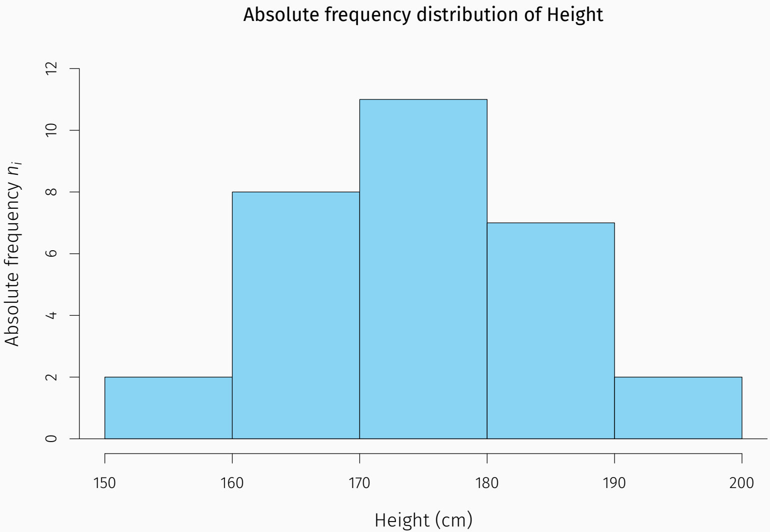
Para cada clase se dibuja una barra de altura la correspondiente frecuencia. A diferencia del diagrama de barras, la anchura del la barra coincide con la anchura de las clases y no hay separación entre dos barras consecutivas.

Dependiendo del tipo de frecuencia representada en el eje Y existen distintos tipos de histogramas.

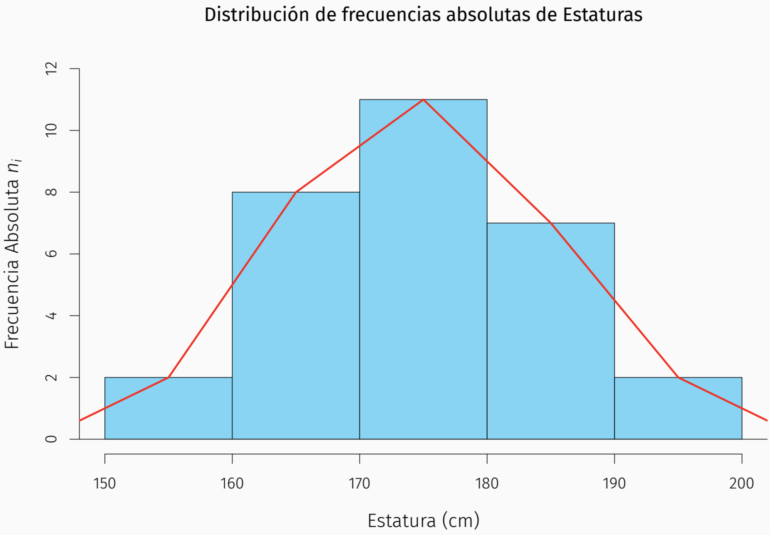
A veces se dibuja un polígono, conocido como polígono de frecuencias, uniendo los puntos más altos de cada barra con segmentos.

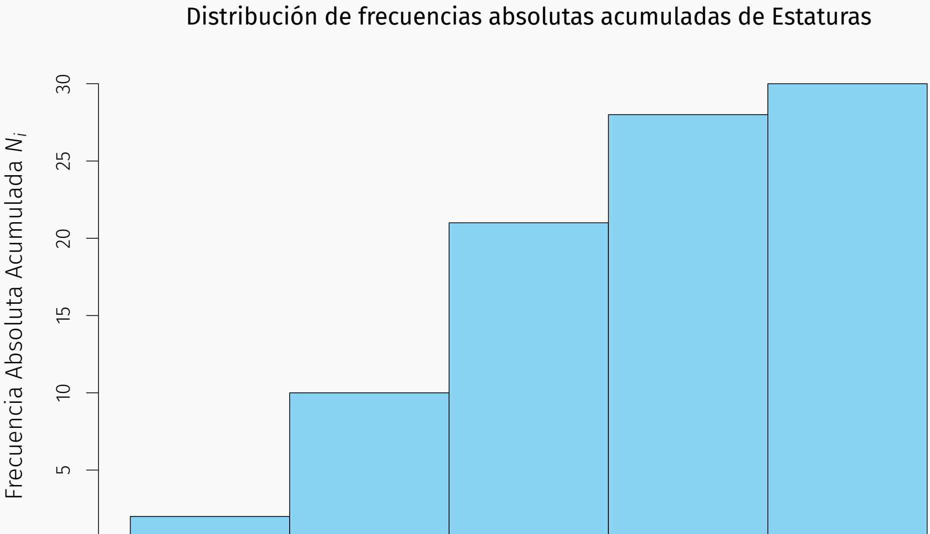
Histograma de frecuencias absolutas

Datos agrupados



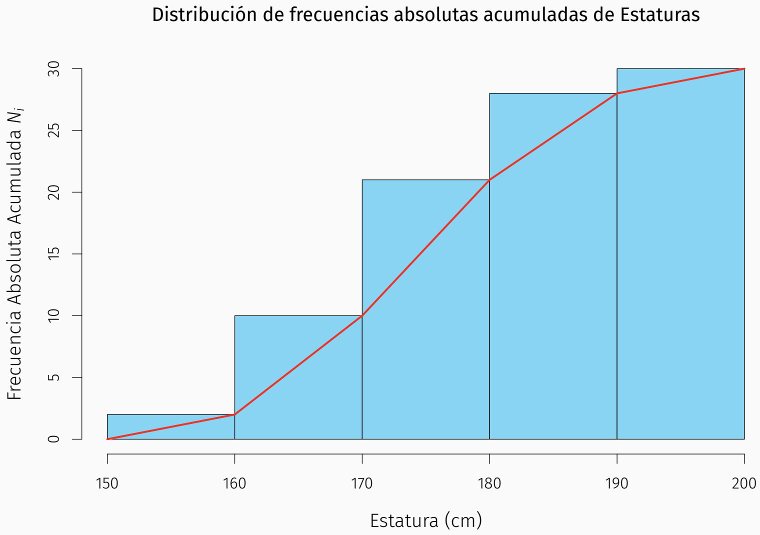
Polígono de frecuencias absolutas
Datos agrupados




Distribución de frecuencias absolutas acumuladas de Estaturas
Frecuencia Absoluta Acumulada N_i
5
10
15
20
25
30

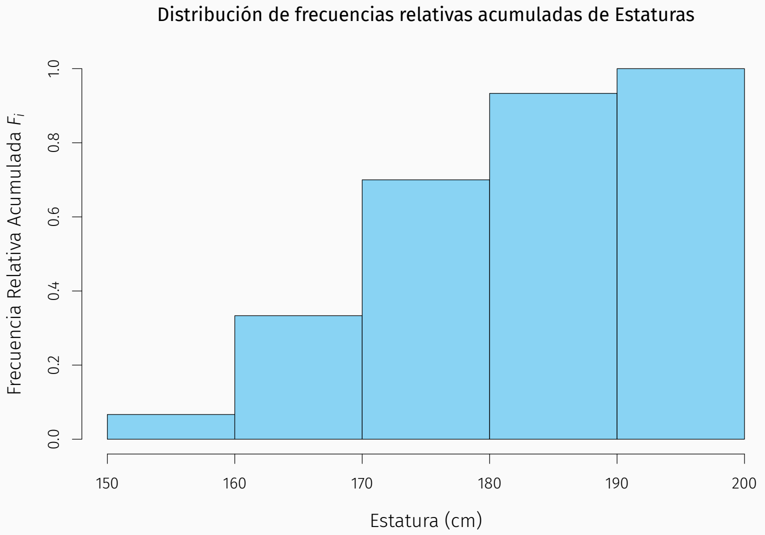
Polígono de frecuencias absolutas acumuladas

Datos agrupados

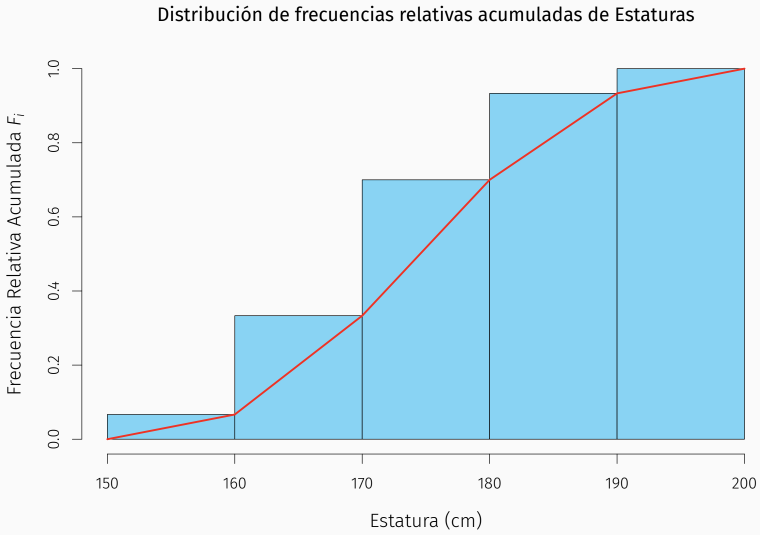


Histograma de frecuencias relativas acumuladas

Datos agrupados



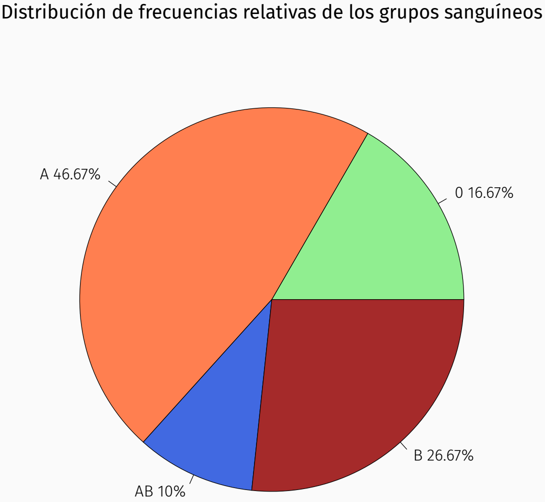
Polígono de frecuencias relativas acumuladas
Datos agrupados



Un diagrama de sectores consiste en un círculo divido en porciones, uno por cada valor o categoría de la variable.

Cada porción se conoce como sector y su ángulo o área es proporcional a la correspondiente frecuencia del valor o categoría.

Los diagramas de sectores pueden representar frecuencias absolutas o relativas, pero no pueden representar frecuencias acumuladas, y se utilizan sobre todo con atributos nominales. Para atributos ordinales o variables cuantitativas es mejor utilizar diagramas de barras o histogramas, ya es más fácil percibir las diferencias en una dimensión (altura de las barras) que en dos dimensiones (áreas de los sectores).



Distribución de frecuencias relativas de los grupos sanguíneos
A 46.67%
0 16.67%
B 26.67%
AB 10%

Datos atípicos

Uno de los principales problemas de las muestras son los **datos atípicos**, que son valores muy distintos de los demás valores en la muestra.

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, y pueden detectarse fácilmente con un diagrama de caja y bigotes (como se verá después).

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:

- Eliminar el dato atípico si es un error.
- Sustituir el dato atípico por el mayor o menor valor de la distribución que no sea atípico, si no es un error pero que no concuerda con el modelo de distribución teórico de la población.
- Dejarl el dato atípico si no es un error y cambiar el modelo de distribución teórico para ajustarse a los datos atípicos.

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.

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.

Media aritmética

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
$\sum$	25	1	44	1.76

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{44}{25} = 1.76 \qquad \bar{x} = \sum x_i f_i = 1.76.$$

Es decir, el número de hijos que mejor representa a la muestra es 1.76

En el ejemplo anterior de las estaturas se tiene

$$\bar{x} = \frac{179 + 173 + \cdots + 187}{30} = 175.07 \text{ cm}.$$

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
$\sum$		30	1	5240	174.67

$$\bar{x} = \frac{\sum x_i n_i}{n} = \frac{5240}{30} = 174.67 \qquad \bar{x} = \sum x_i f_i = 174.67.$$

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 $\frac{n}{2}+1$.

$X =$Estatura

n impar

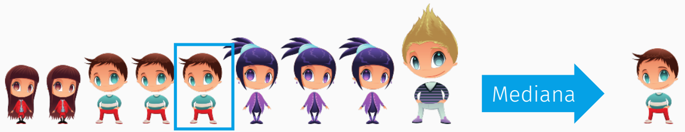
Cálculo de la mediana con datos no agrupados

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 $\frac{n}{2}+1$.

$X =$Estatura

n impar



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 $\frac{n}{2}+1$.

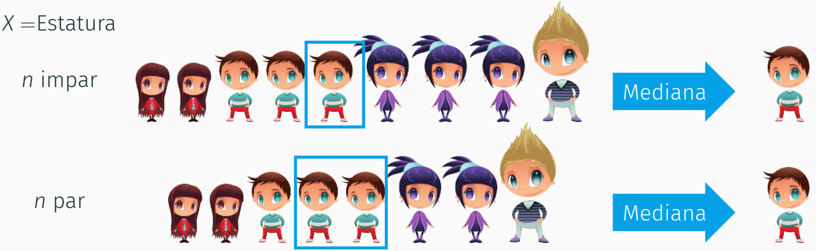
$X =$Estatura

n impar

n par

Con datos no agrupados pueden darse varios casos:

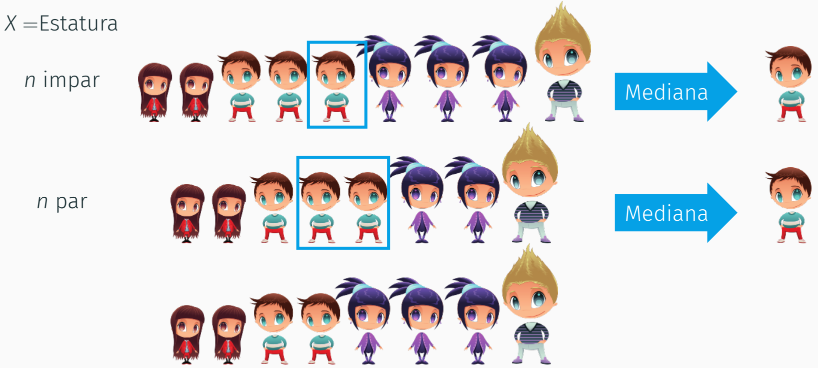
- Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.
- Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$.



Cálculo de la mediana con datos no agrupados

Con datos no agrupados pueden darse varios casos:

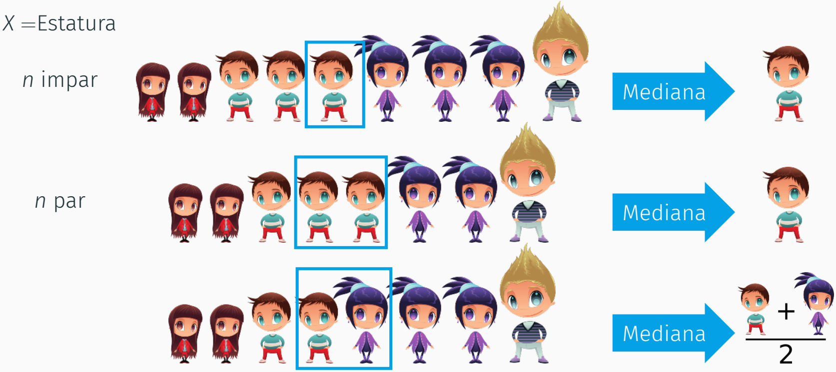
- Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.
- Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$.



Cálculo de la mediana 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 $\frac{n}{2}+1$.



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 $\frac{25+1}{2} = 13$.

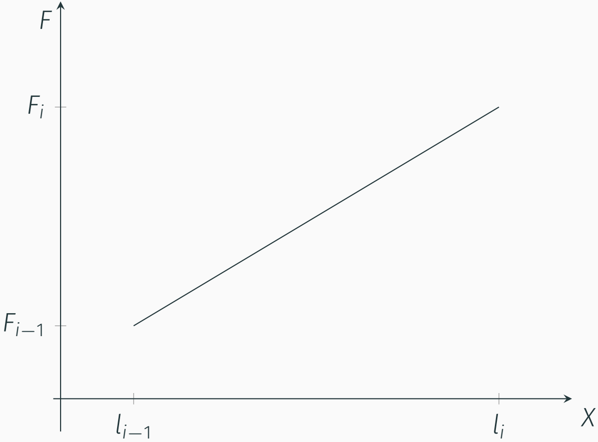
$$0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, \boxed{2}, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4$$

y la mediana es 2 hijos.

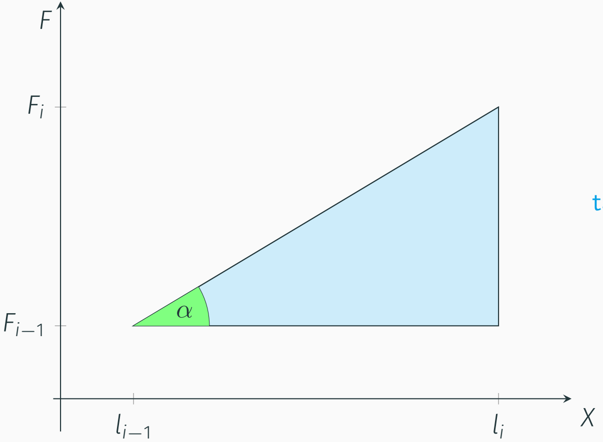
Si se trabaja con la tabla de frecuencias, se debe buscar el primer valor cuya frecuencia absoluta acumulada iguala o supera a 13, que es la posición que le corresponde a la mediana, o bien el primer valor cuya frecuencia relativa acumulada iguala o supera a 0.5:

x_i	n_i	f_i	N_i	F_i
0	2	0.08	2	0.08
1	6	0.24	8	0.32
$\boxed{2}$	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1

Cálculo de la mediana con datos agrupados

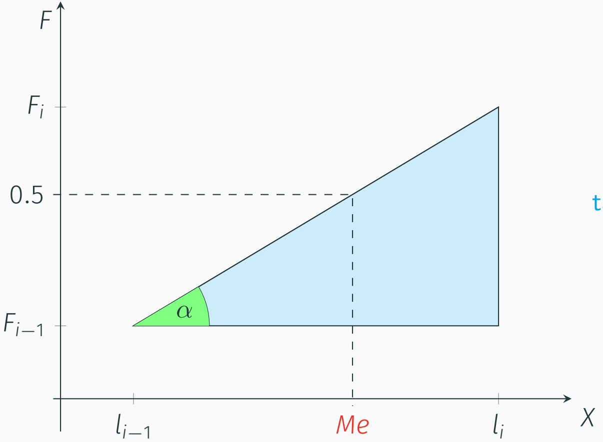


Cálculo de la mediana con datos agrupados



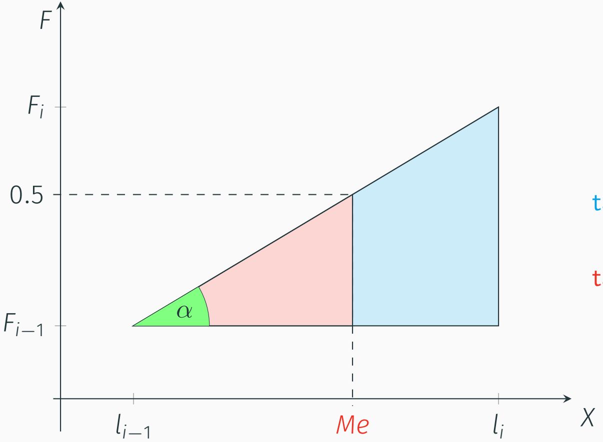
$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

Cálculo de la mediana con datos agrupados



$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

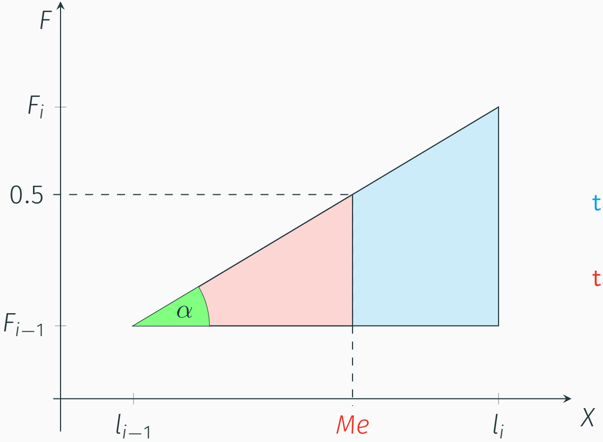
Cálculo de la mediana con datos agrupados



$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

$$\tan(\alpha) = \frac{0.5 - F_{i-1}}{Me - l_{i-1}}$$

Cálculo de la mediana con datos agrupados



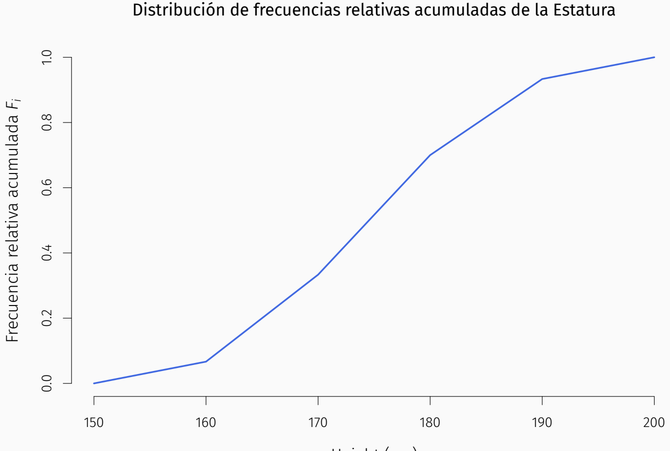
$$\tan(\alpha) = \frac{F_i - F_{i-1}}{l_i - l_{i-1}}$$

$$\tan(\alpha) = \frac{0.5 - F_{i-1}}{Me - l_{i-1}}$$

$$Me = l_i + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(l_i - l_{i-1}) = l_i + \frac{0.5 - F_{i-1}}{f_i} a_i$$

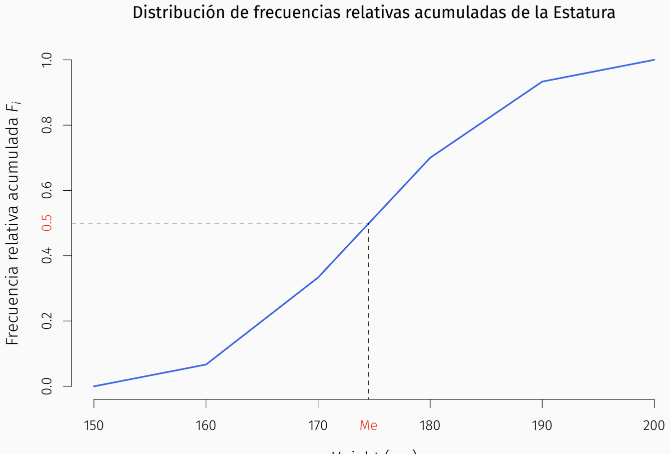
Cálculo de la mediana
Ejemplo con datos agrupados

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 $(170, 180]$.

