

Manual Básico de Estadística

Alfredo Sánchez Alberca (asalber@ceu.es)

Feb 2017

Departamento de Matemática Aplicada y Estadística
CEU San Pablo



CEU
*Universidad
San Pablo*




Términos de la licencia

Esta obra está bajo una licencia Reconocimiento – No comercial – Compartir bajo la misma licencia 2.5 España de Creative Commons. Para ver una copia de esta licencia, visite <http://creativecommons.org/licenses/by-nc-sa/4.0/es/>.

Con esta licencia eres libre de:

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

Bajo las siguientes condiciones:

-  **Reconocimiento.** Debe reconocer los créditos de la obra de la manera especificada por el autor o el licenciador (pero no de una manera que sugiera que tiene su apoyo o apoyan el uso que hace de su obra).
-  **No comercial.** No puede utilizar esta obra para fines comerciales.
-  **Compartir bajo la misma licencia.** Si altera o transforma esta obra, o genera una obra derivada, sólo puede distribuir la obra generada bajo una licencia idéntica a ésta.

- Al reutilizar o distribuir la obra, tiene que dejar bien claro los términos de la licencia de esta obra.
- Estas condiciones pueden no aplicarse si se obtiene el permiso del titular de los derechos de autor.
- Nada en esta licencia menoscaba o restringe los derechos morales del autor.

Índice general

1	Introducción a la Estadística	3
1.1	La estadística como herramienta científica	3
1.2	Población y muestra	3
1.3	Muestreo	7
1.4	Variables estadísticas	8
1.5	Fases del análisis estadístico	10
2	Distribución de frecuencias: Tabulación y gráficos	12
2.1	Distribución de frecuencias	12
2.2	Representaciones gráficas	15
2.3	Estadísticos muestrales	21
2.4	Estadísticos de posición	21
2.5	Estadísticos de dispersión	28
2.6	Estadísticos de forma	34
2.7	Transformaciones de variables	39
3	Regresión y Correlación	43
3.1	Distribución de frecuencias conjunta	43
3.2	Covarianza	46
3.3	Regresión	49
3.4	Recta de regresión	50
3.5	Correlación	53
3.6	Coeficientes de determinación y correlación	54
3.7	Regresión no lineal	57
3.8	Medidas de relación entre atributos	62
4	Teoría de la Probabilidad	66
4.1	Experimentos y sucesos aleatorios	66
4.2	Teoría de conjuntos	67
4.3	Definición de probabilidad	69
4.4	Probabilidad condicionada	71
4.5	Dependencia e independencia de sucesos	71
4.6	Teorema de la probabilidad total	73
4.7	Teorema de Bayes	74
4.8	Tests diagnósticos	75

1 Introducción a la Estadística

1.1 La estadística como herramienta científica

¿Qué es la estadística?

Definición 1 (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.

Pero, ¿por qué es necesaria la Estadística?

La variabilidad de nuestro mundo

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

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

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

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

1.2 Población y muestra

Población estadística

Definición 2 (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 3 (Tamaño poblacional). El número de individuos de una población se conoce como *tamaño poblacional* y se representa como N .

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

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

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

Inconvenientes en el estudio de la población

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

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

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

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

Muestra estadística

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

Definición 4 (Muestra). Una *muestra* es un subconjunto de la población.

Definición 5 (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.

Determinación del tamaño muestral

Muestra pequeña de los píxeles de una imagen

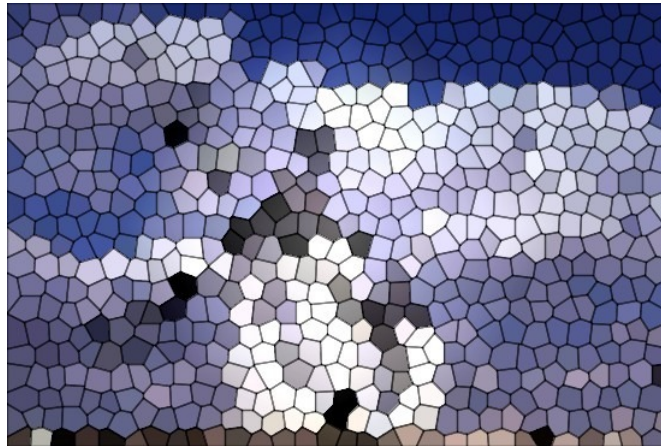
Example. Para entender a qué nos referimos cuando hablamos de un tamaño muestral suficiente para comprender lo que ocurre en la población, podemos utilizar el siguiente símil en que se trata de comprender el motivo que representa una fotografía.

Una fotografía digital está formada por multitud de pequeños puntitos llamados píxeles que se dispone en una enorme tabla de filas y columnas (cuantas más filas y columnas haya se habla de que la foto tiene más resolución). Aquí la población estaría formada por todos y cada uno de los píxeles que forman la foto. Por otro lado cada pixel tiene un color y es la variedad de colores a lo largo de los pixels la que permite formar la imagen de la fotografía.

¿Cuántos píxeles debemos tomar en una muestra para averiguar la imagen de la foto?

La respuesta depende de la variabilidad de colores en la foto. Si todos los píxeles de la foto son del mismo color, entonces un sólo pixel basta para desvelar la imagen. Pero, si la foto tiene mucha variabilidad de colores, necesitaremos muchos más píxeles en la muestra para descubrir el motivo de la foto.

¿Puedes averiguar el motivo de la foto?



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

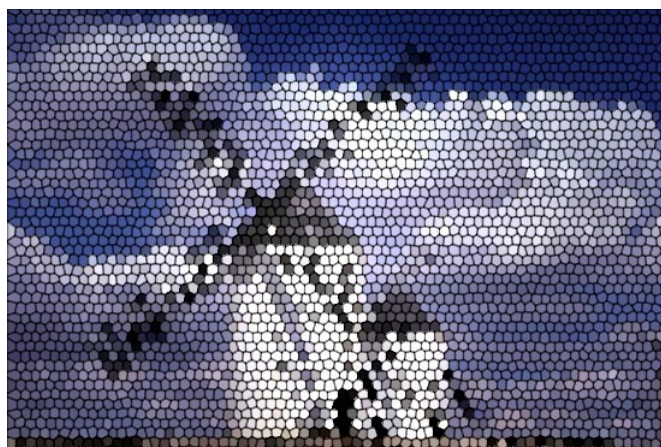
Determinación del tamaño muestral

Muestra mayor de los píxeles de una imagen

Seguramente no has podido averiguar el motivo de la fotografía, porque en este caso el número de píxeles que hemos tomado en la muestra es insuficiente para comprender toda la variabilidad de colores que hay en la foto.

La siguiente imagen contiene una muestra mayor de píxeles.

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



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

Determinación del tamaño muestral

Población completa de los píxeles de una imagen

Y aquí está la población completa



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

Tipos de razonamiento



Tipos de razonamiento

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

Características de la inducción: No garantiza la certeza de las conclusiones (si algo se cumple en la muestra, puede que no se cumpla en la población, así que ¡cuidado con las extrapolaciones!). 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.

1.3 Muestreo

Muestreo

Definición 6 (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.

Muestreo aleatorio simple

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

- Todos los individuos de la población tienen la misma probabilidad de ser elegidos para la muestra.
- La selección de individuos es con reemplazamiento, 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.

1.4 Variables estadísticas

Variables estadísticas y datos

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 7 (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.

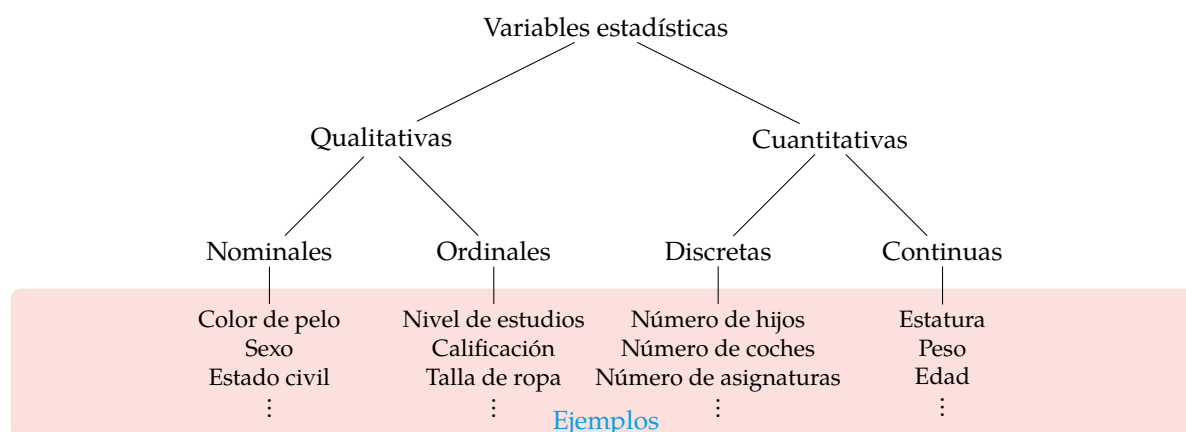


Variables estadísticas y atributos

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 identificar relaciones de asociación entre las variables.

La tabla de datos

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

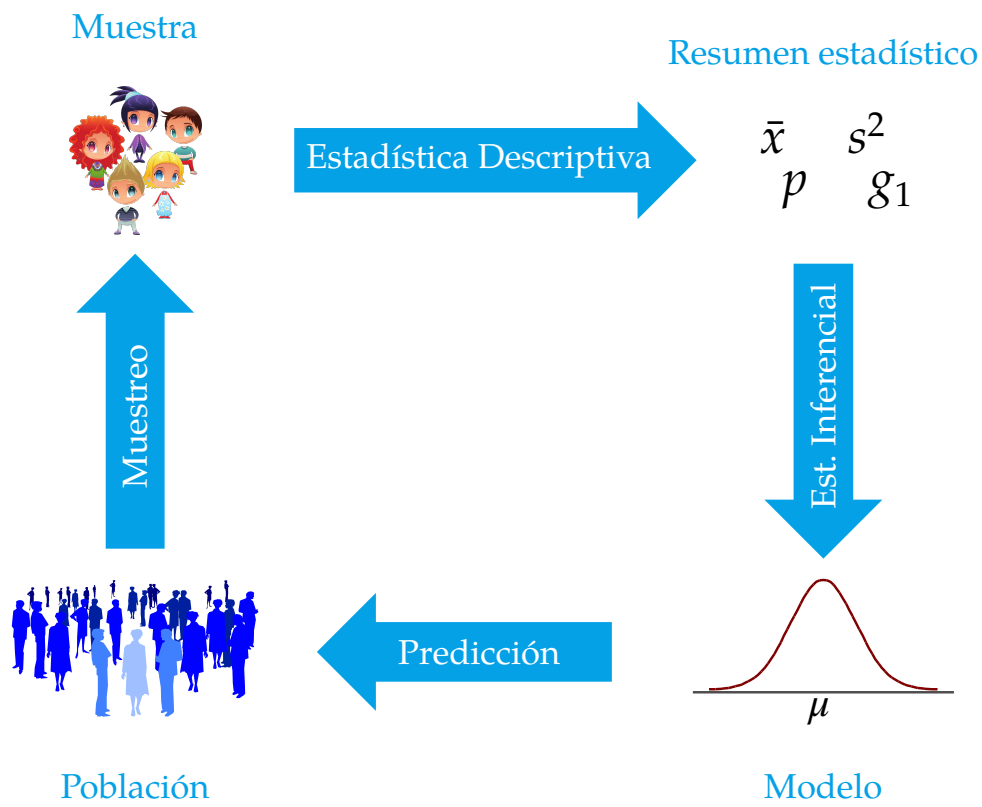
1.5 Fases del análisis estadístico

Fases del análisis estadístico

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



2 Distribución de frecuencias: Tabulación y gráficos

Estadística descriptiva

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!*

Clasificación de la muestra

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

Existen dos formas de clasificar estos valores:

Sin agrupar : Ordenar todos los valores obtenidos en la muestra de menor a mayor (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.

2.1 Distribución de frecuencias

Clasificación de la muestra

X =Estatura



Recuento de frecuencias

X =Estatura

**Frecuencias muestrales**

Definición 8 (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 .

$$N_i = n_1 + \dots + n_i$$

- **Frecuencia relativa acumulada F_i** : Es la proporción de valores en la muestra menores o iguales que x_i .

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

Tabla de frecuencias

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

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

Tabla de frecuencias*Ejemplo de datos sin agrupar*

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
Σ	25	1		

Tabla de frecuencias*Ejemplo de datos agrupados*

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
Σ	30	1		

Construcción de clases

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

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

- El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo \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.

Tabla de frecuencias*Ejemplo con un atributo*

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
Σ	30	1

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

2.2 Representaciones gráficas

Representaciones gráficas

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 variable, dibujadas en unos ejes cartesianos.

