Estadística fundamental
Estadística descriptiva
Máster en Data Science y Big Data (Finanzas)

Pilar Barrios

Diciembre 2021

Índice

1 Introducción a la estadística

¿Qué es la estadística?

Definición

- La estadística es una rama de las matemáticas que se encarga de la recogida, análisis e interpretación de datos.
- La estadística es imprescindible en cualquier disciplina científica o técnica donde se manejen datos, especialmente si son grandes volúmenes de datos: ciencias biosanitarias (química, biología, medicina) y ciencias sociales (economía y psicología).

¿Por qué es necesaria la estadística?

La variabilidad de nuestro mundo

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

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

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

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

Población estadística

Definición (Población)

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

Definición (Tamaño poblacional)

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

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

- Población teórica: conjunto de elementos a los que se quiere extrapolar los resultados del estudio.

- Población estudiada: conjunto de elementos realmente accesibles en el estudio.

Inconvenientes en el estudio de la población

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

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

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

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

Muestra estadística

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

Definición (Muestra)

Una muestra es un subconjunto de la población.

Definición (Tamaño muestral)

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

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

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

Determinación del tamaño muestral

Una de las preguntas más interesantes que surge inmediatamente es:

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

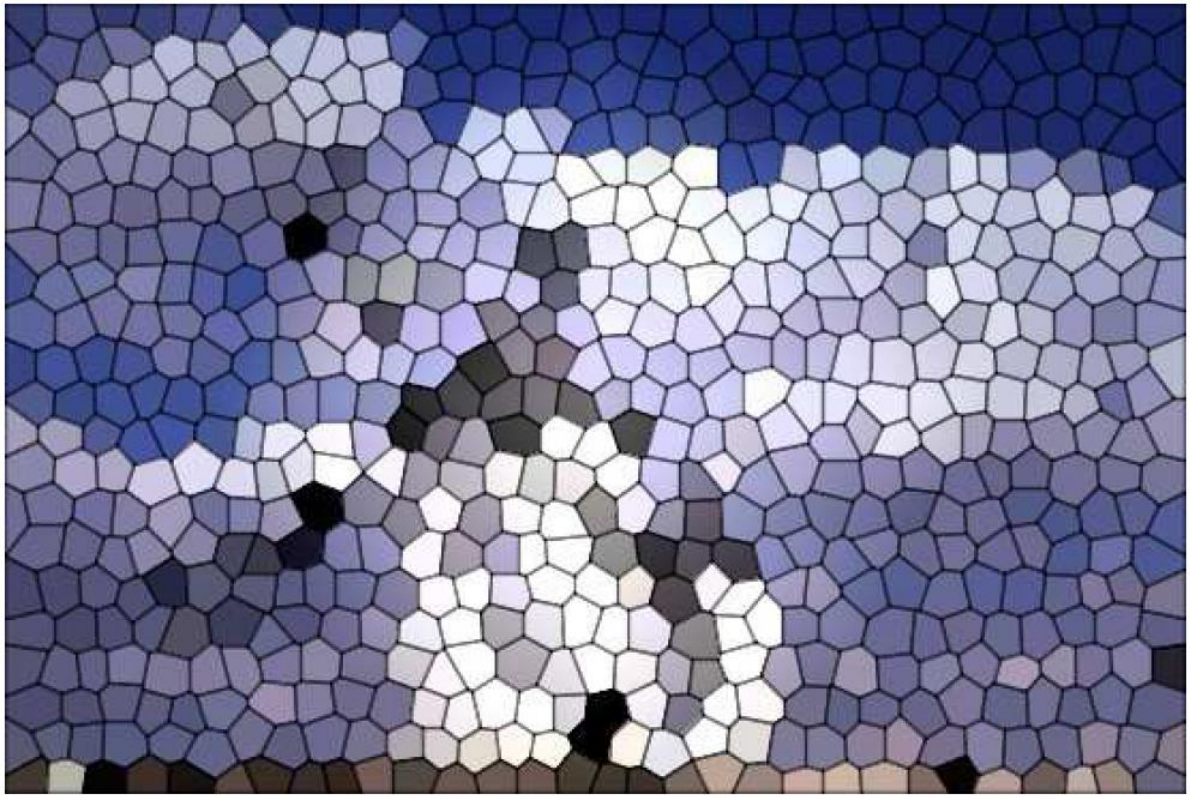
La respuesta depende de muchos factores, entre los que destacan:

- La **variabilidad** de la población.
- La **fiabilidad** deseada para las extrapolaciones que se hagan hacia la población.

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

Si se toma una muestra pequeña de los píxeles de una imagen…



¿De qué imagen se trata?

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

Determinación del tamaño muestral

Si se toma una muestra mayor de los píxeles de una imagen…

¿De qué imagen se trata?

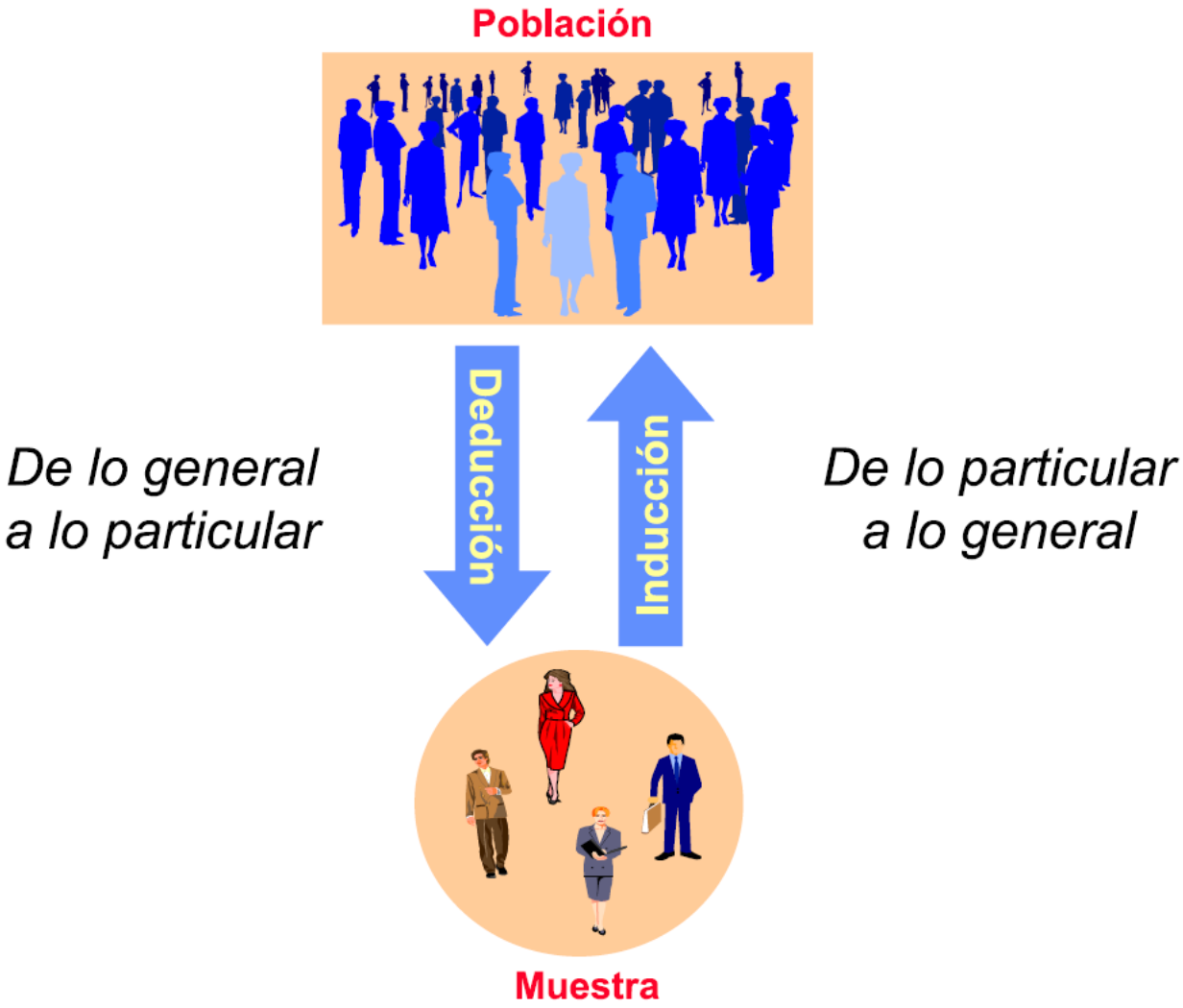
¡Con una muestra mayor es más fácil averiguar el contenido de la imagen!

Determinación del tamaño muestral

Y aquí la población completa de los píxeles de una imagen.

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

Tipos de razonamiento



Tipos de razonamiento

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

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

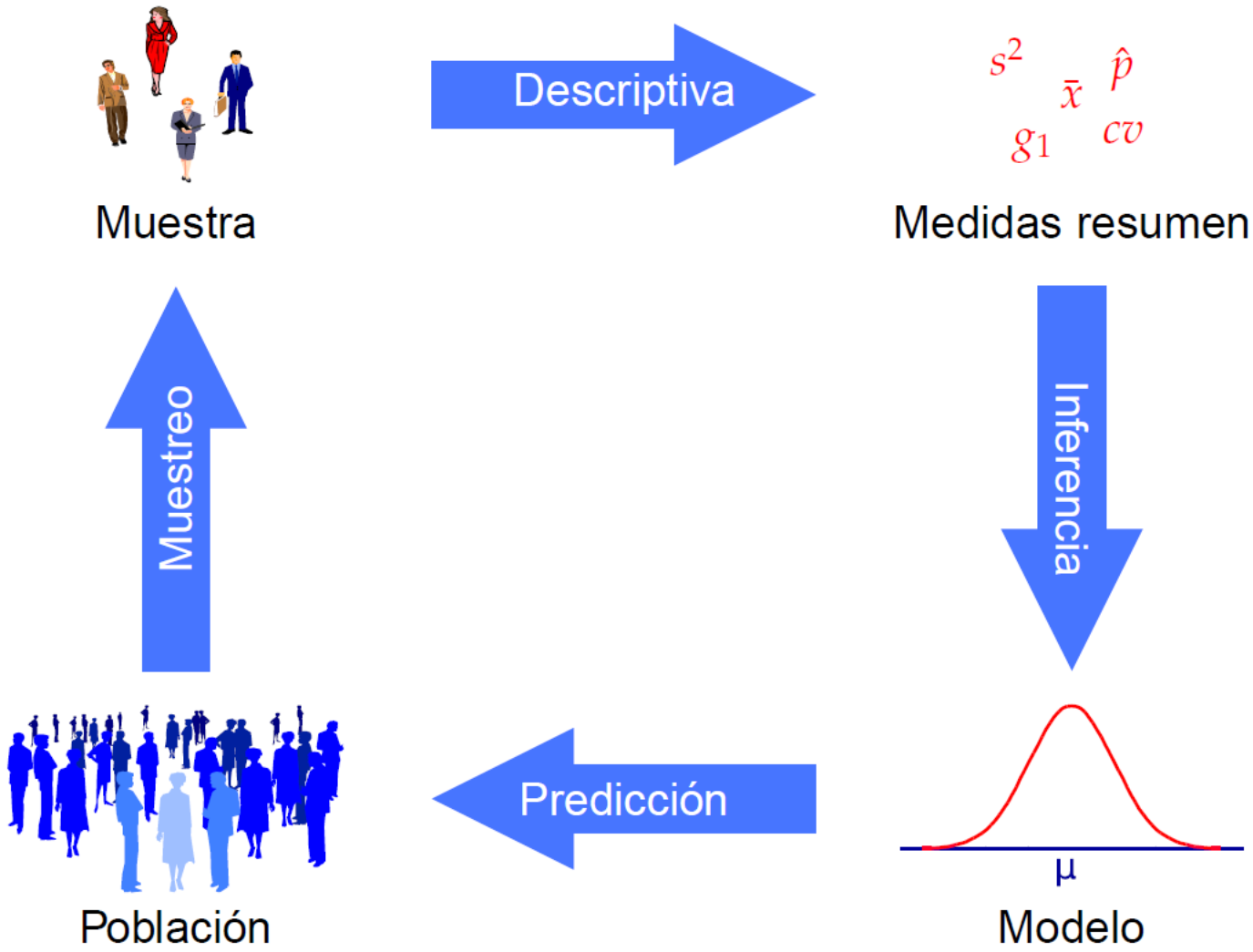
La estadística se apoya fundamentalmente en el razonamiento inductivo ya que utiliza la información obtenida a partir de muestras para sacar conclusiones sobre las poblaciones.

Fases del análisis estadístico

Normalmente un estudio estadístico posee 4 etapas:

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

Fases del análisis estadístico



Muestreo

Definición (Muestreo): El proceso de selección de los elementos que compondrán una muestra se conoce como muestreo.

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

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).

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

- Muestreo no aleatorio: Los individuos se eligen de forma no aleatoria.

  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:

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

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

Estadística descriptiva

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

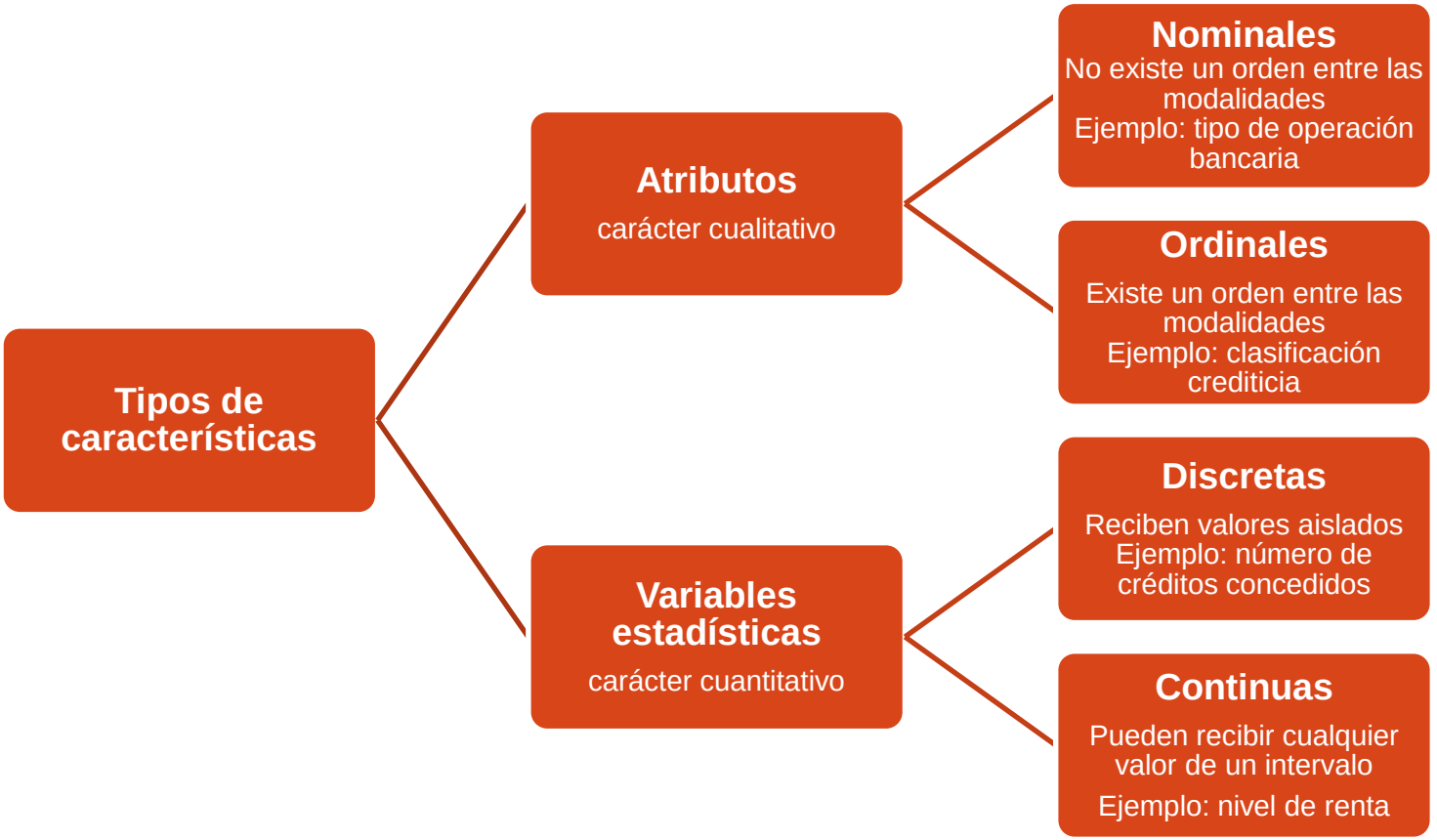
Tras el proceso de muestreo, es la siguiente etapa de todo estudio estadístico y suele consistir en:

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

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

Variables estadísticas y atributos

Las características presentan la siguiente clasificación:



Clasificación de la muestra

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

Existen dos formas de clasificar estos valores:

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

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

Clasificación de la muestra

X = Estatura

Clasificación

Recuento de frecuencias

X = Estatura

Frecuencias

Frecuencias muestrales

Definición (Frecuencias muestrales)

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

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

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

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

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

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

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

Tabla de frecuencias

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

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

Tabla de frecuencias

- Ejemplo de datos sin agrupar:

En una encuesta a 25 familias sobre el número de préstamos solicitados a una entidad bancaria se obtuvieron los siguientes datos:

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

La tabla de frecuencias asociada a esta muestra es:

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

Tabla de frecuencias

- Ejemplo de datos agrupados:

Se ha medido el número de empleados de 30 PYMES obteniendo:

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

La tabla de frecuencias asociada a esta muestra es:

x_i	n_i	f_i	N_i	F_i
$(150, 160]$	2	0,07	2	0,07
$(160, 170]$	8	0,27	10	0,34
$(170, 180]$	11	0,36	21	0,70
$(180, 190]$	7	0,23	28	0,93
$(190, 200]$	2	0,07	30	1
$\sum$	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 las siguientes consideraciones:

1. El número de intervalos no debe ser muy grande ni muy pequeño. Una regla orientativa es tomar un número de intervalos próximo a la raíz cuadrada del tamaño muestral $\sqrt{n}$.