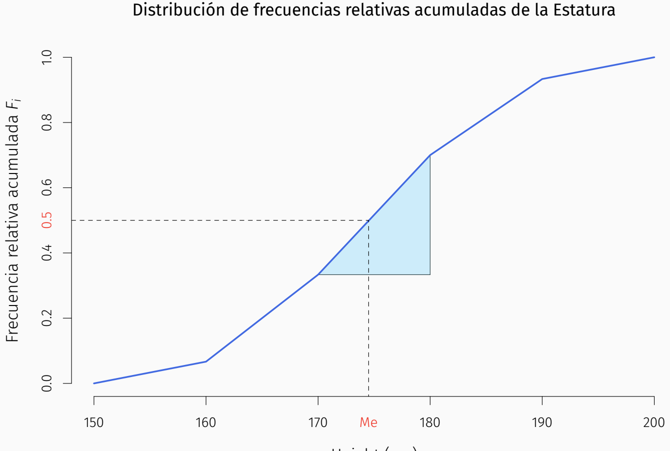


Cálculo de la mediana
Ejemplo con datos agrupados

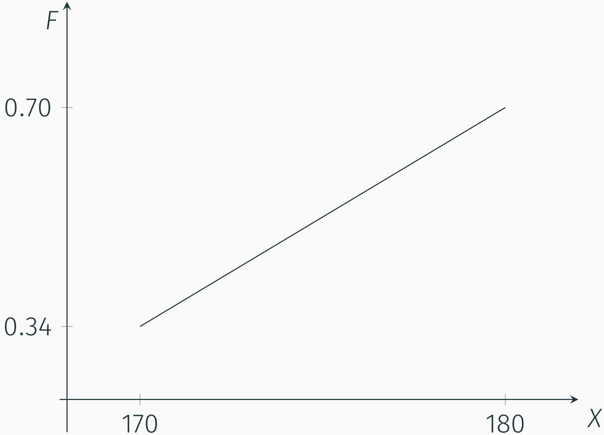
En el ejemplo de las estaturas $n/2 = 30/2 = 15$. Si miramos en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo $(170, 180]$.



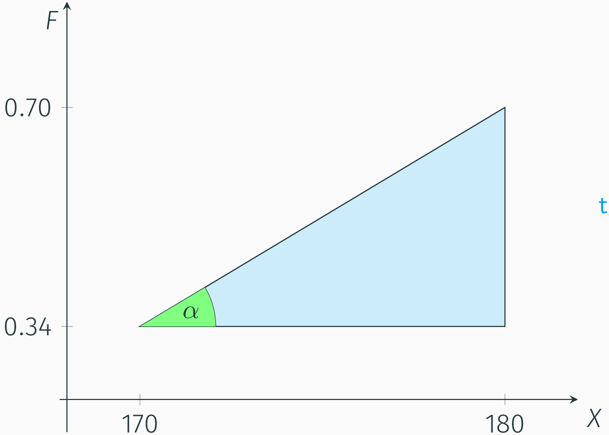
En el ejemplo de las estaturas $n/2 = 30/2 = 15$. Si miramos en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo $(170, 180]$.



Interpolación en el polígono de frecuencias absolutas acumuladas

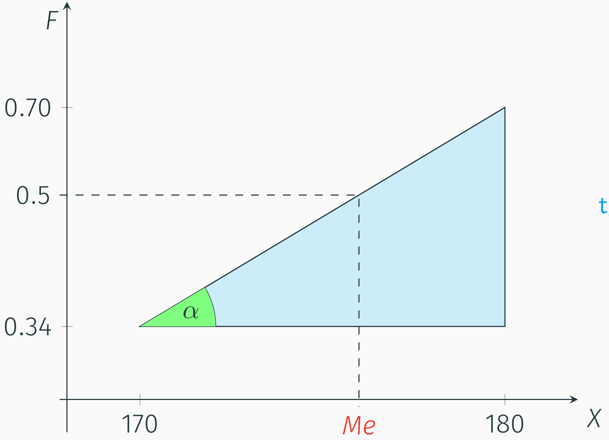


Interpolación en el polígono de frecuencias absolutas acumuladas



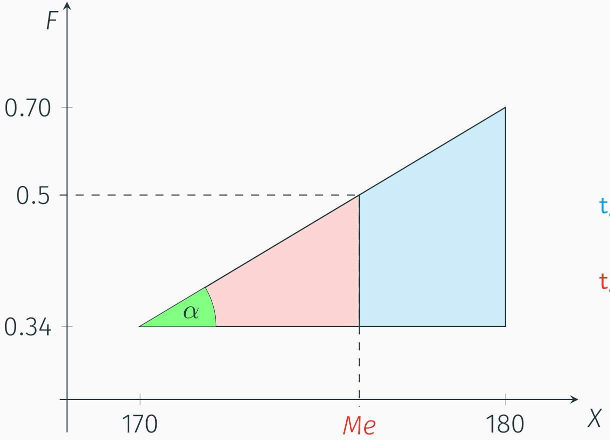
$$\mathrm{tg}(\alpha) = \frac{0.7 - 0.34}{180 - 170}$$

Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{0.7 - 0.34}{180 - 170}$$

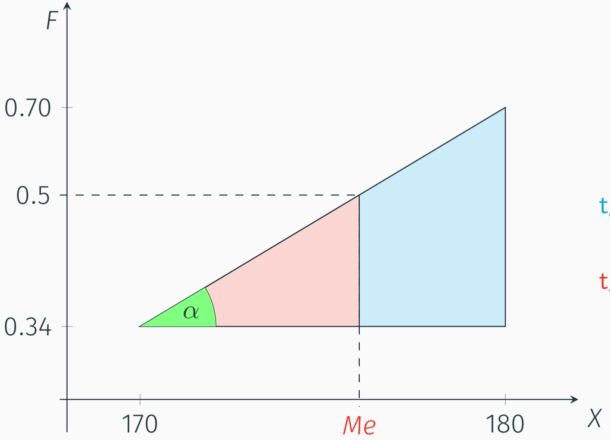
Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{0.7 - 0.34}{180 - 170}$$

$$\mathrm{tg}(\alpha) = \frac{0.5 - 0.34}{Me - 170}$$

Interpolación en el polígono de frecuencias absolutas acumuladas



$$\mathrm{tg}(\alpha) = \frac{0.7 - 0.34}{180 - 170}$$

$$\mathrm{tg}(\alpha) = \frac{0.5 - 0.34}{Me - 170}$$

$$Me = 170 + \frac{0.5 - 0.34}{0.7 - 0.34}(180 - 170) = 170 + \frac{0.16}{0.36}10 = 174.54 \text{ cm}$$

Moda

Definición (Moda muestral Mo)

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.

Moda

Definición (Moda muestral *Mo*)

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.

Definición (Moda muestral *Mo*)

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.

Moda

Definición (Moda muestral Mo)

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
$\boxed{2}$	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
$\boxed{(170,180]}$	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, ¡ojo! 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:

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

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

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.

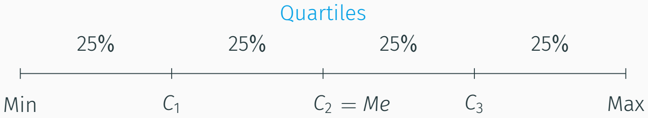
Los más utilizados son:

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) ,..., D_9 (90% acumulado).

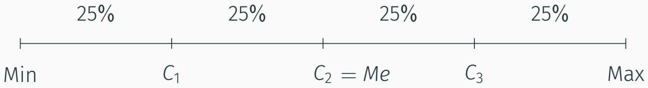
Percentiles: Dividen la distribución en 100 partes iguales.
Hay 99 percentiles: P_1 (1% acumulado),..., P_{99} (99% acumulado).

CUANTILES

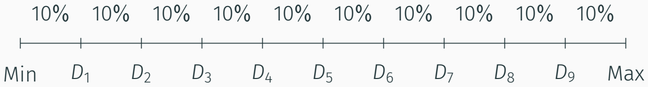


Cuantiles

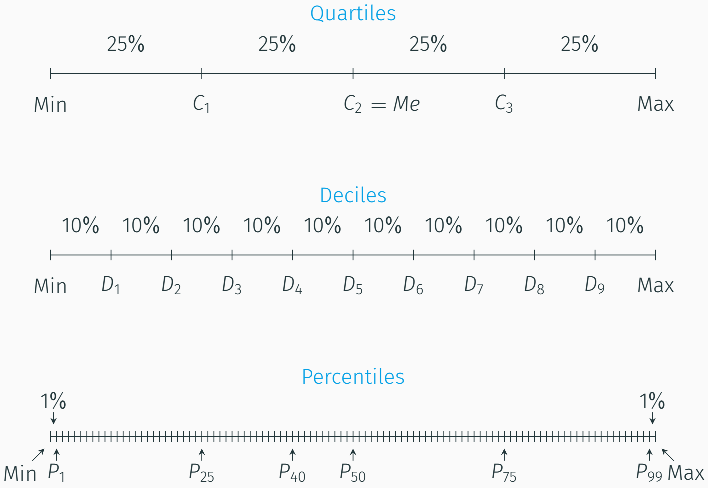
Quartiles



Deciles

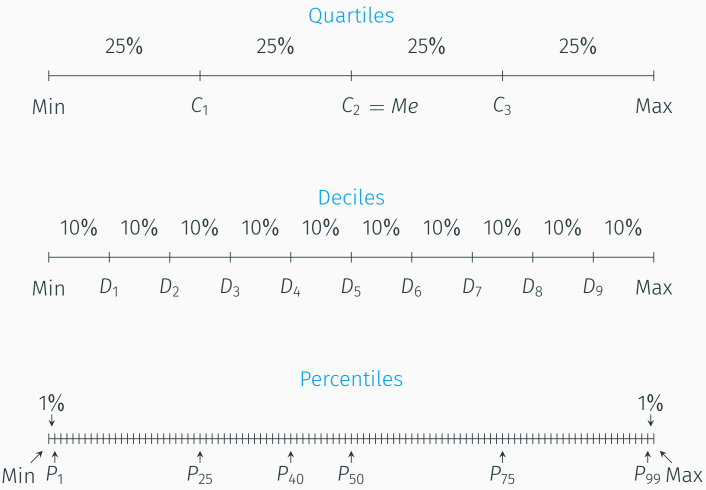


Cuantiles



Quartiles
25% 25% 25% 25%
Min C_1 $C_2 = Me$ C_3 Max
Deciles
10% 10% 10% 10% 10% 10% 10% 10% 10% 10%
Min D_1 D_2 D_3 D_4 D_5 D_6 D_7 D_8 D_9 Max
Percentiles
1% 1%
Min P_1 P_{25} P_{40} P_{50} P_{75} P_{99} Max

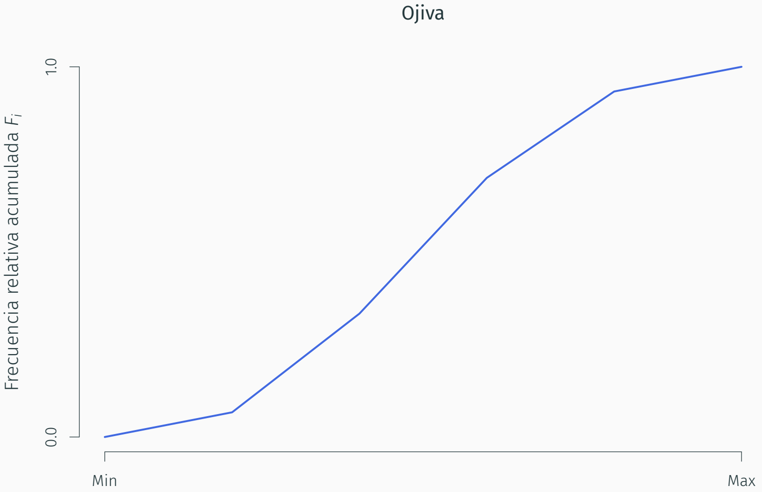
Cuantiles



Observese que hay una correspondencia entre los cuartiles, decilies y percentiles. Por ejemplo, el primer cuartil coincide con el percentil 25, y el cuarto decil coincide con el percentil 40.

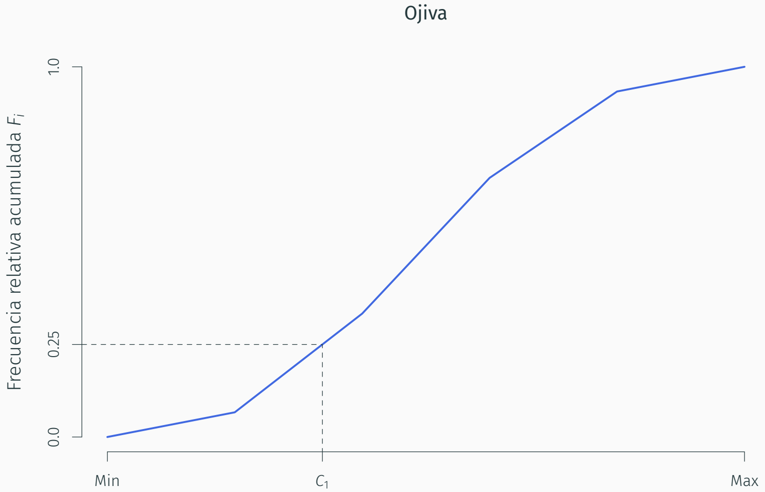
Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



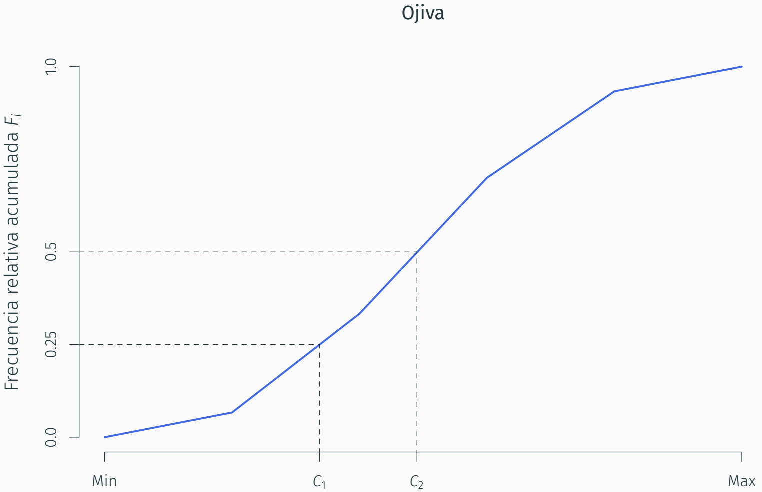
Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



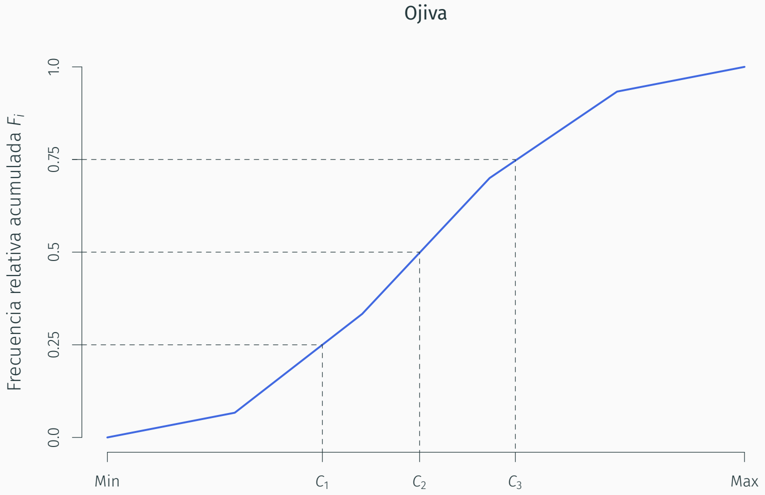
Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



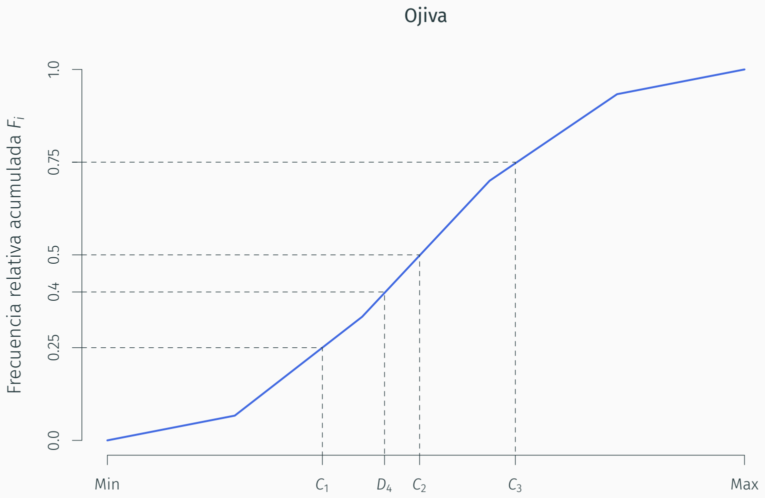
Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



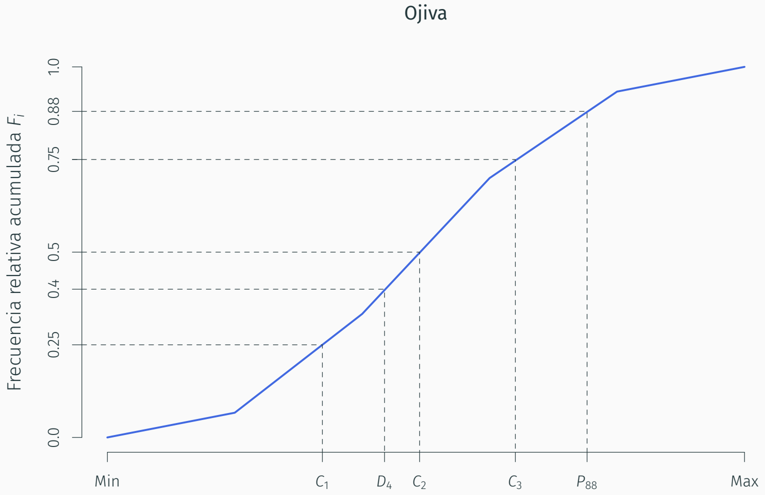
Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



Cálculo de los cuantiles

Los cuantiles se calculan de forma similar a la mediana. La única diferencia es la frecuencia relativa acumulada que le corresponde a cada cuantil.



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

$$\begin{aligned}
F_{C_1} &= 0.25 \Rightarrow C_1 = 1 \text{ hijos},\\
F_{C_2} &= 0.5 \Rightarrow C_2 = 2 \text{ hijos},\\
F_{C_3} &= 0.75 \Rightarrow C_3 = 2 \text{ hijos},\\
F_{D_3} &= 0.3 \Rightarrow D_3 = 1 \text{ hijos},\\
F_{P_{92}} &= 0.92 \Rightarrow P_{92} = 3 \text{ hijos}.
\end{aligned}$$

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.

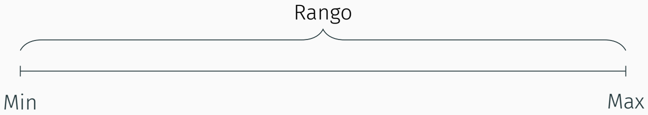
Recorrido

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.

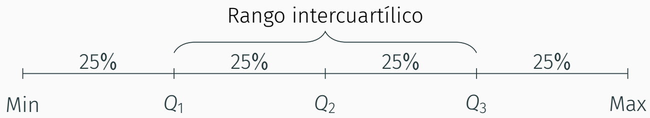


Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

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 mide la variación del 50% de los datos centrales.

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.

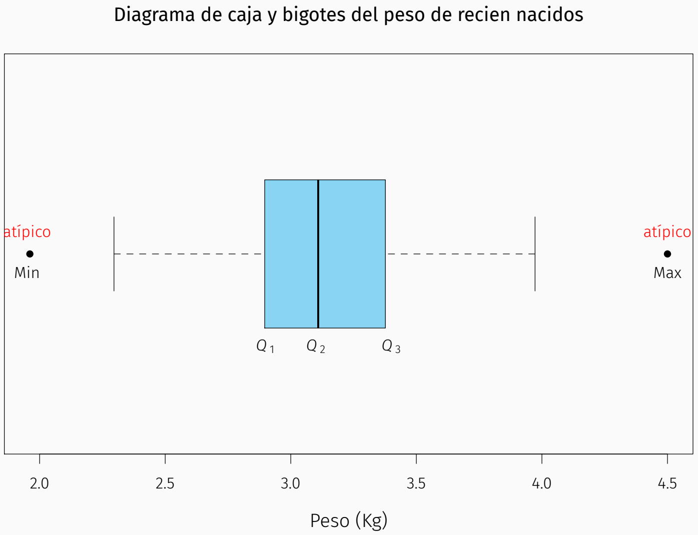


Diagrama de caja y bigotes del peso de recien nacidos
atípico
Min
atípico
Max
Q_1
Q_2
Q_3
2.0
2.5
3.0
3.5
4.0
4.5
Peso (Kg)

Construcción del diagrama de caja y bigotes

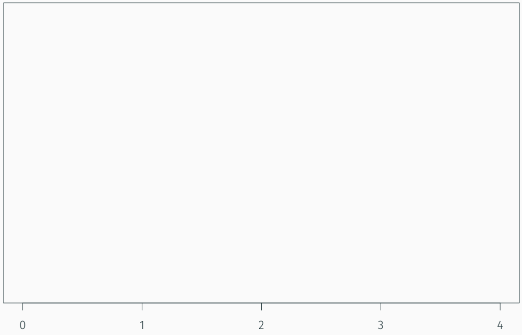
1. Calcular los cuartiles.
2. Dibujar una caja de manera que el extremo inferior caiga sobre el primer cuartil y el extremo superior sobre el tercer cuartil.
3. Dividir la caja con una línea que caiga sobre el segundo cuartil.
4. Para los bigotes inicialmente se determina la posición de los puntos denominados *vallas* v_1 y v_2 restando y sumando respectivamente a primer y tercer cuartil 1.5 veces el rango intercuartílico RI:

   $$v_1 = C_1 - 1.5RI$$
   $$v_2 = C_3 + 1.5RI$$

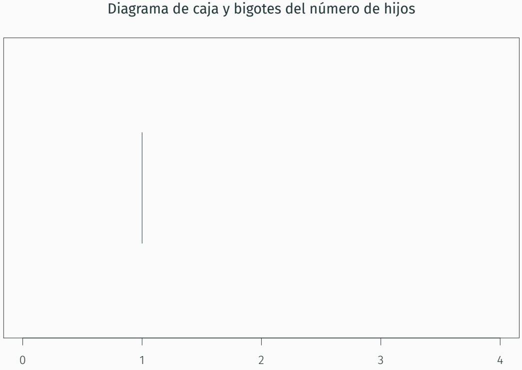
   A partir de las vallas se buscan los valores b_1, que es el mínimo valor de la muestra mayor o igual que v_1, y b_2, que es máximo valor de la muestra menor o igual que v_2. Para el bigote inferior se dibuja un segmento desde el borde inferior de la caja hasta b_1 y para el superior se dibuja un segmento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

Diagrama de caja y bigotes del número de hijos

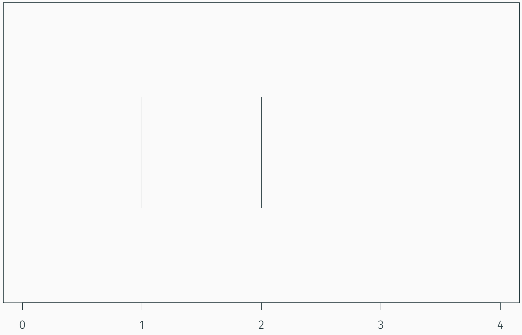


1. Calcular los cuartiles: $C_1 = 1$ hijos

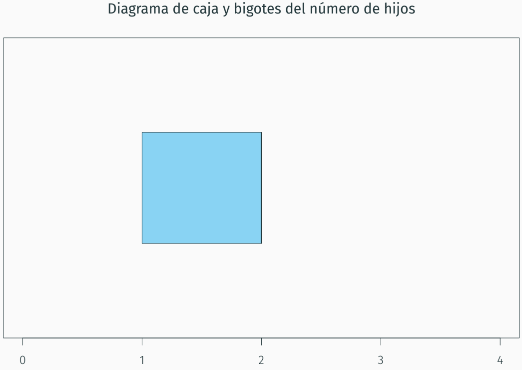


1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.