Habitualmente los 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 variable 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

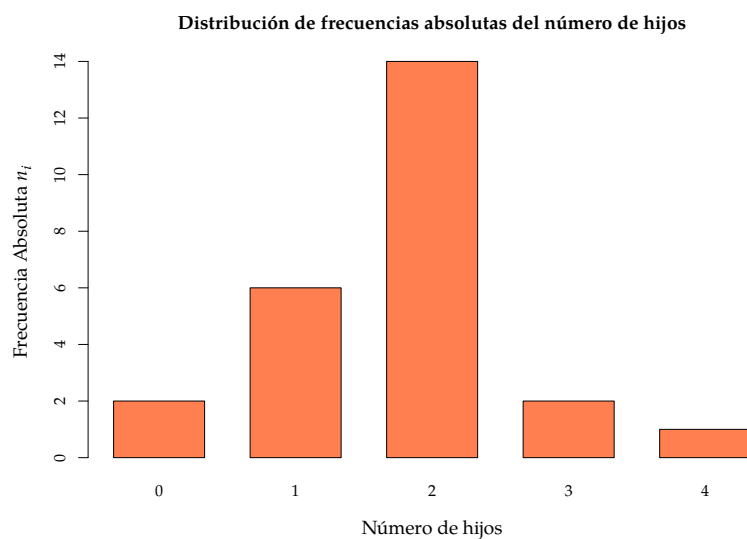


Diagrama de líneas o Polígono de frecuencias absolutas

Datos sin agrupar

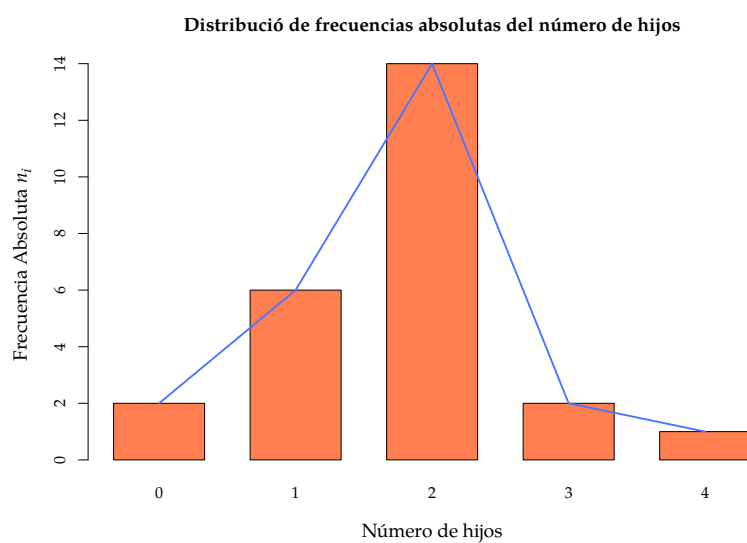


Diagrama de barras de frecuencias acumuladas

Datos sin agrupar

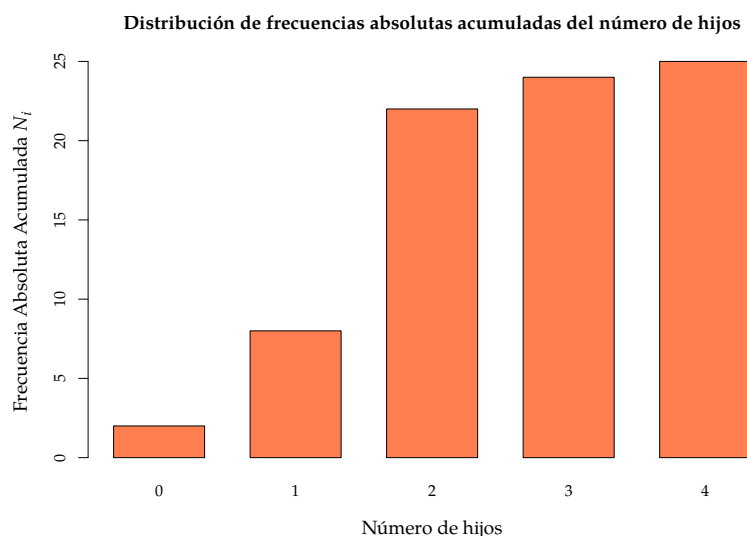
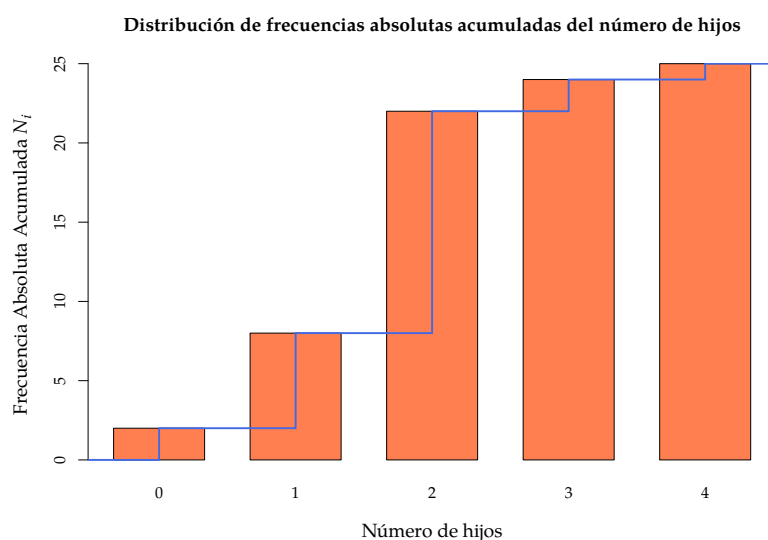


Diagrama de línea o polígono de frecuencias absolutas acumuladas

Datos sin agrupar



Histograma

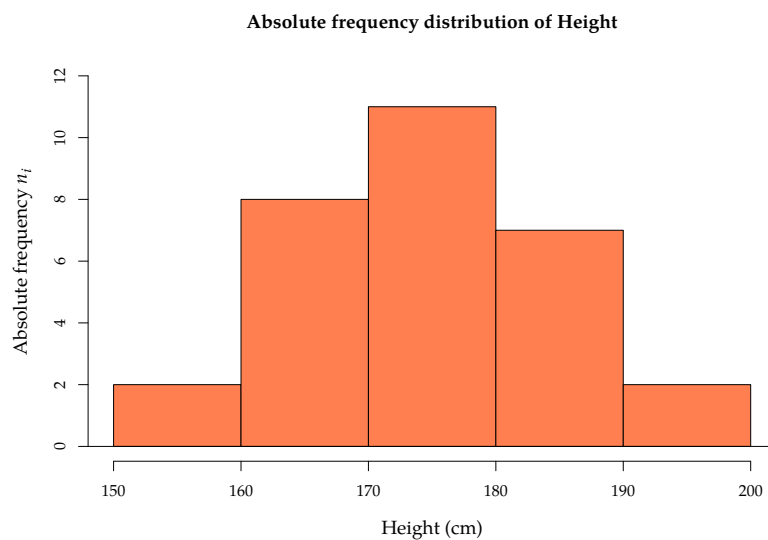
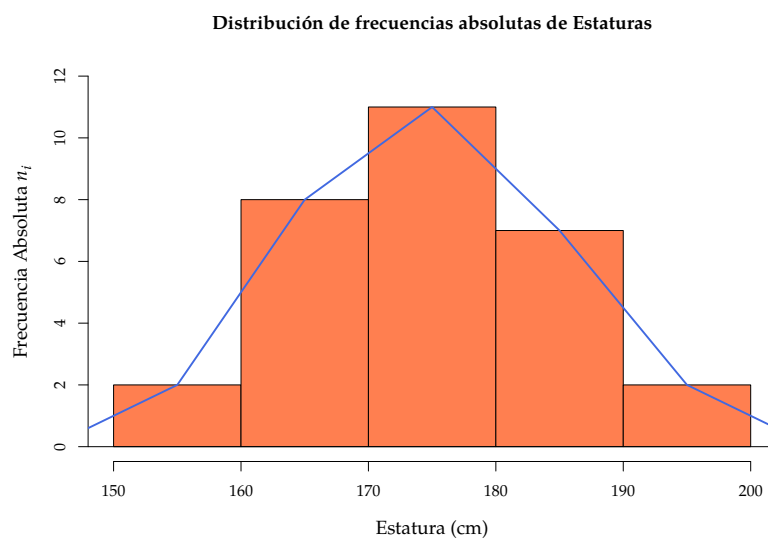
Un **histograma** es similar a un diagrama de barras pero para datos agrupados.

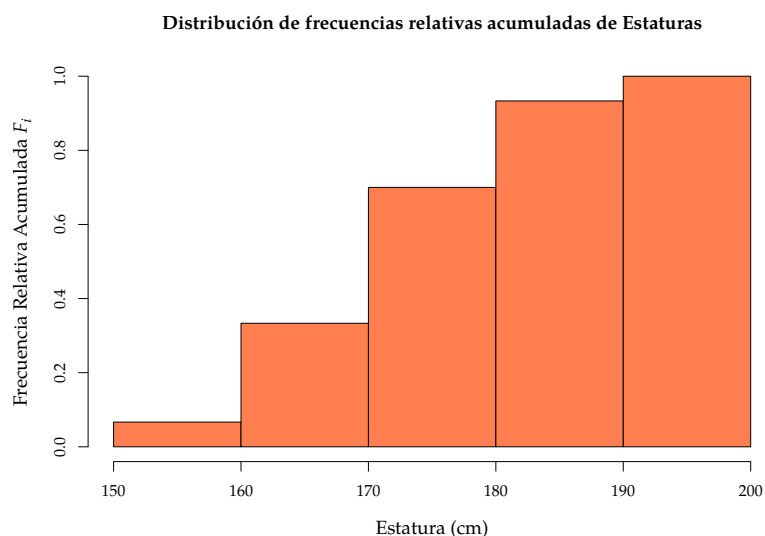
Habitualmente las clases o intervalos de agrupación se representan en el eje X, y las frecuencias en el eje Y.

Para cada clase se dibuja una barra de altura la correspondiente frecuencia. A diferencia del diagrama de barras, la anchura de 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.

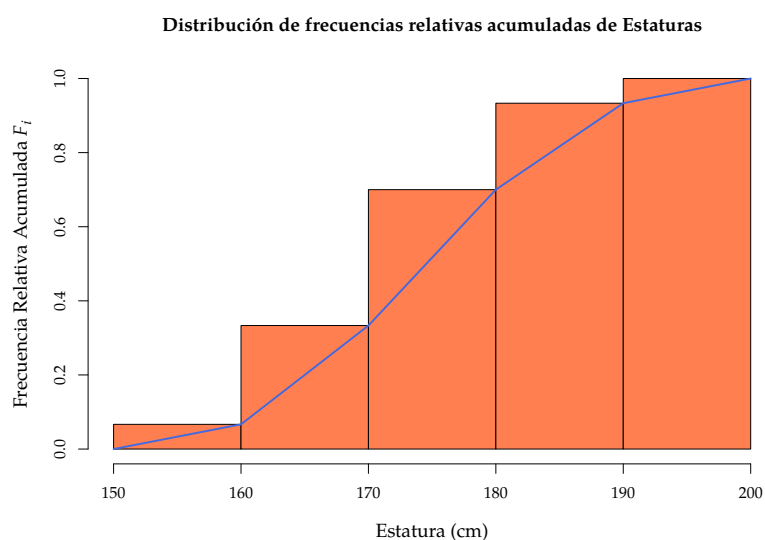
A veces se dibuja un polígono, conocido como **polígono de frecuencias**, uniendo los puntos más altos de cada barra con segmentos.

Histograma de frecuencias absolutas*Datos agrupados***Polígono de frecuencias absolutas***Datos agrupados***Histograma de frecuencias relativas acumuladas***Datos agrupados*



Polígono de frecuencias relativas acumuladas

Datos agrupados



El polígono de frecuencias acumuladas (absolutas o relativas) se conoce como **ojiva**.

Observe que en la ojiva se unen con segmentos los vértices superiores derechos de cada barra, en lugar de los centros, ya que no se consigue acumular la correspondiente frecuencia hasta el final del intervalo.

Diagrama de sectores

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

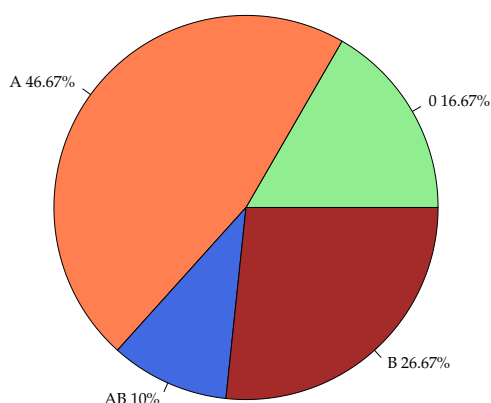
Cada porción se conoce como **sector** y su ángulo o área es proporcional a la correspondiente frecuencia del valor o categoría.

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).

Diagrama de sectores

Atributos

Distribución de frecuencias relativas de los grupos sanguíneos



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.
- Dejar el dato atípico si no es un error y cambiar el modelo de distribución teórico para ajustarse a los datos atípicos.

2.3 Estadísticos muestrales

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.

2.4 Estadísticos de posición

Estadísticos de posición

Pueden ser de dos tipos:

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

- Media aritmética
- Mediana
- Moda

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

- Cuantiles: Cuartiles, Deciles, Percentiles.