2. 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.

3. El valor más pequeño debe caer dentro del primer intervalo y el más grande dentro del último.

Tabla de frecuencias

- Ejemplo con un atributo:

Las calificaciones crediticias (*ratings*) de una muestra de 30 empresas son:

A, AA, AA, A, AAA, D, D, A, AA, AA, A, A, A, A,
AAA, A, A, A, AA, D, AA, AA, AA, A, A, A, D, A, AAA, D

La tabla de frecuencias asociada a esta muestra es:

x_i	n_i	f_i
AAA	5	0,16
AA	14	0,47
A	8	0,27
D	3	0,10
$\sum$	30	1

¿Por qué en este caso no se calculan las frecuencias acumuladas?

Representaciones gráficas

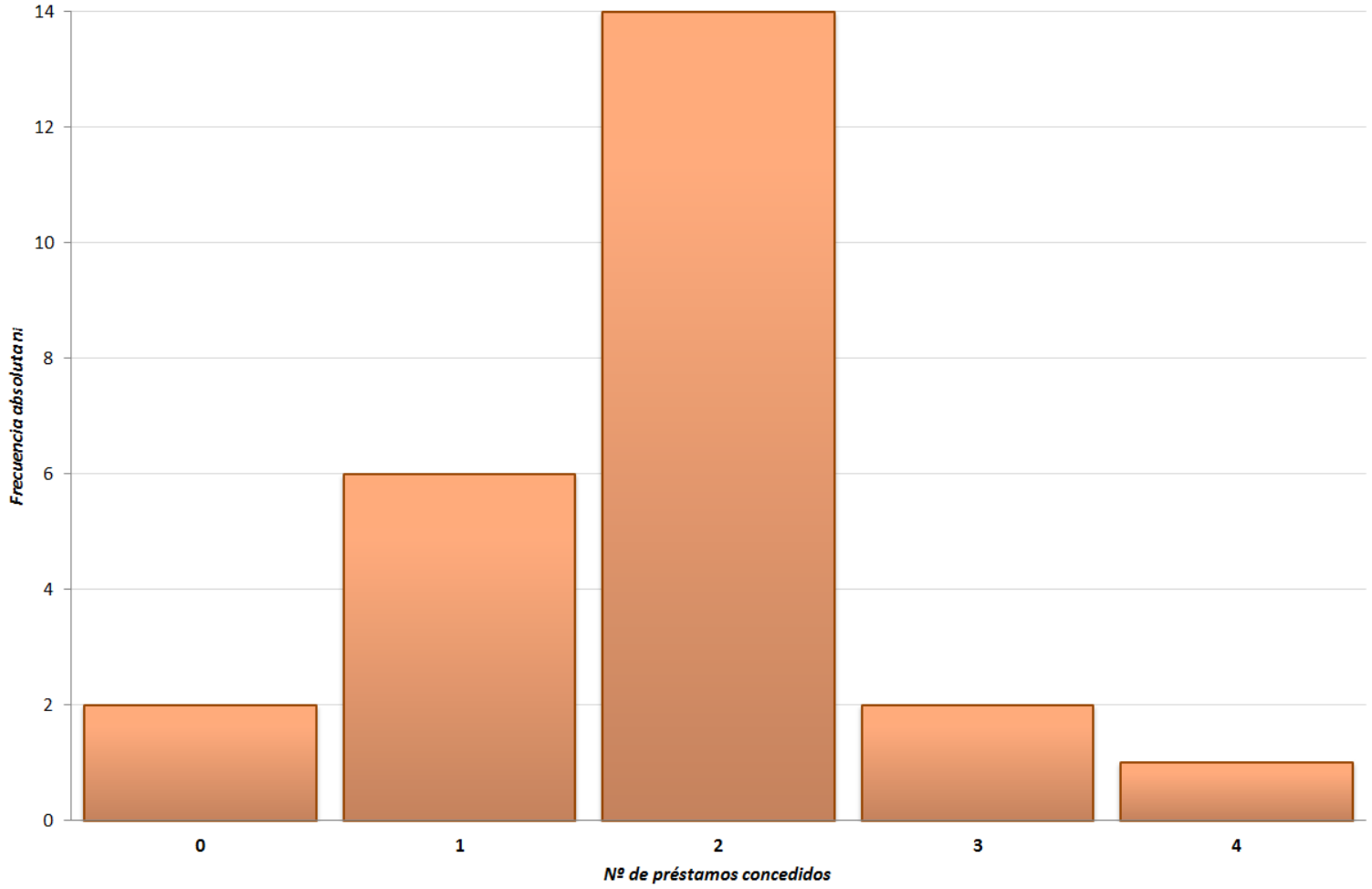
También es habitual representar la distribución muestral de frecuencias de forma gráfica. Dependiendo del tipo de variable y de si se han agrupado o no los datos se utilizan distintos tipos de gráficos:

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

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

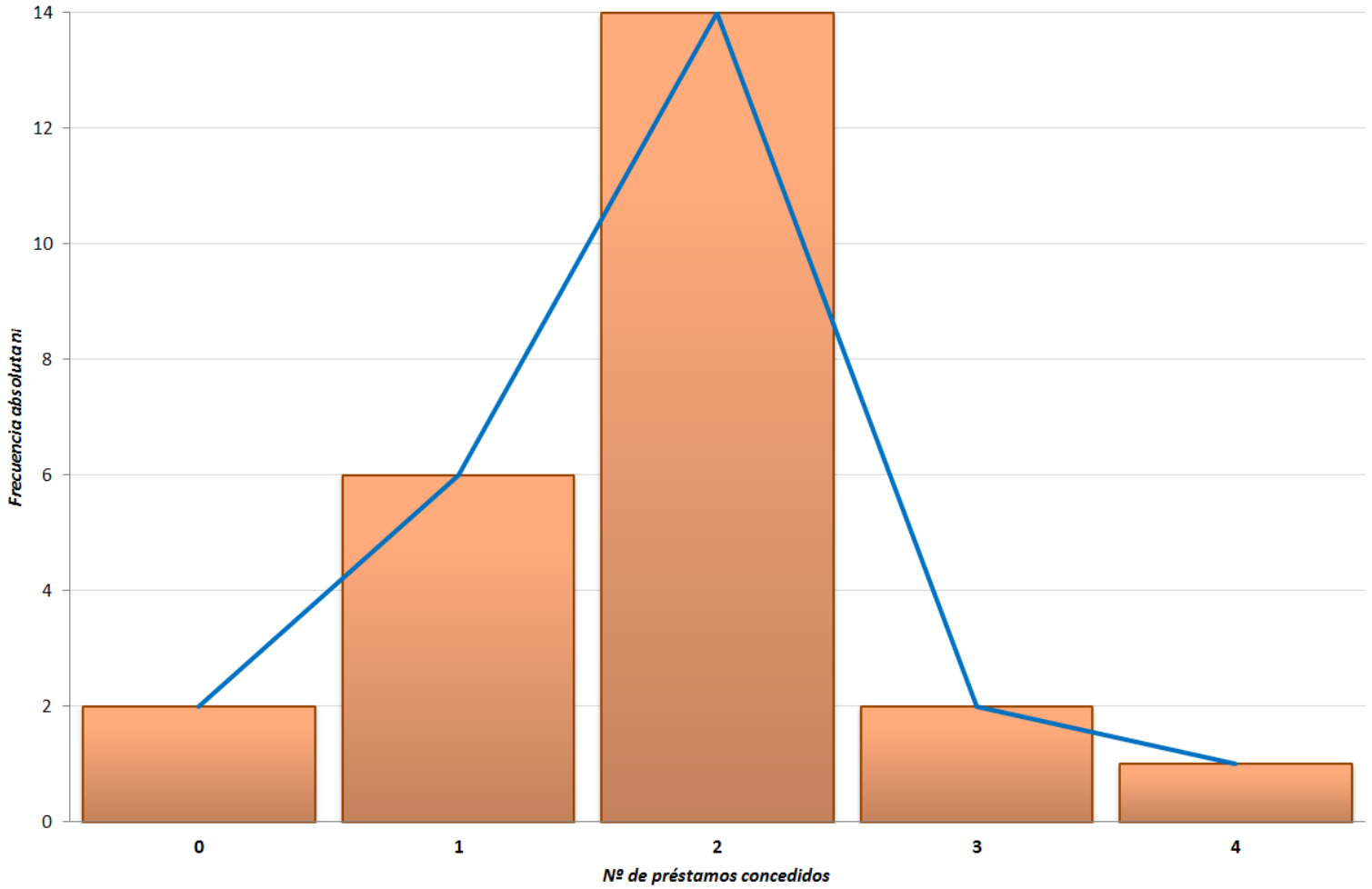
Representaciones gráficas

- **Diagrama de barras de frecuencias absolutas (datos sin agrupar)**



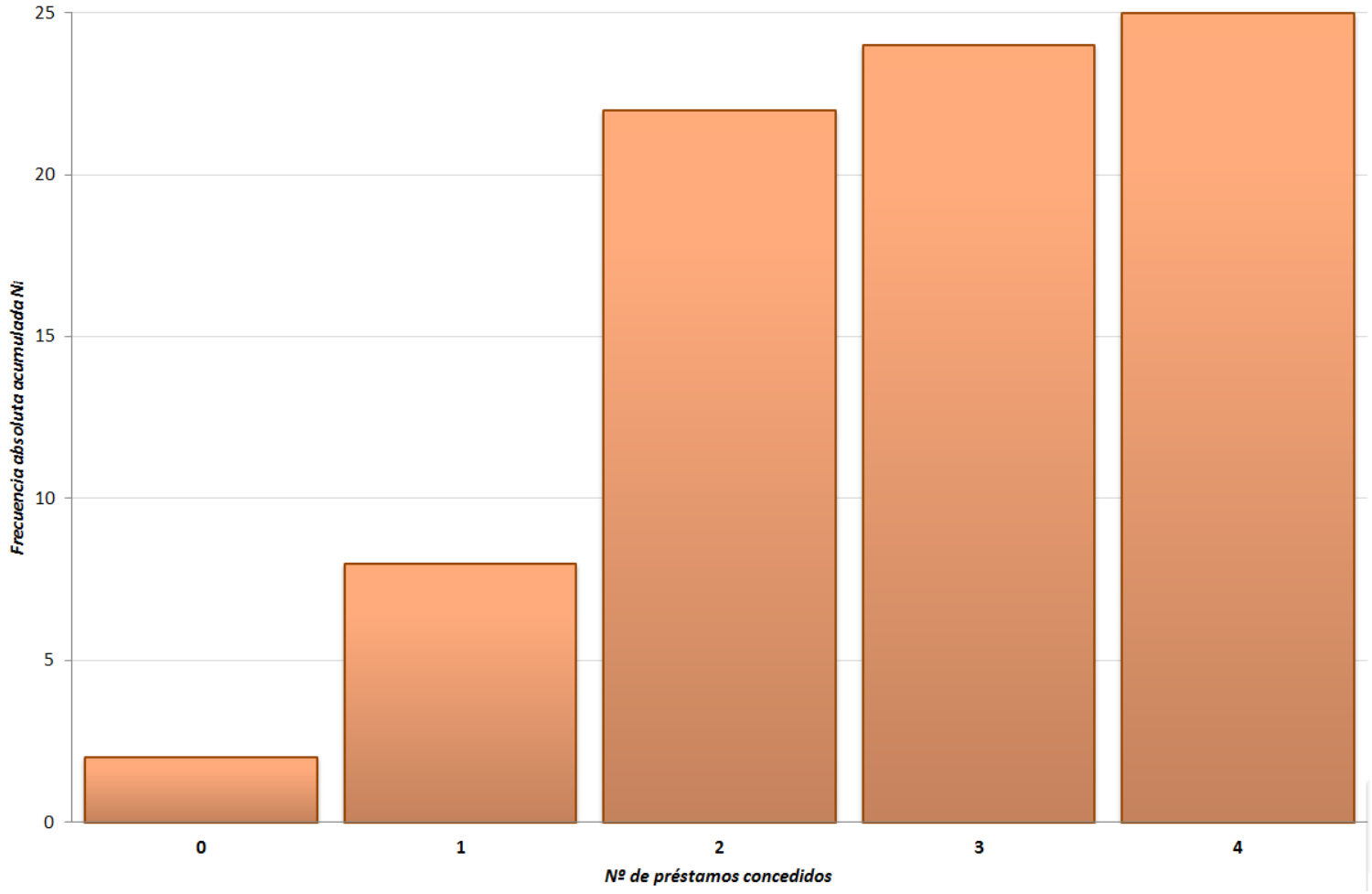
Representaciones gráficas

- **Polígono de frecuencias absolutas (datos sin agrupar)**



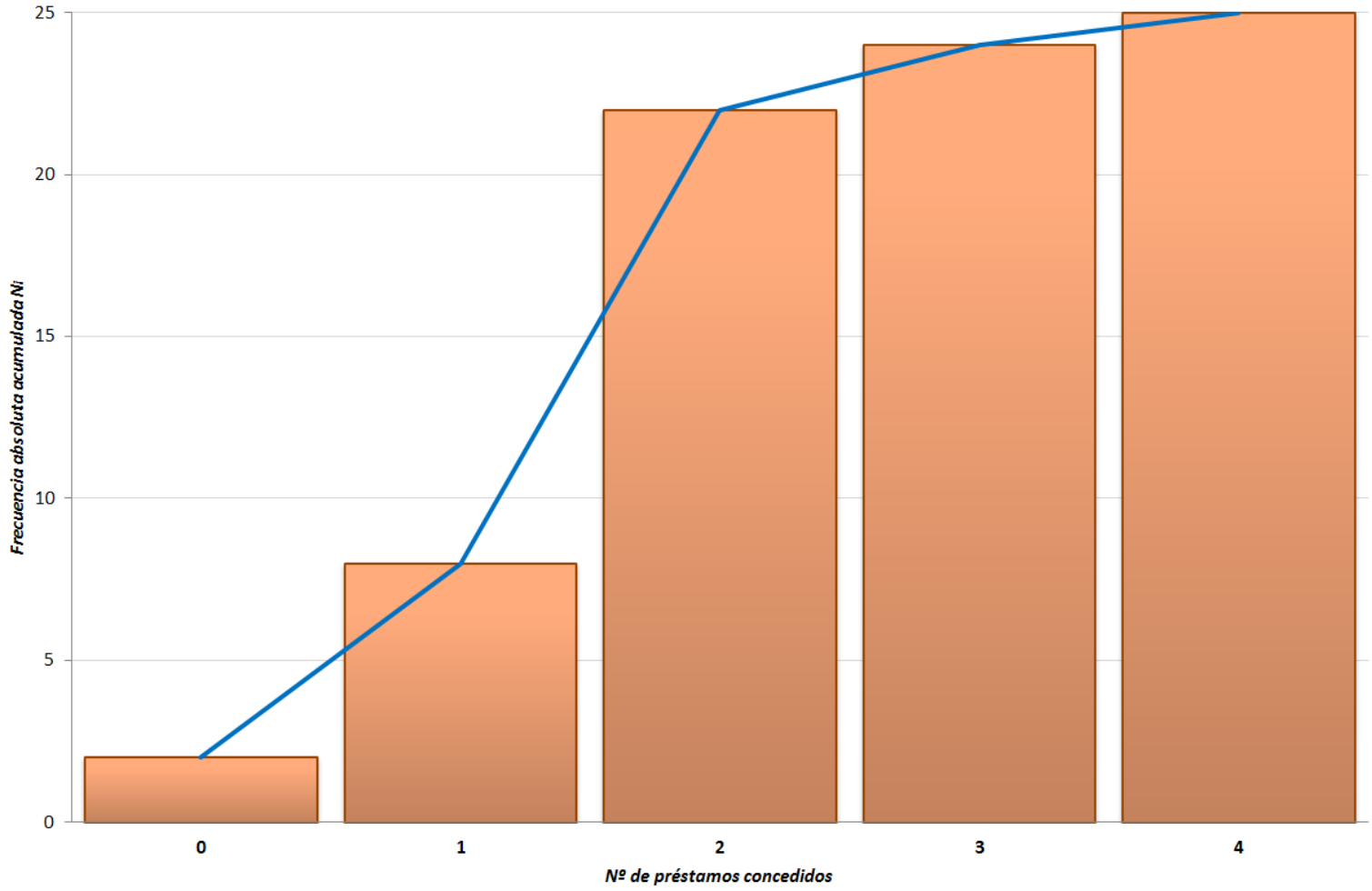
Representaciones gráficas

- **Diagrama de barras de frecuencias absolutas acumuladas (datos sin agrupar)**



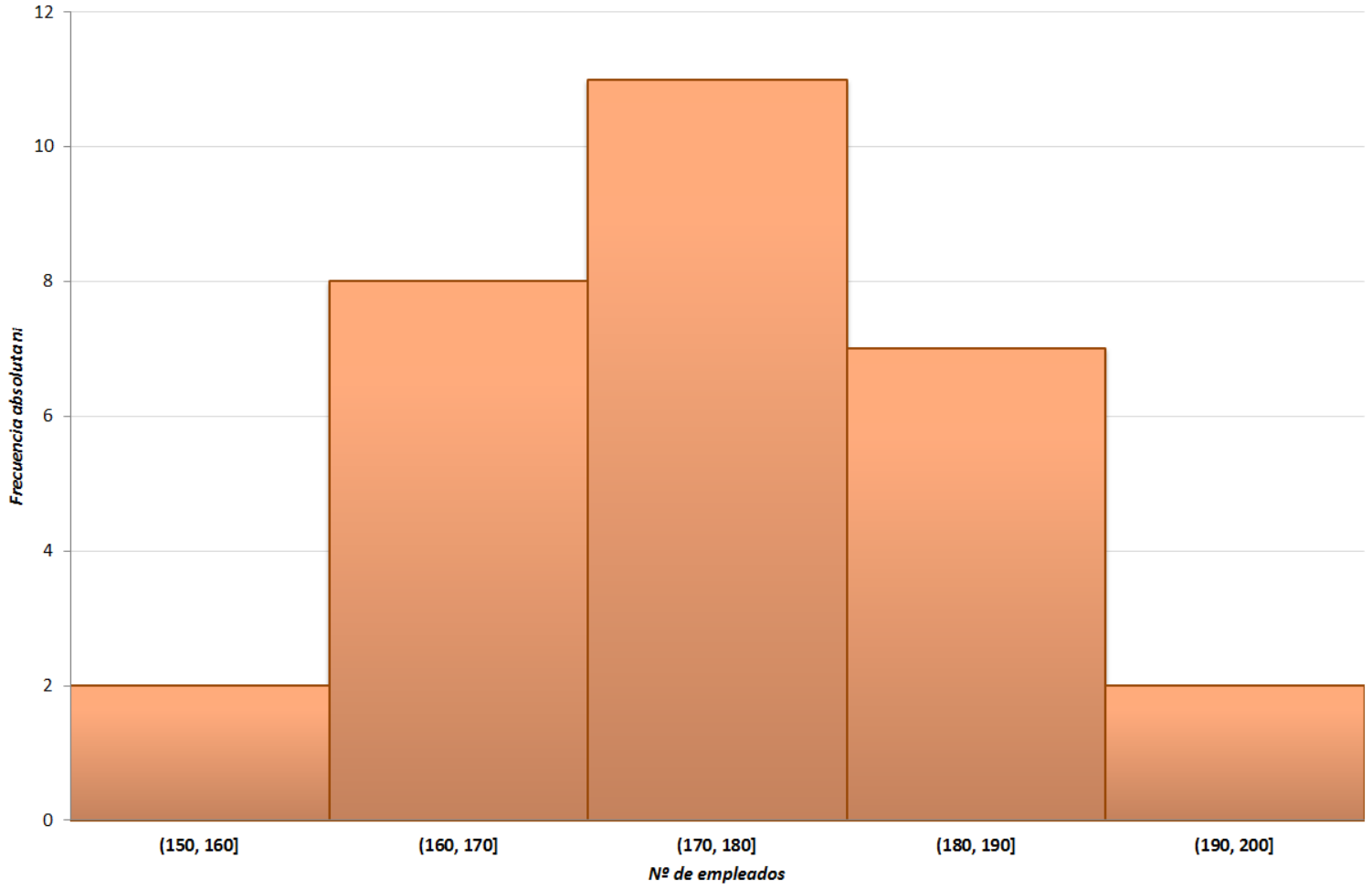
Representaciones gráficas

- **Polígono de frecuencias absolutas acumuladas (datos sin agrupar)**



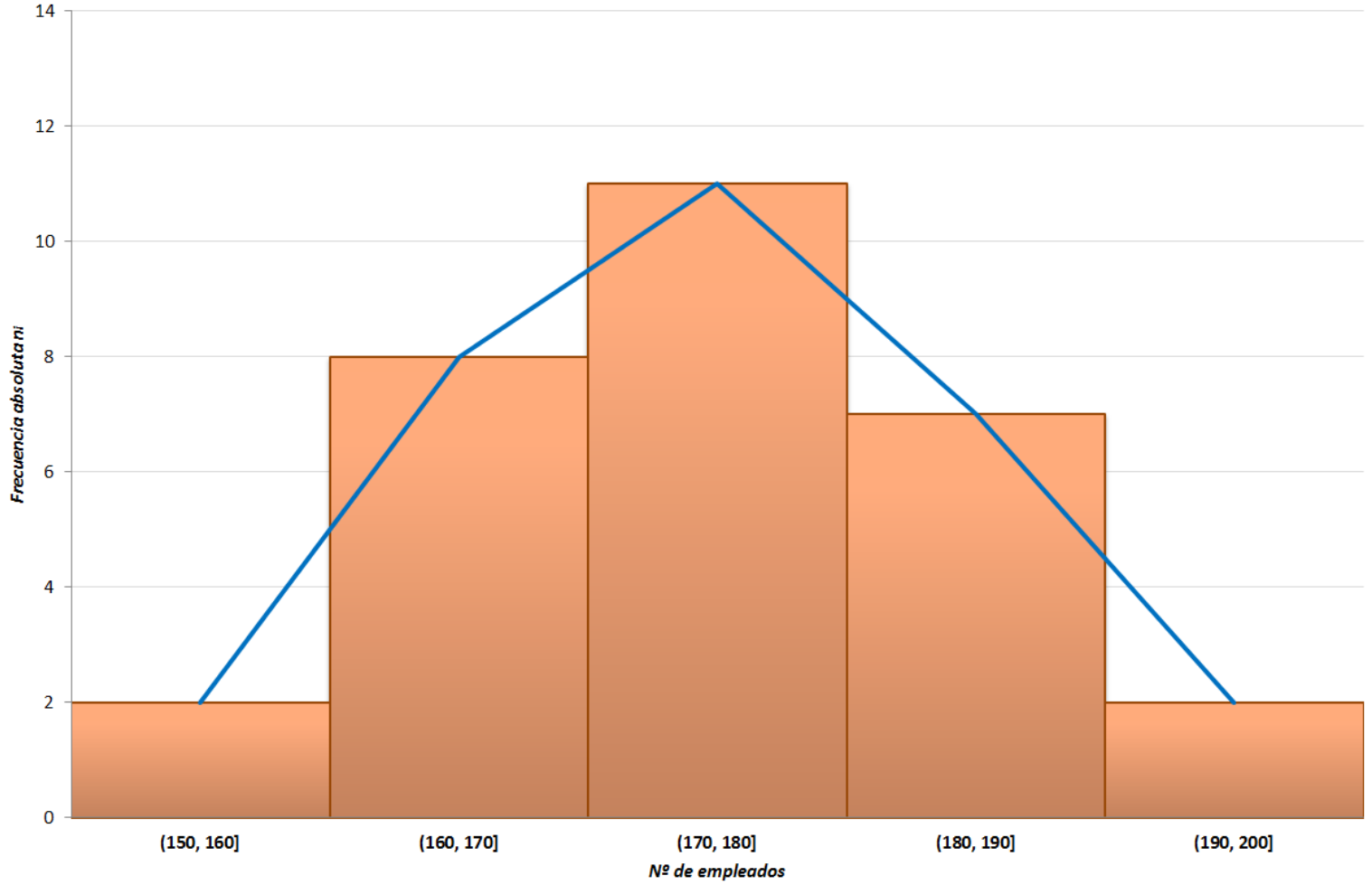
Representaciones gráficas

- **Diagrama de barras de frecuencias absolutas (datos agrupados)**



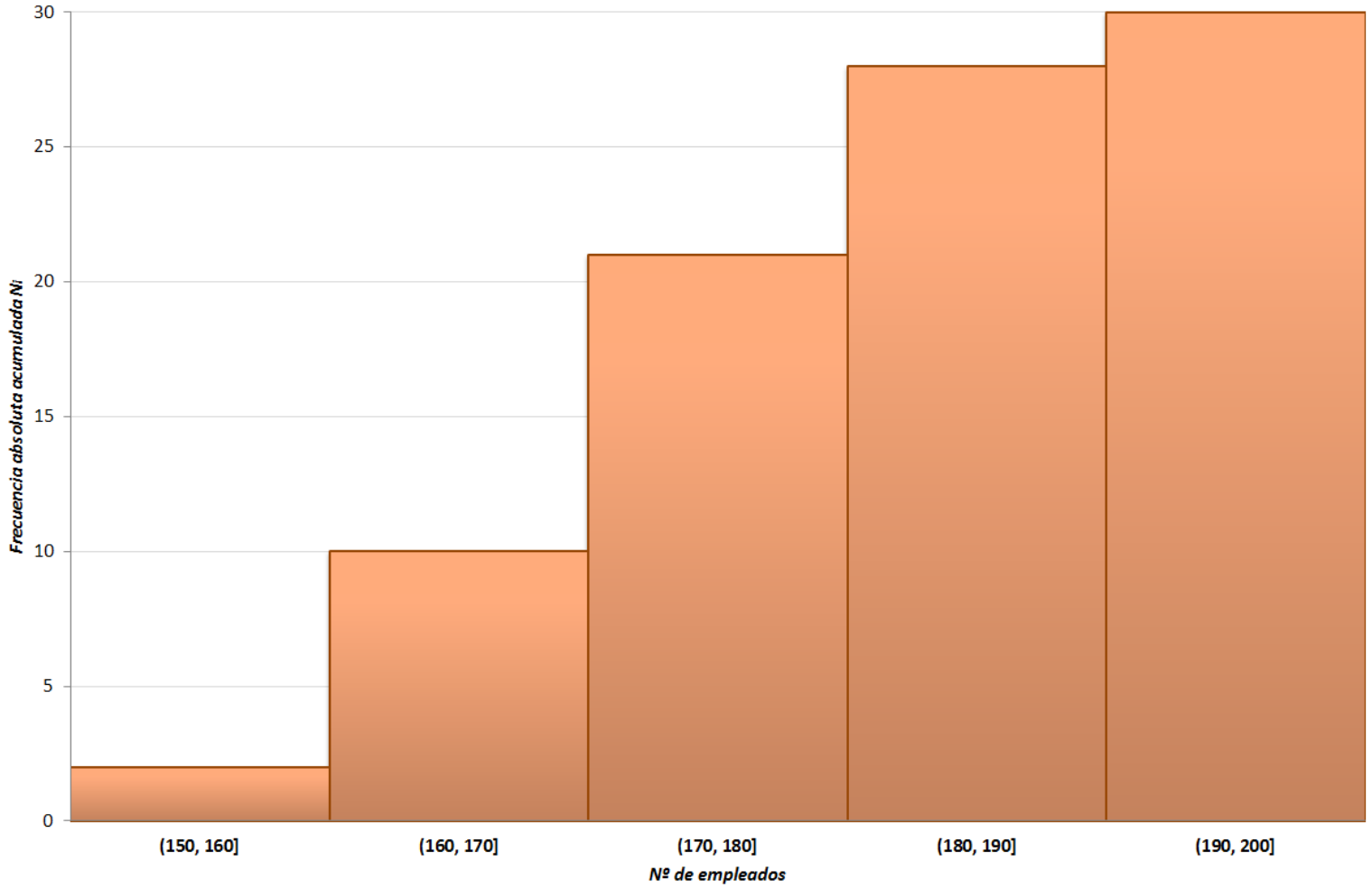
Representaciones gráficas

- **Polígono de frecuencias absolutas (datos agrupados)**



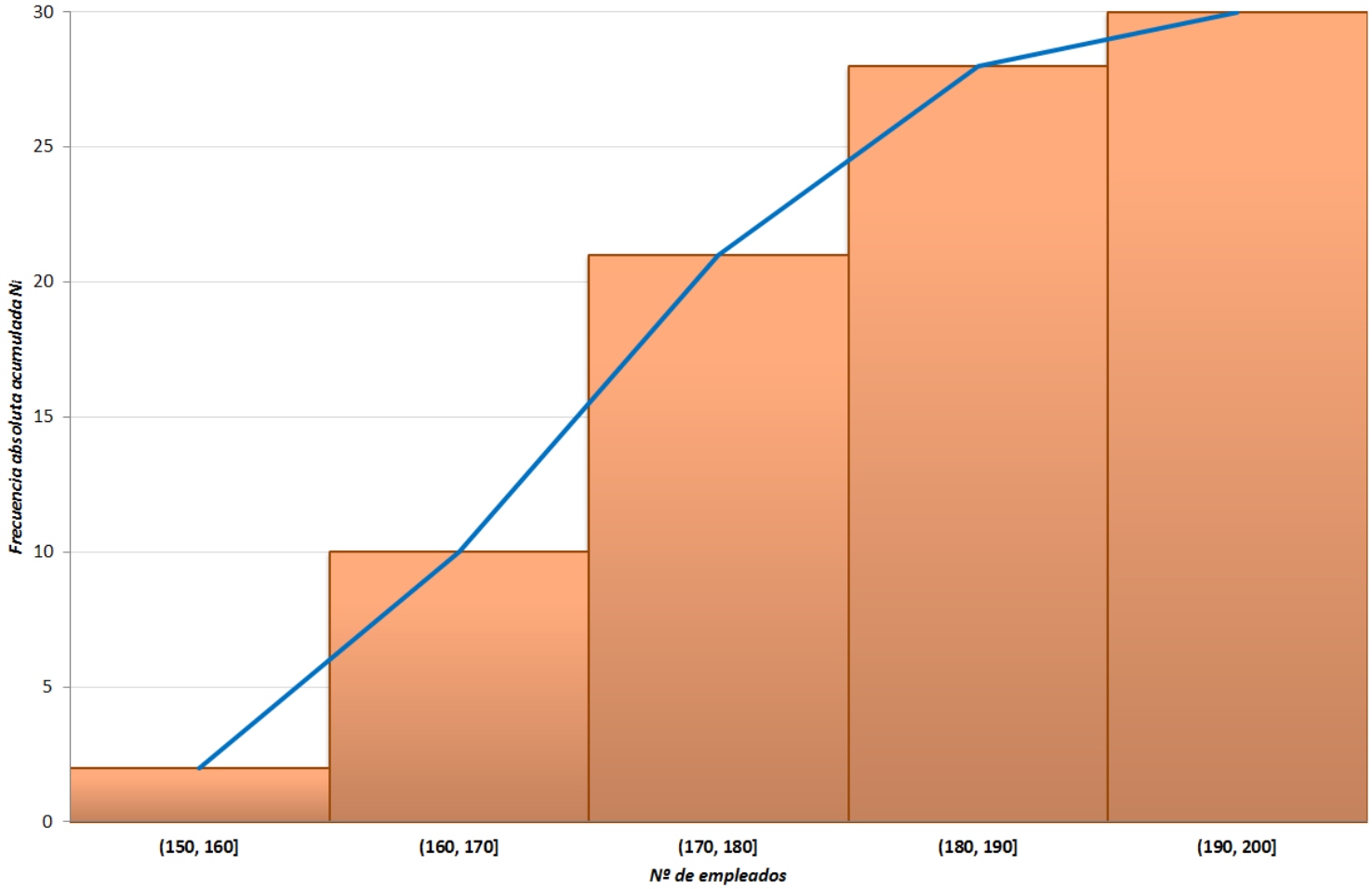
Representaciones gráficas

- **Diagrama de barras de frecuencias absolutas acumuladas (datos agrupados)**



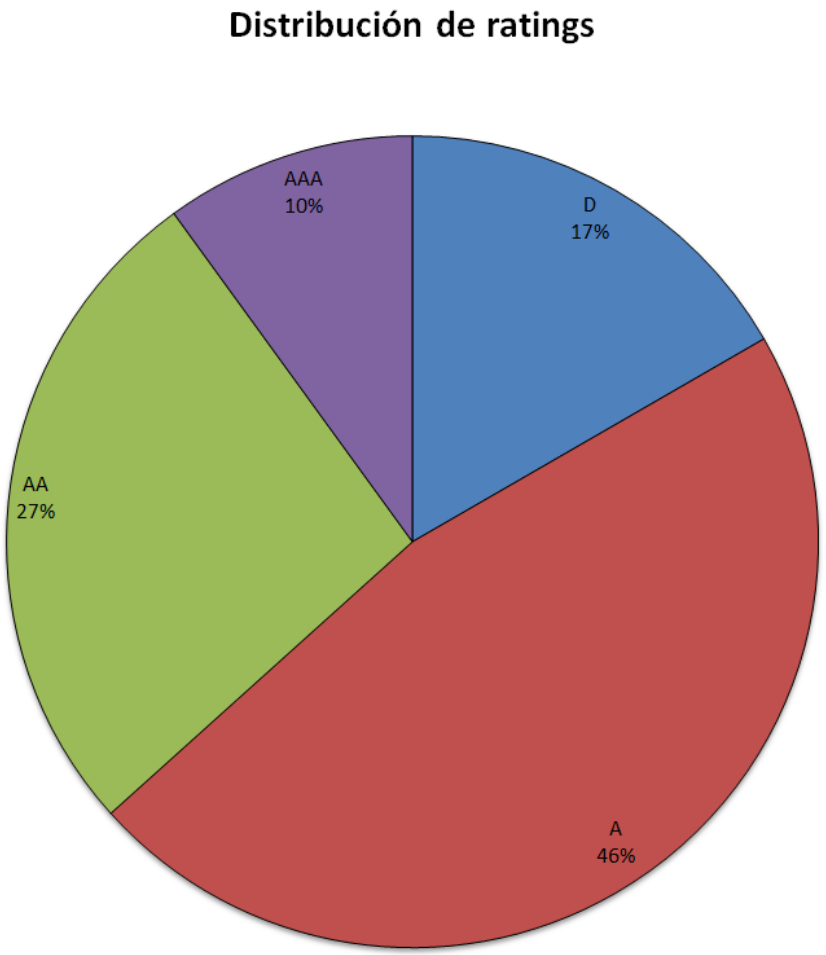
Representaciones gráficas

- **Polígono de frecuencias absolutas acumuladas (datos agrupados)**



Representaciones gráficas

- **Diagrama de sectores (atributos)**



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** 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 distribuyen en la distribución.
- Estadísticos de dispersión: Miden la heterogeneidad de los datos.
- Estadísticos de forma: Miden aspectos de la representación gráfica de los datos, como la simetría o el apuntamiento (curtosis).

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 las 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 y Percentiles

Estadísticos de posición: Media aritmética

- Definición (Media aritmética muestral $\bar{x}$)

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

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

A partir de la tabla de frecuencias puede calcularse como:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i n_i}{n} = \sum_{i=1}^{n} x_i f_i$$

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

¡Ojo! No puede calcularse para atributos

Estadísticos de posición: Media aritmética

En el ejemplo anterior del número de préstamos concedidos se tiene

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

o bien, desde la tabla de frecuencias:

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