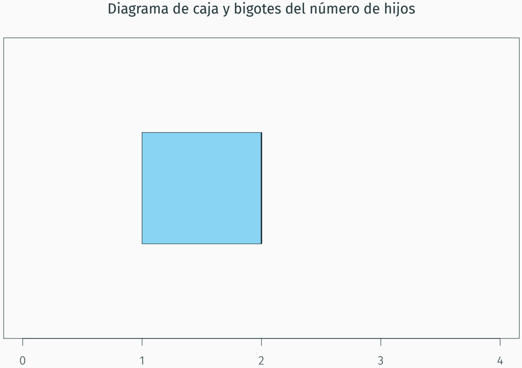
Diagrama de caja y bigotes del número de hijos



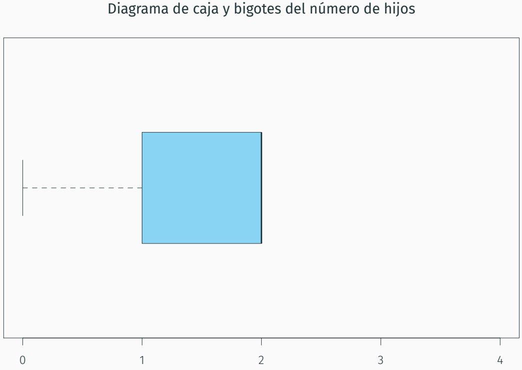
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.



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 $v_2 = 2 + 1.5 * 1 = 3.5$.



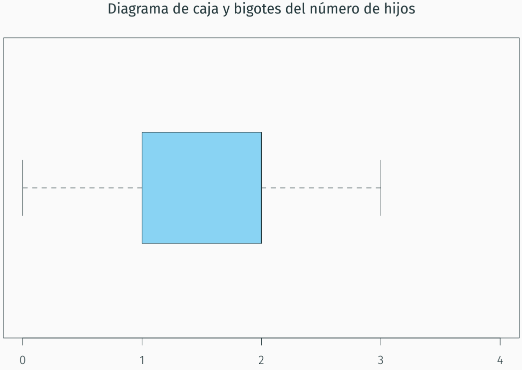
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1.5 * 1 = -0.5$ y $v_2 = 2 + 1.5 * 1 = 3.5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos



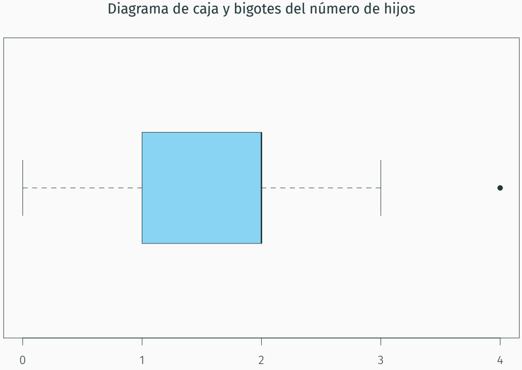
Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

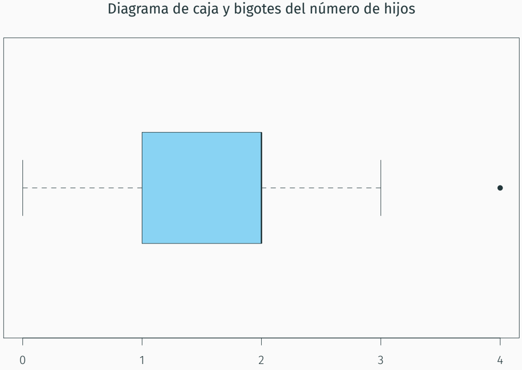
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1.5 * 1 = -0.5$ y $v_2 = 2 + 1.5 * 1 = 3.5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos y $b_1 = 3$ hijos.



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 $b_1 = 3$ hijos.
5. Dibujar los datos atípicos: 4 hijos.



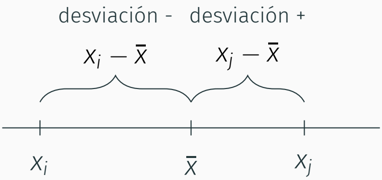
1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1.5 * 1 = -0.5$ y $v_2 = 2 + 1.5 * 1 = 3.5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos y $b_1 = 3$ hijos.
5. Dibujar los datos atípicos: 4 hijos.



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.

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.

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. Cuando la varianza o la desviación típica son pequeñas, los datos de la muestra están concentrados en torno a la media, y la media es una buena medida de representatividad. Por contra, cuando la varianza o la desviación típica son grandes, los datos de la muestra están alejados de la media, y la media ya no representa tan bien.

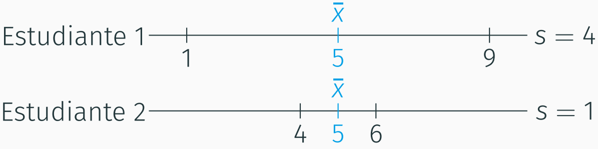
$$\begin{array}{lcl} \textit{Desviación típica pequeña} & \Rightarrow & \textit{Media representativa} \\ \textit{Desviación típica grande} & \Rightarrow & \textit{Media no representativa} \end{array}$$

Estudiante 1  $s = 4$

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. Cuando la varianza o la desviación típica son pequeñas, los datos de la muestra están concentrados en torno a la media, y la media es una buena medida de representatividad. Por contra, cuando la varianza o la desviación típica son grandes, los datos de la muestra están alejados de la media, y la media ya no representa tan bien.

$$\begin{array}{lcl} \textit{Desviación típica pequeña} & \Rightarrow & \textit{Media representativa} \\ \textit{Desviación típica grande} & \Rightarrow & \textit{Media no representativa} \end{array}$$



¿En qué caso es más representativa la media?

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
$\sum$	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.

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
$\sum$		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 $s = \sqrt{102.06} = 10.1$ 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.

Tanto la varianza como la desviación típica tienen unidades y eso dificulta a veces su interpretación, especialmente cuando se compara la dispersión de variables con diferentes unidades.

Por este motivo, es también común utilizar la siguiente medida de dispersión que no tiene unidades.

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.

En el caso del número de hijos, como $\bar{x} = 1.76$ hijos y $s = 0.8616$ hijos, se tiene que el coefiente 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

$$cv = \frac{s}{|\bar{x}|} = \frac{10.1}{|174.67|} = 0.06.$$

Esto significa que 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.

Son medidas que describen la forma de la distribución.

Los aspectos más relevantes son:

Simetría: Mide la simetría de la distribución de frecuencias en torno a la media.
El estadístico más utilizado es el *Coeficiente de Asimetría de Fisher*.

Apuntamiento: Mide el apuntamiento o el grado de concentración de valores en torno a la media de la distribución de frecuencias.
El estadístico más utilizado es el *Coeficiente de Apuntamiento o Curtosis*.

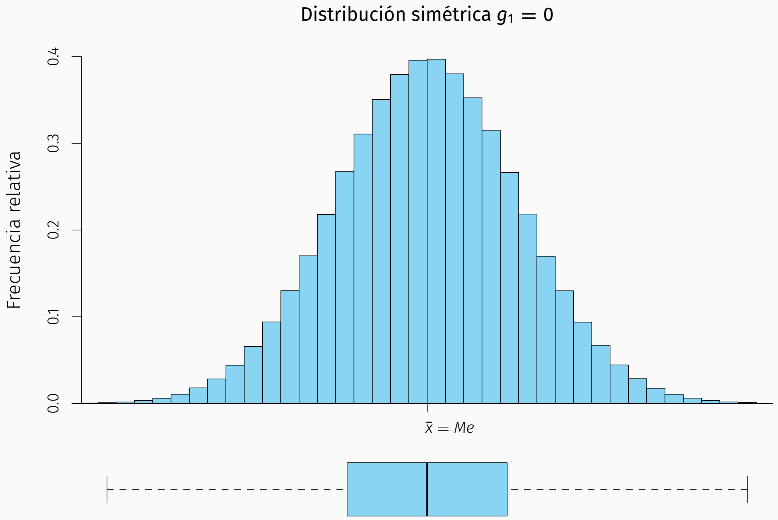
Definición (Coeficiente de asimetría muestral g_1)

El *coeficiente de asimetría muestral* de una variable X es 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.

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

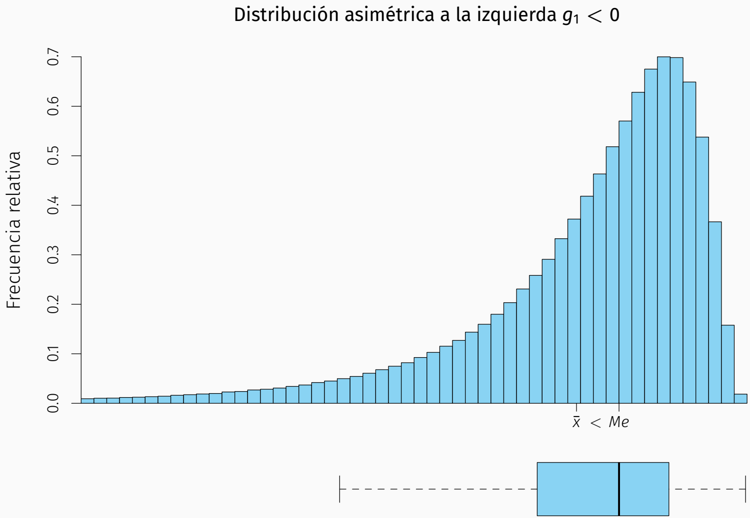
Mide el grado de simetría de los valores de la muestra con respecto a la media muestra, es decir, cuantos valores de la muestra están por encima o por debajo de la media y cómo de alejados de esta.

- $g_1 = 0$ indica que hay el mismo número de valores por encima y por debajo de la media e igualmente alejados de ella (simétrica).
- $g_1 < 0$ indica que la mayoría de los valores son mayores que la media, pero los valores menores están más alejados de ella (asimétrica a la izquierda).
- $g_1 > 0$ indica que la mayoría de los valores son menores que la media,


Distribución simétrica $g_1 = 0$
Frecuencia relativa
0.0
0.1
0.2
0.3
0.4
$\bar{x} = Me$

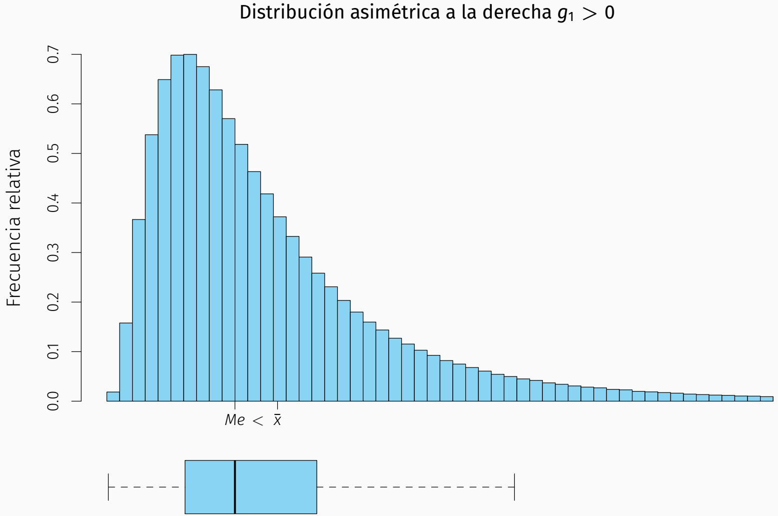
Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la izquierda



Coeficiente de asimetría

Ejemplo de distribución asimétrica hacia la derecha



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
$\sum$		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.

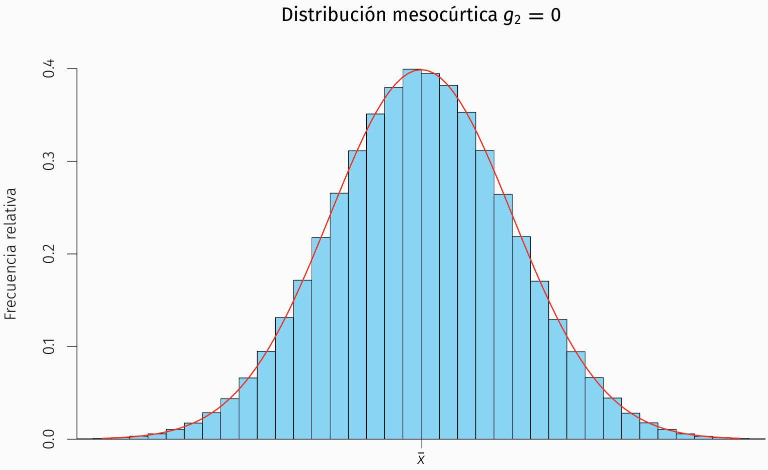
Definición (Coeficiente de apuntamiento muestral g_2)

El *coeficiente de apuntamiento muestral* de una variable X es 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.

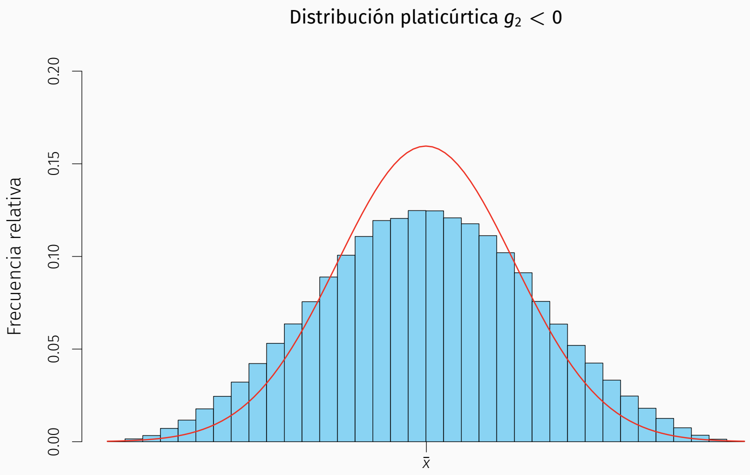
$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i/n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

El coeficiente de apuntamiento mide la concentración de valores en torno a la media y la longitud de las colas de la distribución. Se toma como referencia la distribución normal

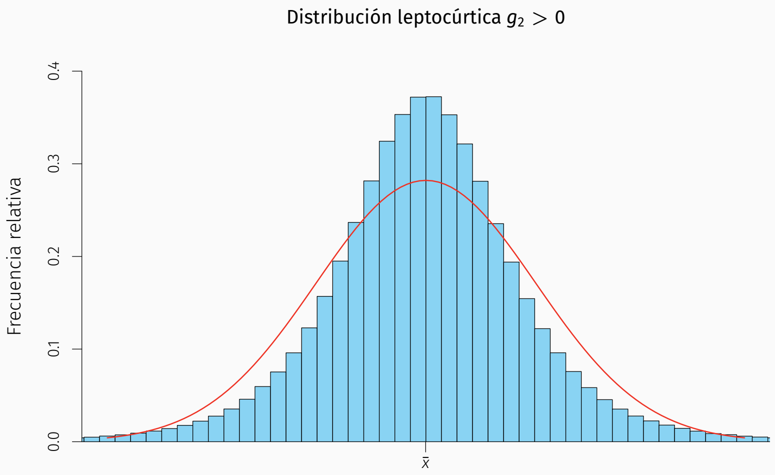
- $g_2 = 0$ indica que la distribución tienen un apuntamiento normal (*mesocúrtica*).
- $g_2 < 0$ indica que la distribución tiene menos apuntamiento de lo normal (*platicúrtica*).
- $g_2 > 0$ indica que la distribución tiene más apuntamiento de lo



Distribución mesocúrtica $g_2 = 0$
Frecuencia relativa
0.0
0.1
0.2
0.3
0.4
$\bar{x}$



Distribución platicúrtica $g_2 < 0$
Frecuencia relativa
0.00
0.05
0.10
0.15
0.20
$\bar{x}$



Distribución leptocúrtica $g_2 > 0$
Frecuencia relativa
0.0
0.1
0.2
0.3
0.4
$\bar{x}$

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
$\sum$		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.

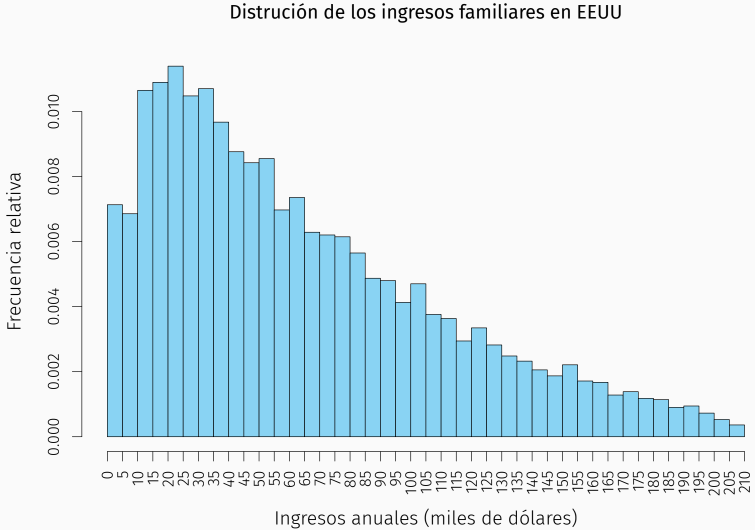
Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

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.

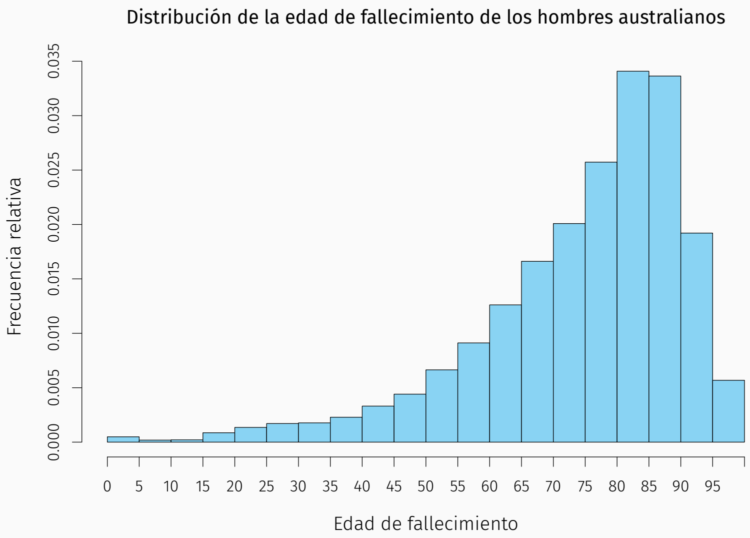
Distribución asimétrica a la derecha no normal

Ingresos por familia



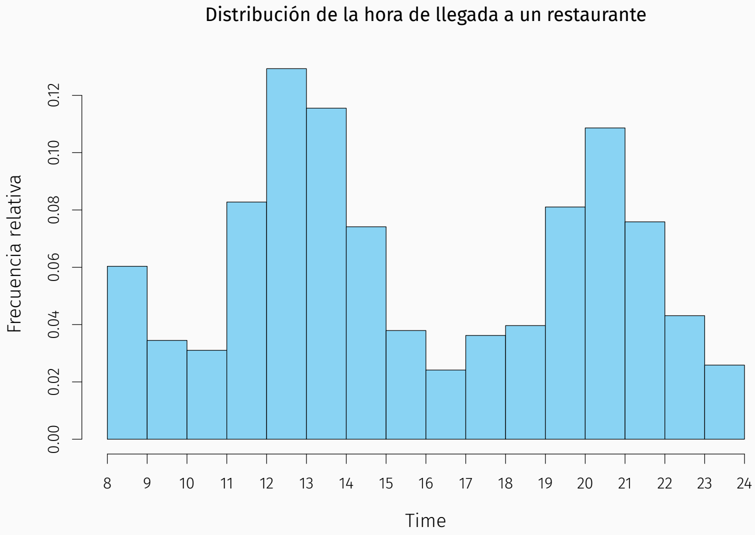
Distribución asimétrica a la izquierda no normal

Edad de fallecimiento



Distribución bimodal no normal

Hora de llegada de los clientes de un restaurante



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

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

$$1.75\text{m}, 1.65\text{m}, 1.80\text{m},$$

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

$$175\text{cm}, 165\text{cm}, 180\text{cm},$$

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

$$10\text{cm}, 0\text{cm}, 15\text{cm},$$

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:

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:

$$\bar{y} = a + b\bar{x},$$
$$s_y = |b|s_x$$

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.