Media aritmética

Definición 9 (Media aritmética muestral \bar{x}). La *media aritmética muestral* de una variable X es la suma de los valores observados en la muestra dividida por el tamaño muestral

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

A partir de la tabla de frecuencias puede calcularse como:

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

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

¡Ojo! No puede calcularse para atributos.

Cálculo de la media aritmética

Ejemplo con datos no agrupados

En el ejemplo anterior del número de hijos tenemos

$$\bar{x} = \frac{1 + 2 + 4 + 2 + 2 + 2 + 3 + 2 + 1 + 1 + 0 + 2 + 2}{25} + \frac{0 + 2 + 2 + 1 + 2 + 2 + 3 + 1 + 2 + 2 + 1 + 2}{25} = \frac{44}{25} = 1.76 \text{ hijos.}$$

o bien, desde la tabla de frecuencias

x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
0	2	0.08	0	0
1	6	0.24	6	0.24
2	14	0.56	28	1.12
3	2	0.08	6	0.24
4	1	0.04	4	0.16
Σ	25	1	44	1.76

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

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

Cálculo de la media aritmética

Ejemplo con datos agrupados

En el ejemplo anterior de las estaturas se tiene

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

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

X	x_i	n_i	f_i	$x_i n_i$	$x_i f_i$
(150, 160]	155	2	0.07	310	10.33
(160, 170]	165	8	0.27	1320	44.00
(170, 180]	175	11	0.36	1925	64.17
(180, 190]	185	7	0.23	1295	43.17
(190, 200]	195	2	0.07	390	13
Σ		30	1	5240	174.67

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

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

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 10 (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:

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

Mediana

Definición 11 (Mediana muestral Me). La *mediana muestral* de una variable X es el valor de la variable que, una vez ordenados los valores de la muestra de menor a mayor, deja el mismo número de valores por debajo y por encima de él.

La mediana cumple $N_{Me} = n/2$ y $F_{Me} = 0.5$.

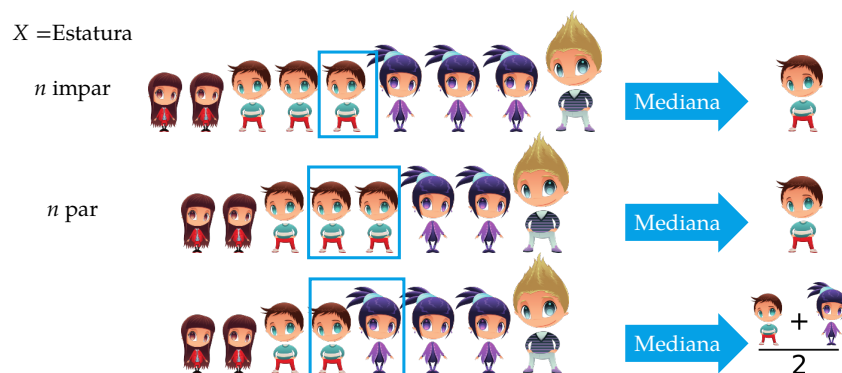
El cálculo de la mediana se realiza de forma distinta según se hayan agrupado los datos o no.

¡Ojo! No puede calcularse para atributos nominales.

Cálculo de la mediana con datos no agrupados

Con datos no agrupados pueden darse varios casos:

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



Cálculo de la mediana

Ejemplo con datos no agrupados

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

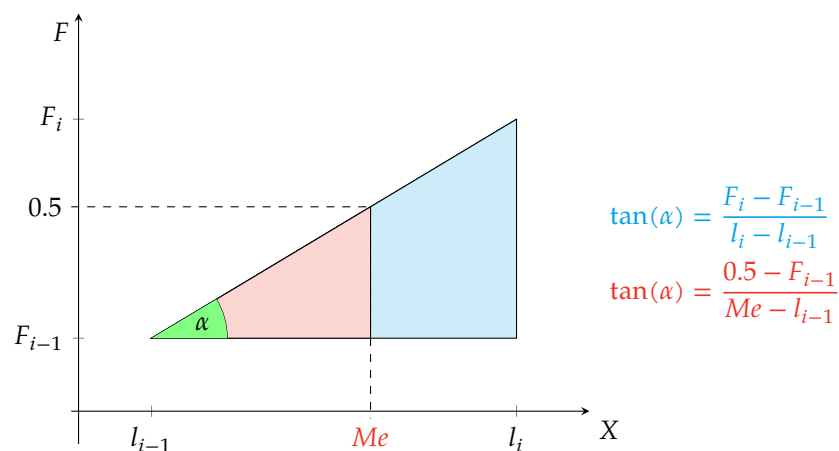
0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3, 3, 4

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
2	14	0.56	22	0.88
3	2	0.08	24	0.96
4	1	0.04	25	1
Σ	25	1		

Cálculo de la mediana con datos agrupados

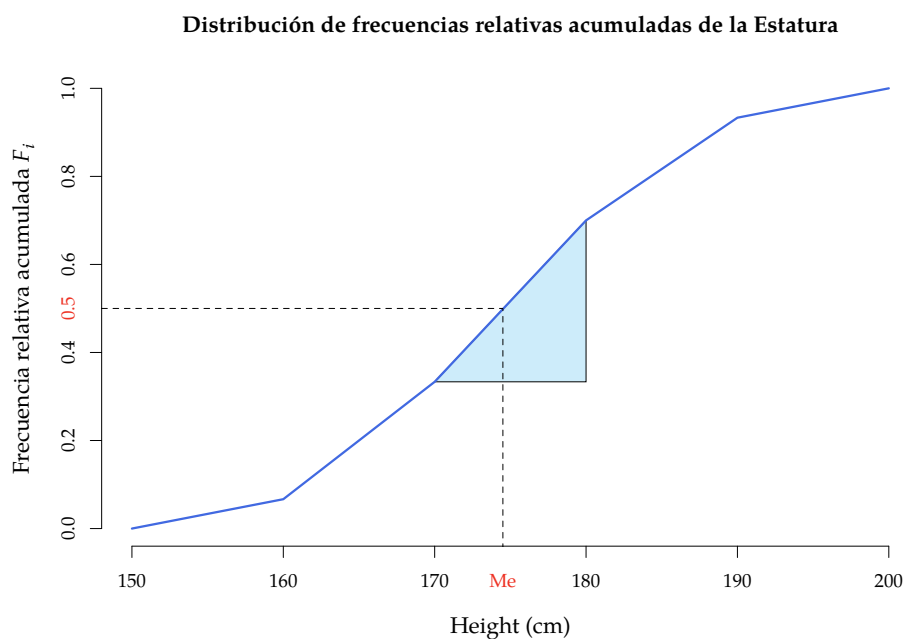


$$Me = l_i + \frac{0.5 - F_{i-1}}{F_i - F_{i-1}}(l_i - l_{i-1}) = l_i + \frac{0.5 - F_{i-1}}{f_i}a_i$$

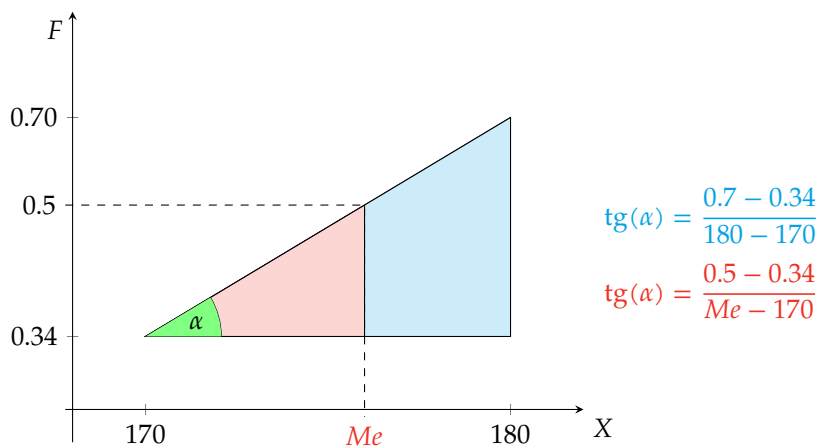
Cálculo de la mediana

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].



Interpolación en el polígono de frecuencias absolutas acumuladas



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

Moda

Definición 12 (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
2	14
3	2
4	1

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

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

¿Qué estadístico de tendencia central usar?

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

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

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

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

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

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

¿Qué representante de la muestra tomarías?

Cuantiles

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

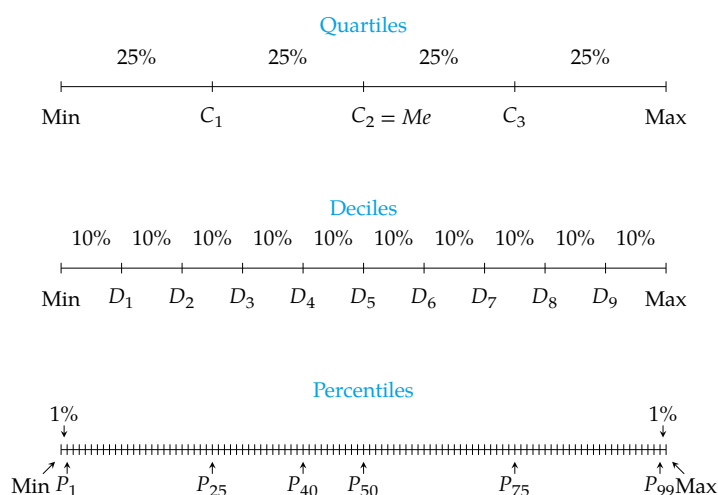
Los más utilizados son:

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

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

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

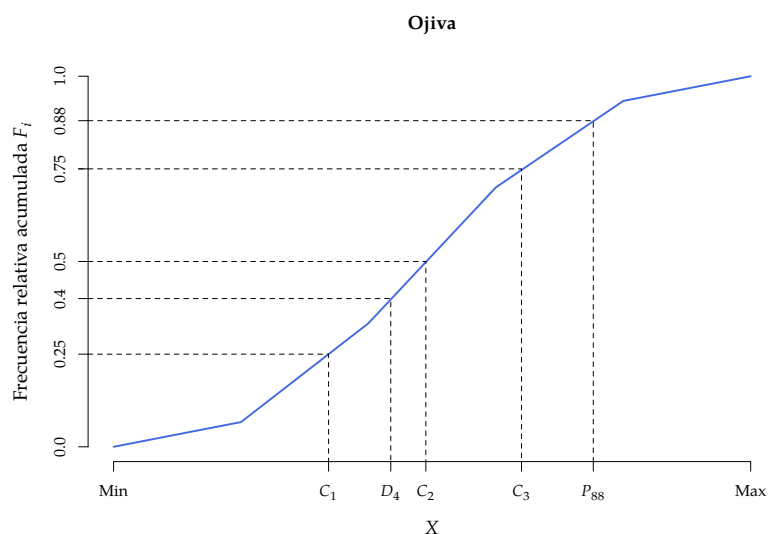
Cuantiles



Observese que hay una correspondencia entre los cuartiles, deciles 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 cuartil.



Cálculo de los cuantiles

Ejemplo con datos no agrupados

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

x_i	F_i
0	0.08
1	0.32
2	0.88
3	0.96
4	1

$$F_{C_1} = 0.25 \Rightarrow C_1 = 1 \text{ hijos,}$$

$$F_{C_2} = 0.5 \Rightarrow C_2 = 2 \text{ hijos,}$$

$$F_{C_3} = 0.75 \Rightarrow C_3 = 2 \text{ hijos,}$$

$$F_{D_3} = 0.3 \Rightarrow D_3 = 1 \text{ hijos,}$$

$$F_{P_{92}} = 0.92 \Rightarrow P_{92} = 3 \text{ hijos.}$$

2.5 Estadísticos de dispersión

Estadísticos de dispersión

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.

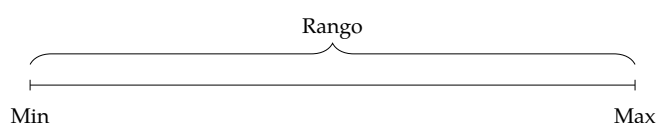
- Desviación Típica.
- Coeficiente de Variación.

Recorrido

Definición 13 (Recorrido muestral Re). El *recorrido muestral* de una variable X se define como la diferencia entre el máximo y el mínimo de los valores en la muestra.

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

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

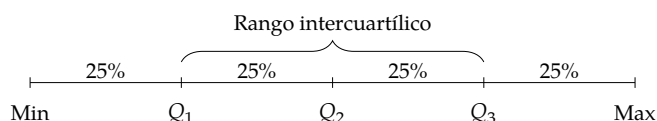


Rango intercuartílico

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

Definición 14 (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.