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

Es decir, el número de préstamos concedidos que mejor representa al muestra es 1,76 préstamos

Estadísticos de posición: Media aritmética

En el ejemplo anterior del número de empleados se tiene

$$\bar{x} = \frac{179 + 173 + \cdots + 187}{30} = 175{,}07$$

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

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

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

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

Estadísticos de posición: Mediana

- Definición (Mediana muestral Me)

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

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

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

¡Ojo! No debe calcularse para atributos nominales

Estadísticos de posición: Mediana

- Cálculo de la mediana con datos no agrupados

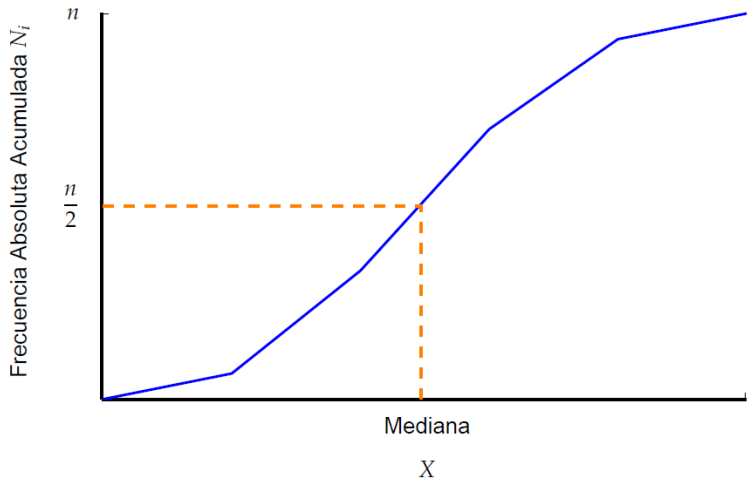
Con datos no agrupados pueden darse varios casos:

a) Tamaño muestral impar: La mediana es el valor que ocupa la posición $\frac{n+1}{2}$.

b) Tamaño muestral par: La mediana es la media de los valores que ocupan las posiciones $\frac{n}{2}$ y $\frac{n}{2}+1$

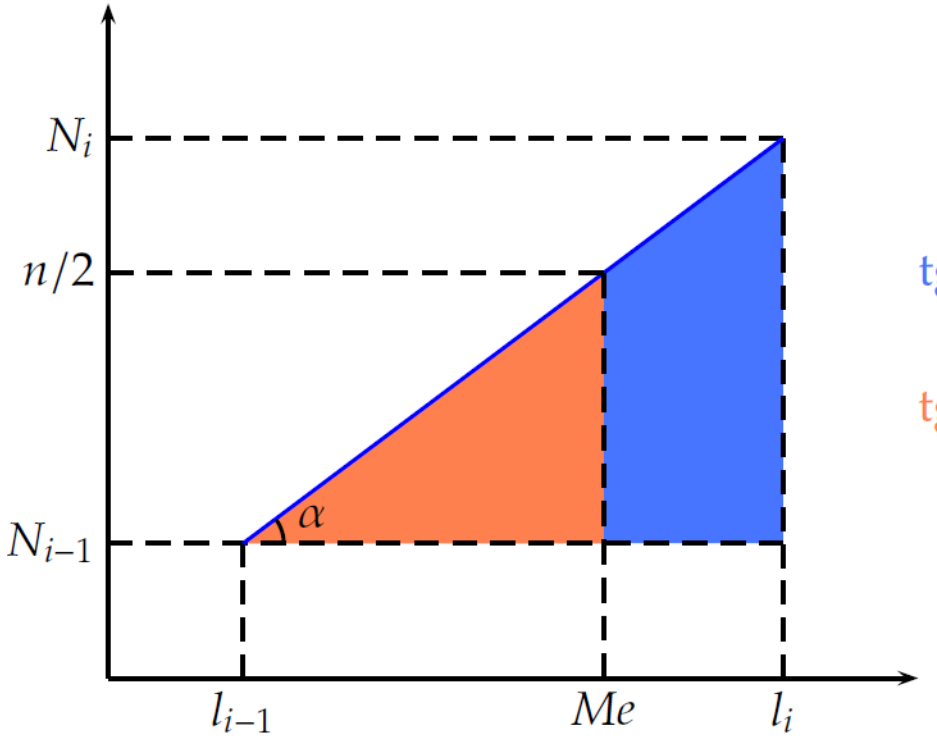
- Cálculo de la mediana con datos agrupados

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



Estadísticos de posición: Mediana

- Interpolación en el polígono de frecuencias absolutas acumuladas



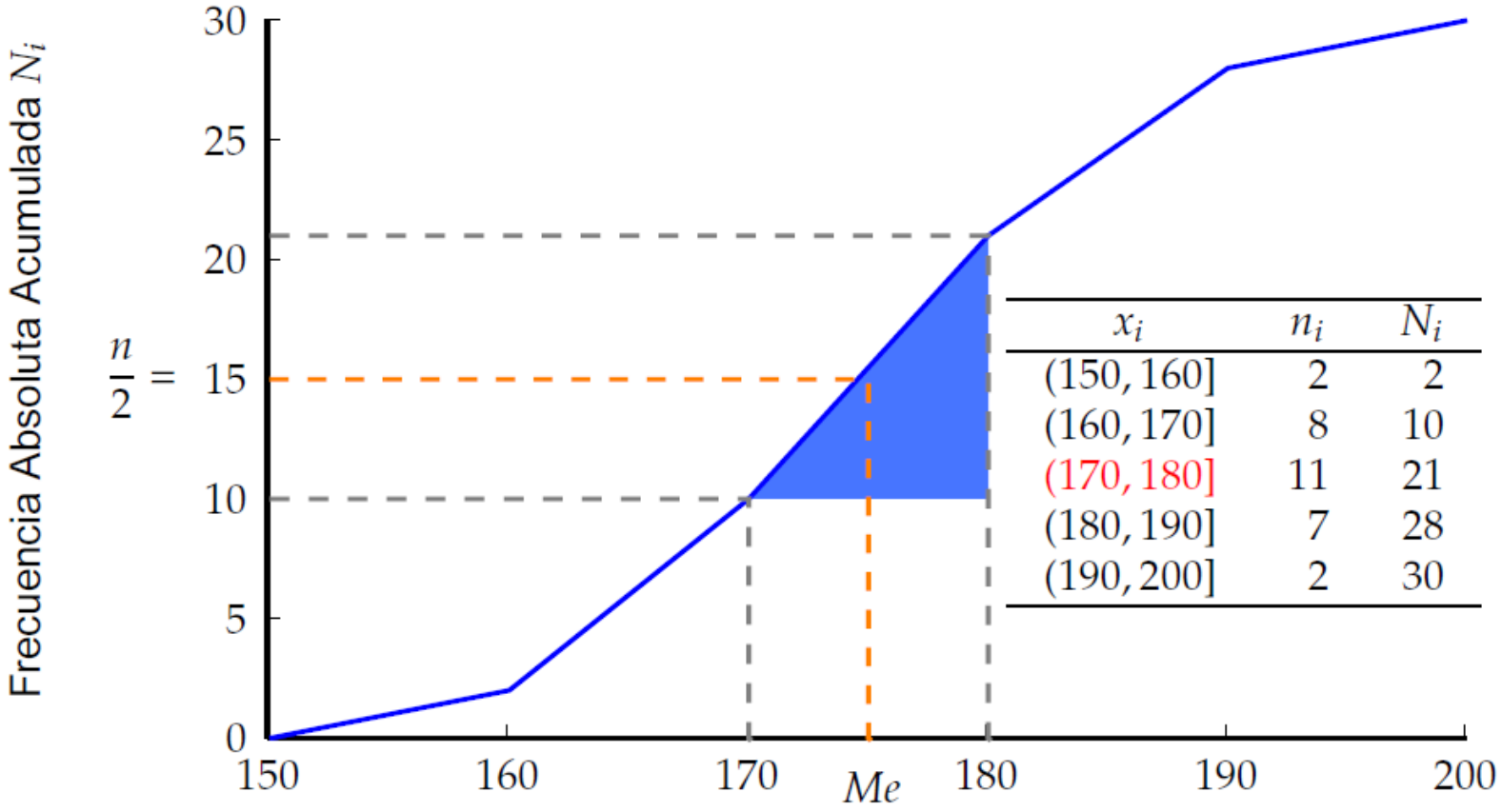
$$\text{tg}(\alpha) = \frac{N_i - N_{i-1}}{l_i - l_{i-1}}$$

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

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

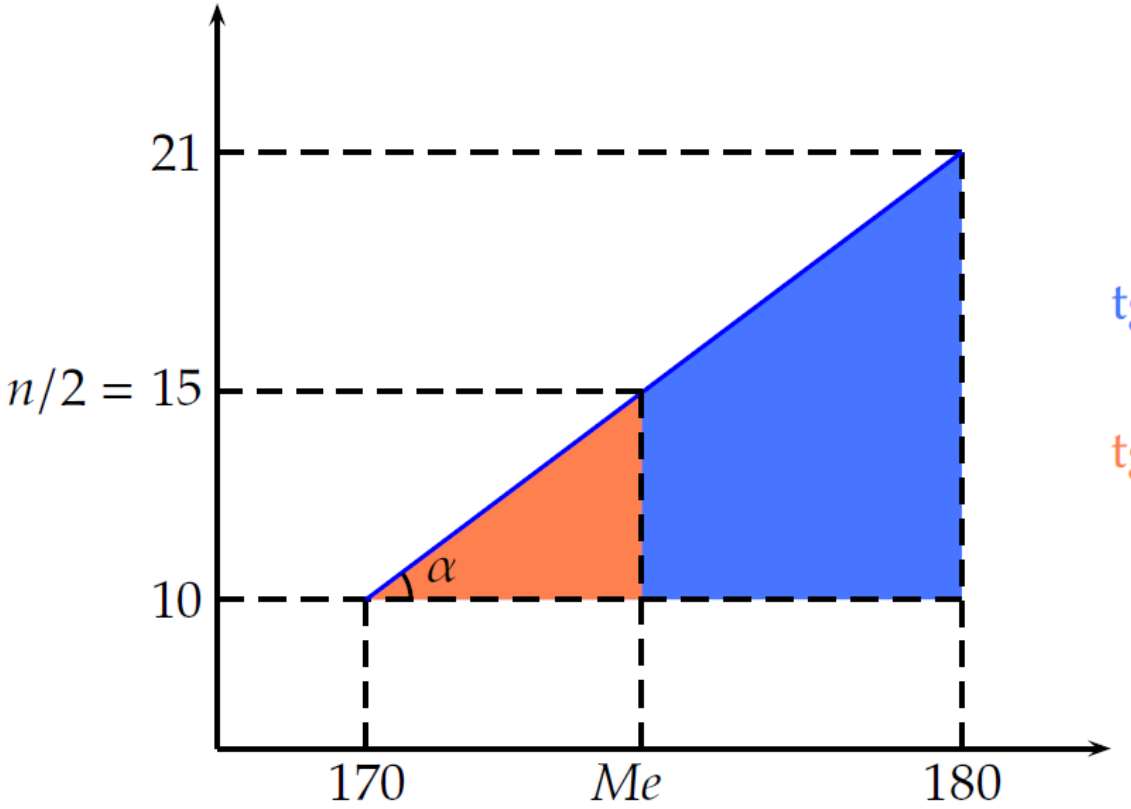
Estadísticos de posición: Mediana

En el ejemplo del número de empleados $^{n}/_{2} = {}^{30}/_{2} = 15$. Si se mira en el polígono de frecuencias acumuladas comprobamos que la mediana caerá en el intervalo (170, 180]



x_i	n_i	N_i
(150, 160]	2	2
(160, 170]	8	10
(170, 180]	11	21
(180, 190]	7	28
(190, 200]	2	30

Estadísticos de posición: Mediana



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

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

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

Estadísticos de posición: Moda

- Definición (Moda muestral Mo)

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

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

En ocasiones puede haber más de una moda.