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

Para cada valor x_i de la muestra, la *puntuación típica* es el valor que resulta de aplicarle la transformación de tipicación

$$z_i = \frac{x_i - \bar{x}}{s_x}.$$

La puntuación típica es el número de desviaciones típicas que un valor está por encima o por debajo de la media, y es útil para evitar la dependencia de una variable respecto de las unidades de medida empleadas.

Los valores tipificados se conocen como puntuaciones típicas y miden el

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	$\bar{y} = 5$	$s_y = 3.16$

¿Ha tenido el mismo rendimiento el cuarto alumno en la asignatura X que el tercero en la asignatura Y?

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

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
$\sum$	3	14	12	13	8

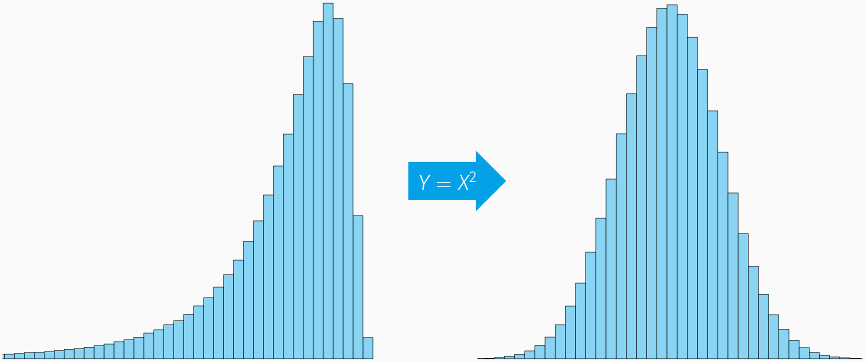
El mejor alumno sería el segundo.

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
$\sum$	-2.76	1.26	0.45	1.5	-0.45

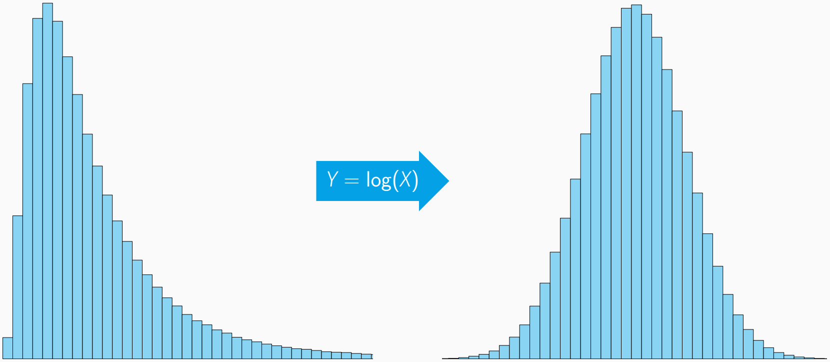
Transformaciones no lineales

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.



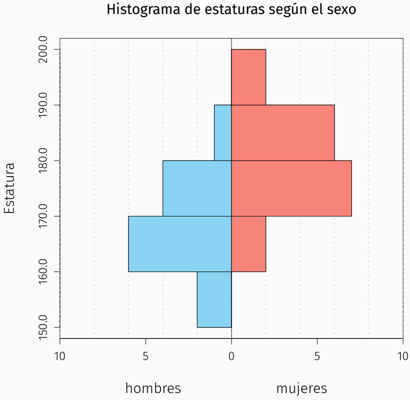
Variables clasificadoras o factores

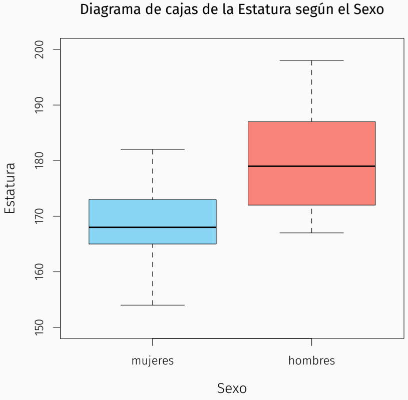
En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos correspondientes a las categorías de otra variable conocida como **variable clasificadora** o **factor**.

Dividiendo la muestra de estaturas según el sexo se obtienen dos submuestras:

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.

Comparación de distribuciones según los niveles de un factor





Regresión y Correlación

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, en un *estudio de dependencia* se considera una **variable dependiente** Y que se supone relacionada con otras variables $X_1, \ldots, X_n$ 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*.

Frecuencias conjuntas

Al estudiar la dependencia simple entre dos variables X e Y, no se pueden estudiar sus distribuciones por separado, sino que hay que estudiar la distribución conjunta de la **variable bidimensional** (X, Y), cuyos valores son los pares (x_i, y_j) donde el primer elemento es un valor X y el segundo uno de Y.

Definición (Frecuencias muestrales conjuntas)

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 veces que el par (x_i, y_j) aparece en la muestra.
- Frecuencia relativa f_{ij}: Es la proporcion de veces que el par (x_i, y_j) aparece en la muestra.

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

¡Atención! Para las variables bidimensionales no tienen sentido las

Distribución de frecuencias bidimensional

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

La estatura (en cm) y el peso (en Kg) de una muestra de 30 estudiantes es:

(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).

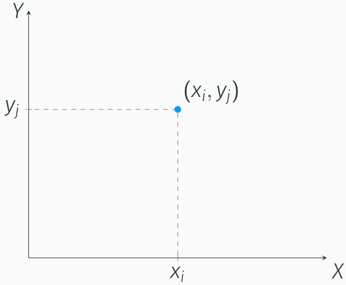
La tabla de frecuencias bidimensional es

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

La distribución de frecuencias conjunta de una variable bidimensional puede representarse gráficamente mediante un **diagrama de dispersión**, donde los datos se representan como una colección de puntos en un plano cartesiano.

Habitualmente la variable independiente se representa en el eje X y la variable dependiente en el eje Y. Por cada par de valores (x_i, y_j) en la muestra se dibuja un punto en el plano con esas coordenadas.



El resultado es un conjunto de puntos que se conoce como *nube de puntos*

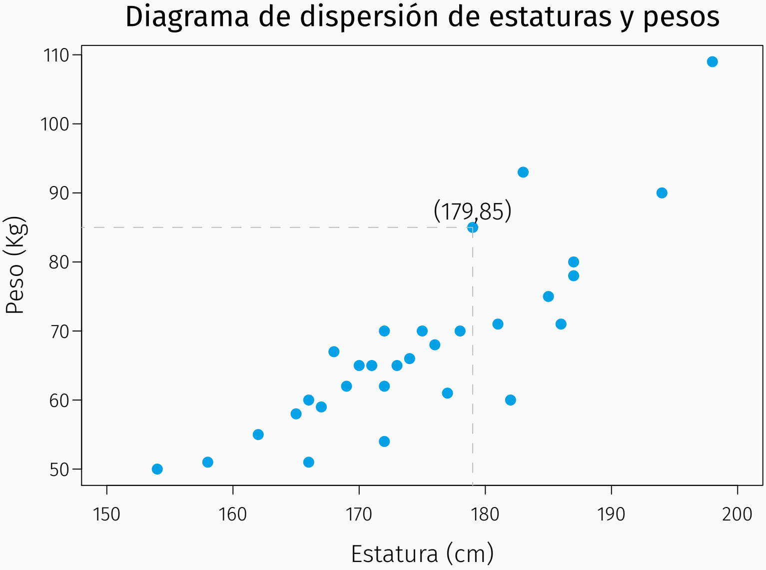
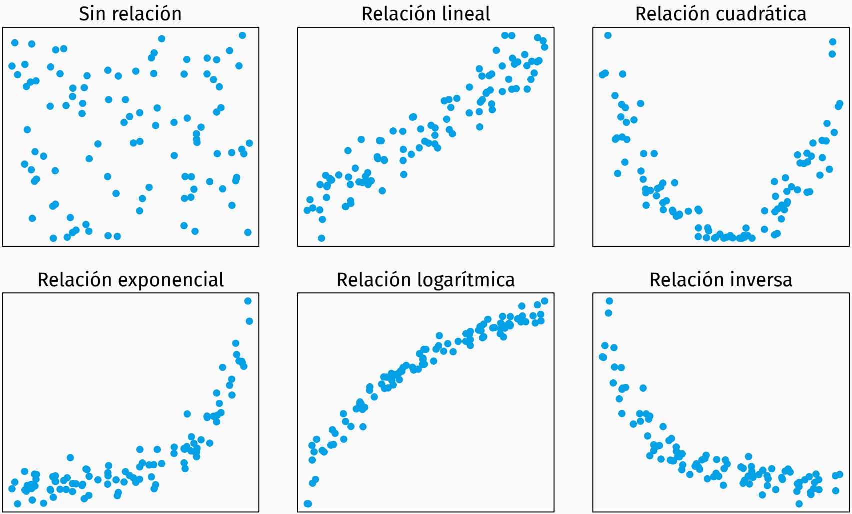


Diagrama de dispersión de estaturas y pesos
110
100
90
80
70
60
50
Peso (Kg)
(179,85)
150
160
170
180
190
200
Estatura (cm)

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_{x_1}
$\vdots$	$\vdots$	$\vdots$	$+$	$\vdots$	$\vdots$	$\vdots$
x_i	n_{i1}	$+$	n_{ij}	$+$	n_{iq}	n_{x_i}
$\vdots$	$\vdots$	$\vdots$	$+$	$\vdots$	$\vdots$	$\vdots$
x_p	n_{p1}	$\cdots$	n_{pj}	$\cdots$	n_{pq}	n_{x_p}
n_y	n_{y_1}	$\cdots$	n_{y_j}	$\cdots$	n_{y_q}	n

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

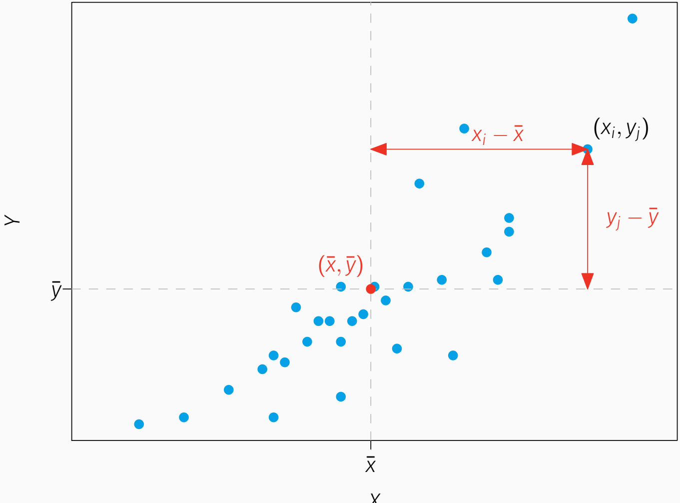
X/Y	$[50,60)$	$[60,70)$	$[70,80)$	$[80,90)$	$[90,100)$	$[100,110)$	n_x
$(150,160]$	2	0	0	0	0	0	2
$(160,170]$	4	4	0	0	0	0	8
$(170,180]$	1	6	3	1	0	0	11
$(180,190]$	0	1	4	1	1	0	7
$(190,200]$	0	0	0	0	1	1	2
n_y	7	11	7	2	2	1	30

y los estadísticos correspondientes son

$$
\begin{array}{lll}
\bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 & s_x = 10.1 \text{ cm} \\
\bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 & s_y = 12.82 \text{ Kg}
\end{array}
$$

Desviaciones respecto de las medias

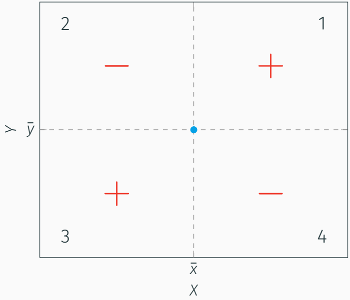
Para analizar la relación entre dos variables cuantitativas es importante hacer un estudio conjunto de las desviaciones respecto de la media de cada variable.



Signo de las desviaciones respecto de las medias

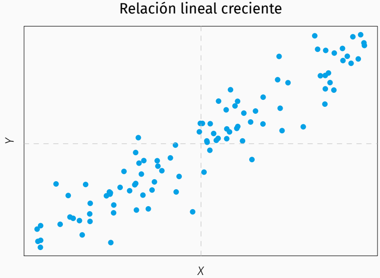
Si dividimos la nube de puntos del diagrama de dispersión en 4 cuadrantes centrados en el punto de medias $(\bar{x}, \bar{y})$, el signo de las desviaciones será:

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



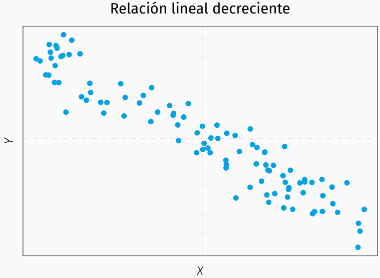
Estudio de las desviaciones respecto de las medias

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

Usando el producto de las desviaciones respecto de las medias surge el siguiente estadístico.

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.

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

$$\bar{x} = 174.67 \text{ cm} \qquad \bar{y} = 69.67 \text{ Kg}$$

la covarianza vale

$$\begin{aligned} s_{xy} &= \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \cdots + 195 \cdot 105 \cdot 1}{30} - 174.67 \cdot 69.67 \\ &= \frac{368200}{30} - 12169.26 = 104.07 \text{ cm} \cdot \text{ Kg}, \end{aligned}$$

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 dos variables, sino explicarla mediante alguna función matemática

$$y = f(x)$$

que permita predecir la variable dependiente para cada valor de la independiente.

La regresion es la parte de la Estadística encargada de construir esta función, que se conoce como función de regresión o odelo 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	$y = a_0 + a_1x + \cdots + a_nx^n$
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, es decir, la función que mejor predice la variable dependiente.

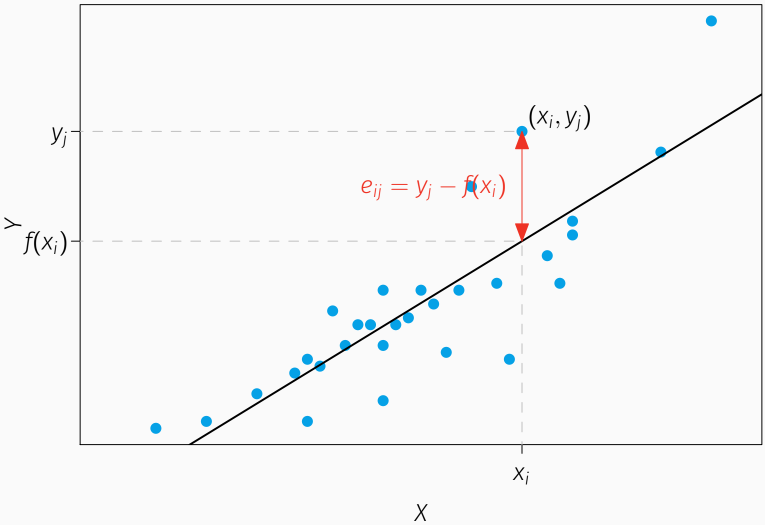
El objetivo es encontrar la función de regresión que haga mínimas las distancias entre los valores de la variable dependiente observados en la muestra, y los predichos por la función de regresión. Estas distancias se conocen como *residuos* o *errores predictivos*.

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:

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

Residuos o errores predictivos en Y



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.

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:

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

Tras resolver el sistema se obtienen los valores

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

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

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.

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

$$\begin{array}{lll} \bar{x} = 174.67 \text{ cm} & s_x^2 = 102.06 \text{ cm}^2 & s_x = 10.1 \text{ cm} \\ \bar{y} = 69.67 \text{ Kg} & s_y^2 = 164.42 \text{ Kg}^2 & s_y = 12.82 \text{ Kg} \\ & s_{xy} = 104.07 \text{ cm}\cdot \text{ Kg} & \end{array}$$

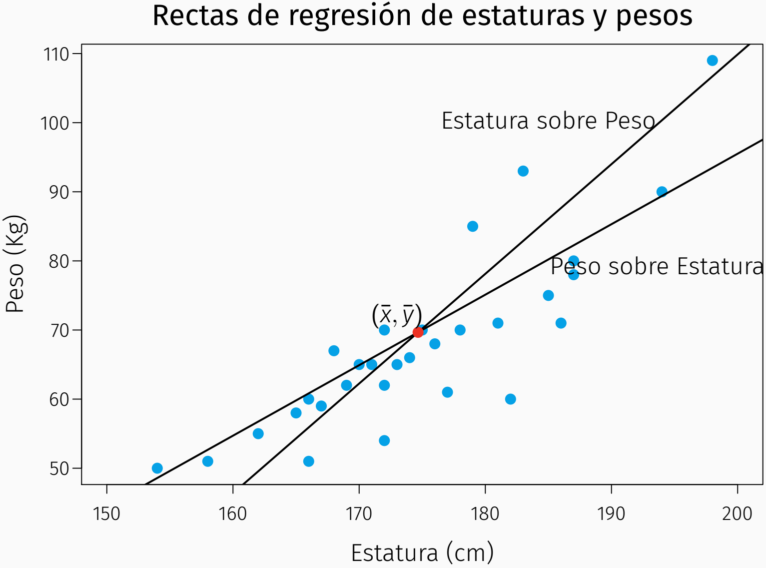
Entonces, la recta de regresión del peso sobre la estatura es

$$y = \bar{y} + \frac{s_{xy}}{s_x^2}(x - \bar{x}) = 69.67 + \frac{104.07}{102.06}(x - 174.67) = -108.49 + 1.02x.$$

De igual modo, si tomamos la estatura como variable dependiente, la recta de regresión de la estatura sobre el peso es

$$x = \bar{x} + \frac{s_{xy}}{s_y^2}(y - \bar{y}) = 174.67 + \frac{104.07}{164.42}(y - 69.67) = 130.78 + 0.63y.$$

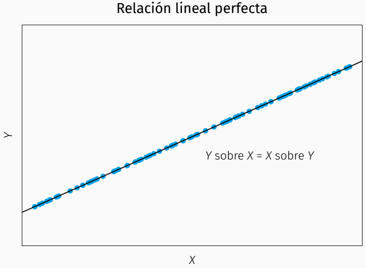
¡Obsérvese que ambas rectas de regresión son diferentes!



Rectas de regresión de estaturas y pesos
Estatura sobre Peso
Peso sobre Estatura
$(\bar{x}, \bar{y})$
Peso (Kg)
Estatura (cm)
110
100
90
80
70
60
50
150
160
170
180
190
200

Habitualmente, las rectas de regresión Y sobre X y de X sobre Y no coinciden, pero siempre se cortan en el punto de medias $(\bar{x}, \bar{y})$.

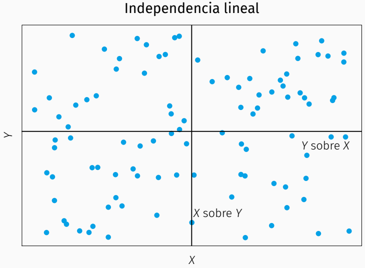
Si entre las variables la relación lineal es perfecta, entonces ambas rectas coinciden ya que sus residuos son nulos.



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

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

y se cortan perpendicularmente



El parámetro más importante de una recta de regresión es su pendiente.

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.

Refleja el crecimiento de la variable dependiente en relación a la independiente según la recta de regresión. En concreto da el número de unidades que aumenta o disminuye la variable dependiente por cada unidad que aumenta la variable independiente.

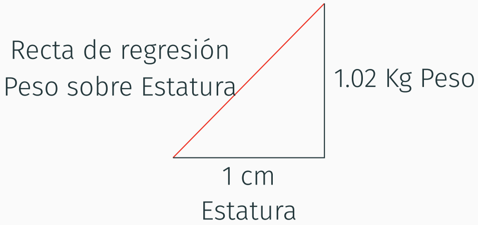
En el ejemplo de las estaturas y los pesos, la recta de regresión del peso sobre la estatura era

$$y = -108.49 + 1.02x,$$

de manera que el coeficiente de regresión del peso sobre la estatura es

$$b_{yx} = 1.02 \text{Kg/cm}.$$

Esto significa que, según la recta de regresión del peso sobre la estatura, por cada cm más de estatura, la persona pesará 1.02 Kg más.



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**.

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)

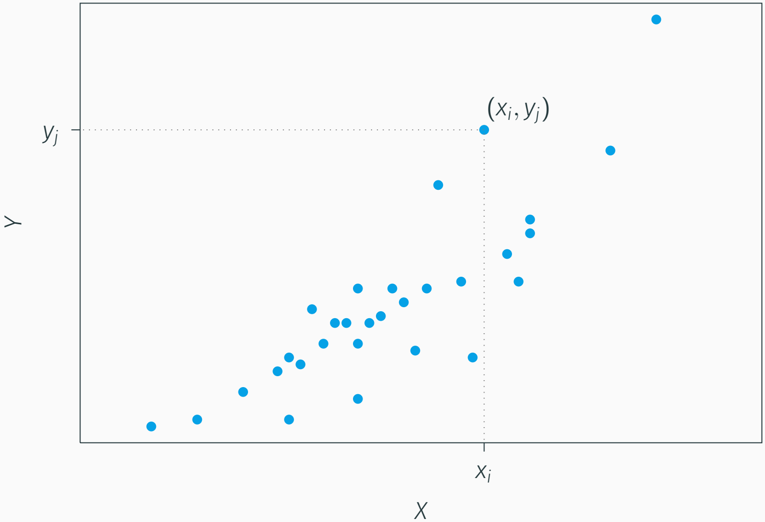
Dado un modelo de regresión simple $y = f(x)$ de una variable bidimensional (X, Y), su *varianza residual muestral* es el promedio de los cuadrados de los residuos para los valores de la muestra,

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

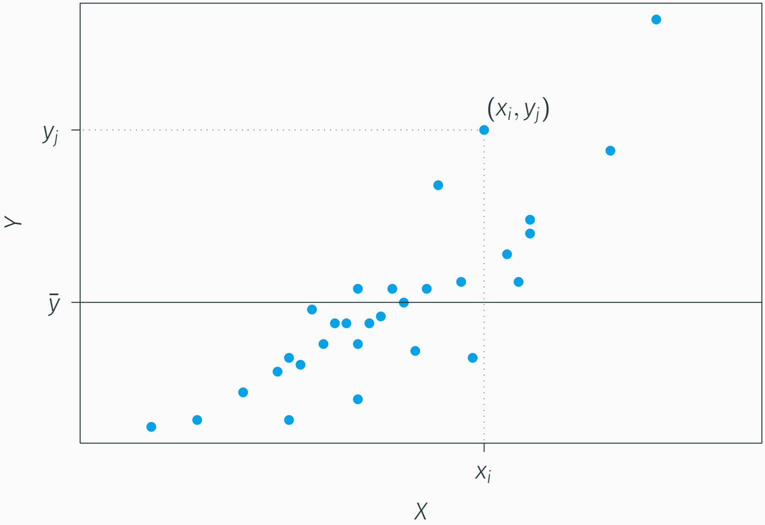
Cuanto más alejados estén los puntos de la curva de regresión, mayor será la varianza residual y menor la dependencia.