Diagrama de caja y bigotes

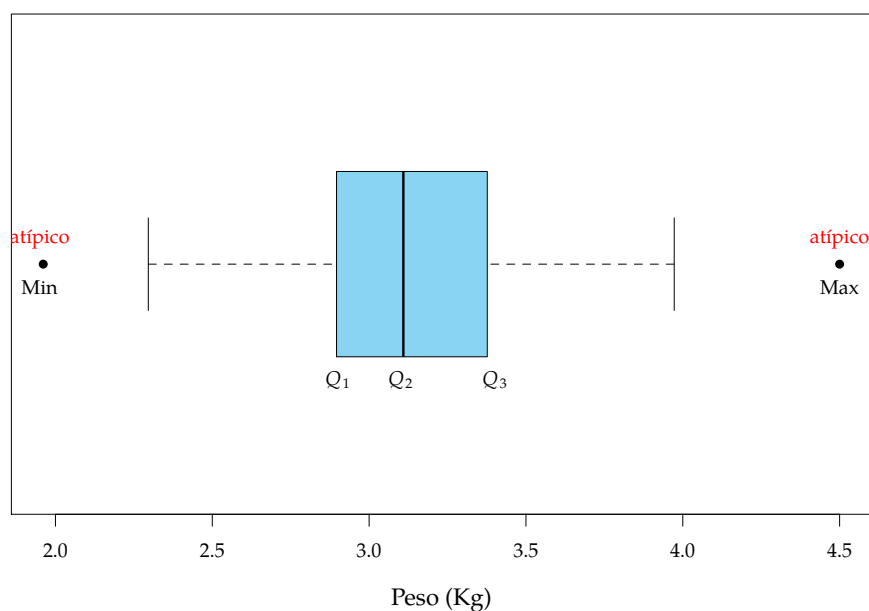
La dispersión de una variable suele representarse gráficamente mediante un **diagrama de caja y bigotes**, que consiste en una caja sobre un eje X donde el borde inferior de la caja es el primer cuartil, y el borde superior el tercer cuartil, y por tanto, la anchura de la caja es el rango intercuartílico.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

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

Diagrama de caja y bigotes

Ejemplo con pesos de recién nacidos

Diagrama de caja y bigotes del peso de recién nacidos



Construcción del diagrama de caja y bigotes

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

$$v_1 = C_1 - 1.5RI$$

$$v_2 = C_3 + 1.5RI$$

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

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

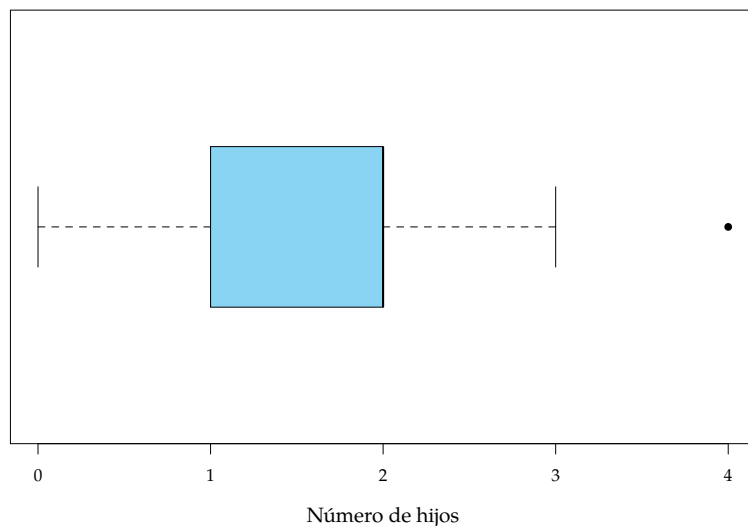
Construcción del diagrama de caja y bigotes

Ejemplo del número de hijos

1. Calcular los cuartiles: $C_1 = 1$ hijos y $C_3 = 2$ hijos.
2. Dibujar la caja.
3. Calcular las vallas: $v_1 = 1 - 1.5 * 1 = -0.5$ y $v_2 = 2 + 1.5 * 1 = 3.5$.
4. Dibujar los bigotes: $b_1 = 0$ hijos y $b_2 = 3$ hijos.

5. Dibujar los datos atípicos: 4 hijos.

Diagrama de caja y bigotes del número de hijos

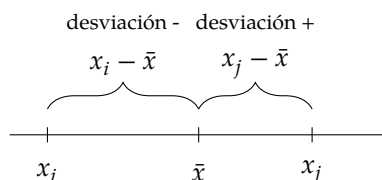


ξ

Desviaciones respecto de la media

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

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



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

Varianza y desviación típica

Definición 15 (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:

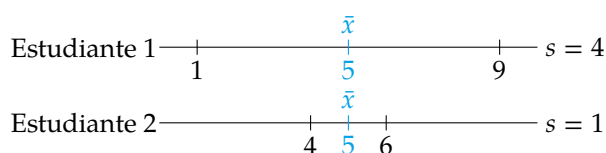
Definición 16 (Desviación típica s). La *desviación típica muestral* de una variable X se define como la raíz cuadrada positiva de su varianza muestral.

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

Interpretación de la varianza y la desviación típica

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

Desviación típica pequeña \Rightarrow *Media representativa*
Desviación típica grande \Rightarrow *Media no representativa*



¿En qué caso es más representativa la media?

Cálculo de la varianza y la desviación típica

Ejemplo con datos no agrupados

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

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

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{96}{25} - 1.76^2 = 0.7424 \text{ hijos}^2.$$

Y la desviación típica es $s = \sqrt{0.7424} = 0.8616$ hijos.

Comparado este valor con el recorrido, que va de 0 a 4 hijos se observa que no es demasiado grande por lo que se puede concluir que no hay mucha dispersión y en consecuencia la media de 1.76 hijos representa bien a los matrimonios de la muestra.

Cálculo de la varianza y la desviación típica

Ejemplo con datos agrupados

En el ejemplo de las estaturas, al ser datos agrupados, el cálculo se realiza igual que antes pero

tomando como valores de la variable las marcas de clase.

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

$$s^2 = \frac{\sum x_i^2 n_i}{n} - \bar{x}^2 = \frac{918350}{30} - 174.67^2 = 102.06 \text{ cm}^2.$$

Y la desviación típica es $s = \sqrt{102.06} = 10.1 \text{ cm}$.

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

Coefficiente de variación

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 17 (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.

El coeficiente de variación es muy útil para comparar la dispersión de distribuciones de variables diferentes, incluso si las variables tienen unidades diferentes.

¡Ojo! No tiene sentido cuando la media muestral vale 0 o valores próximos.

Coefficiente de variación

Ejemplo

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

$$cv = \frac{s}{|\bar{x}|} = \frac{0.8616}{|1.76|} = 0.49.$$

En el caso de las estaturas, como $\bar{x} = 174.67 \text{ cm}$ y $s = 10.1 \text{ cm}$, 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.

2.6 Estadísticos de forma

Estadísticos de forma

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 *Coefficiente 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 *Coefficiente de Apuntamiento o Curtosis*.

Coefficiente de asimetría

Definición 18 (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.

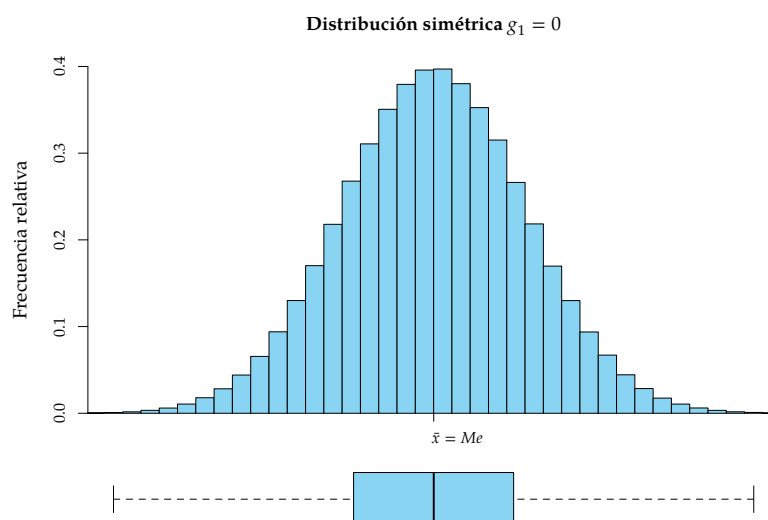
$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{\sum (x_i - \bar{x})^3 f_i}{s^3}$$

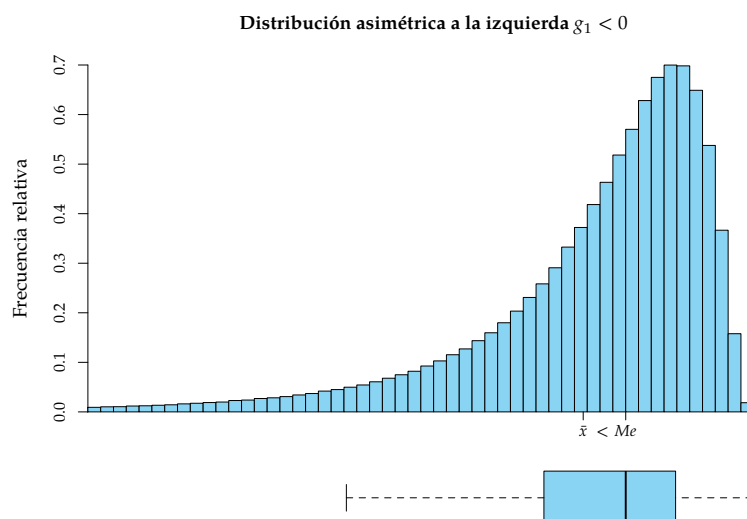
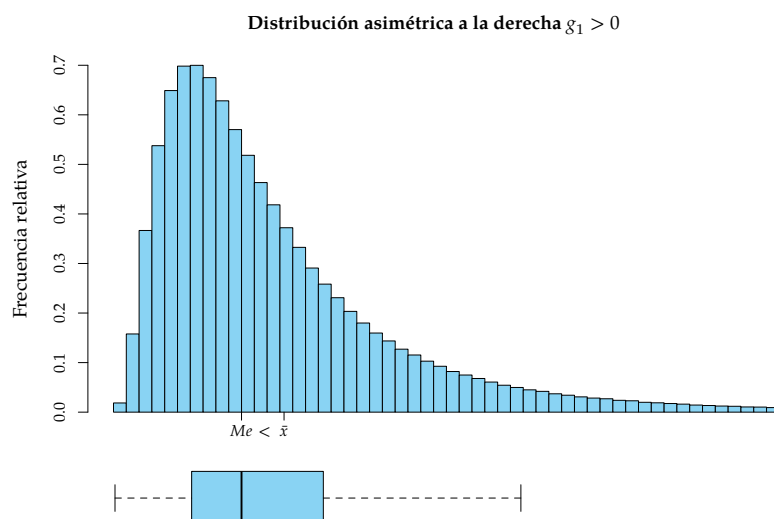
Mide el grado de simetría de los valores de la muestra con respecto a la media muestral, es decir, cuantos valores de la muestra están por encima o por debajo de la media y cómo de alejados de esta.

- $g_1 = 0$ 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, pero los valores mayores están más alejados de ella (asimétrica a la derecha).

Coefficiente de asimetría

Ejemplo de distribución simétrica



Coeficiente de asimetría*Ejemplo de distribución asimétrica hacia la izquierda***Coeficiente de asimetría***Ejemplo de distribución asimétrica hacia la derecha***Cálculo del coeficiente de asimetría***Ejemplo con datos agrupados*

Siguiendo con el ejemplo de las estaturas, podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con los cubos de las desviaciones a la media

$\bar{x} = 174.67$ cm:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^3 n_i$
(150, 160]	155	2	-19.67	-15221.00
(160, 170]	165	8	-9.67	-7233.85
(170, 180]	175	11	0.33	0.40
(180, 190]	185	7	10.33	7716.12
(190, 200]	195	2	20.33	16805.14
Σ		30		2066.81

$$g_1 = \frac{\sum (x_i - \bar{x})^3 n_i / n}{s^3} = \frac{2066.81/30}{10.1^3} = 0.07.$$

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

Coefficiente de apuntamiento o curtosis

Definición 19 (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.