- Cálculo de la moda

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

Y en el ejemplo del número de empleados también puede observarse en la tabla de frecuencias que la clase modal es $Mo = (170, 180]$.

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

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

¿Qué estadístico de tendencia central utilizar?

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

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

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

Por ejemplo, consideramos la siguiente muestra del número de préstamos concedidos de 7 familias:

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

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

¿Qué representante de la muestra tomarías?

Otros estadísticos de posición: 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.

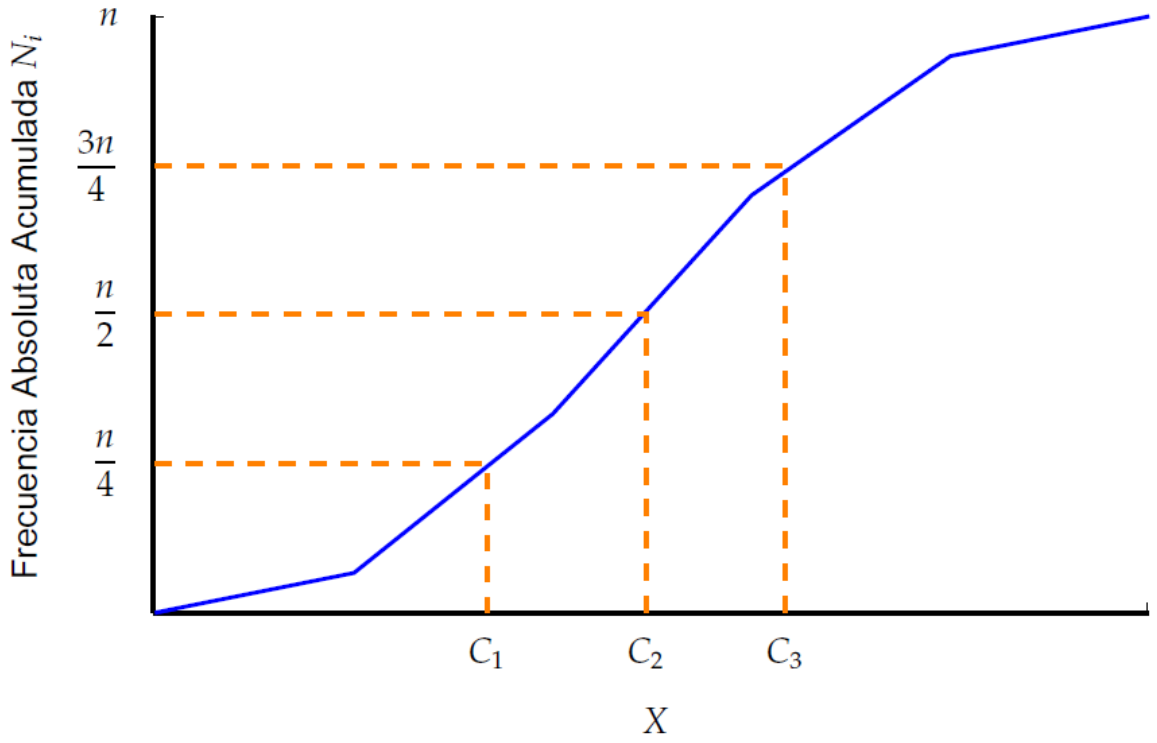
Los más utilizados son:

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

Otros estadísticos de posición: Cuantiles

- Cálculo de los cuantiles:

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



Estadísticos de dispersión

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

Para las variables cuantitativas, las más utilizadas son:

- Recorrido
- Rango intercuartílico
- Varianza
- Desviación típica
- Coeficiente de variación

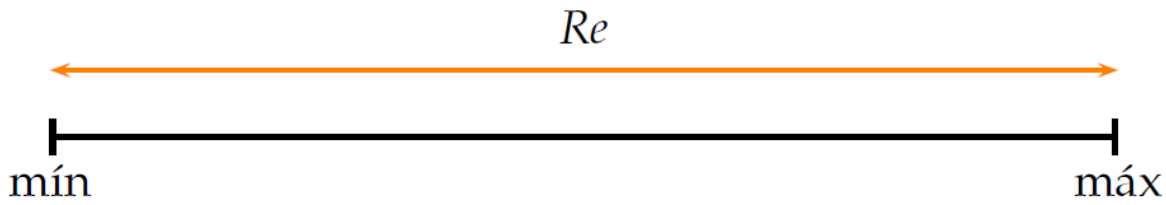
Estadísticos de dispersión: Recorrido

- Definición (Recorrido muestral Re)

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



Estadísticos de dispersión: Rango intercuartílico

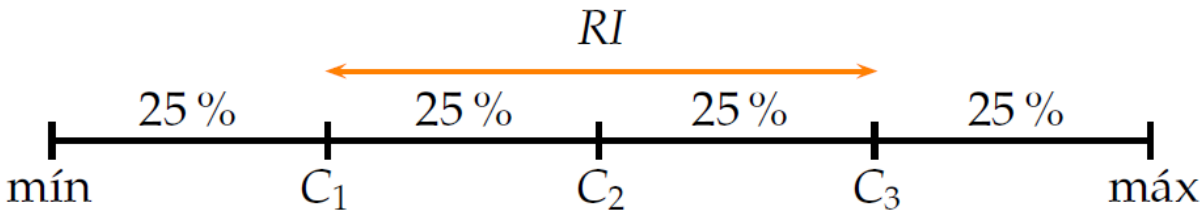
Para evitar el problema de los datos atípicos en el recorrido, se puede utilizar el primer y tercer cuartil en lugar del mínimo y el máximo.

- Definición (Rango intercuartílico muestral RI)

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

$$RI = C_3 - C_1$$

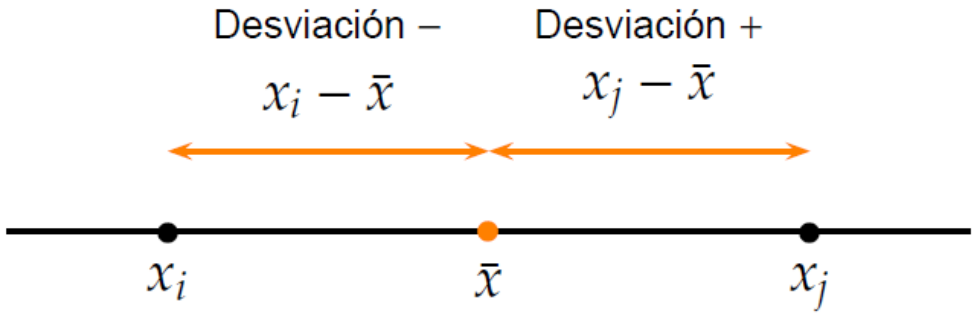
El rango intercuartílico da una idea de la variación que hay en el 50% de los datos centrales.



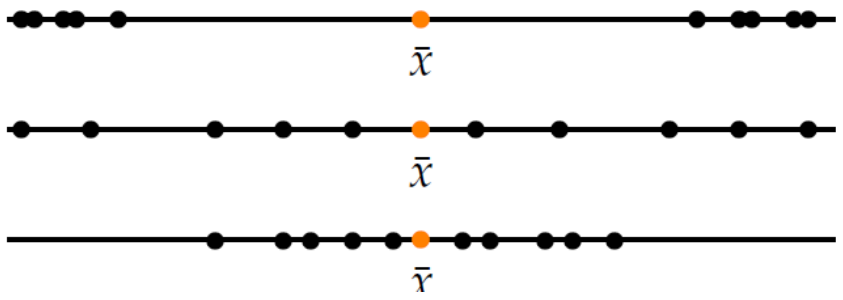
Estadísticos de dispersión: Varianza y desviación

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 considerar la distancia de cada valor a la media. A este valor se le llama desviación respecto de la media.



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



Estadísticos de dispersión: Varianza y desviación

- Definición (Varianza muestral 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.

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 n_i}{n} = \sum_{i=1}^{n}(x_i - \bar{x})^2 f_i$$

También puede calcularse de manera más sencilla mediante la fórmula

$$S^2 = \frac{\sum_{i=1}^{n} x_i^2 n_i}{n} - \bar{x}^2 = \sum_{i=1}^{n} x_i^2 f_i - \bar{x}^2$$

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

- Definición (Desviación típica muestral S)

La desviación típica muestral de una variable X se define como la raíz cuadrada positiva de su varianza muestral $S = \sqrt{S^2}$

Estadísticos de dispersión: Varianza y desviación

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

- Cálculo de la varianza y la desviación típica (datos agrupados)

En el ejemplo del número de empresas se puede calcular la varianza a partir de la tabla de frecuencias añadiendo una nueva columna con los cuadrados de los valores:

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

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

Y la desviación típica es $S = \sqrt{102{,}06} = 10{,}1$

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

Estadísticos de dispersión: Coeficiente de variación

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

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

- Definición (Coeficiente de variación muestral cv)

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

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

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

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

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

Estadísticos de forma:

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

Los aspectos más relevantes son:

- Simetría:

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

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

- Apuntamiento:

Miden el apuntamiento de la distribución de frecuencias.

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

Estadísticos de forma: Coeficiente de asimetría

- Definición (Coeficiente de asimetría muestral g_1)

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

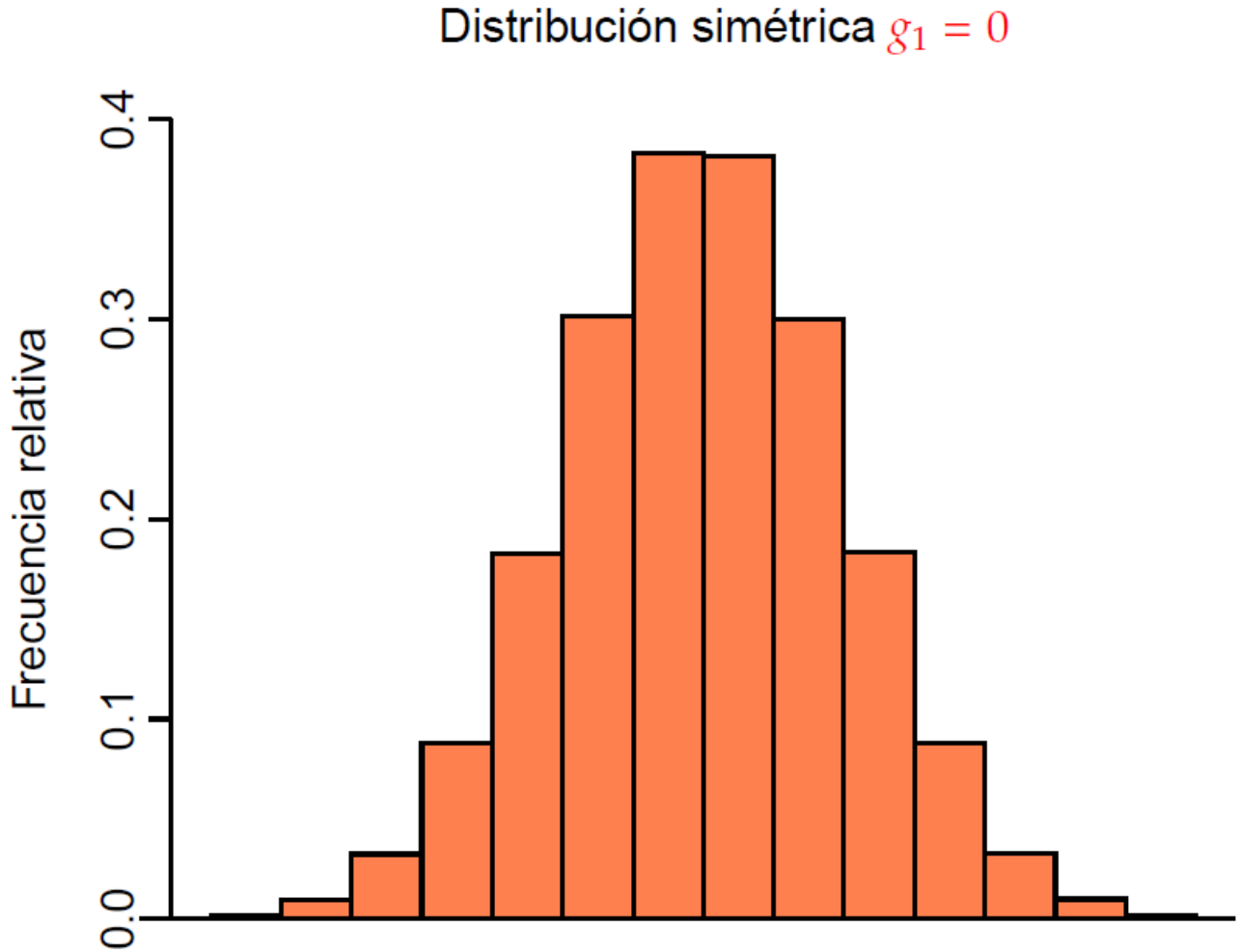
$$g_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 \, n_i/n}{S^3} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3 \, f_i}{S^3}$$