Cuando la relación lineal es perfecta los residuos se anulan y la varianza residual vale cero. Por contra, cuando no existe relación, los residuos coinciden con las desviaciones de la media, y la varianza residual es igual a la varianza de la variable dependiente.

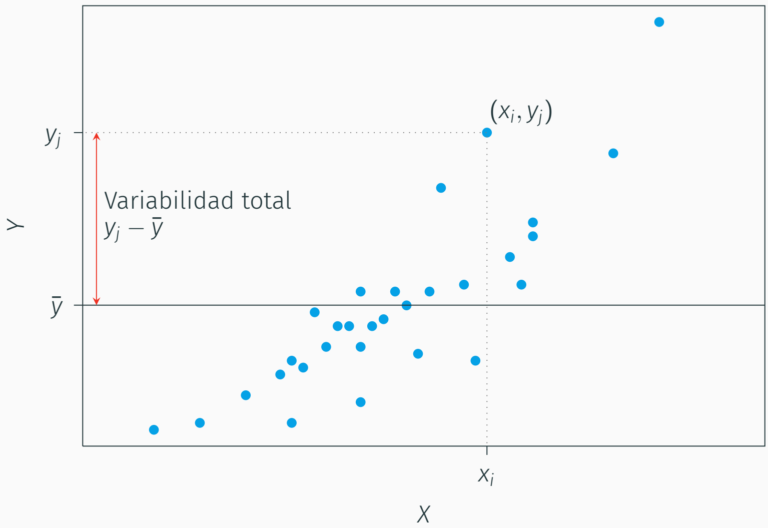
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



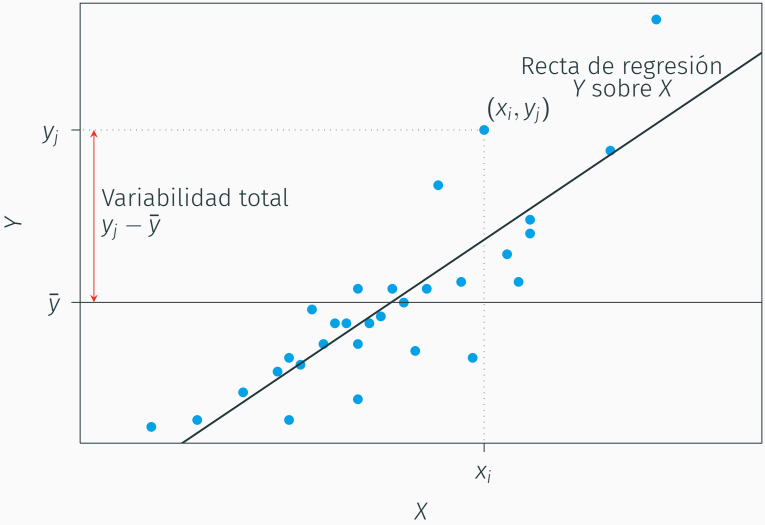
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



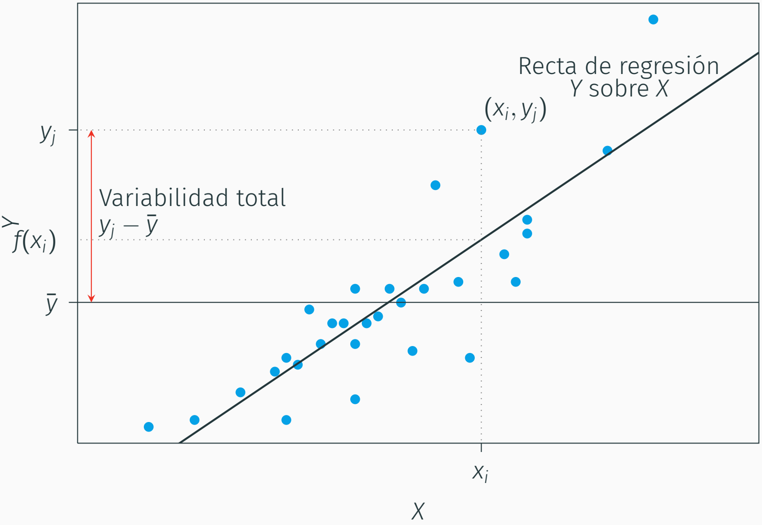
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



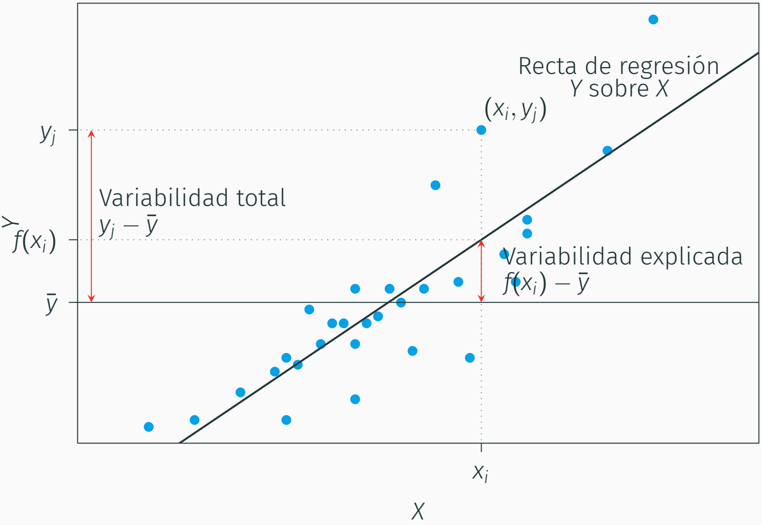
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



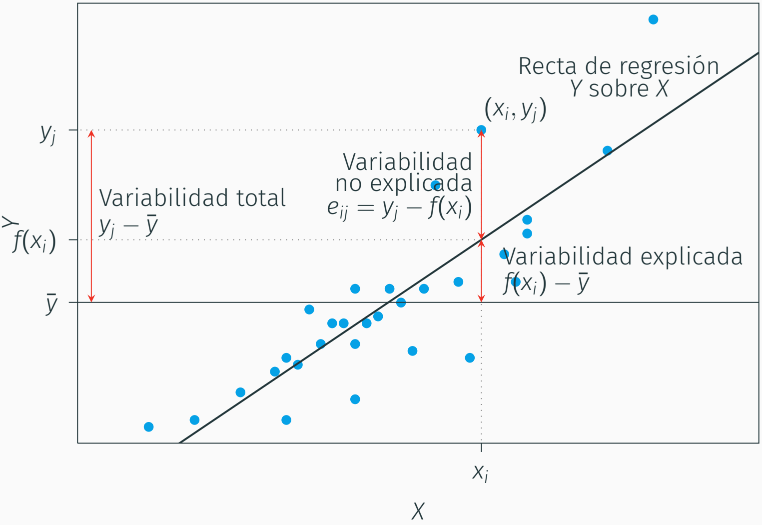
Descomposición de la variabilidad total: Variabilidad explicada y no explicada



Descomposición de la variabilidad total: Variabilidad explicada y no explicada



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

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

Como la varianza residual puede tomar valores entre 0 y s_y^2, se tiene que

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

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.

Coeficiente 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

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

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.

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, r tomará valores entre -1 y 1:

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

El coeficiente de correlación lineal no sólo mide mide el grado de dependencia lineal sino también su dirección (creciente o decreciente):

- 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.

En el ejemplo de las estaturas y los pesos se tenía

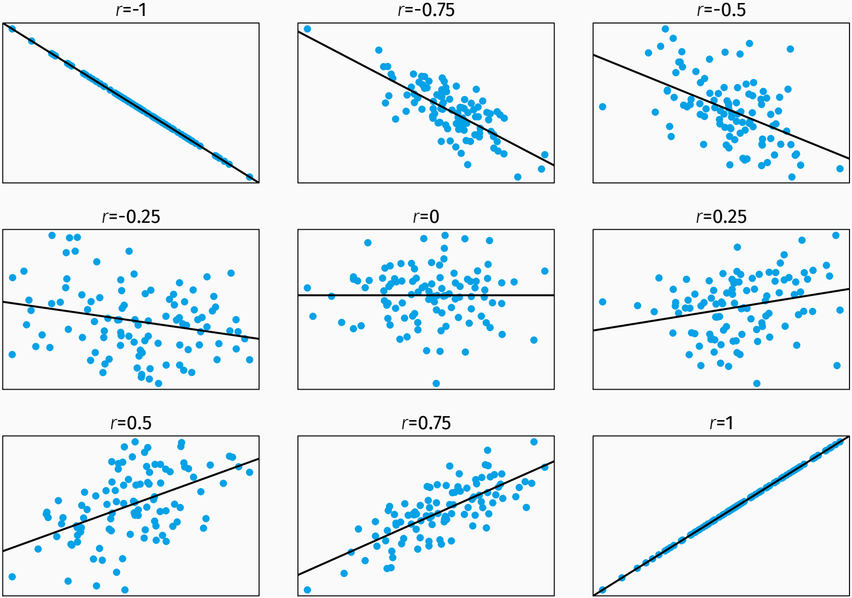
$$\begin{aligned} \bar{x} &= 174.67 \text{ cm} & s_x^2 &= 102.06 \text{ cm}^2 \\ \bar{y} &= 69.67 \text{ Kg} & s_y^2 &= 164.42 \text{ Kg}^2 \\ s_{xy} &= 104.07 \text{ cm}\cdot\text{Kg} \end{aligned}$$

De modo que el coeficiente de correlación lineal vale

$$r = \frac{s_{xy}}{s_x s_y} = \frac{104.07 \text{ cm}\cdot\text{Kg}}{10.1 \text{ cm} \cdot 12.82 \text{ Kg}} = +0.8.$$

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

Distintos grados de correlación



Fiabilidad de las predicciones de un modelo de regresión

Aunque el coeficiente de determinación o el de correlación determinan la bondad de ajuste de un modelo de regresión, existen otros factores que influyen en la fiabilidad de las predicciones de un modelo de regresión:

- 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.

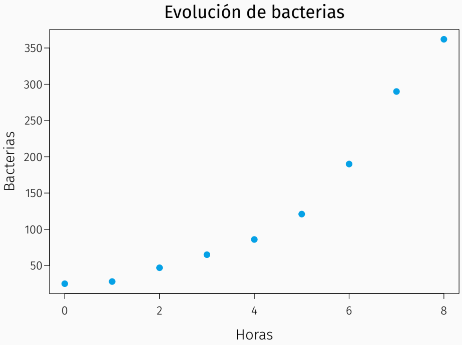
Transformación de modelos de regresión no lineales

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

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

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

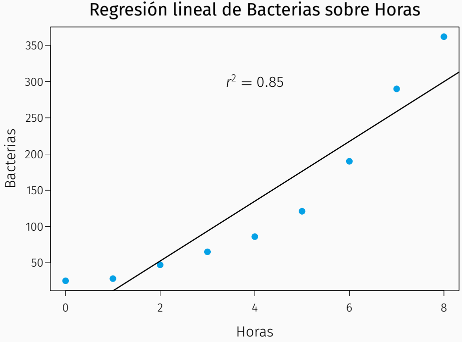
El diagrama de dispersión asociado es



Relación 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?

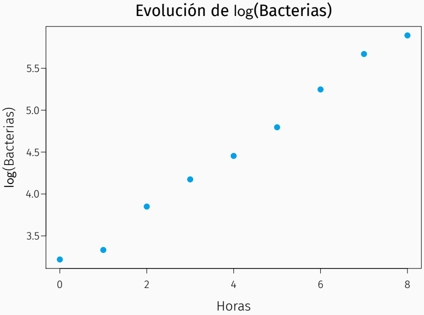
Ajuste de un modelo de regresión exponencial

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

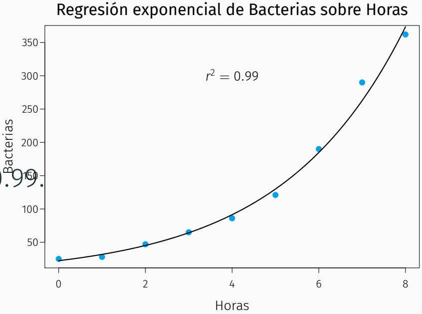


Ahora sólo queda calcular la recta de regresión del logaritmo de Bacterias sobre Horas

$$\text{Log Bacterias} = 3.107 + 0.352\,\text{Horas}.$$

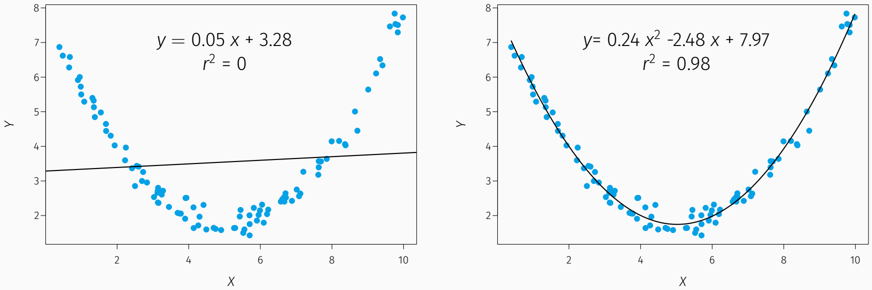
Y deshaciendo el cambio de variable, se obtiene el modelo exponencial

$$\text{Bacterias} = e^{3.107+0.352\,\text{Horas}}, \text{ con } r^2 = 0.99.$$



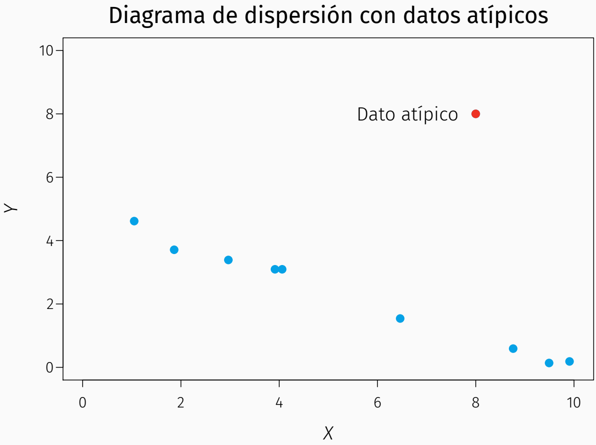
Como se puede apreciar, el modelo exponencial se ajusta mucho mejor que el modelo lineal.

Es importante señalar que cada modelo de regresión tiene su propio coeficiente de determinación. Así, un coeficiente de determinación cercano a cero 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.



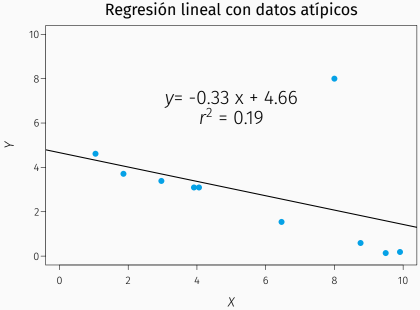
Datos atípicos en regresión

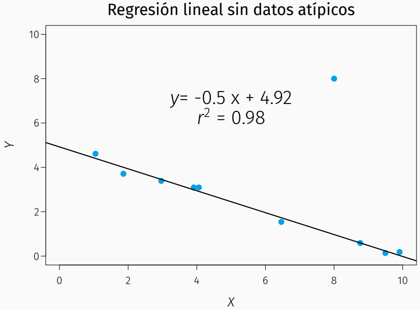
Los *datos atípicos* en un estudio de regresión son los puntos que claramente no siguen la tendencia del resto de los puntos en el diagrama de dispersión, incluso si los valores del par no se pueden considerar atípicos para cada variable por separado.



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.





Relaciones entre atributos

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.

Coeficiente 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 ce 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 perfercta.
- 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
$\sum$			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
$\sum$			0	8.5

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6\sum d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 8.5}{5(5^2-1)} = 0.58.$$

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.

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
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	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.

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\Fuma	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

$$\chi^2 = \frac{(12-15.2)^2}{15.2} + \frac{(28-24.8)^2}{24.8} + \frac{(26-22.8)^2}{22.8} + \frac{(34-37.2)^2}{37.2} = 1.81,$$

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

Introducción

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, y que para obtener conclusiones válidas para la población la muestra debe ser representativa de esta. Por esta razón, para garantizar la representatividad de la muestra, esta debe extraerse *aleatoriamente*, es decir, 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 un experimento que cumple dos condiciones:

1. El conjunto de posibles resultados es conocido.
2. No se puede predecir con absoluta certeza el resultado del experimento.

Ejemplo. Un ejemplo típico de experimentos aleatorios son los juegos de azar. El lanzamiento de un dado, por ejemplo, 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.

Definición (Espacio muestral)

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

Algunos ejemplos de espacios muestrales son:

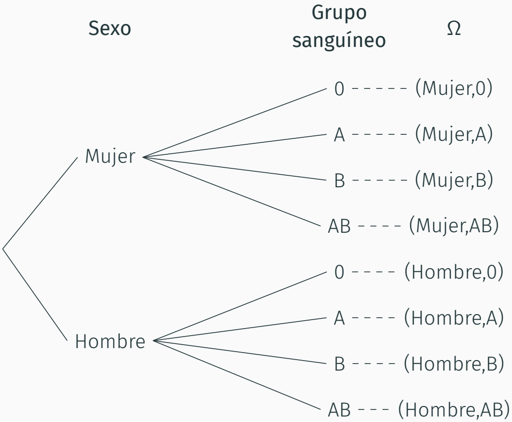
- Lanzamiento de una moneda: $\Omega = \{c, x\}$.
- Lanzamiento de un dado: $\Omega = \{1, 2, 3, 4, 5, 6\}$.
- Grupo sanguíneo de un individuo seleccionado al azar: $\Omega = \{\mathrm{A}, \mathrm{B}, \mathrm{AB}, 0\}$.
- Estatura de un individuo seleccionado al azar: $\Omega = \mathbb{R}^+$.

En experimentos donde se mide más de una variable, la determinación del espacio muestral puede resultar compleja. En tales casos es recomendable utilizar un diagrama de árbol para construir el espacio muestral.

En un diagrama de árbol cada variable se representa en un nivel del árbol y cada posible valor de la variable como una rama.

Diagrama de árbol

Ejemplo de sexo y grupo sanguíneo



Definición (Suceso aleatorio)

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

Existen distintos tipos de sucesos:

Suceso imposible: Es el suceso vacío $\emptyset$. Este suceso nunca ocurre.

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

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

Suceso seguro: Es el suceso que contiene el propio espacio muestral Ω. Este suceso siempre ocurre.

Definición (Espacio de sucesos)

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

Ejemplo. Dado el espacio muestral $\Omega = \{a, b, c\}$, su espacio de sucesos es

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

Operaciones entre sucesos

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

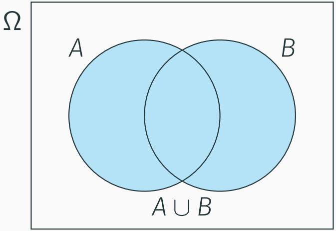
- Unión.
- Intersección.
- Complementario.
- Diferencia.

Unión de sucesos

Definición (Suceso unión)

Dados dos sucesos $A, B \subseteq \Omega$, 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,

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

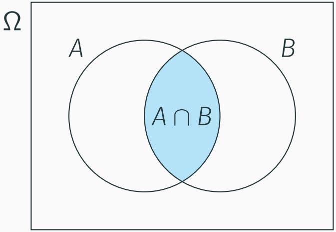


El suceso unión $A \cup B$ ocurre siempre que ocurre A o B.

Definición (Suceso intersección)

Dados dos sucesos $A, B \subseteq \Omega$, 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,

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



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

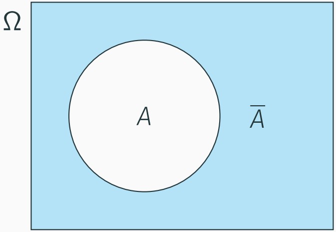
Diremos que dos sucesos son **incompatibles** si su intersección es vacía.

Contrario de un suceso

Definición (Suceso contrario)

Dado suceso $A \subseteq \Omega$, se llama *suceso contrario o complementario* de A, y se denota $\overline{A}$, al suceso formado por los elementos de Ω que no pertenecen a A, es decir,

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

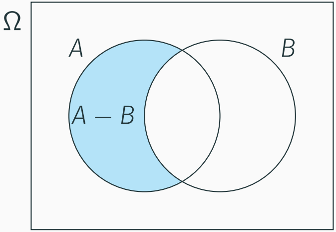


El suceso contrario $\overline{A}$ ocurre siempre que no ocurre A.

Definición (Suceso diferencia)

Dados dos sucesos $A, B \subseteq \Omega$, 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,

$$A - B = \{x \,|\, x \in A \text{ y } x \notin B\} = A \cap \overline{B}.$$



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

Dado el espacio muestral correspondiente al lanzamiento de un dado $\Omega = \{1, 2, 3, 4, 5, 6\}$ y los sucesos $A = \{2, 4, 6\}$ y $B = \{1, 2, 3, 4\}$,

- La unión de A y B es $A \cup B = \{1, 2, 3, 4, 6\}$.
- La intersección de A y B es $A \cap B = \{2, 4\}$.
- El contrario de A es $\overline{A} = \{1, 3, 5\}$.
- Los eventos A y $\overline{A}$ son incompatibles.
- La diferencia de A y B es $A - B = \{6\}$, y la diferencia de B y A es $B - A = \{1, 3\}$.

Álgebra de sucesos

Dados los sucesos $A, B, C \in \mathcal{P}(\Omega)$, 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 (Probabilidad — Laplace)

Dado un espacio muestral Ω de un experimento aleatorio donde todos los elementos de Ω son equiprobables, la *probabilidad* de un suceso $A \subseteq \Omega$ es el cociente entre el número de elementos de A y el número de elementos de Ω

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

Esta definición es ampliamente utilizada, aunque tiene importantes restricciones:

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

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

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)

Dado un espacio muestral Ω de un experimento aleatorio reproducible, la *probabilidad* de un suceso $A \subseteq \Omega$ es la frecuencia relativa del suceso A en infinitas repeticiones del experimento

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

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

Definición axiomática de probabilidad

Definición (Kolmogórov)

Dado un espacio muestral Ω de un experimento aleatorio, una función de *probabilidad* es una aplicación que asocia a cada suceso $A \subseteq \Omega$ un número real $P(A)$, conocido como probabilidad de A, que cumple los siguientes axiomas:

1. La probabilidad de un suceso cualquiera es positiva o nula,
$$P(A) \geq 0.$$
2. La probabilidad del suceso seguro es igual a la unidad,
$$P(\Omega) = 1.$$
3. La probabilidad de la unión de dos sucesos incompatibles ($A \cap B = \emptyset$) es igual a la suma de las probabilidades de cada uno de ellos,
$$P(A \cup B) = P(A) + P(B).$$

Consecuencias de los axiomas de probabilidad

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

Consecuencias de los axiomas de probabilidad

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

1. $P(\overline{A}) = 1 - P(A)$.