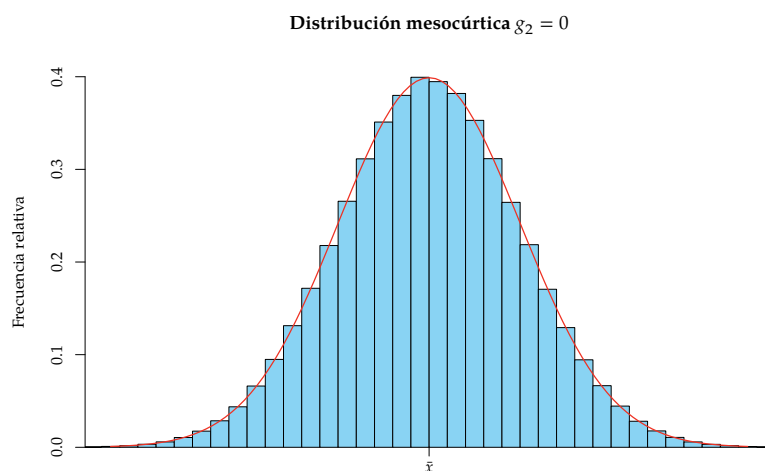
$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{\sum (x_i - \bar{x})^4 f_i}{s^4} - 3$$

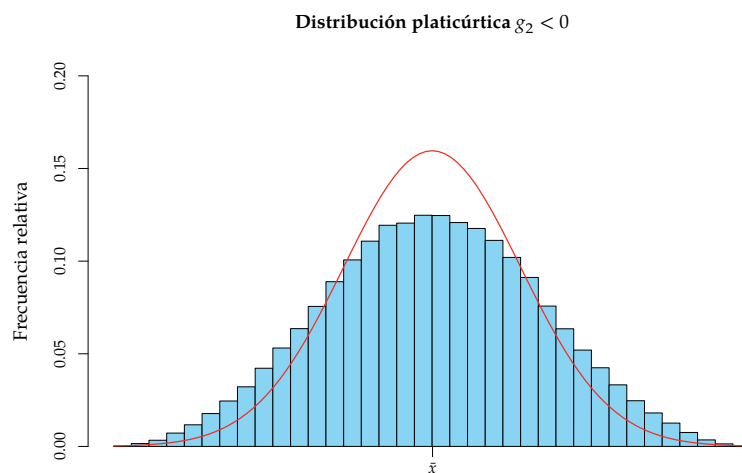
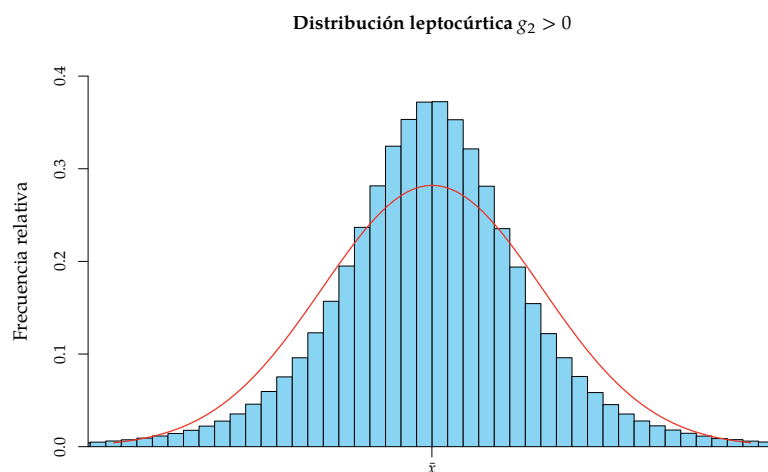
El coeficiente de apuntamiento mide la concentración de valores en torno a la media y la longitud de las colas de la distribución. Se toma como referencia la distribución normal

- $g_2 = 0$ 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 normal (*leptocúrtica*).

Coefficiente de apuntamiento o curtosis

Ejemplo de distribución mesocúrtica



Coefficiente de apuntamiento o curtosis*Ejemplo de distribución platicúrtica***Coefficiente de apuntamiento o curtosis***Ejemplo de distribución leptocúrtica***Cálculo del coeficiente de apuntamiento***Ejemplo con datos agrupados*

De nuevo para el ejemplo de las estaturas podemos calcular el coeficiente de asimetría a partir de la tabla de frecuencias añadiendo una nueva columna con las desviaciones a la media $\bar{x} = 174.67$ cm

elevadas a la cuarta:

X	x_i	n_i	$x_i - \bar{x}$	$(x_i - \bar{x})^4 n_i$
(150, 160]	155	2	-19.67	299396.99
(160, 170]	165	8	-9.67	69951.31
(170, 180]	175	11	0.33	0.13
(180, 190]	185	7	10.33	79707.53
(190, 200]	195	2	20.33	341648.49
Σ		30		790704.45

$$g_2 = \frac{\sum (x_i - \bar{x})^4 n_i / n}{s^4} - 3 = \frac{790704.45 / 30}{10.1^4} - 3 = -0.47.$$

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

Interpretación de los coeficientes de asimetría y apuntamiento

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

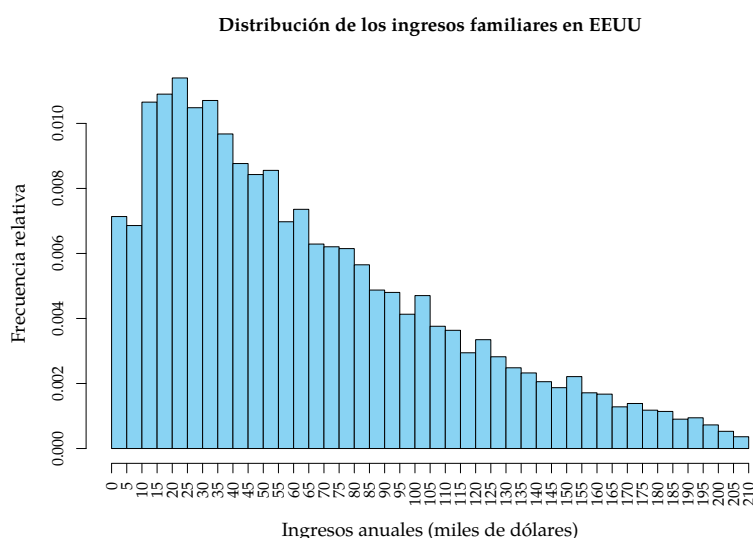
Las poblaciones normales se caracterizan por ser simétricas y mesocúrticas, de manera que, tanto el coeficiente de asimetría como el de apuntamiento pueden utilizarse para contrastar si los datos de la muestra provienen de una población normal.

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

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

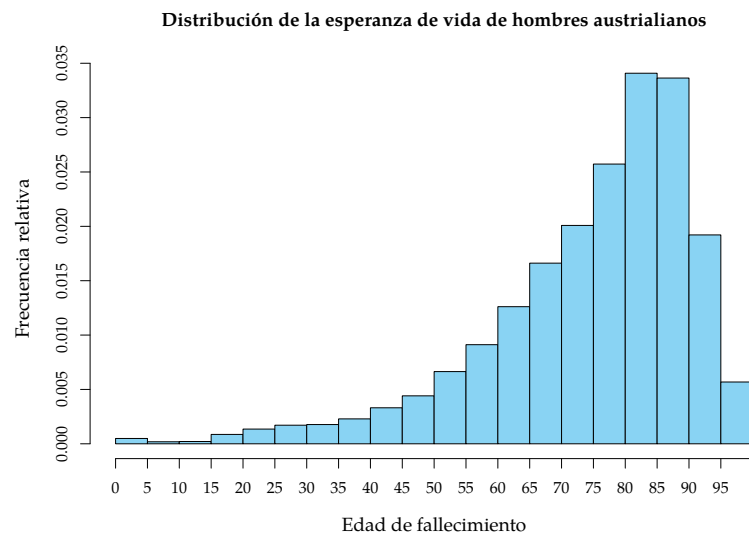
Distribución asimétrica a la derecha no normal

Ingresos por familia



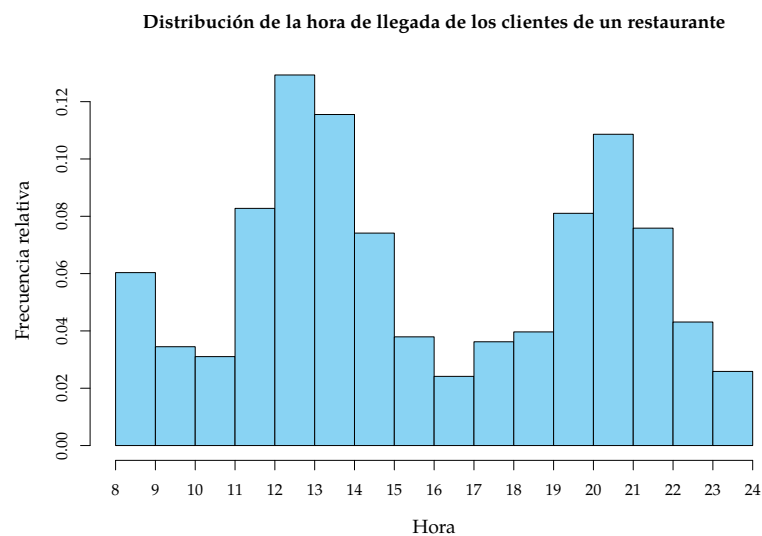
Distribución asimétrica a la izquierda no normal

Edad de fallecimiento



Distribución bimodal no normal

Hora de llegada de los clientes de un restaurante



2.7 Transformaciones de variables

Transformaciones de variables

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

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

1.75m, 1.65m, 1.80m,

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

175cm, 165cm, 180cm,

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

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

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:

$$Y = 100X - 165$$

Transformaciones lineales

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

$$Y = a + bX.$$

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

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

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

Transformación de tipificación y puntuaciones típicas

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

Definición 20 (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 tipificación

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

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 número de desviaciones típicas que dista de la media cada observación, lo cual es útil para comparar variables con distintas unidades.

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

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

Transformación de tipificación y puntuaciones típicas

Ejemplo

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

Alumno:	1	2	3	4	5		
X :	2	5	4	8	6	$\bar{x} = 5$	$s_x = 2$
Y :	1	9	8	5	2	$\bar{y} = 5$	$s_y = 3.16$

¿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:

$$\begin{array}{rcccccc} X : & -1.5 & 0 & -0.5 & 1.5 & 0.5 \\ Y : & -1.26 & 1.26 & 0.95 & 0 & -0.95 \end{array}$$

Es decir, el alumno que tiene un 8 en X está 1.5 veces la desviación típica por encima de la media de su grupo, mientras que el alumno que tiene un 8 en Y sólo está 0.95 desviaciones típicas por encima de su media. Así pues, el primer alumno tuvo un rendimiento superior al segundo.

Transformación de tipificación y puntuaciones típicas

Ejemplo

Siguiendo con el ejemplo anterior

¿Cuál es el mejor alumno?

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

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

El mejor alumno sería el segundo.

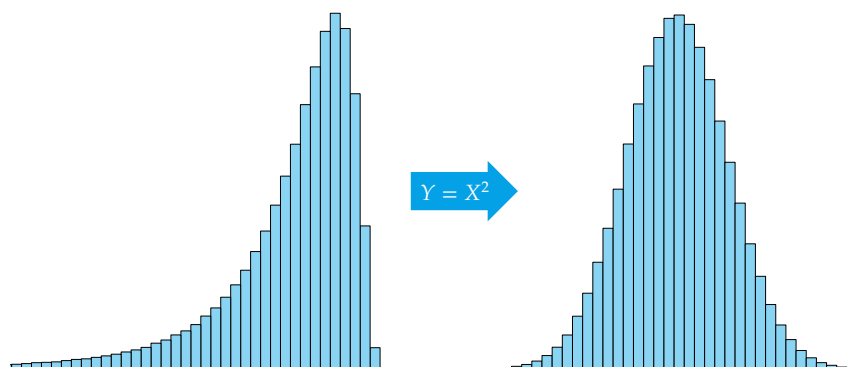
Pero si se considera el rendimiento relativo tomando las puntuaciones típicas se tiene:

Alumno:	1	2	3	4	5
X :	-1.5	0	-0.5	1.5	0.5
Y :	-1.26	1.26	0.95	0	-0.95
Σ	-2.76	1.26	0.45	1.5	-0.45

Y el mejor alumno sería el cuarto.

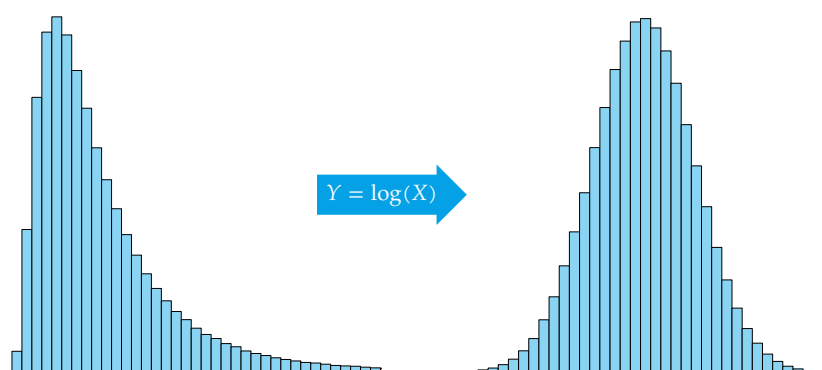
Transformaciones no lineales

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



Transformaciones no lineales

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



Variables clasificadoras o factores

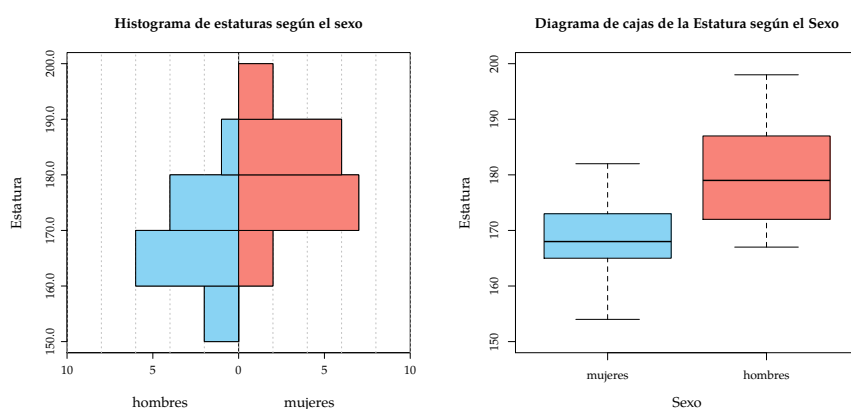
En ocasiones interesa describir el comportamiento de una variable, no para toda la muestra, sino para distintos grupos de individuos correspondientes a las categorías de otra variable conocida como **variable clasificadora** o **factor**.