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

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

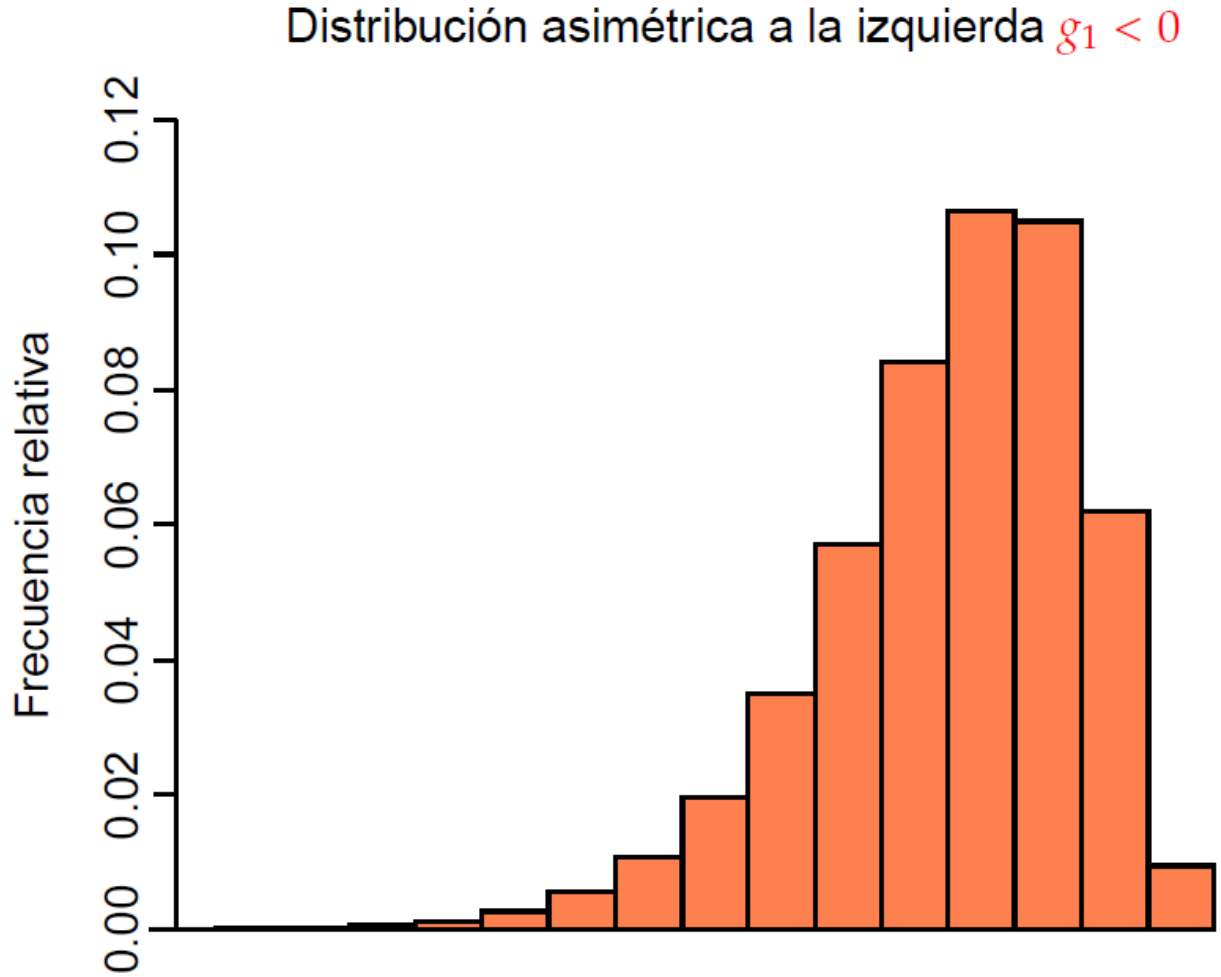
Estadísticos de forma: Coeficiente de asimetría

- Ejemplo de distribución simétrica



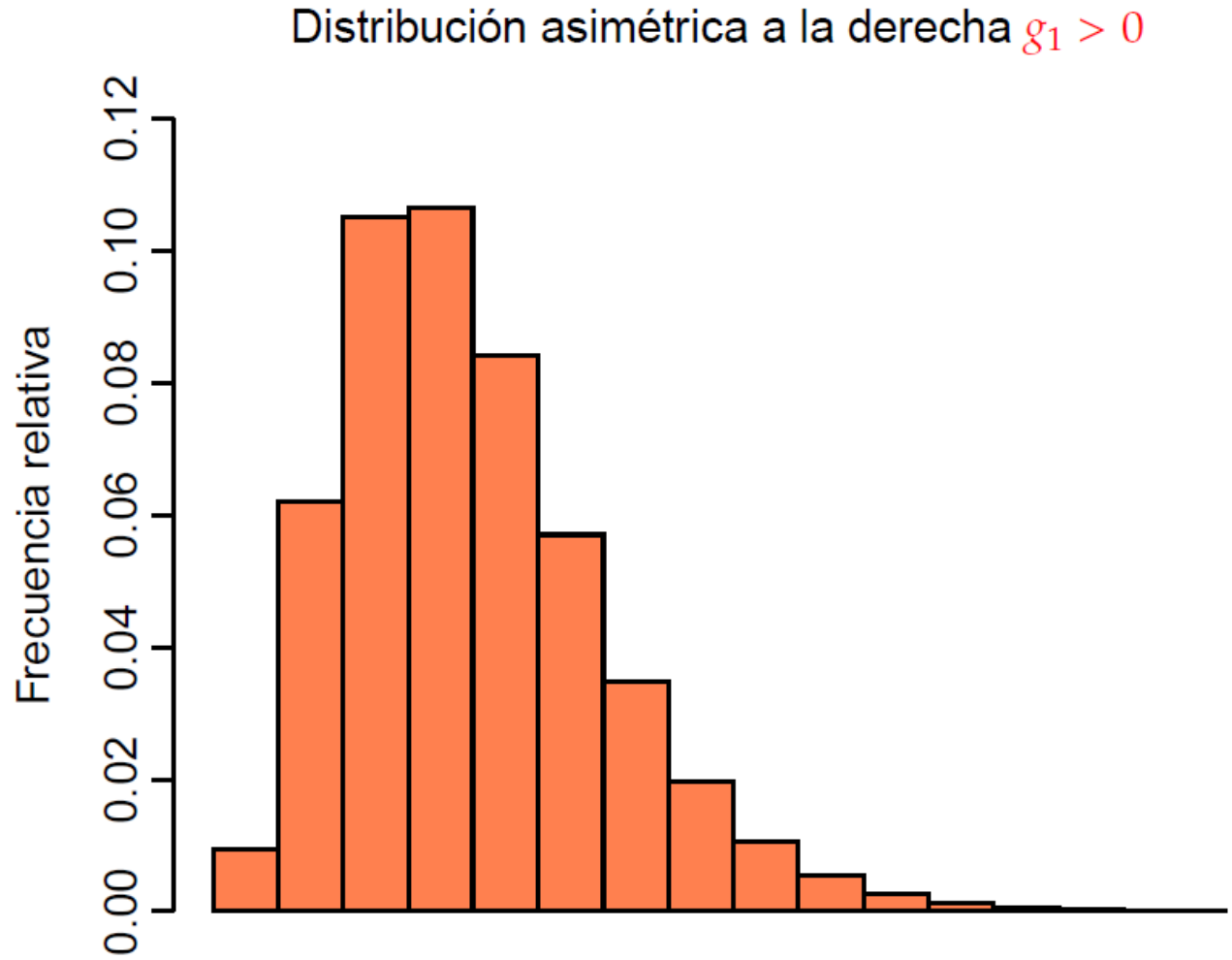
Estadísticos de forma: Coeficiente de asimetría

- Ejemplo de distribución asimétrica hacia la izquierda



Estadísticos de forma: Coeficiente de asimetría

- Ejemplo de distribución asimétrica hacia la derecha



Estadísticos de forma: Coeficiente de asimetría

Cálculo del coeficiente de asimetría

Siguiendo con el ejemplo del número de empleados se puede 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 $\bar{x} = 174{,}67$:

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

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

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

Estadísticos de forma: Coeficiente de apuntamiento o curtosis

Definición (Coeficiente de apuntamiento muestral g_2)

El coeficiente de apuntamiento muestral de una variable X se define como el promedio de las desviaciones de los valores de la muestra respecto de la media muestral, elevadas a la cuarta, dividido entre la desviación típica a la cuarta.

$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \, n_i/n}{S^4} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \, f_i}{S^4}$$

En la práctica se mide la curtosis con respecto a la distribución Normal ($g_2 = 3$):

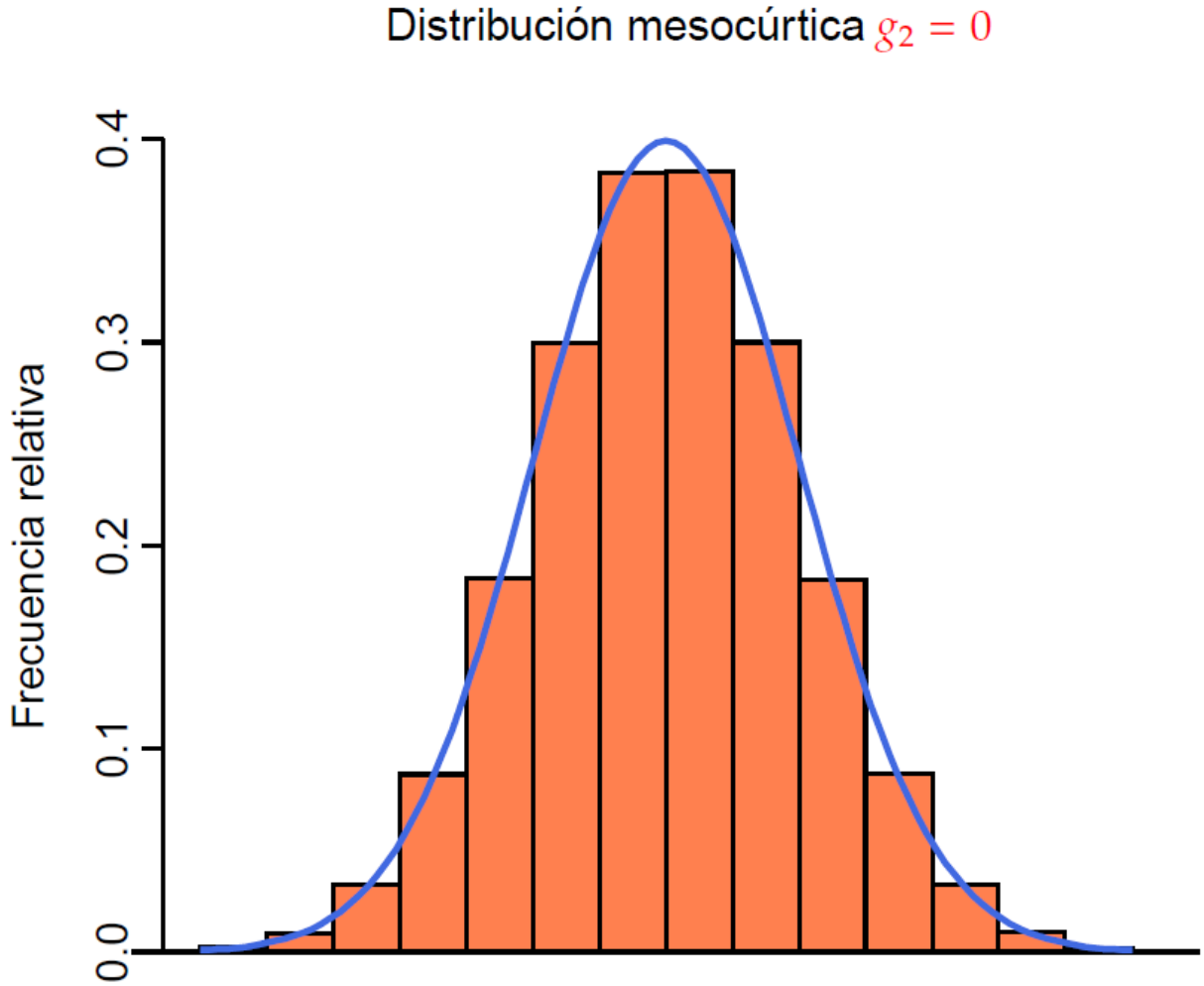
$$g_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \, n_i/n}{S^4} - 3 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4 \, f_i}{S^4} - 3$$

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

- $g_2 = 0$ indica que la distribución tiene 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).

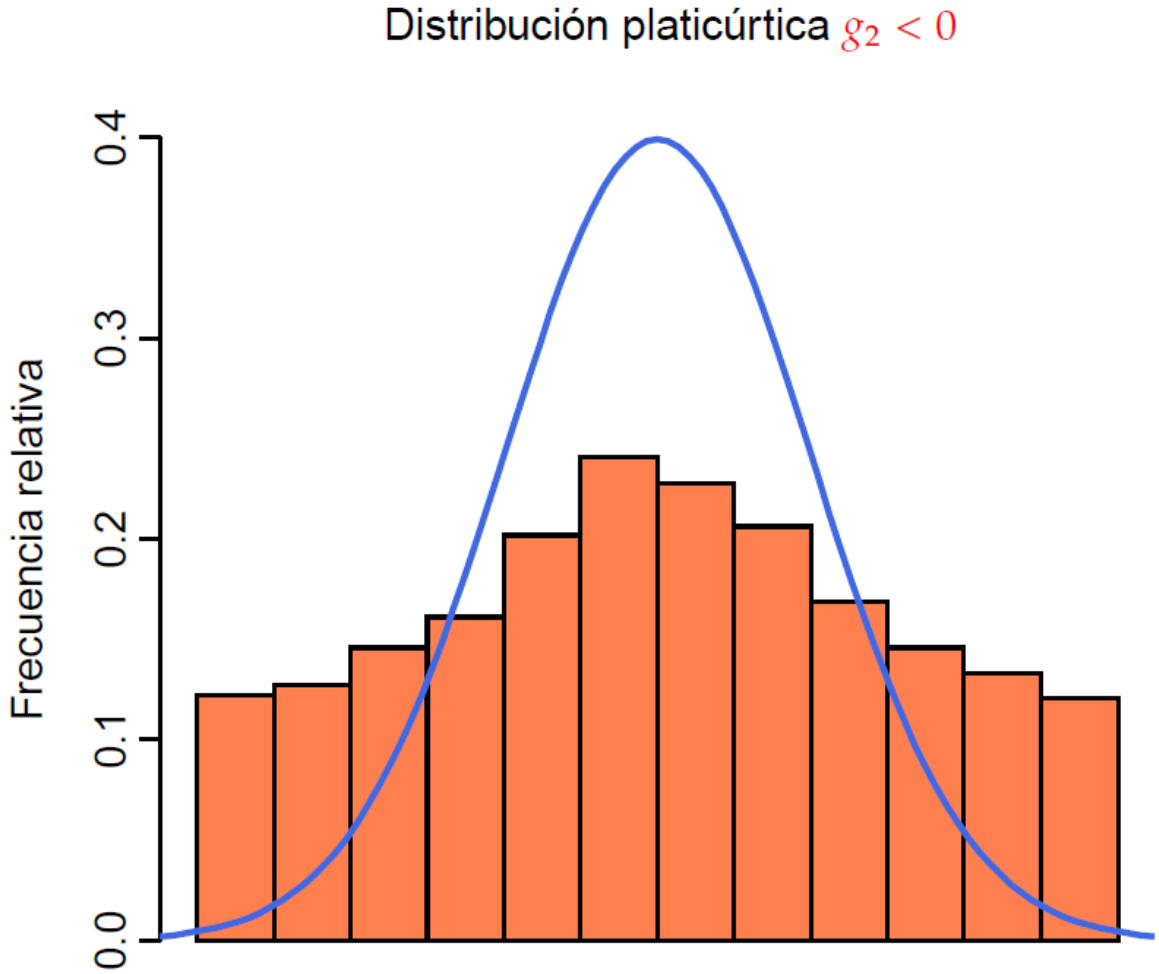
Estadísticos de forma: Coeficiente de apuntamiento o curtosis