Consecuencias de los axiomas de probabilidad

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

1. $P(\overline{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.

Consecuencias de los axiomas de probabilidad

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

1. $P(\overline{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. Si $A \subseteq B$ entonces $P(A) \leq P(B)$.

Consecuencias de los axiomas de probabilidad

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

1. $P(\overline{A}) = 1 - P(A)$.
2. $P(\emptyset) = 0$.
3. Si $A \subseteq B$ entonces $P(A) \leq P(B)$.
4. $P(A) \leq 1$.

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

1. $P(\overline{A}) = 1 - P(A)$.
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
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Consecuencias de los axiomas de probabilidad

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

1. $P(\overline{A}) = 1 - P(A)$.
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 $e_1, e_2, ..., e_n$, entonces
$$P(A) = \sum_{i=1}^{n} P(e_i).$$

Interpretación de la probabilidad

Como ha quedado claro en los axiomas anteriores, la probabilidad de un evento A es un número real $P(A)$ que está siempre entre 0 y 1.

En cierto modo, este número expresa la verosimilitud del evento, es decir, la confianza que hay en que ocurra A en el experimento. Por tanto, también nos da una medida de la incertidumbre sobre el suceso.

- La mayor incertidumbre corresponde a $P(A) = 0.5$ (Es tan probable que ocurra A como que no ocurra).
- La menor incertidumbre corresponde a $P(A) = 1$ (A sucederá con absoluta certeza) y $P(A) = 0$ (A no sucederá con absoluta certeza).

Cuando $P(A)$ está más próximo a 0 que a 1, la confianza en que no ocurra A es mayor que la de que ocurra A. Por el contrario, cuando $P(A)$ está más próximo a 1 que a 0, la confianza en que ocurra A es mayor que la de que no ocurra A.

En algunas ocasiones, es posible que tengamos alguna información sobre el experimento antes de su realización. Habitualmente esa información se da en forma de un suceso B del mismo espacio muestral que sabemos que es cierto antes de realizar el experimento.

En tal caso se dice que el suceso B es un suceso *condicionante*, y la probabilidad de otro suceso A se conoce como **probabilidad condicionada** y se expresa

$$P(A|B).$$

Esto debe leerse como *probabilidad de A dado B* o *probabilidad de A bajo la condición de B*.

Experimentos condicionados

Los condicionantes suelen cambiar el espacio muestral del experimento y por tanto las probabilidades de sus sucesos.

Supongamos que tenemos una muestra de 100 hombres y 100 mujeres con las siguientes frecuencias

	No fumadores	Fumadores
Mujeres	80	20
Hombres	60	40

Entonces, usando la definición frecuentista de probabilidad, la probabilidad de que una persona elegida al azar sea fumadora es

$$P(\text{Fumadora}) = \frac{60}{200} = 0.3.$$

Experimentos condicionados

Los condicionantes suelen cambiar el espacio muestral del experimento y por tanto las probabilidades de sus sucesos.

Supongamos que tenemos una muestra de 100 hombres y 100 mujeres con las siguientes frecuencias

	No fumadores	Fumadores
Mujeres	80	20
Hombres	60	40

Entonces, usando la definición frecuentista de probabilidad, la probabilidad de que una persona elegida al azar sea fumadora es

$$P(\text{Fumadora}) = \frac{60}{200} = 0.3.$$

Sin embargo, si se sabe que la persona elegida es mujer, entonces la muestra se reduce a la primera fila, y la probabilidad de ser fumadora es

$$P(\text{Fumadora}|\text{Mujer}) = \frac{20}{100} = 0.2.$$

Definición (Probabilidad condicionada)

Dado un espacio muestral Ω de un experimento aleatorio, y dos dos sucesos $A, B \subseteq \Omega$, 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

$$P(\text{Fumadora}|\text{Mujer}) = \frac{P(\text{Fumadora} \cap \text{Mujer})}{P(\text{Mujer})} = \frac{20/200}{100/200} = \frac{20}{100} = 0.2.$$

Probabilidad del suceso intersección

A partir de la definición de probabilidad condicionada es posible obtener la fórmula para calcular la probabilidad de la intersección de dos sucesos.

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

Ejemplo. En una población hay un 30% de fumadores y se sabe que el 40% de los fumadores tiene cáncer de pulmón. La probabilidad de que una persona elegida al azar sea fumadora y tenga cáncer de pulmón es

$$P(\text{Fumadora} \cap \text{Cáncer}) = P(\text{Fumadora})P(\text{Cáncer}|\text{Fumadora}) = 0.3 \times 0.4 = 0.12.$$

En ocasiones, la ocurrencia del suceso condicionante no cambia la probabilidad original del suceso principal.

Definición (Sucesos independientes)

Dado un espacio muestral Ω de un experimento aleatorio, dos sucesos $A, B \subseteq \Omega$ son *independientes* si la probabilidad de A no se ve alterada al condicionar por B, y viceversa, es decir,

$$P(A|B) = P(A) \quad \text{and} \quad P(B|A) = P(B),$$

si $P(A) \neq 0$ y $P(B) \neq 0$.

Esto significa que la ocurrencia de uno evento no aporta información relevante para cambiar la incertidumbre sobre el otro.

Cuando dos eventos son independientes, la probabilidad de su intersección es igual al producto de sus probabilidades,

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

Espacio probabilístico

Definición (Espacio probabilístico)

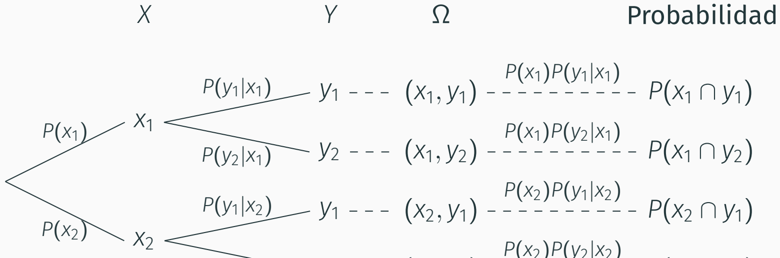
Un *espacio probabilístico* de un experimento aleatorio es una terna $(\Omega, \mathcal{F}, P)$ donde

- Ω es el espacio muestral del experimento.
- $\mathcal{F}$ es un un conjunto de sucesos del experimento.
- P es una función de probabilidad.

Si conocemos la probabilidad de todos los elementos de Ω, entonces podemos calcular la probabilidad de cualquier suceso en $\mathcal{F}$ y se puede construir fácilmente el espacio probabilístico.

Para determinar la probabilidad de cada suceso elemental se puede utilizar un diagrama de árbol, mediante las siguientes reglas:

1. Para cada nodo del árbol, etiquetar la rama que conduce hasta él con la probabilidad de que la variable en ese nivel tome el valor del nodo, condicionada por los sucesos correspondientes a sus nodos antecesores en el árbol.
2. La probabilidad de cada suceso elemental en las hojas del árbol es el producto de las probabilidades de las ramas que van desde la raíz a la hoja del árbol.

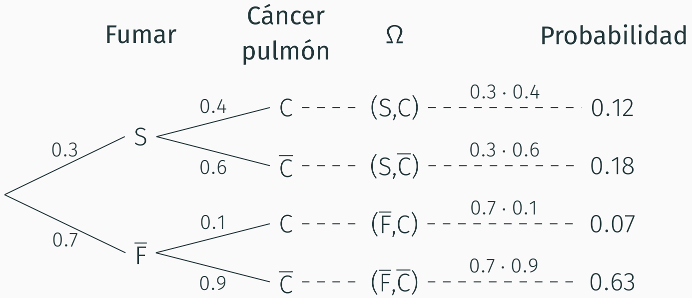


Árboles de probabilidad con variables dependientes

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%.

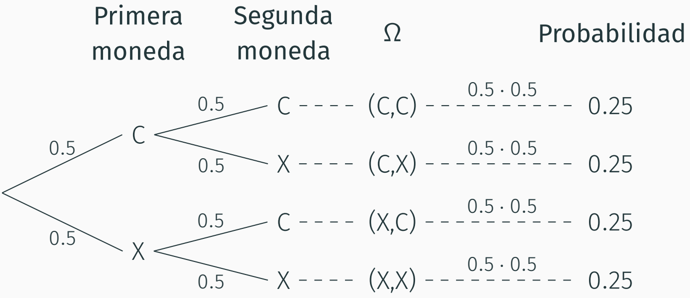
El espacio probabilístico del experimento aleatorio que consiste en elegir una persona al azar y medir las variables Fumar y Cáncer de pulmón se muestra a continuación.



Árboles de probabilidad con variables independientes

Ejemplo de independencia en el lanzamiento de dos monedas

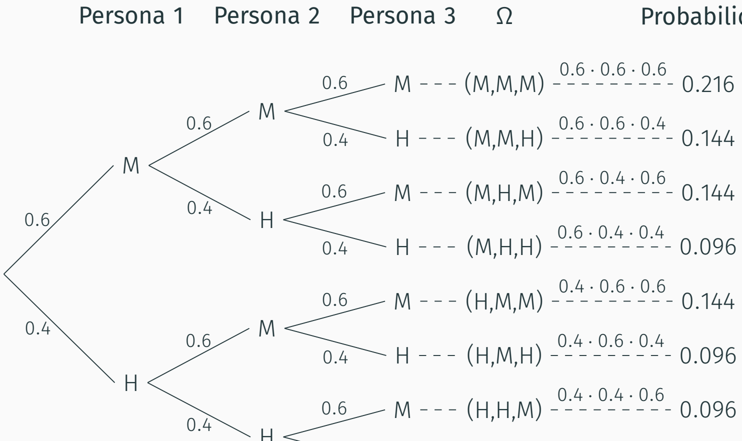
El árbol de probabilidad asociado al experimento aleatorio que consiste en el lanzamiento de dos monedas se muestra a continuación.



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 árbol de probabilidad que se muestra a continuación.



Sistema completo de sucesos

Definición (Sistema completo de sucesos)

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

1. La unión de todos es el espacio muestral: $A_1 \cup \cdots \cup A_n = \Omega$.
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 $A_1, \ldots, A_n$ y un suceso B de un espacio muestral Ω, la probabilidad de cualquier suceso B del espacio muestral se puede calcular mediante la fórmula

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

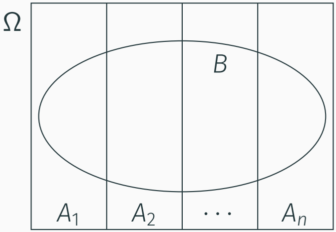
La demostración del teorema es sencilla, ya que al ser $A_1, \ldots, A_n$ un sistema completo tenemos

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

y como estos sucesos son incompatibles entre sí, se tiene

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$

$$= P(A_1)P(B/A_1) + \cdots + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$



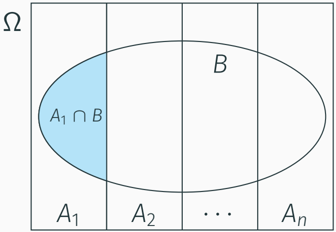
La demostración del teorema es sencilla, ya que al ser $A_1, \ldots, A_n$ un sistema completo tenemos

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

y como estos sucesos son incompatibles entre sí, se tiene

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$

$$= P(A_1)P(B/A_1) + \cdots + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$



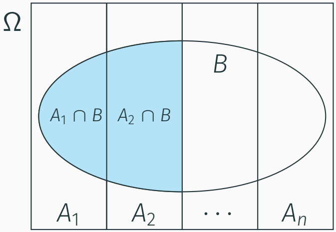
La demostración del teorema es sencilla, ya que al ser $A_1, \ldots, A_n$ un sistema completo tenemos

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

y como estos sucesos son incompatibles entre sí, se tiene

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$

$$= P(A_1)P(B/A_1) + \cdots + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$



Teorema de la probabilidad total

Demostración

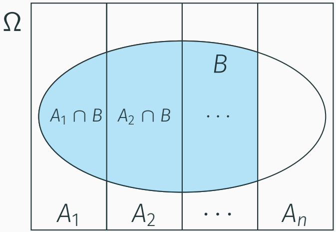
La demostración del teorema es sencilla, ya que al ser $A_1, \ldots, A_n$ un sistema completo tenemos

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

y como estos sucesos son incompatibles entre sí, se tiene

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$

$$= P(A_1)P(B/A_1) + \cdots + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$

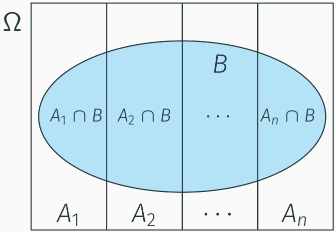


La demostración del teorema es sencilla, ya que al ser $A_1, \ldots, A_n$ un sistema completo tenemos

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

y como estos sucesos son incompatibles entre sí, se tiene

$$P(B) = P((B \cap A_1) \cup \cdots \cup (B \cap A_n)) = P(B \cap A_1) + \cdots + P(B \cap A_n) =$$
$$= P(A_1)P(B/A_1) + \cdots + P(A_n)P(B/A_n) = \sum_{i=1}^{n} P(A_i)P(B/A_i).$$



Un determinado síntoma S puede ser originado por una enfermedad E pero también lo pueden presentar las personas sin la enfermedad. Sabemos que la prevalencia de la enfermedad E es 0.2. Además, se sabe que el 90% de las personas con la enfermedad presentan el síntoma, mientras que sólo el 40% de las personas sin la enfermedad lo presentan.

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

Para responder a la pregunta se puede aplicar el teorema de la probabilidad total usando el sistema completo $\{E, \overline{E}\}$:

$$P(S) = P(E)P(S|E) + P(\overline{E})P(S|\overline{E}) = 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.5.$$

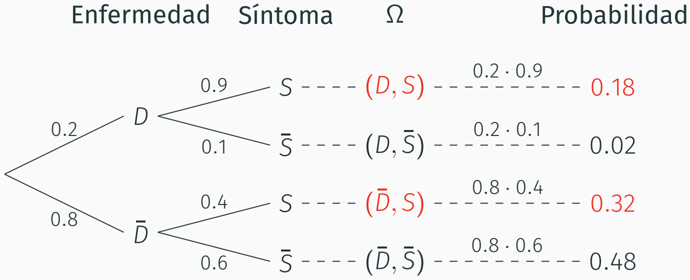
Es decir, la mitad de la población tendrá el síntoma.

¡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 espacio probabilístico del experimento.



$$P(S) = P(E, S) + P(\bar{E}, S) = P(E)P(S|E) + P(\bar{E})P(S|\bar{E})$$
$$= 0.2 \cdot 0.9 + 0.8 \cdot 0.4 = 0.18 + 0.32 = 0.5.$$

Teorema de Bayes

Los sucesos de un sistema completo de sucesos $A_1, \cdots, A_n$ 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 $A_1, \ldots, A_n$ y un suceso B de un espacio muestral Ω y otro suceso B del mismo espacio muestral, la probabilidad de cada suceso A_i $i = 1, \ldots, n$ condicionada por B puede calcularse con la siguiente fórmula

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

En el ejemplo anterior, una pregunta más interesante es qué diagnosticar a una persona que presenta el síntoma.

En este caso se puede interpretar E y $\overline{E}$ como las dos posibles hipótesis para el síntoma S. Las probabilidades a priori para ellas son $P(E) = 0.2$ y $P(\overline{E}) = 0.8$. Esto quiere decir que si no se dispone de información sobre el síntoma, el diagnóstico será que la persona no tiene la enfermedad.

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,

$$P(E|S) \text{ y } P(\overline{E}|S)$$

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

$$P(E|S) = \frac{P(E)P(S|E)}{P(E)P(S|E) + P(\bar{E})P(S|\bar{E})} = \frac{0.2 \cdot 0.9}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.18}{0.5} = 0.36,$$

$$P(\bar{E}|S) = \frac{P(\bar{E})P(S|\bar{E})}{P(E)P(S|E) + P(\bar{E})P(S|\bar{E})} = \frac{0.8 \cdot 0.4}{0.2 \cdot 0.9 + 0.8 \cdot 0.4} = \frac{0.32}{0.5} = 0.64.$$

Como se puede ver la probabilidad de tener la enfermedad ha aumentado. No obstante, la probabilidad de no tener la enfermedad sigue siendo mayor que la de tenerla, y por esta razón el diagnóstico seguirá siendo que no tiene la enfermedad.

En este caso se dice que el síntoma *S no es determinante* a la hora de diagnosticar la enfermedad.

Epidemiología

Una de las ramas de la Medicina que hace un mayor uso de la probabilidad es la **Epidemiología**, que estudia la distribución y las causas de las enfermedades en las poblaciones, identificando factores de riesgos para las enfermedades de cara a la atención médica preventiva.

En Epidemiología interesa la frecuencia de un *suceso médico E* (típicamente una enfermedad como la gripe, un factor de riesgo como fumar o un factor de protección como vacunarse) que se mide mediante una variable nominal con dos categorías (ocurrencia o no del suceso).

Hay diferentes medidas relativas a la frecuencia de un suceso médico. Las más importantes son:

- Prevalencia
- Incidencia
- Riesgo relativo
- Odds ratio

Prevalencia

Definición (Prevalencia)

La *prevalencia* de un suceso médico E es la proporción de una población que está afectada por el suceso.

$$\text{Prevalencia}(E) = \frac{\text{Nº individuos afectados por } E}{\text{Tamaño poblacional}}$$

A menudo, la prevalencia se estima mediante una muestra como la frecuencia relativa de los individuos afectados por el suceso en la muestra. Es también común expresarla est frecuencia como un porcentaje.

Ejemplo. Para estimar la prevalencia de la gripe se estudió una muestra de 1000 personas de las que 150 presentaron gripe. Así, la prevalencia de la gripe es aproximadamente 150/1000=0.15, es decir, un 15%.

La incidencia mide la probabilidad de ocurrencia de un suceso médico en una población durante un periodo de tiempo específico. La incidencia puede medirse como una proporción acumulada o como una tasa.

Definición (Incidencia acumulada)

La *incidencia acumulada* de un suceso médico E es la proporción de individuos que experimentaron el evento en un periodo de tiempo, es decir, el número de nuevos casos afectados por el evento en el periodo de tiempo, divido por el tamaño de la población inicialmente en riesgo de verse afectada.

$$R(E) = \frac{\text{Nº de nuevos casos con } E}{\text{Tamaño de la población en riesgo}}$$

Ejemplo. Una población contenía inicialmente 1000 personas sin gripe y después de dos años se observó que 160 de ellas sufrieron gripe. La incidencia acumulada de la gripe es 160 casos pro 1000 personas por dos años, es decir, 16% en dos años.

Definición (Tasa de incidencia)

La *tasa de incidencia* o *riesgo absoluto* de un suceso médico E es el número de nuevos casos afectados por el evento divido por la población en riesgo y por el número de unidades temporales del periodo considerado.

$$R(E) = \frac{\text{Nº nuevos casos con } E}{\text{Tamaño población en riesgo} \times \text{Nº unidades de tiempo}}$$

Ejemplo. Una población contenía inicialmente 1000 personas sin gripe y después de dos años se observó que 160 de ellas sufrieron gripe. Si se considera el año como intervalo de tiempo, la tasa de incidencia de la gripe es 160 casos dividida por 1000 personas y por dos años, es decir, 80 casos por 1000 personas-año o 8% de personas al año.

Prevalencia vs Incidencia

La prevalencia no debe confundirse con la incidencia. La prevalencia indica cómo de extendido está el suceso médico en una población, sin preocuparse por cuándo los sujetos se han expuesto al riesgo o durante cuánto tiempo, mientras que la incidencia se fija en el riesgo de verse afectado por el suceso en un periodo concreto de tiempo.

Así, la prevalencia se calcula en estudios transversales en un momento temporal puntual, mientras que para medir la incidencia se necesita un estudio longitudinal que permita observar a los individuos durante un periodo de tiempo.

La incidencia es más útil cuando se pretende entender la causalidad del suceso: por ejemplo, si la incidencia de una enfermedad en una población aumenta, seguramente hay un factor de riesgo que lo está promoviendo.

Cuando la tasa de incidencia es aproximadamente constante en la duración del suceso, la prevalencia es aproximadamente el producto de la incidencia por la duración media del suceso, es decir,

$$\text{prevalencia} = \text{incidencia} \times \text{duración}$$

Comparación de riesgos

Para determinar si un factor o característica está asociada con el suceso médico es necesario comparar el riesgo del suceso en dos poblaciones, una expuesta a el factor y la otra no. El grupo expuesto a el factor se conoce como *grupo tratamiento* o *grupo experimental* y el grupo no expuesto como *grupo control*.

Habitualmente los casos observados para cada grupo se representan en una tabla de 2×2 como la siguiente:

	Suceso E	No suceso $\overline{E}$
Grupo tratamiento (expuestos)	a	b
Grupo control (no expuestos)	c	d

Riesgo atribuible o diferencia de riesgos RA

Definición (Riesgo atribuible)

El *riesgo atribuible* o *diferencia de riesgo* de un suceso médico E para los individuos expuestos a un factor es la diferencia entre los riesgos absolutos de los grupos tratamiento y control.

$$RA(E) = R_T(E) - R_C(E) = \frac{a}{a+b} - \frac{c}{c+d}.$$

El riesgo atribuible es el riesgo de un suceso que es debido específicamente al factor de interés.

Obsérvese que el riesgo atribuible puede ser positivo, cuando el riesgo del grupo tratamiento es mayor que el del grupo control, o negativo, de lo contrario.

Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

	Gripe E	No gripe $\overline{E}$
Grupo tratamiento (vacunados)	20	480
Grupo control (No vacunados)	80	420

El riesgo atribuible de contraer la gripe cuando se es vacunado es

$$AR(D) = \frac{20}{20+480} - \frac{80}{80+420} = -0.12.$$

Esto quiere decir que el riesgo de contraer la gripe es un 12% menor en vacunados que en no vacunados.

Definición (Riesgo relativo)

El *riesgo relativo* de un suceso médico E para los individuos expuestos a un factor es el cociente entre las proporciones de individuos afectados por el suceso en un periodo de tiempo de los grupos tratamiento y control. Es decir, el cociente entre las incidencias de grupo tratamiento y el grupo control.

$$RR(D) = \frac{\text{Riesgo grupo tratamiento}}{\text{Riesgo grupo control}} = \frac{R_T(E)}{R_C(E)} = \frac{a/(a+b)}{c/(c+d)}$$

El riesgo relativo compara el riesgo de desarrollar un suceso médico entre el grupo tratamiento y el grupo control.

- $RR = 1 \Rightarrow$ No hay asociación entre el suceso y la exposición al factor.
- $RR < 1 \Rightarrow$ La exposición al factor disminuye el riesgo del suceso.
- $RR > 1 \Rightarrow$ La exposición al factor aumenta el riesgo del suceso.