Variables clasificadoras

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



3 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, \dots, X_n llamadas **variables independientes**.

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

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

3.1 Distribución de frecuencias conjunta

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 21 (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 proporción 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 frecuencias acumuladas.

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	\dots	y_j	\dots	y_q
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1q}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_i	n_{i1}	\dots	n_{ij}	\dots	n_{iq}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
x_p	n_{p1}	\dots	n_{pj}	\dots	n_{pq}

Distribución de frecuencias bidimensional

Ejemplo con estaturas y pesos

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).

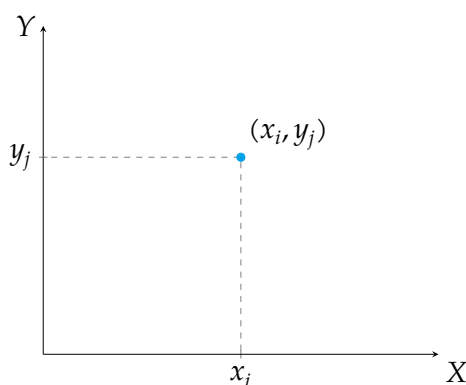
La tabla de frecuencias bidimensional es

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.

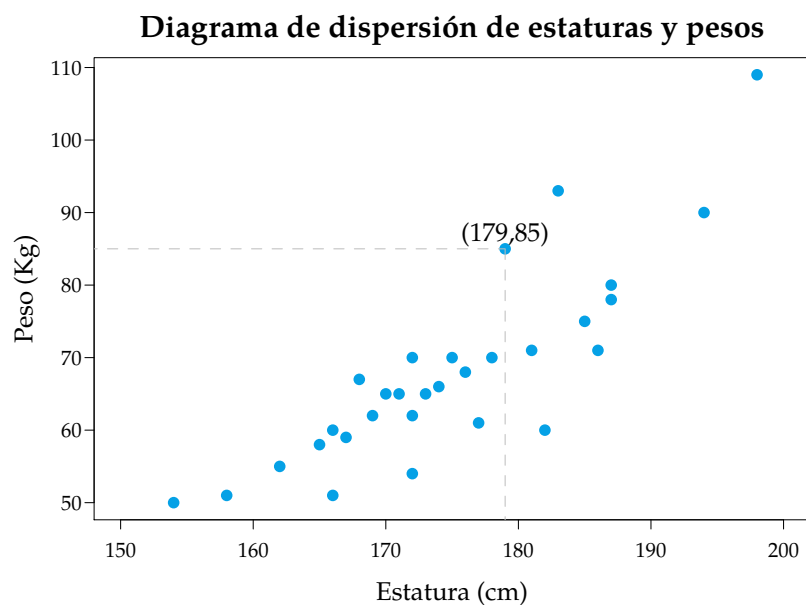
Habitualmente la variable independiente se representa en el eje X y la variable dependiente en el eje Y. Por cada par de valores (x_i, y_j) en la muestra se dibuja un punto en el plano con esas coordenadas.



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

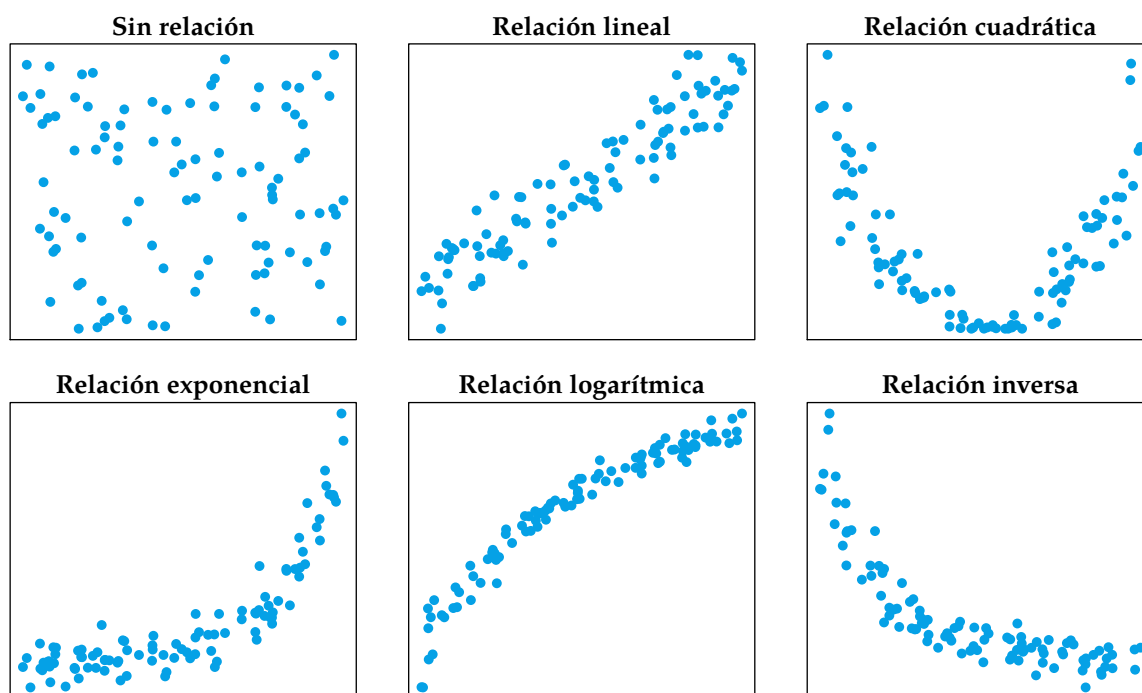
¡Ojo! No tiene sentido cuando alguna de las variables es un atributo.

Diagrama de dispersión



Interpretación del diagrama de dispersión

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



Distribuciones marginales

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

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

$X \backslash Y$	y_1	\dots	y_j	\dots	y_q	n_x
x_1	n_{11}	\dots	n_{1j}	\dots	n_{1q}	n_{x_1}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_i	n_{i1}	$\rightarrow +$	n_{ij}	$\rightarrow +$	n_{iq}	n_{x_i}
\vdots	\vdots	\vdots	$\downarrow +$	\vdots	\vdots	\vdots
x_p	n_{p1}	\dots	n_{pj}	\dots	n_{pq}	n_{x_p}
n_y	n_{y_1}	\dots	n_{y_j}	\dots	n_{y_q}	n

Distribuciones marginales

Ejemplo con estaturas y pesos

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

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

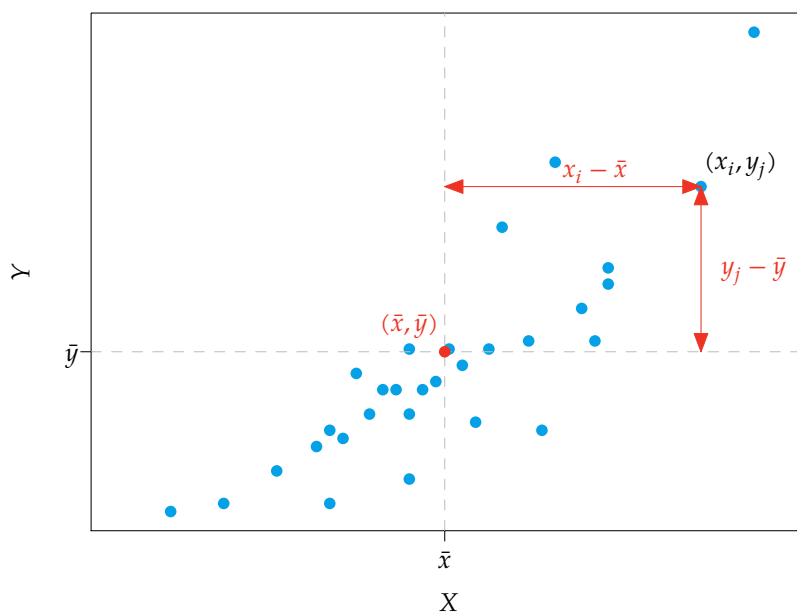
y los estadísticos correspondientes son

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

3.2 Covarianza

Desviaciones respecto de las medias

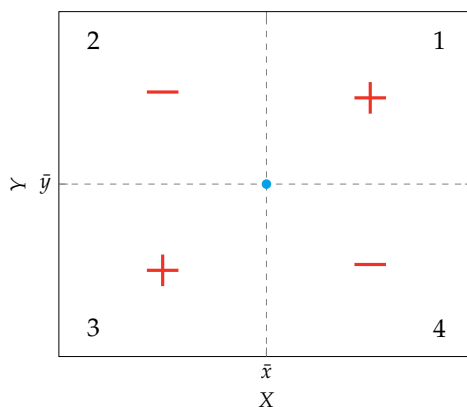
Para analizar la relación entre dos variables cuantitativas es importante hacer un estudio conjunto de las desviaciones respecto de la media de cada variable.



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:

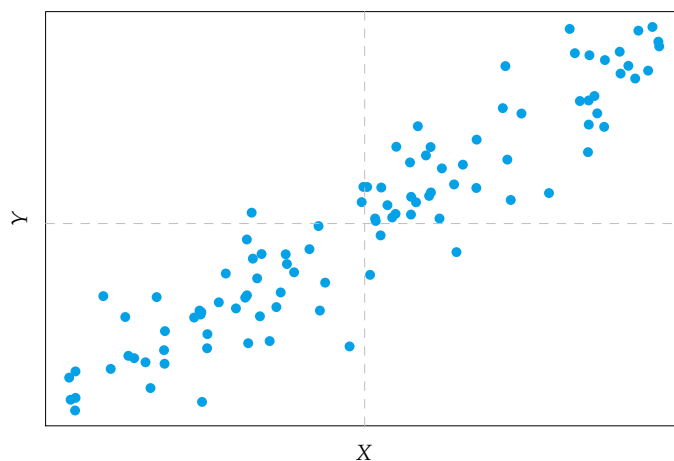
Cuadrante	$(x_i - \bar{x})$	$(y_j - \bar{y})$	$(x_i - \bar{x})(y_j - \bar{y})$
1	+	+	+
2	-	+	-
3	-	-	+
4	+	-	-



Estudio de las desviaciones respecto de las medias

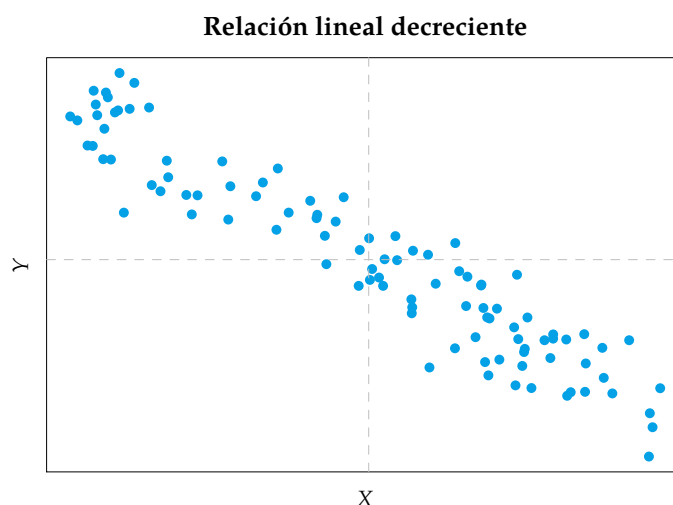
Si la relación entre las variables es *lineal y creciente*, entonces la mayor parte de los puntos estarán en los cuadrantes 1 y 3 y la suma de los productos de desviaciones será positiva.