- Ejemplo de distribución mesocúrtica



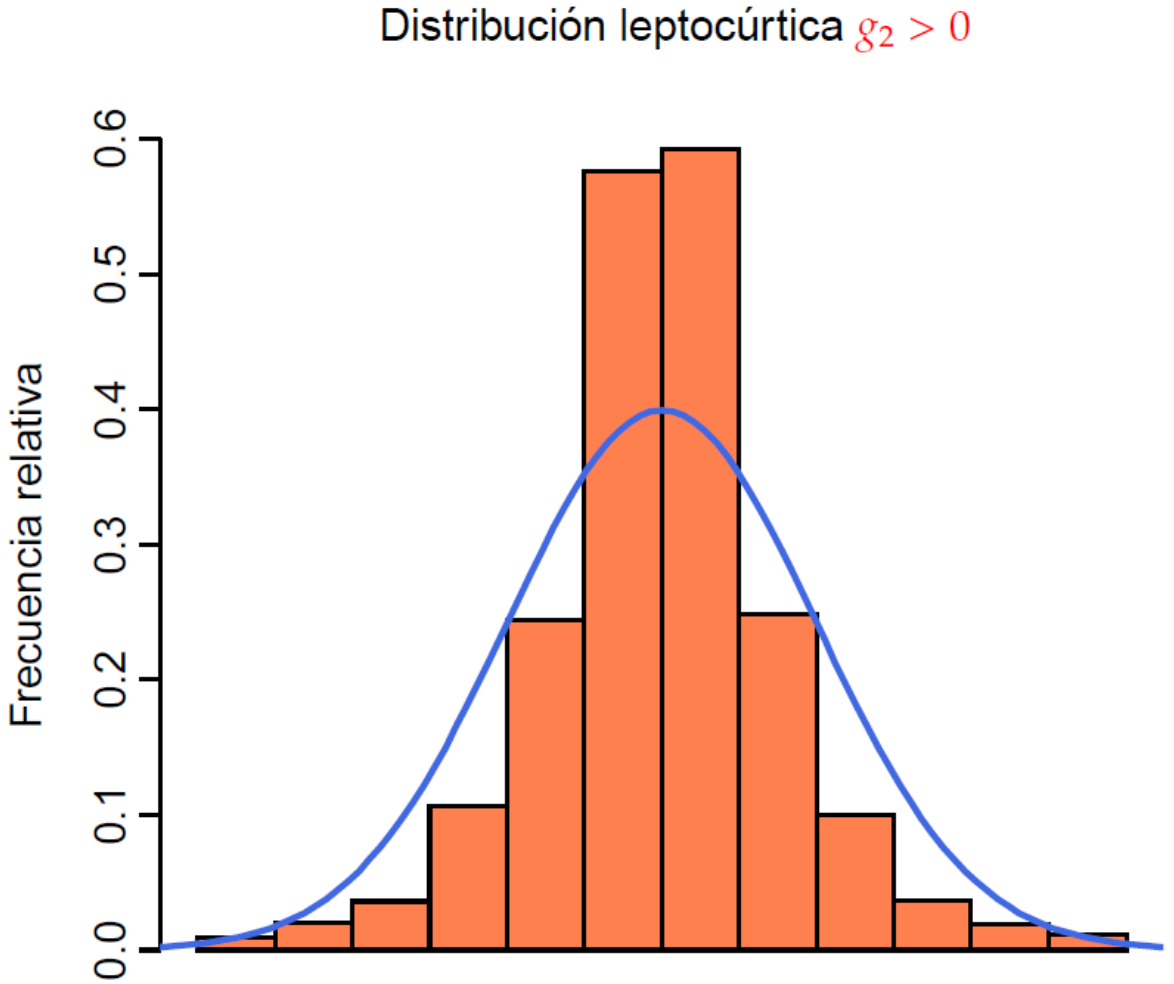
Estadísticos de forma: Coeficiente de apuntamiento o curtosis

- Ejemplo de distribución platicúrtica



Estadísticos de forma: Coeficiente de apuntamiento o curtosis

- Ejemplo de distribución leptocúrtica



Estadísticos de forma: Coeficiente de apuntamiento o curtosis

Cálculo del coeficiente de apuntamiento

De nuevo con el ejemplo del número de empleados se puede calcular el coeficiente de curtosis a partir de la tabla de frecuencias añadiendo una nueva columna con las desviaciones a la media $\bar{x} = 174{,}67$ elevadas a la cuarta:

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

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

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

Interpretación de los coeficientes de asimetría y apuntamiento

Muchas de las pruebas estadísticas solo pueden aplicarse a poblaciones normales (inferencia estadística).

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

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

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

Transformaciones de variables

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

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

$$y = \alpha + \beta x$$

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

$$\bar{y} = \alpha + \beta \bar{x},$$

$$S_y = |\beta| S_x$$

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

Transformaciones de variables

Una de las transformaciones lineales más habituales es la tipificación:

- Definición (Variable tipificada)

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

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

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

Los valores tipificados se conocen como puntuaciones típicas y miden el número de desviaciones típicas que dista cada observación de la media, 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:

$$\bar{z} = 0 \qquad S_z = 1$$

Transformaciones de variables

La calificación crediticia de 5 empresas proporcionada por las agencias de *rating* X e Y son:

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

¿Tienen el mismo rendimiento las empresas con una puntuación de 8?

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

X:	$-1{,}5$	0	$-0{,}5$	$1{,}5$	$0{,}5$
Y:	$-1{,}26$	$1{,}26$	$0{,}95$	0	$-0{,}95$

Es decir, la empresa que tiene una calificación de 8 según X está 1,5 veces la desviación típica por encima de la media de su grupo, mientras que la empresa que tiene un 8 según Y sólo está 0,95 desviaciones típicas por encima de su media. Así pues, la primera empresa tuvo un rendimiento superior a la segunda.

Transformaciones de variables

Siguiendo con el ejemplo anterior

¿Cuál es la empresa que posee mejor calificación crediticia?

Si simplemente sumamos las calificaciones de cada agencia crediticia tenemos:

	1	2	3	4	5
X:	2	5	4	8	6
Y:	1	9	8	5	2
$\sum$	3	14	12	13	8

La empresa con mejor calificación global es la segunda.

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

	1	2	3	4	5
X:	−1,5	0	−0,5	1,5	0,5
Y:	−1,26	1,26	0,95	0	−0,95
$\sum$	−2,76	1,26	0,45	1,5	−0,45

La empresa con mejor calificación es la cuarta.

Relaciones entre variables

Hasta ahora se ha visto cómo 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 la probabilidad de *default* de una empresa, se deberían incluir todas las variables con las que podría tener relación: valor contable de la compañía (*book value to equity*), beneficios netos no distribuidos, tasa de dividendos, beneficio por acción, ingresos netos, etc.

El objetivo de la estadística en este caso es dar medidas del grado y del tipo de relación entre dichas variables.

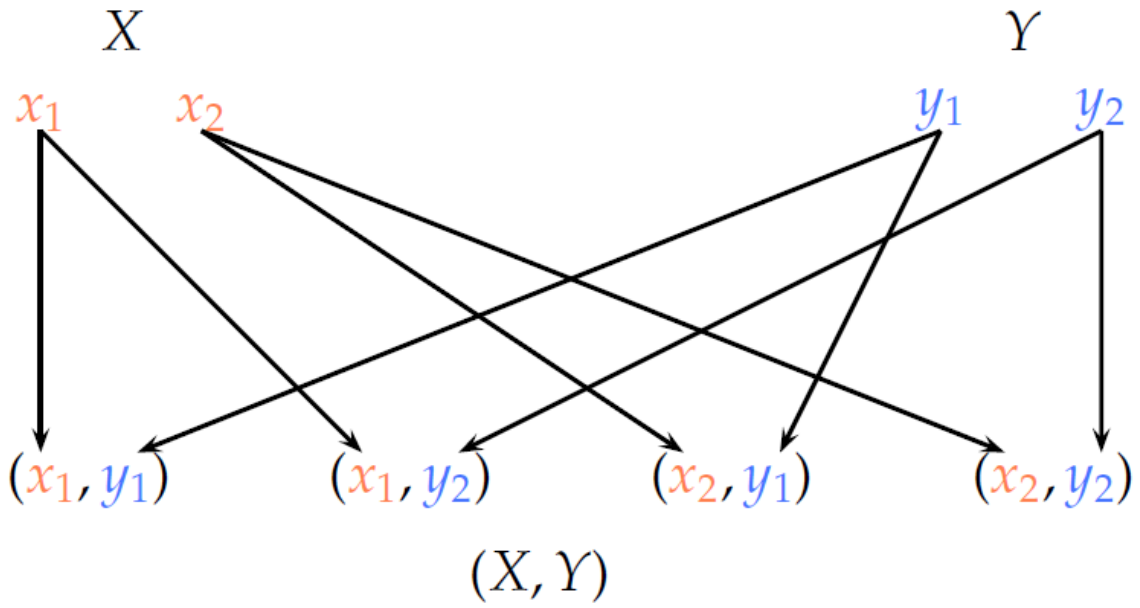
Generalmente, se considera una **variable dependiente** Y que se supone relacionada con otras variables $X_1, \dots, X_n$ llamadas **variables independientes**.

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

Variables bidimensionales

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

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



Frecuencias de una variable bidimensional

- Definición (Frecuencias muestrales de una variable bidimensional)

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

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

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

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

¡Ojo! Comprobar si tiene sentido para las variables bidimensionales calcular las frecuencias acumuladas.

Distribución de frecuencias bidimensional

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

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

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

Distribución de frecuencias bidimensional

- Ejemplo bidimensional

Se ha registrado el número de empleados de 2 filiales de 30 empresas multinacionales:

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

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

Diagrama de dispersión

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

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

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

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

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

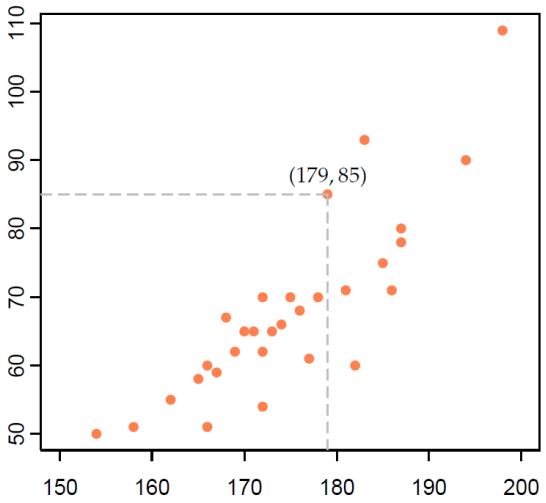
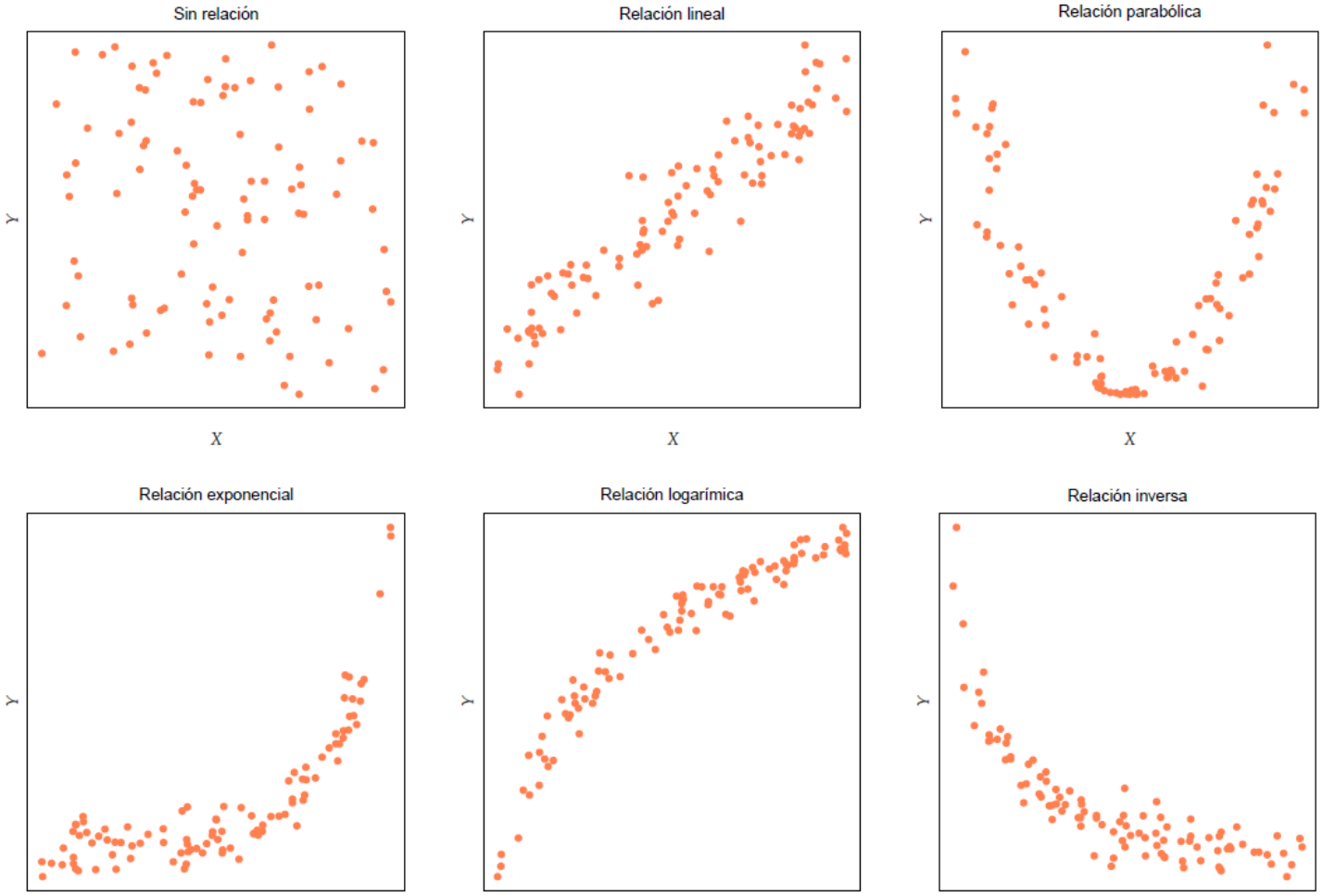


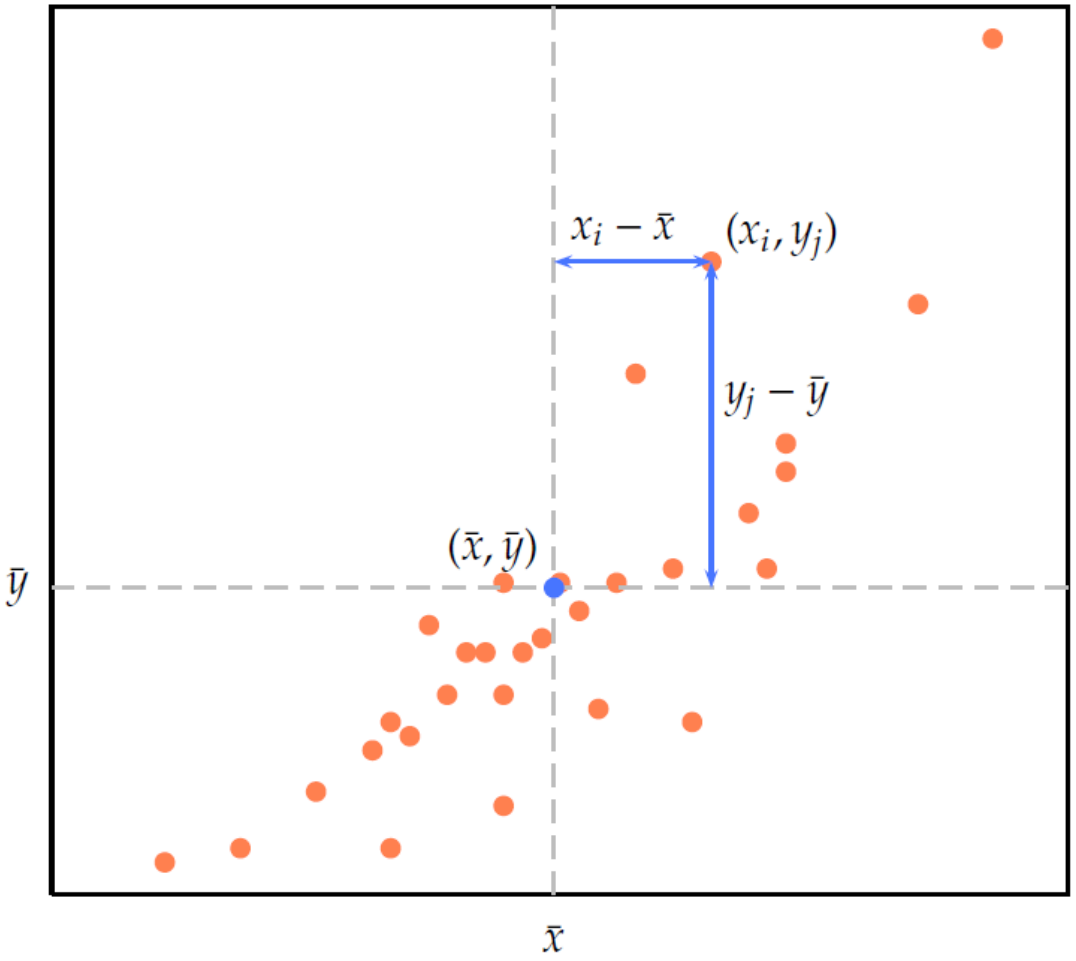
Diagrama de dispersión

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



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.