Cuanto más lejos de 1, más fuerte es la asociación.

Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

	Gripe E	No gripe $\overline{E}$
Grupo tratamiento (vacunados)	20	480
Grupo control (No vacunados)	80	420

El riesgo relativo de contraer la gripe cuando se es vacunado es

$$RR(D) = \frac{20/(20+480)}{80/(80+420)} = 0.25.$$

Así, la probabilidad de contraer la gripe en los individuos vacunados fue la cuarta parte de la de contraerla en el caso de no haberse vacunado, es

Una forma alternativa de medir el riesgo de un suceso médico es el *odds*.

Definición (Odie)

El *odds* de un suceso médico E en una población es el cociente entre el número de individuos que adquirieron el suceso y los que no en un periodo de tiempo.

$$ODDS(E) = \frac{\text{Nº nuevos casos con } E}{\text{Nº casos sin } E} = \frac{P(E)}{P(\overline{E})}$$

A diferencia de la incidencia, que es una proporción menor o igual que 1, el odds puede ser mayor que 1. No obstante es posible convertir el odds en una probabilidad con al fórmula

$$P(E) = \frac{ODDS(E)}{ODDS(E) + 1}$$

Ejemplo Una población contenía inicialmente 1000 personas sin gripe. Después de un año 160 de ellas tuvieron gripe. Entonces el odds de la gripe

Odds ratio OR

Definición (Odds ratio)

El *odds ratio* o la *oportunidad relativa* de un suceso médico E para los individuos expuestos a un factor es el cociente entre los odds del sucesos de los grupos tratamiento y control.

$$OR(E) = \frac{\text{Odds en grupo tratamiento}}{\text{Odds en grupo control}} = \frac{a/b}{c/d} = \frac{ad}{bc}$$

El odds ratio compara los odds de un suceso médico entre el grupo tratamiento y control. La interpretación es similar a la del riesgo relativo:

- $OR = 1 \Rightarrow$ No existe asociación entre el suceso y la exposición al factor.
- $OR < 1 \Rightarrow$ La exposición al factor disminuye el riesgo del suceso.
- $OR > 1 \Rightarrow$ La exposición al factor aumenta el riesgo del suceso.

Cuanto más lejos de 1, más fuerte es la asociación.

Para determinar la efectividad de una vacuna contra la gripe, una muestra de 1000 personas sin gripe fueron seleccionadas al comienzo del año. La mitad de ellas fueron vacunadas (grupo tratamiento) y la otra mitad recibieron un placebo (grupo control). La tabla siguiente resume los resultados al final del año.

	Gripe D	No gripe $\overline{D}$
Grupo tratamiento (vacunados)	20	480
Grupo control (No vacunados)	80	420

El odds ratio de sufrir la gripe para los individuos vacunados es

$$OR(D) = \frac{20/480}{80/420} = 0.21875.$$

Esto quiere decir que el odds de sufrir la gripe frente a no sufrirla en los vacunados es casi un quinto del de los no vacunados, es decir, que

Riesgo relativo vs Odds ratio

El riesgo relativo y el odds ratio son dos medidas de asociación pero su interpretación es ligeramente diferente. Mientras que el riesgo relativo expresa una comparación de riesgos entre los grupos tratamiento y control, el odds ratio expresa una comparación de odds, que no es lo mismo que el riesgo. Así, un odds ratio de 2 *no* significa que el grupo tratamiento tiene el doble de riesgo de adquirir el suceso.

La interpretación del odds ratio es un poco más enrevesada porque es contrafactual, y nos da cuántas veces es más frecuente el suceso en el grupo tratamiento en comparación con el control, asumiendo que en el grupo control es tan frecuente que ocurra el suceso como que no.

La ventaja del odds ratio es que no depende de la prevalencia o la incidencia del suceso, y debe usarse siempre que el número de individuos que presenta el suceso se selecciona arbitrariamente en ambos grupos, como ocurre en los estudios casos-control.

Para determinar la asociación entre el cáncer de pulmón y fumar se tomaron dos muestras (la segunda con el doble de individuos sin cáncer) obteniendo los siguientes resultados:

Sample 1

	Cáncer	No cáncer
Fumadores	60	80
No fumadores	40	320

$$RR(D) = \frac{60/(60+80)}{40/(40+320)} = 3.86.$$

$$OR(D) = \frac{60/80}{40/320} = 6.$$

Sample 2

	Cáncer	No cáncer
Fumadores	60	160
No fumadores	40	640

$$RR(D) = \frac{60/(60+160)}{40/(40+640)} = 4.64.$$

$$OR(D) = \frac{60/160}{40/640} = 6.$$

Así, cuando cambia la incidencia o prevalencia de un suceso (cáncer de pulmón) el riesgo relativo cambia, mientras que el odds ratio no.

Riesgo relativo vs Odds ratio

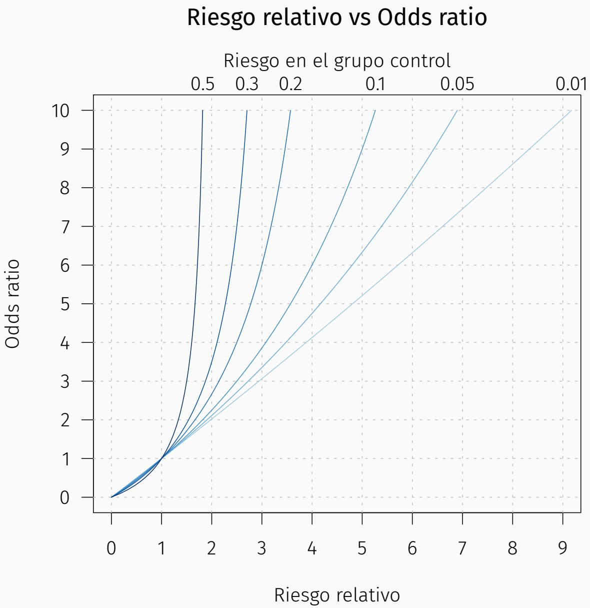
La relación entre el riesgo relativo y el odds ratio viene dada por la siguiente fórmula

$$RR = \frac{OR}{1 - R_0 + R_0 * OR} = OR * \frac{1 - R_1}{1 - R_0},$$

donde R_C and R_T son la prevalencia o la incidencia en los grupos control y tratamiento respectivamente.

El odds ratio siempre sobrestima el riesgo relativo cuando este es mayor que 1 y lo subestima cuando es menor que 1. No obstante, con sucesos médicos raros (con una prevalencia o incidencia baja) el riesgo relativo y el odds ratio son casi iguales.

Riesgo relativo vs Odds ratio



Tests diagnósticos

En Epidemiología es común el uso de test para diagnosticar enfermedades.

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

	Presencia enfermedad E	Ausencia enfermedad $\overline{D}$
Test positivo $+$	Verdadero positivo VP	Falso positivo FP
Test negativo $-$	Falso negativo FN	Verdadero Negativo VN

Sensibilidad y especificidad de un test diagnóstico

La fiabilidad de un test diagnóstico depende de las siguientes probabilidades.

Definición (Sensibilidad)

La *sensibilidad* de un test diagnóstico es la proporción de resultados positivos del test en personas con la enfermedad,

$$P(+|E) = \frac{VP}{VP + FN}$$

Definición (Especificidad)

La *especificidad* de un test diagnóstico es la proporción de resultados negativos del test en personas sin la enfermedad,

$$P(-|\overline{E}) = \frac{VN}{VN + FP}$$

Interpretación de la sensibilidad y la especificidad

Normalmente existe un balance entre la sensibilidad y la especificidad.

Un test con una alta sensibilidad detectará la enfermedad en la mayoría de las personas enfermas, pero también dará más falsos positivos que un test menos sensible. De este modo, un resultado positivo en un test con una gran sensibilidad no es muy útil para confirmar la enfermedad, pero un resultado negativo es útil para descartar la enfermedad, ya que raramente da resultados negativos en personas con la enfermedad.

Por otro lado, un test con una alta especificidad descartará la enfermedad en la mayoría de las personas sin la enfermedad, pero también producirá más falsos negativos que un test menos específico. Así, un resultado negativo en un test con una gran especificidad no es útil para descartar la enfermedad, pero un resultado positivo es muy útil para confirmar la enfermedad, ya que raramente da resultados positivos en personas sin la enfermedad.

Interpretación de la sensibilidad y la especificidad

Decidir entre un test con una gran sensibilidad o un test con una gran especificidad depende del tipo de enfermedad y el objetivo del test. En general, utilizaremos un test sensible cuando:

- La enfermedad es grave y es importante detectarla.
- La enfermedad es curable.
- Los falsos positivos no provocan traumas serios.

Y utilizaremos un test específico cuando:

- La enfermedad es importante pero difícil o imposible de curar.
- Los falsos positivos pueden provocar traumas serios.
- El tratamiento de los falsos positivos puede tener graves consecuencias.

Valores predictivos de un test diagnóstico

Pero el aspecto más importante de un test diagnóstico es su poder predictivo, que se mide con las siguientes probabilidades a posteriori.

Definición (Valor predictivo positivo VPP)

El *valor predictivo positivo* de un test diagnóstico es la proporción de personas con la enfermedad entre las personas con resultado positivo en el test,

$$P(E|+) = \frac{VP}{VP + FP}$$

Definición (Valor predictivo negativo VPN)

El *valor predictivo negativo* de un test diagnóstico es la proporción de personas sin la enfermedad entre las personas con resultado negativo en el test,

$$P(\overline{E}|-) = \frac{VN}{VN + FN}$$

Interpretación de los valores predictivos

Los valores predictivos positivo y negativo permiten confirmar o descartar la enfermedad, respectivamente, si alcanzan al menos el umbral de 0.5.

$$\begin{aligned} VPP > 0.5 \quad &\Rightarrow \quad \text{Diagnosticar la enfermedad} \\ VPN > 0.5 \quad &\Rightarrow \quad \text{Diagnosticar la no enfermedad} \end{aligned}$$

No obstante, estas probabilidades dependen de la proporción de personas con la enfermedad en la población $P(E)$ que se conoce como **prevalencia** de la enfermedad. Pueden calcularse a partir de la sensibilidad y la especificidad del test diagnóstico usando el teorema de Bayes.

$$VPP = P(E|+) = \frac{P(E)P(+|E)}{P(E)P(+|E) + P(\bar{E})P(+|\bar{E})}$$

$$VPN = P(\bar{E}|-) = \frac{P(\bar{E})P(-|\bar{E})}{P(E)P(-|E) + P(\bar{E})P(-|\bar{E})}$$

Así, con enfermedades frecuentes, el valor predictivo positivo aumenta, y con enfermedades raras, el valor predictivo negativo aumenta.

Un test diagnóstico para la gripe se ha aplicado a una muestra aleatoria de 1000 personas. Los resultados aparecen resumidos en la siguiente tabla.

	Presencia de gripe E	Ausencia de gripe $\overline{E}$
Test $+$	95	90
Test $-$	5	810

Según esta muestra, la prevalencia de la gripe puede estimarse como

$$P(E) = \frac{95+5}{1000} = 0.1.$$

La sensibilidad del test diagnóstico es

$$P(+|E) = \frac{95}{95+5} = 0.95.$$

Y la especificidad es

$$P(-|\overline{E}) = \frac{810}{90+810} = 0.9.$$

El valor predictivo positivo del test es

$$VPP = P(E|+) = \frac{95}{95+90} = 0.5135.$$

Como este valor es mayor que 0.5, eso significa que se diagnosticará la gripe si el resultado del test es positivo. No obstante, la confianza en el diagnóstico será baja, ya que el valor es poco mayor que 0.5.

Por otro lado, el valor predictivo negativo es

$$VPN = P(\overline{E}|-) = \frac{810}{5+810} = 0.9939.$$

Como este valor es casi 1, eso significa que es casi seguro que no se tiene la gripe cuando el resultado del test es negativo.

Así, se puede concluir que este test es muy potente para descartar la gripe, pero no lo est tanto para confirmarla.

Razón de verosimilitud de un test diagnóstico

La siguientes medidas también se derivan de la sensibilidad y la especificidad de un test diagnóstico.

Definición (Razón de verosimilitud positiva $RV+$)

La *razón de verosimilitud positiva* de un test diagnóstico es el cociente entre la probabilidad de un resultado positivo en personas con la enfermedad y personas sin la enfermedad, respectivamente.

$$RV+ = \frac{P(+|E)}{P(+|\overline{E})} = \frac{\text{Sensibilidad}}{1 - \text{Especificidad}}$$

Definición (Razón de verosimilitud negativa $RV-$)

La *razón de verosimilitud negativa* de un test diagnóstico es el cociente entre la probabilidad de un resultado negativo en personas con la enfermedad y personas sin la enfermedad, respectivamente.

$$RV- = \frac{P(-|E)}{P(-|\overline{E})} = \frac{1 - \text{Sensibilidad}}{\text{Especificidad}}$$

Interpretación de las razones de verosimilitud

La razón de verosimilitud positiva puede interpretarse como el número de veces que un resultado positivo es más probable en personas con la enfermedad que en personas sin la enfermedad.

Por otro lado, la razón de verosimilitud negativa puede interpretarse como el número de veces que un resultado negativo es más probable en personas con la enfermedad que en personas sin la enfermedad.

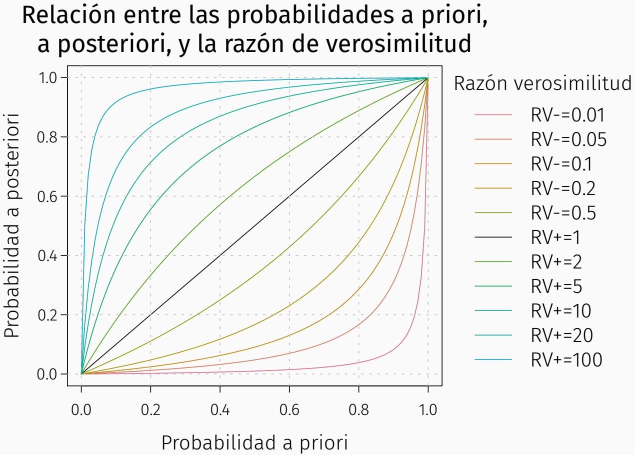
Las probabilidades a posteriori pueden calculares a partir de las probabilidades a priori usando las razones de verosimilitud

$$P(E|+) = \frac{P(E)P(+|E)}{P(E)P(+|E) + P(\bar{E})P(+|\bar{E})} = \frac{P(E)RV+}{1 - P(E) + P(E)RV+}$$

Así,

- Una razón de verosimilitud positiva mayor que 1 aumenta la probabilidad de la enfermedad.
- Una razón de verosimilitud positiva menor que 1 disminuye la probabilidad de la enfermedad.

Interpretación de las razones de verosimilitud



Variables Aleatorias

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.

$$X : \Omega \to \mathbb{R}$$

Al conjunto de posibles valores que puede tomar la variable aleatoria se le llama *rango* o *recorrido* de la variable, y se representa por $\mathrm{Ran}(X)$.

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.

Ejemplo. La variable X que mide el resultado del lanzamiento de un dado

Las variables aleatorias se clasifican en dos tipos:

Discretas (VAD): Toman valores aislados y su rango es numerable. Ejemplo. Número de hijos, número de cigarrillos, número de asignaturas aprobadas, etc.

Continuas (VAC): Toman valores en un intervalo real y su rango es no numerable
Ejemplo. Peso, estatura, edad, nivel de colesterol, etc.

Los modelos probabilísticos de cada tipo de variables tienen características diferenciadas y por eso se estudiarán por separado. En este capítulo se estudian los modelos probabilísticos de las variables aleatorias discretas.

Distribución de probabilidad de una variable aleatoria 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 de la variable 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 de la variable la probabilidad de que la variable tome un valor menor o igual que x_i,

Distribución de probabilidad de una variable aleatoria discreta

Al rango de la variable, junto a su función de probabilidad o de distribución, se le llama **Distribución de probabilidad** de la variable, y se suele representar en forma de tabla

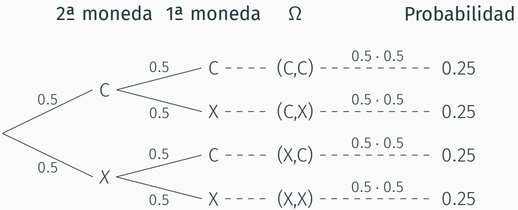
X	x_1	x_2	$\cdots$	x_n	$\sum$
$f(x)$	$f(x_1)$	$f(x_2)$	$\cdots$	$f(x_n)$	1
$F(x)$	$F(x_1)$	$F(x_2)$	$\cdots$	$F(x_n) = 1$	

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 del espacio probabilístico del experimento se muestra a continuación.



y según esto, su distribución de probabilidad es

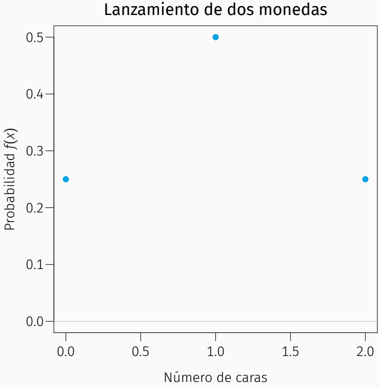
X	0	1	2
$f(x)$	0.25	0.5	0.25
$F(x)$	0.25	0.75	1

$$F(x) = \begin{cases} 0 & \text{si } x < 0 \\ 0.25 & \text{si } 0 \leq x < 1 \\ 0.75 & \text{si } 1 \leq x < 2 \\ 1 & \text{si } x \geq 2 \end{cases}$$

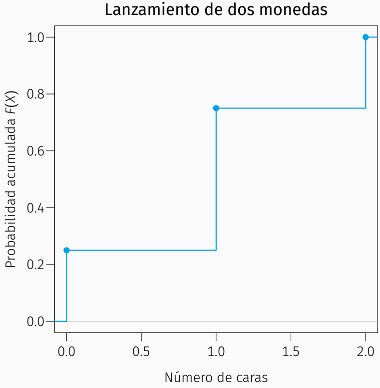
Gráficos de la distribución de probabilidad

Ejemplo del lanzamiento de dos monedas

Función de probabilidad



Función de distribución



Estadísticos poblacionales

Al igual que para describir las variables en las muestras se utilizan estadísticos descriptivos muestrales, para describir las 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 son[1]:

- 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:

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

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

En el experimento aleatorio 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

$$\mu = \sum_{i=1}^{n} x_i f(x_i) = 0 \cdot 0.25 + 1 \cdot 0.5 + 2 \cdot 0.25 = 1 \text{ cara},$$

$$\sigma^2 = \sum_{i=1}^{n} x_i^2 f(x_i) - \mu^2 = (0^0 \cdot 0.25 + 1^2 \cdot 0.5 + 2^2 \cdot 0.25) - 1^2 = 0.5 \text{ caras}^2,$$

$$\sigma = +\sqrt{0.5} = 0.71 \text{ caras}.$$

Modelos de distribución de probabilidad

De acuerdo al tipo de experimento en el que se mide la variable aleatoria, existen diferentes modelos de distribución de probabilidad. Los más importantes son

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

Distribución Uniforme $U(a, b)$

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

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

Se dice que una variable aleatoria X sigue un *modelo de distribución uniforme* de parámetros a, b, y se nota $X \sim U(a, b)$, si su rango es $Ran(X) = \{a, a+1, \ldots, b\}$, y su función de probabilidad vale

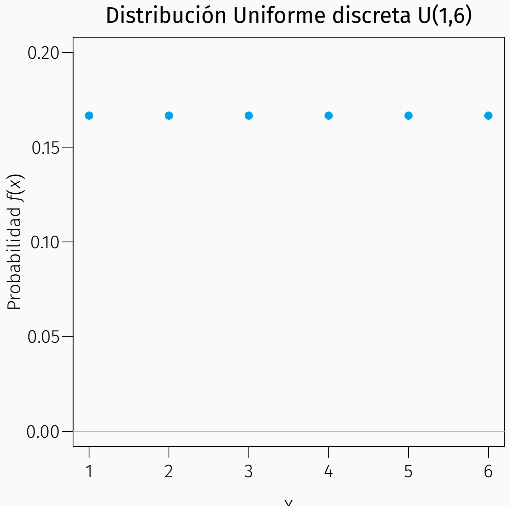
$$f(x) = \frac{1}{b-a+1}.$$

Obsérvese que a y b son el mínimo y el máximo del rango respectivamente.

Su media y varianza valen

$$\mu = \sum_{i=0}^{b-a} \frac{a+i}{b-a+1} = \frac{a+b}{2} \qquad \sigma^2 = \sum_{i=0}^{b-a} \frac{(a+i-\mu)^2}{b-a+1} = \frac{(b-a+1)^2-1}{12}$$

La variable que mide el número obtenido al lanzar un dado sigue un modelo de distribución uniforme $U(1, 6)$.



Distribución Binomial

La distribución binomial corresponde a una variable aleatoria medida en 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 conocidos como *Éxito* o *Fracaso*.
- Los ensayos son independientes.
- La probabilidad de éxito es idéntica para todos los ensayos y vale $P(\text{Éxisto}) = p$.

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, y se nota $X \sim B(n,p)$, si su recorrido es $Ran(X) = \{0, 1, \ldots, n\}$ y su función de probabilidad vale

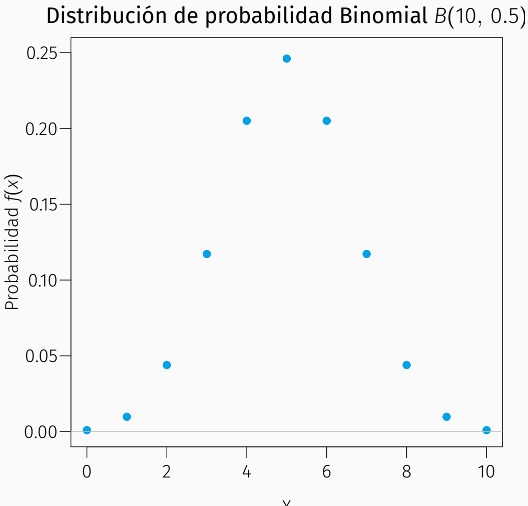
$$f(x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

Obsérvese que n es conocido como el número de repeticiones de la prueba o ensayo y p como la probabilidad de Éxito en cada repetición.

Su media y varianza valen

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

La variable que mide el número de caras obtenidos al lanzar 10 veces una moneda sigue un modelo de distribución binomial $B(10, 0.5)$.