Relación lineal creciente



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

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

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

También puede calcularse de manera más sencilla mediante la fórmula

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

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

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

Cálculo de la covarianza

Ejemplo con estaturas y pesos

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

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

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

la covarianza vale

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

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

3.3 Regresión

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 **regresión** 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	$y = a_0 + a_1x + \dots + a_nx^n$
Potencial	$y = a \cdot x^b$
Exponencial	$y = a \cdot e^{bx}$
Logarítmica	$y = a + b \log x$
Inverso	$y = a + \frac{b}{x}$
Curva S	$y = e^{a + \frac{b}{x}}$

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

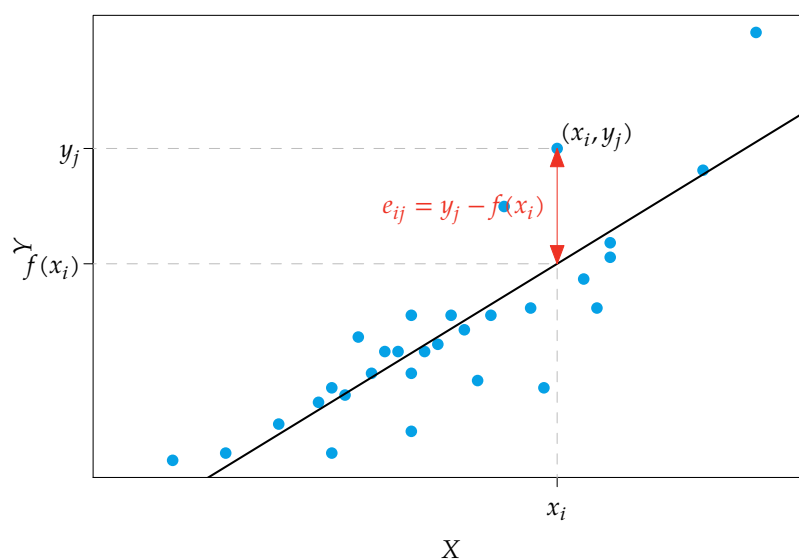
Residuos o errores predictivos

Una vez elegida la familia de curvas que mejor se adapta a la nube de puntos, se determina, dentro de dicha familia, la curva que mejor se ajusta a la distribución, 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 23 (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 **Ajuste de mínimos cuadrados**

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

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

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

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

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

3.4 Recta de regresión**Cálculo de la recta de regresión***Método de mínimos cuadrados*

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

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

Tras resolver el sistema se obtienen los valores

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

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

Recta de regresión

Definición 24 (Recta de regresión). Dada una variable bidimensional (X, Y) , la *recta de regresión* de Y sobre X es

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

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

Cálculo de la recta de regresión

Ejemplo con estaturas y pesos

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

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

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

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

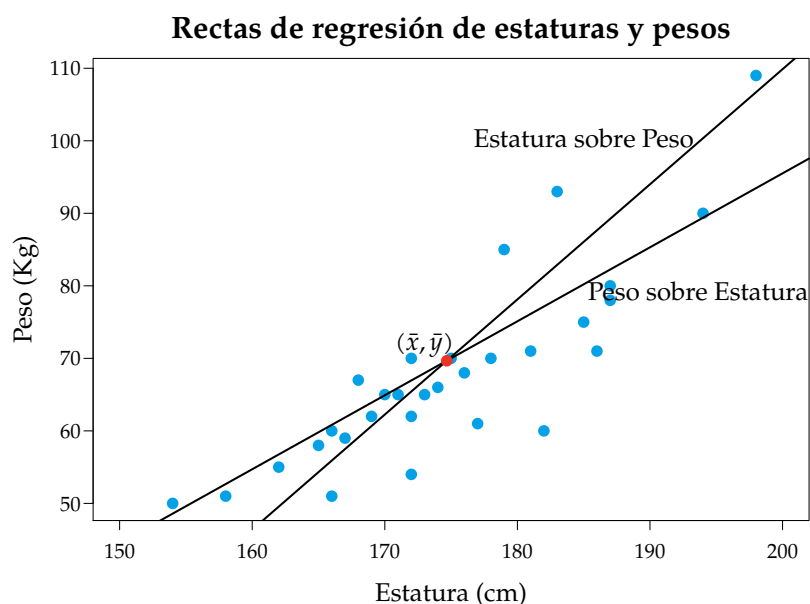
De igual modo, si tomamos la estatura como variable dependiente, la recta de regresión de la estatura sobre el peso es

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

¡Obsérvese que ambas rectas de regresión son diferentes!

Rectas de regresión

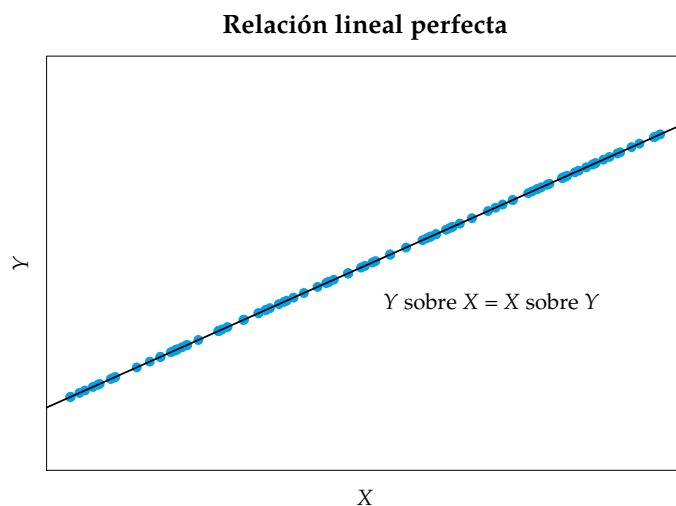
Ejemplo de estaturas y pesos



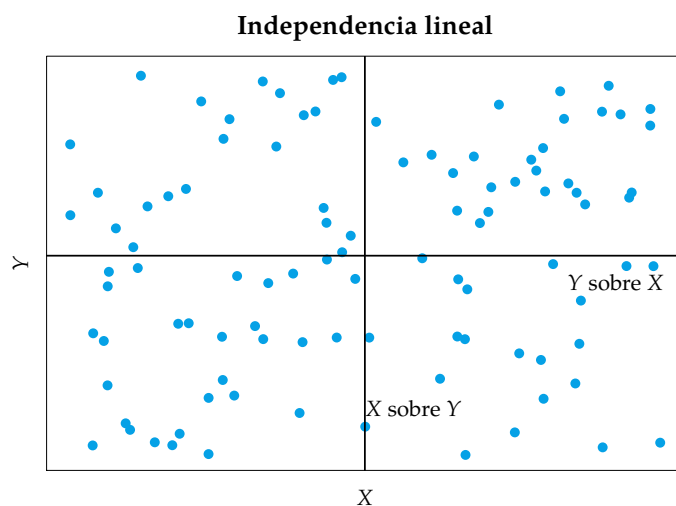
Posición relativa de las rectas de regresión

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 $y = \bar{y}$, $x = \bar{x}$, y se cortan perpendicularmente



Coefficiente de regresión

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

Definición 25 (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.

Regression coefficient

Example of heights and weights

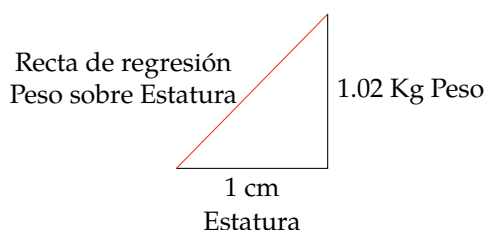
En el ejemplo de las estaturas y los pesos, la recta de regresión del peso sobre la estatura era

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

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.



Predicciones con las rectas de regresión

Ejemplo con estaturas y pesos

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

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

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

$$y = 1.02 \cdot 180 - 108.49 = 75.11 \text{ Kg.}$$

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

$$x = 0.63 \cdot 79 + 130.78 = 180.55 \text{ cm.}$$

Ahora bien, ¿qué fiabilidad tienen estas predicciones?

3.5 Correlación

Correlación

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 26 (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,

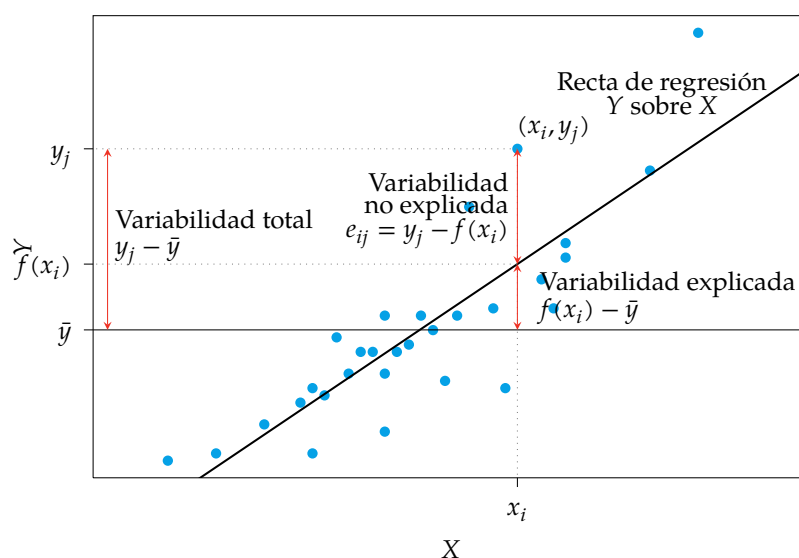
$$s_{ry}^2 = \frac{\sum e_{ij}^2 n_{ij}}{n} = \frac{\sum (y_j - f(x_i))^2 n_{ij}}{n}.$$

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.

$$0 \leq s_{ry}^2 \leq s_y^2$$

Descomposición de la variabilidad total: Variabilidad explicada y no explicada



3.6 Coeficientes de determinación y correlación

Coeficiente de determinación

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

Definición 27 (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

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

Cálculo del coeficiente de determinación lineal

Ejemplo de estaturas y pesos

En el ejemplo de las estaturas y pesos se tenía

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

De modo que el coeficiente de determinación lineal vale

$$r^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{(104.07 \text{ cm Kg})^2}{102.06 \text{ cm}^2 \cdot 164.42 \text{ Kg}^2} = 0.65.$$

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.

Coeficiente de correlación lineal

Definición 28 (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 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.

Coeficiente de correlación lineal

Ejemplo

En el ejemplo de las estaturas y los pesos se tenía

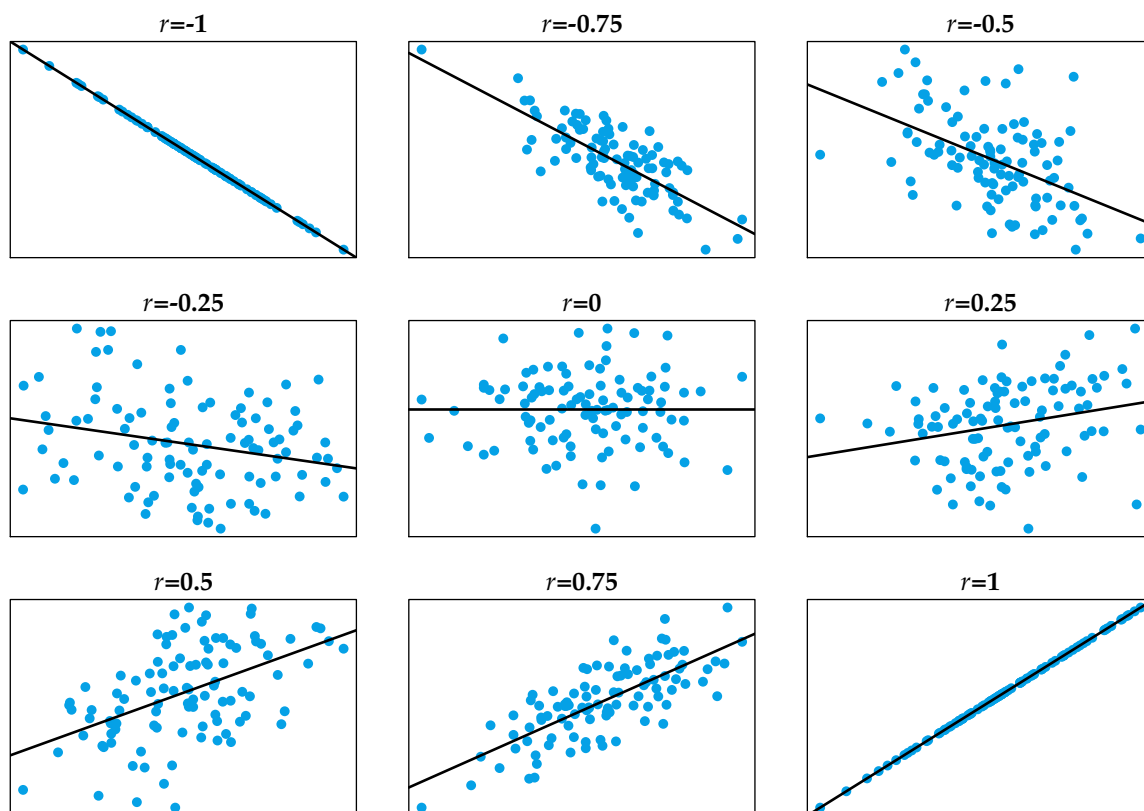
$$\begin{aligned}\bar{x} &= 174.67 \text{ cm} & s_x^2 &= 102.06 \text{ cm}^2 \\ \bar{y} &= 69.67 \text{ Kg} & s_y^2 &= 164.42 \text{ Kg}^2 \\ s_{xy} &= 104.07 \text{ cm} \cdot \text{Kg}\end{aligned}$$

De modo que el coeficiente de correlación lineal vale

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

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.