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

$$\begin{aligned} F(2) &= f(0) + f(1) + f(2) = \\ &= \binom{10}{0} 0.5^0 (1 - 0.5)^{10-0} + \binom{10}{1} 0.5^1 (1 - 0.5)^{10-1} + \binom{10}{2} 0.5^2 (1 - 0.5)^{10-2} \\ &= 0.0547. \end{aligned}$$

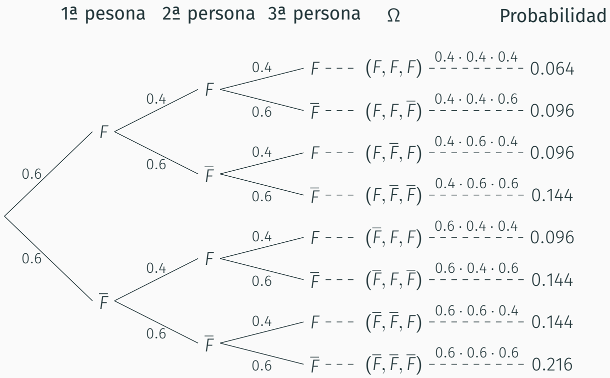
- Y el número esperado de caras es

$$\mu = 10 \cdot 0.5 = 5 \text{ caras.}$$

Distribución Binomial $B(n, p)$

Ejemplo de una muestra aleatoria con reemplazamiento

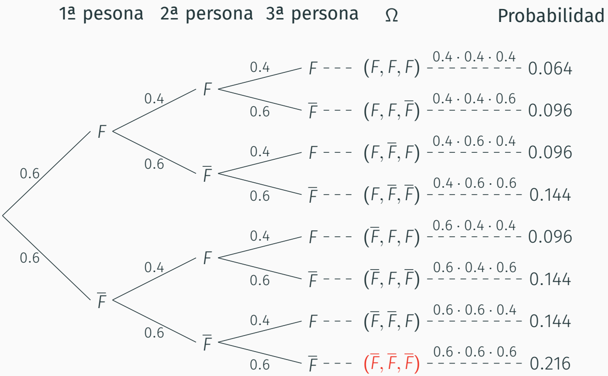
En una población hay un 40% de fumadores. La variable X que mide el número de fumadores en una muestra aleatoria con reemplazamiento de 3 personas sigue un modelo de distribución binomial $X \sim B(3,\ 0.6)$.



Distribución Binomial $B(n, p)$

Ejemplo de una muestra aleatoria con reemplazamiento

En una población hay un 40% de fumadores. La variable X que mide el número de fumadores en una muestra aleatoria con reemplazamiento de 3 personas sigue un modelo de distribución binomial $X \sim B(3,\ 0.6)$.

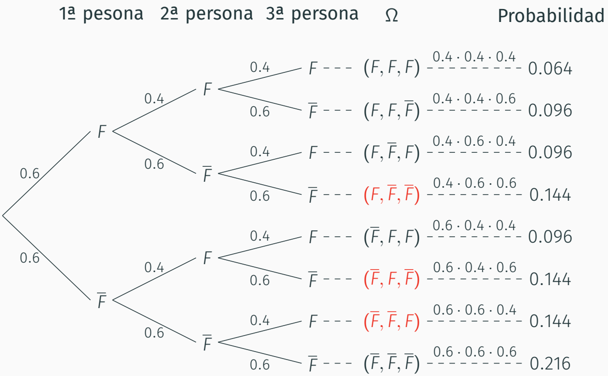


$f(0) = \binom{3}{0} 0.4^0 (1 - 0.4)^{3-0} = 0.6^3,$

Distribución Binomial $B(n, p)$

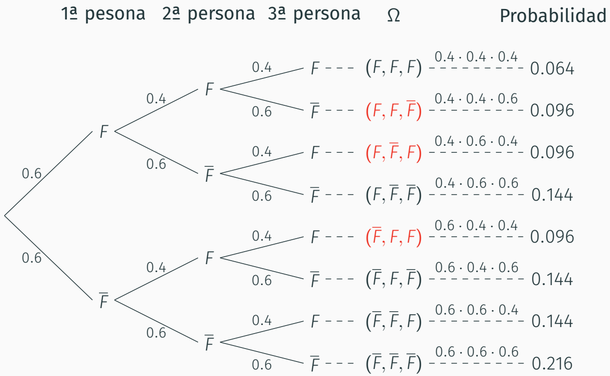
Ejemplo de una muestra aleatoria con reemplazamiento

En una población hay un 40% de fumadores. La variable X que mide el número de fumadores en una muestra aleatoria con reemplazamiento de 3 personas sigue un modelo de distribución binomial $X \sim B(3,\ 0.6)$.



$f(0) = \binom{3}{0} 0.4^0 (1 - 0.4)^{3-0} = 0.6^3,$ $\qquad f(1) = \binom{3}{1} 0.4^1 (1 - 0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

En una población hay un 40% de fumadores. La variable X que mide el número de fumadores en una muestra aleatoria con reemplazamiento de 3 personas sigue un modelo de distribución binomial $X \sim B(3,\ 0.6)$.



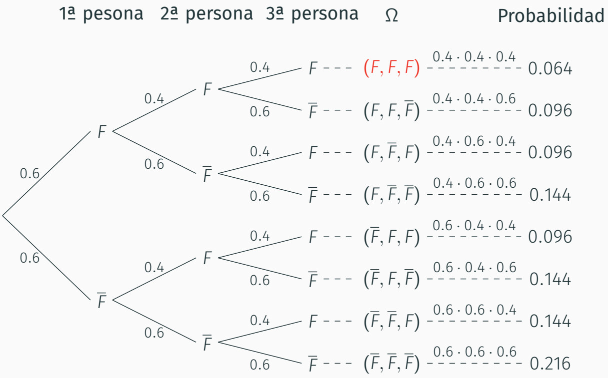
$f(0) = \binom{3}{0} 0.4^0 (1-0.4)^{3-0} = 0.6^3,$ $\qquad f(1) = \binom{3}{1} 0.4^1 (1-0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

$f(2) = \binom{3}{2} 0.4^2 (1-0.4)^{3-2} = 3 \cdot 0.4^2 \cdot 0.6,$

DISTRIBUCIÓN BINOMIAL $B(n, p)$

EJEMPLO DE UNA MUESTRA ALEATORIA CON REEMPLAZAMIENTO

En una población hay un 40% de fumadores. La variable X que mide el número de fumadores en una muestra aleatoria con reemplazamiento de 3 personas sigue un modelo de distribución binomial $X \sim B(3,\ 0.6)$.



$f(0) = \binom{3}{0} 0.4^0 (1-0.4)^{3-0} = 0.6^3,$ $\quad f(1) = \binom{3}{1} 0.4^1 (1-0.4)^{3-1} = 3 \cdot 0.4 \cdot 0.6^2,$

$f(2) = \binom{3}{2} 0.4^2 (1-0.4)^{3-2} = 3 \cdot 0.4^2 \cdot 0.6,$ $\quad f(3) = \binom{3}{3} 0.4^3 (1-0.4)^{3-3} = 0.4^3.$

La distribución de Poisson corresponde a una variable medida en un experimento aleatorio con las siguientes características:

- El experimento consiste en observar la aparición de sucesos puntuales en un intervalo fijo de tiempo o espacio. Por ejemplo, número de nacimientos en un mes, número de correos electrónicos en una hora, número de glóbulos rojos en un volumen de sangre, etc.
- Los sucesos ocurren independientemente.
- El experimento produce, a largo plazo, el mismo número medio de sucesos puntuales λ en el intervalo considerado.

En estas circunstancias, la variable aleatoria X que mide el número de ocurrencias del suceso en el intervalo considerado 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, y se nota $X \sim P(\lambda)$, si su rango es $Ran(X) = \{0, 1, \ldots, \infty\}$, y su función de probabilidad vale

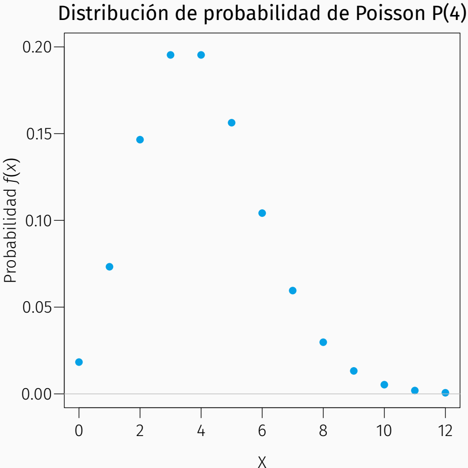
$$f(x) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

Obsérvese que λ es el número medio de sucesos en el intervalo unidad de referencia, y cambiará si se cambia la amplitud del intervalo.

Su media y varianza valen

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

En una ciudad hay 4 nacimientos diarios por término medio. La variable aleatoria X que mide el número de nacimientos diarios en la ciudad sigue un modelo de distribución de probabilidad de Poisson $X \sim P(4)$.



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 nacimientos es
$$f(5) = e^{-4}\frac{4^5}{5!} = 0.1563.$$
- La probabilidad de que un día se produzcan menos de 2 nacimientos 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 nacimiento es
$$P(X > 1) = 1 - P(X \leq 1) = 1 - F(1) = 1 - 0.0916 = 0.9084.$$

El modelo de distribución de Poisson surge a partir del modelo de distribución Binomial, cuando el número de repeticiones del ensayo tiende a infinito y la probabilidad de Éxito tiende a cero.

Teorema (Ley de los casos raros)

La distribución Binomial $X \sim B(n,p)$ tiende a la distribución de Poisson $P(\lambda)$, con $\lambda = n \cdot p$, cuando n tiende a infinito y p tiende a cero, es decir,

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

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 una muestra 50 personas, ¿cuál es la probabilidad de que haya más de 2 personas con reacción adversa?

La variable que mide el número de personas con reacción adversa en la muestra 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 utilizar la distribución de Poisson $P(50 \cdot 0.04) = P(2)$ para realizar los cálculos.

$$
\begin{aligned}
P(X > 2) &= 1 - P(X \leq 2) = 1 - f(0) - f(1) - f(2) = 1 - e^{-2}\frac{2^0}{0!} - e^{-2}\frac{2^1}{1!} - e^{-2}\frac{2^2}{2!} = \\
&= 1 - 5e^{-2} = 0.3233.
\end{aligned}
$$

Variables aleatorias continuas

Variables aleatorias continuas

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 la precisión 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.*

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 de probabilidad)

La *función de densidad de probabilidad* 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,

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

- Las probabilidad de que X tome un valor entre a y b es igual al área que queda por debajo de la función de densidad y el eje de abcisas limitada por a y b, es decir,

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)\, dx$$

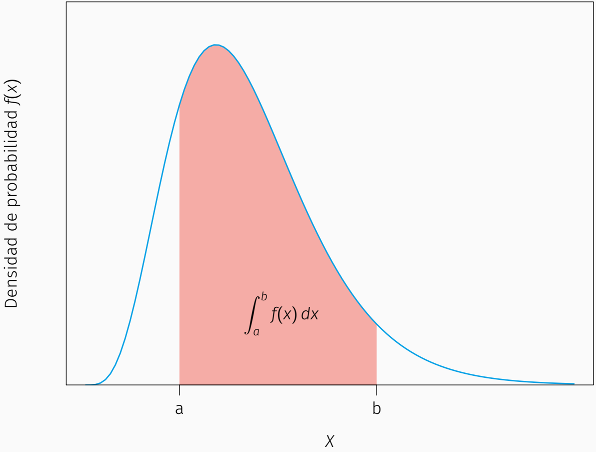
Al igual que para las variables discretas, también tiene sentido medir probabilidades acumuladas por debajo de un determinado valor.

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 a, es decir,

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

Cálculo de probabilidades como áreas



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

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.

Como es evidente que la función es no negativa, solo que da por comprobar que el área por debajo de ella es 1.

$$\int_{-\infty}^{\infty} f(x)\, dx = \int_{-\infty}^{0} f(x)\, dx + \int_{0}^{\infty} f(x)\, dx = \int_{-\infty}^{0} 0\, dx + \int_{0}^{\infty} e^{-x}\, dx =$$
$$= \left[-e^{-x}\right]_{0}^{\infty} = -e^{-\infty} + e^{0} = 1.$$

Calculemos la probabilidad de que la variable tome un valor entre 0 y 2.

$$P(0 \leq X \leq 2) = \int_{0}^{2} f(x)\, dx = \int_{0}^{2} e^{-x}\, dx = \left[-e^{-x}\right]_{0}^{2} = -e^{-2} + e^{0} = 0.8646.$$

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 rango de la variable.

Los más importantes son:

- Media o esperanza matemática:

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

- Varianza:

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

- Desviación típica:

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

Sea X una variable aleatoria con la siguiente función de densidad:

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

Su media es

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx = \int_{-\infty}^{0} xf(x)\,dx + \int_{0}^{\infty} xf(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} xe^{-x}\,dx =$$
$$= \left[-e^{-x}(1+x)\right]_0^{\infty} = 1.$$

y su varianza vale

$$\sigma^2 = \int_{-\infty}^{\infty} x^2 f(x)\,dx - \mu^2 = \int_{-\infty}^{0} x^2 f(x)\,dx + \int_{0}^{\infty} x^2 f(x)\,dx - \mu^2 =$$
$$= \int_{-\infty}^{0} 0\,dx + \int_{0}^{\infty} x^2 e^{-x}\,dx - \mu^2 = \left[-e^{-x}(x^2+2x+2)\right]_0^{\infty} - 1^2 = 2e^0 - 1 = 1.$$

Modelos de distribución continuos

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. Los más importantes son:

- 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 $U(a, b)$)

Una variable aleatoria continua X, sigue un modelo de distribución de probabilidad *uniforme*, y se nota $X \sim U(a, b)$, si su rango es $\mathrm{Ran}(X) = [a, b]$ y su función de densidad es

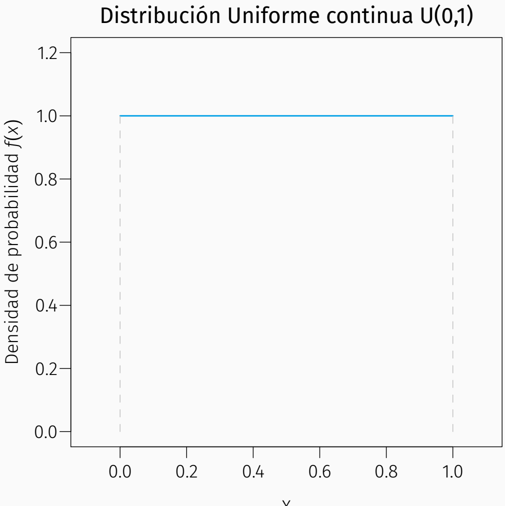
$$f(x) = \frac{1}{b-a} \quad \forall x \in [a, b]$$

Obsérvese que a y b son el mínimo y el máximo del rango respectivamente, y que la función de densidad es constante.

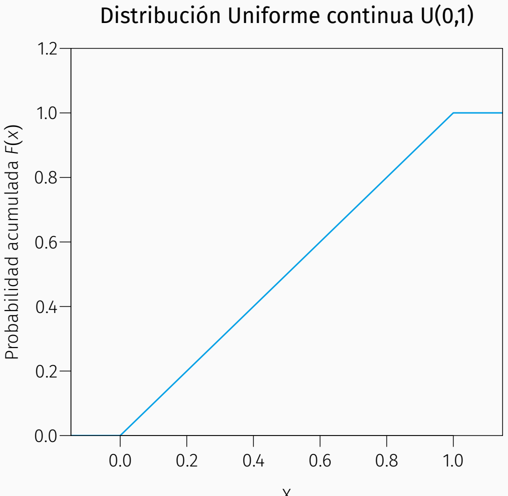
Su media y varianza valen

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

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



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



Cálculo de probabilidades con una Uniforme continua

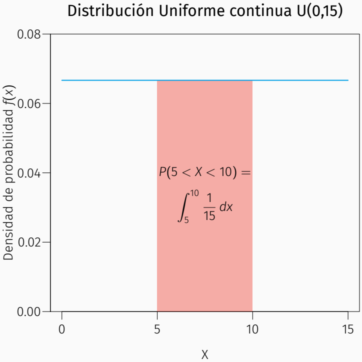
Ejemplo de espera de un autobús

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 equipobrable. Así pues, la probabilidad de esperar entre 5 y 10 minutos es

$$P(5 \leq X \leq 10) = \int_5^{10} \frac{1}{15}\, dx = \left[\frac{X}{15}\right]_5^{10} =$$

$$= \frac{10}{15} - \frac{5}{15} = \frac{1}{3}.$$



Además, el tiempo medio de espera será $\mu = \frac{0+15}{2} = 7.5$ minutos.

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 de probabilidad Normal $N(\mu, \sigma)$)

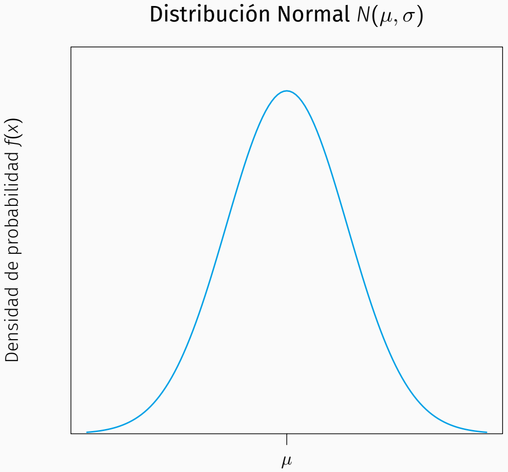
Una variable aleatoria continua X sigue un modelo de distribución *normal* de parámetros μ y σ, y se nota $X \sim N(\mu, \sigma)$, si su rango es $\mathbb{R}$ y su función de densidad vale

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

Los dos parámetros μ y σ son la media y la desviación típica de la población respectivamente.

Función de densidad de la Normal

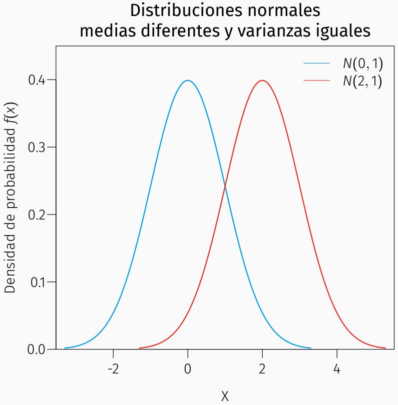
La gráfica de la función de densidad de la distribución normal tiene forma de una especie de campana, conocida como *campana de Gauss* (en honor a su descubridor).

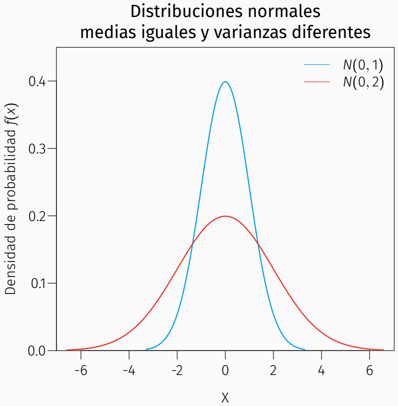


Función de densidad de la Normal

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

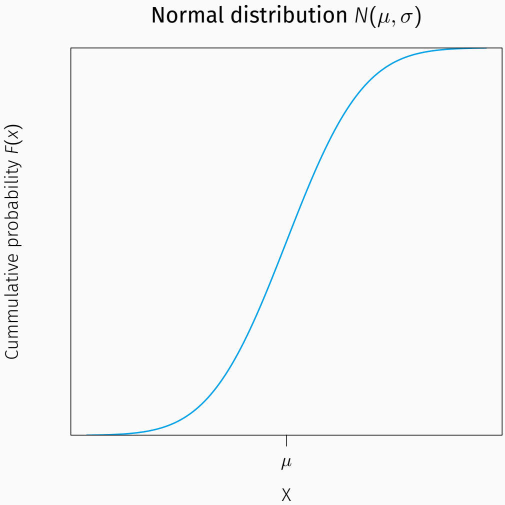
- La media μ determina dónde está centrada la campana.
- La desviación típica σ determina la anchura de la campana.





Función de distribución de la Normal

La gráfica de la función de distribución tiene forma de S.



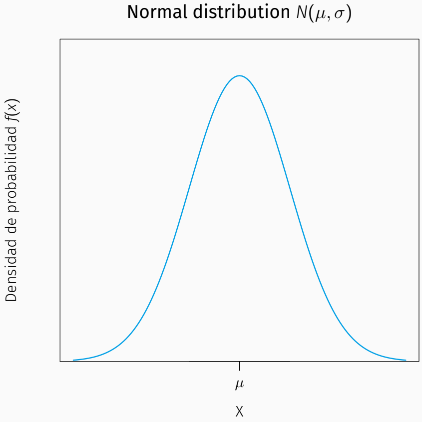
Propiedades de la distribución Normal

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

$$\mu = Me = Mo.$$

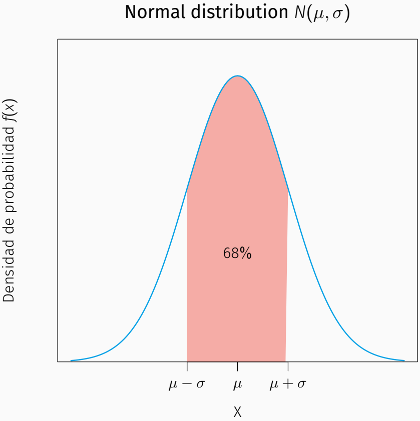
- Tiende asintóticamente a 0 cuando x tiende a $\pm\infty$.

Propiedades de la distribución Normal



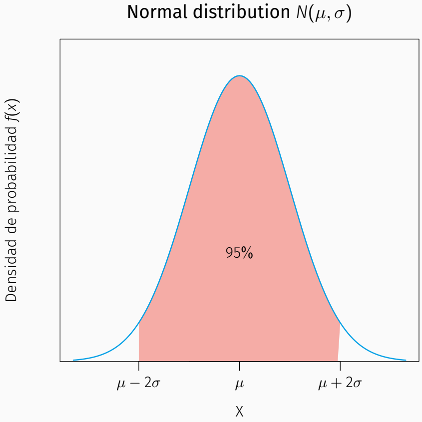
Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$



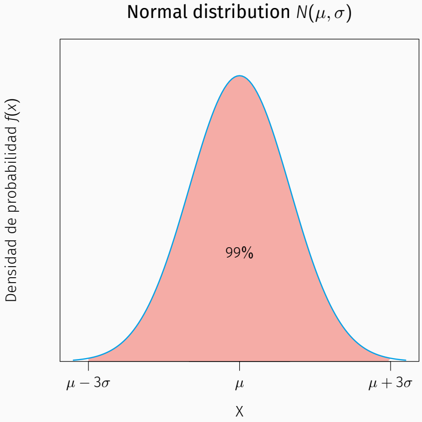
Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95,$$



Propiedades de la distribución Normal

$$P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68,$$
$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) = 0.95,$$
$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.99.$$



es decir, casi la totalidad de los individuos de la población presentarán

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.

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.