3.7 Regresión no lineal

Regresión no lineal

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 $z = \log y$:

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

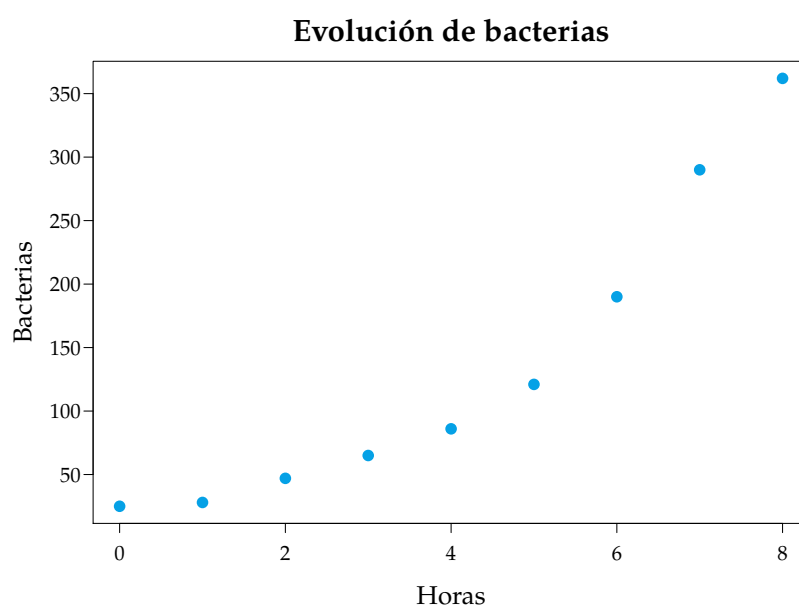
Relación exponencial

Evolución del número de bacterias de un cultivo

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



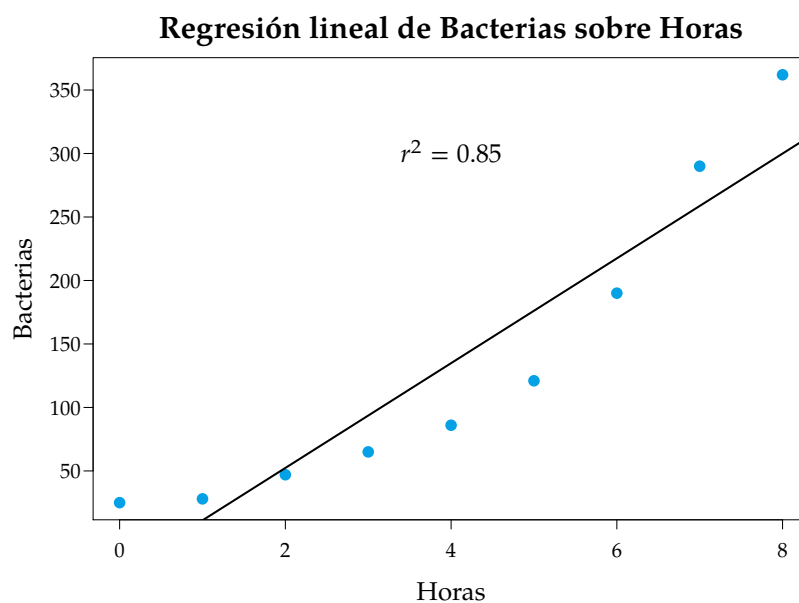
Relación exponencial

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?

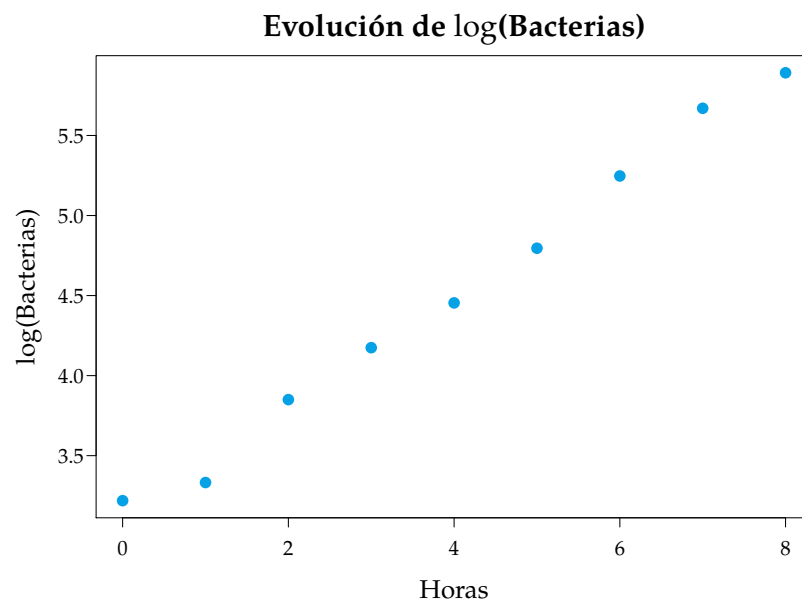
Ajuste de un modelo de regresión exponencial

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



Ajuste de un modelo de regresión exponencial

Evolución del número de bacterias de un cultivo

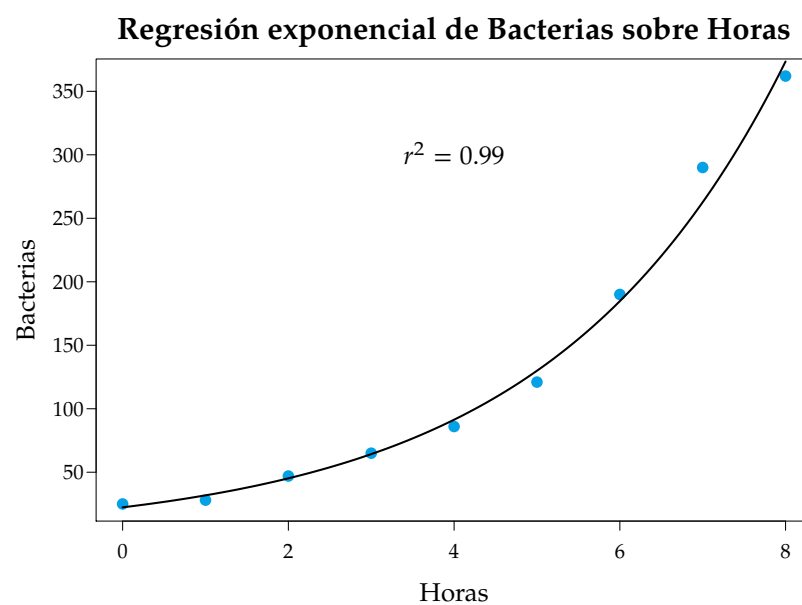
Ahora sólo queda calcular la recta de regresión del logaritmo de Bacterias sobre Horas

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

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

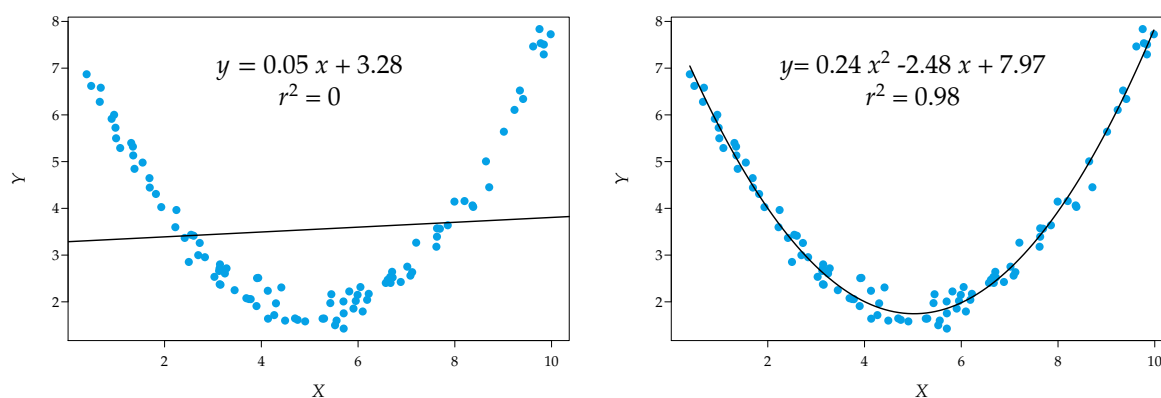
$$\text{Bacterias} = e^{3.107+0.352 \text{ Horas}}, \text{ con } r^2 = 0.99.$$

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



Interpretación de un coeficiente de determinación pequeño

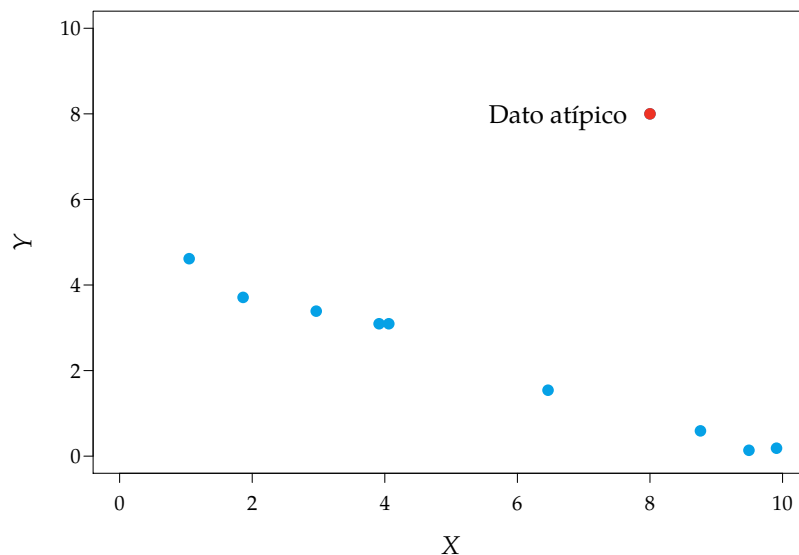
Es importante señalar que cada modelo de regresión tiene su propio coeficiente de determinación. Así, un coeficiente de determinación cercano a cero significa que no existe relación entre las variables del tipo planteado por el modelo, pero *eso no quiere decir que las variables sean independientes*, ya que puede existir relación de otro tipo.



Datos atípicos en regresión

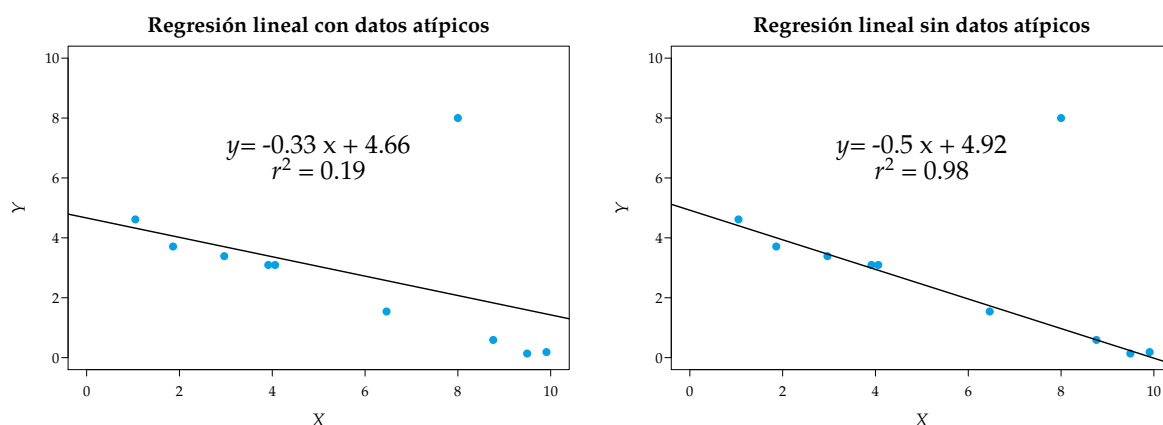
Los *datos atípicos* en un estudio de regresión son los puntos que claramente no siguen la tendencia del resto de los puntos en el diagrama de dispersión, incluso si los valores del par no se pueden considerar atípicos para cada variable por separado.

Diagrama de dispersión con datos atípicos



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.



3.8 Medidas de relación entre atributos

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 de correlación de Spearman, y puede demostrarse fácilmente que puede calcularse a partir de la siguiente fórmula

Definición 29 (Coeficiente de correlación de Spearman). Dada una muestra de n individuos en los que se han medido dos atributos ordinales X e Y , el coeficiente de correlación de Spearman se define como:

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

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

Interpretación del coeficiente de correlación de Spearman

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

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

de manera que:

- Si $r_s = 0$ entonces no existe relación entre los atributos ordinales.

- Si $r_s = 1$ entonces los órdenes de los atributos coinciden y existe una relación directa perfecta.
- Si $r_s = -1$ entonces los órdenes de los atributos están invertidos y existe una relación inversa perfecta.

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

Cálculo del coeficiente de correlación de Spearman

Ejemplo

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

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

El coeficiente de correlación de Spearman para esta muestra es

$$r_s = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)} = 1 - \frac{6 \cdot 8}{5(5^2 - 1)} = 0.6,$$

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

Cálculo del coeficiente de correlación de Spearman

Ejemplo con empates

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

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

Alumnos	X	Y	d_i	d_i^2
Alumno 1	2	3	-1	1
Alumno 2	5	4	1	1
Alumno 3	1	1.5	-0.5	0.25
Alumno 4	3.5	1.5	2	4
Alumno 5	3.5	5	-1.5	2.25
Σ			0	8.5

El coeficiente de correlación de Spearman para esta muestra es

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

Relación entre atributos nominales

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

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

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

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

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

Frecuencias teóricas o esperadas

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

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

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

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

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) .

Coefficiente chi-cuadrado χ^2

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

Definición 30 (Coeficiente chi-cuadrado χ^2). Dada una muestra de tamaño n en la que se han medido dos atributos X e Y , se define el coeficiente χ^2 como

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

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

Por ser suma de cuadrados, se cumple que

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

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

Cálculo del coeficiente chi-cuadrado χ^2 *Ejemplo*

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

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

se obtienen las siguientes frecuencias esperadas:

Sexo\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

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

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

Coeficiente de contingencia

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

Para evitar estos problemas se suele utilizar el siguiente estadístico:

Definición 31 (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.