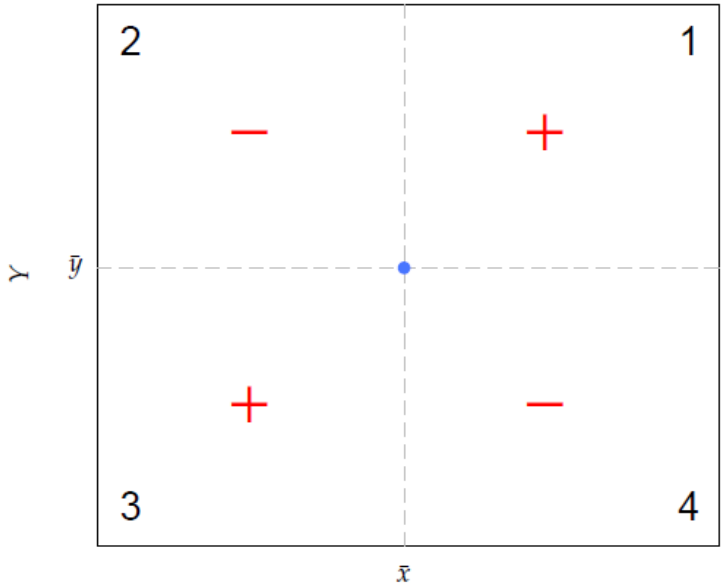


Desviaciones respecto de las medias

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

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

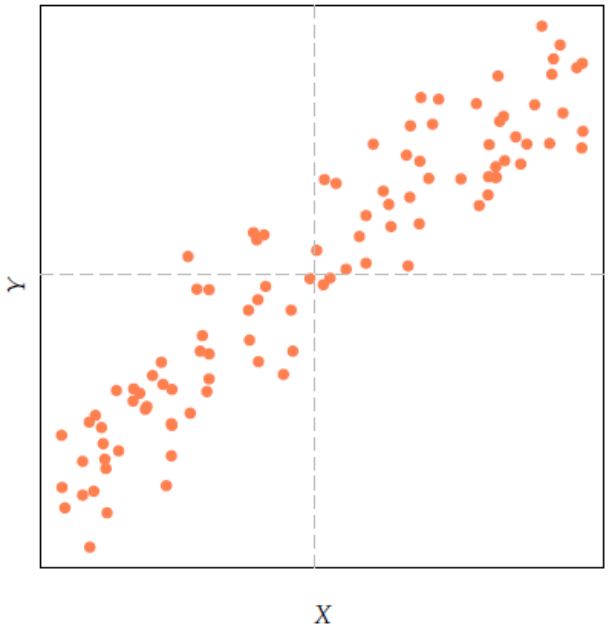
Signo del producto de desviaciones



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.

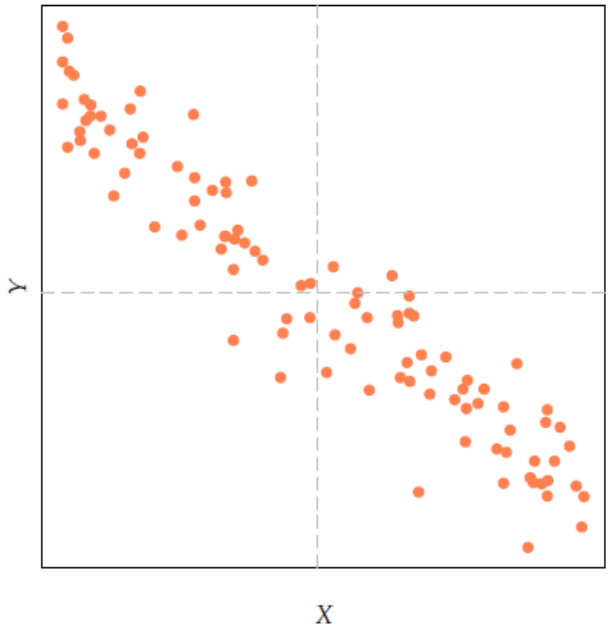
Relación lineal creciente



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

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.

Relación lineal decreciente



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

Covarianzas

Definición (Covarianza muestral)

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

$$S_{xy} = \frac{\sum_{i,j}(x_i - \bar{x})(y_j - \bar{y})n_{ij}}{n}$$

También puede calcularse de manera más sencilla mediante la fórmula:

$$S_{xy} = \frac{\sum_{i,j} x_i y_j n_{ij}}{n} - \bar{x}\bar{y}.$$

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 una relación lineal entre las variables.

Covarianzas

En el ejemplo del número de empleados en dos filiales distintas de 30 multinacionales, 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

$$\bar{x} = 174{,}67 \qquad \bar{y} = 69{,}67$$

La covarianza vale

$$s_{xy} = \frac{\sum x_i y_j n_{ij}}{n} - \bar{x}\bar{y} = \frac{155 \cdot 55 \cdot 2 + 165 \cdot 55 \cdot 4 + \cdots + 195 \cdot 105 \cdot 1}{30} - 174{,}67 \cdot 69{,}67$$

$$= \frac{368200}{30} - 12169{,}26 = 104{,}07$$

Lo que indica que existe una relación lineal creciente entre el número de empleados entre filiales.

Coeficiente de correlación de Pearson

- Definición (Coeficiente de correlación muestral)

El coeficiente de correlación mide el grado de dependencia según la relación planteada por el modelo. Toma valores entre -1 y 1:

$$-1 \leq \rho_{xy} \leq 1$$

- Si $\rho_{xy} = 0$ no existe una relación entre las variables.
- Si $\rho_{xy} = 1$ existe una relación perfecta y creciente entre las variables.
- Si $\rho_{xy} = -1$ existe una relación perfecta y decreciente entre las variables.

$$\rho_{xy} = \frac{S_{xy}}{S_x S_y}$$

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.

- Definición (Coeficiente de correlación de Spearman)

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

$$\rho_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

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

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:

$$-1 \leq \rho_s \leq 1,$$

de manera que:

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

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

Coeficiente de correlación de Spearman

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

	X	Y	d_i	d_i^2
1	2	3	−1	1
2	5	4	1	1
3	1	2	−1	1
4	3	1	2	4
5	4	5	−1	1
$\sum$			0	8

El coeficiente de correlación de Spearman para esta muestra es:

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

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

Coeficiente Chi-cuadrado

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

- Definición (Coeficiente Chi-cuadrado χ^2)

Dada una muestra de 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 las variables son independientes, y crece a medida que aumenta la dependencia entre las variables.

Coeficiente Chi-cuadrado

Una empresa de seguros está interesada en realizar un estudio para ver si existe relación entre el sexo y haber tenido un siniestro. La tabla de contingencia resultante es

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

se obtienen las siguientes frecuencias esperadas:

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

El coeficiente Chi-cuadrado es igual a

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

Lo que indica que no existe gran relación entre el sexo y haber tenido un accidente.

Coeficiente de contingencia

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

Para evitar estos problemas se suele utilizar el siguiente estadístico:

- Definición (Coeficiente de contingencia)

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

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

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.

Coeficiente de contingencia

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 $\sqrt{(k-1)k}$

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 x 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 ese valor, se puede concluir que no existe demasiada relación entre tener un siniestro y el sexo de la persona.

4 | Otras representaciones usuales de variables

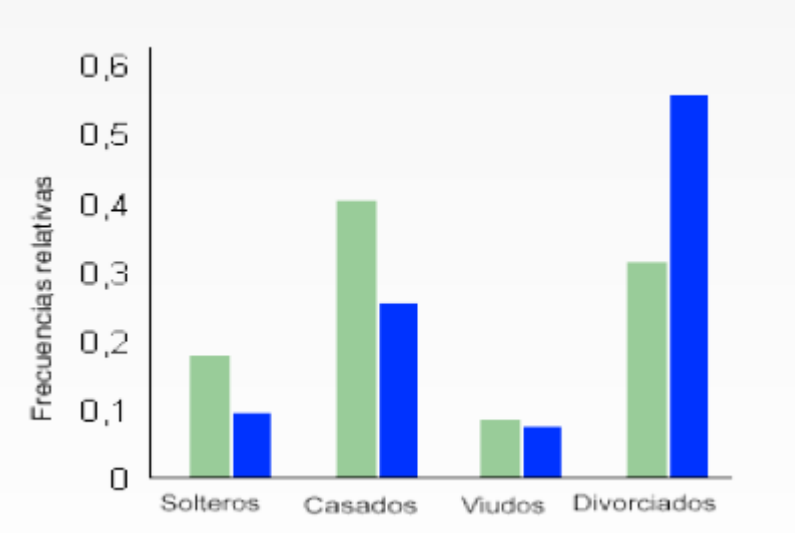
4. Otras representaciones usuales de variables

- **Variables cuantitativas:**
  - Discretas
    - Diagrama de barras
    - Diagrama de escalera
    - Polígono de frecuencias
  - Continuas
    - Histograma
    - Gráficos de sectores
    - Gráficos de línea
    - Box and whisker plot
- **Variables cualitativas**
    - Diagrama de barras
    - Diagrama de escalera
    - Gráficos de sectores
    - Gráfico de tallo y hojas (stem and leaf plot)

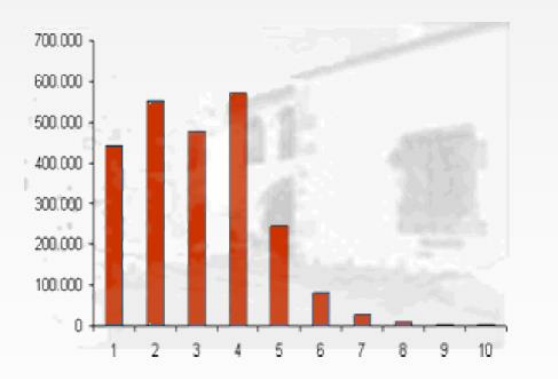
4. Otras representaciones usuales de variables

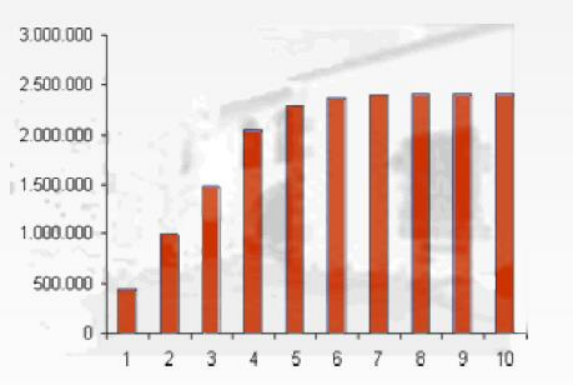
Diagrama de barras

Estado civil de los solicitantes de un préstamo



Distribución de las viviendas según el número de personas que habitan en ellas





4. Otras representaciones usuales de variables

Gráfico de tallo y hojas (stem and leaf plot)

Cada dato representa su valor y a la vez ocupa su espacio de forma que obtenemos simultáneamente la representación de los datos y el perfil de la distribución de la variable.

Cantidad solicitada en un préstamo (miles de €)

145	147	149	152	153	154	154	156	157	158
162	162	162	163	163	164	164	165	167	167
168	169	169	170	171	171	172	173	174	174
175	176	176	178	179	180	181	183	185	186

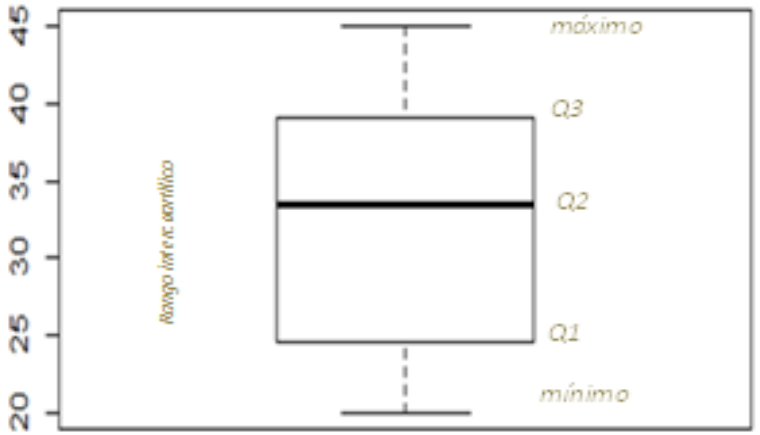
Tallo	Hoja
14	5 7 9
15	2 3 4 4 6 7 8
16	1 2 2 3 3 4 4 5 7 7 8 9 9
17	0 1 1 2 3 4 4 5 6 6 8 9
18	0 1 3 5 6

4. Otras representaciones usuales de variables

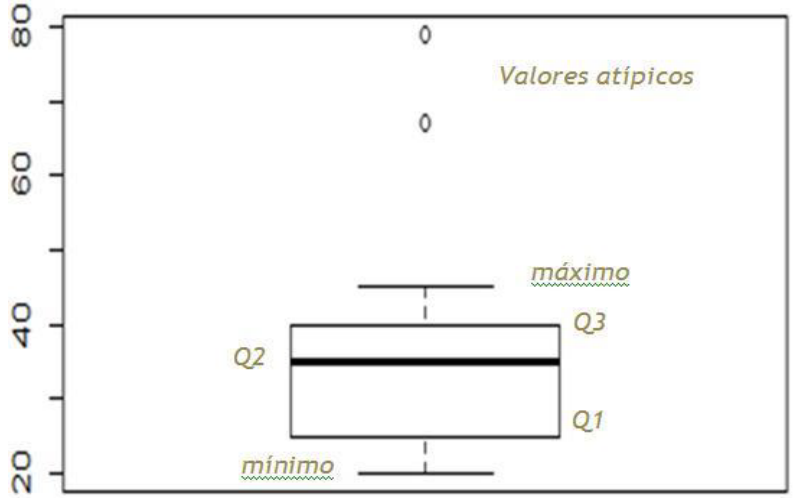
Box and whisker plot

Representación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría de una variable.

Cantidad concedida en préstamos personales (miles de €) durante el mes de septiembre



Cantidad concedida en préstamos personales (miles de €) durante el mes de agosto



4. Otras representaciones usuales de variables

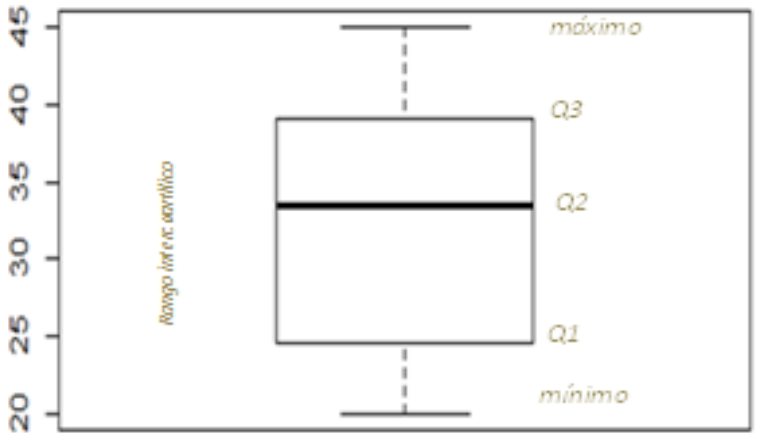
Box and whisker plot

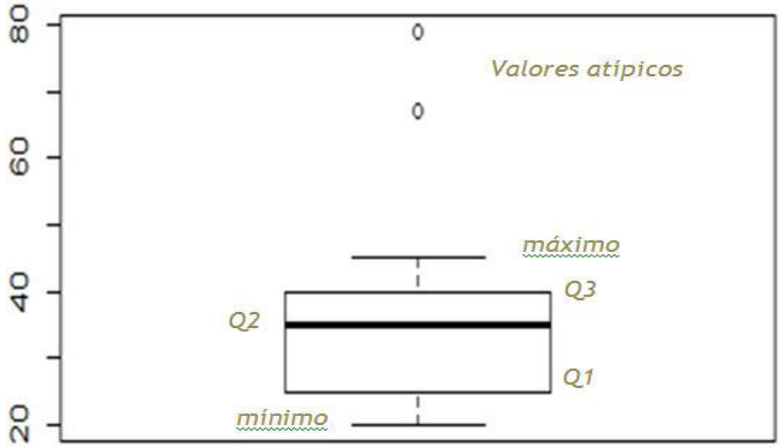
Rango intercuartílico: El 50% de las observaciones se encuentran en el intervalo (Q1-Q3)

Box (Caja): Mediana Q2 junto con los dos cuartiles Q1 y Q3

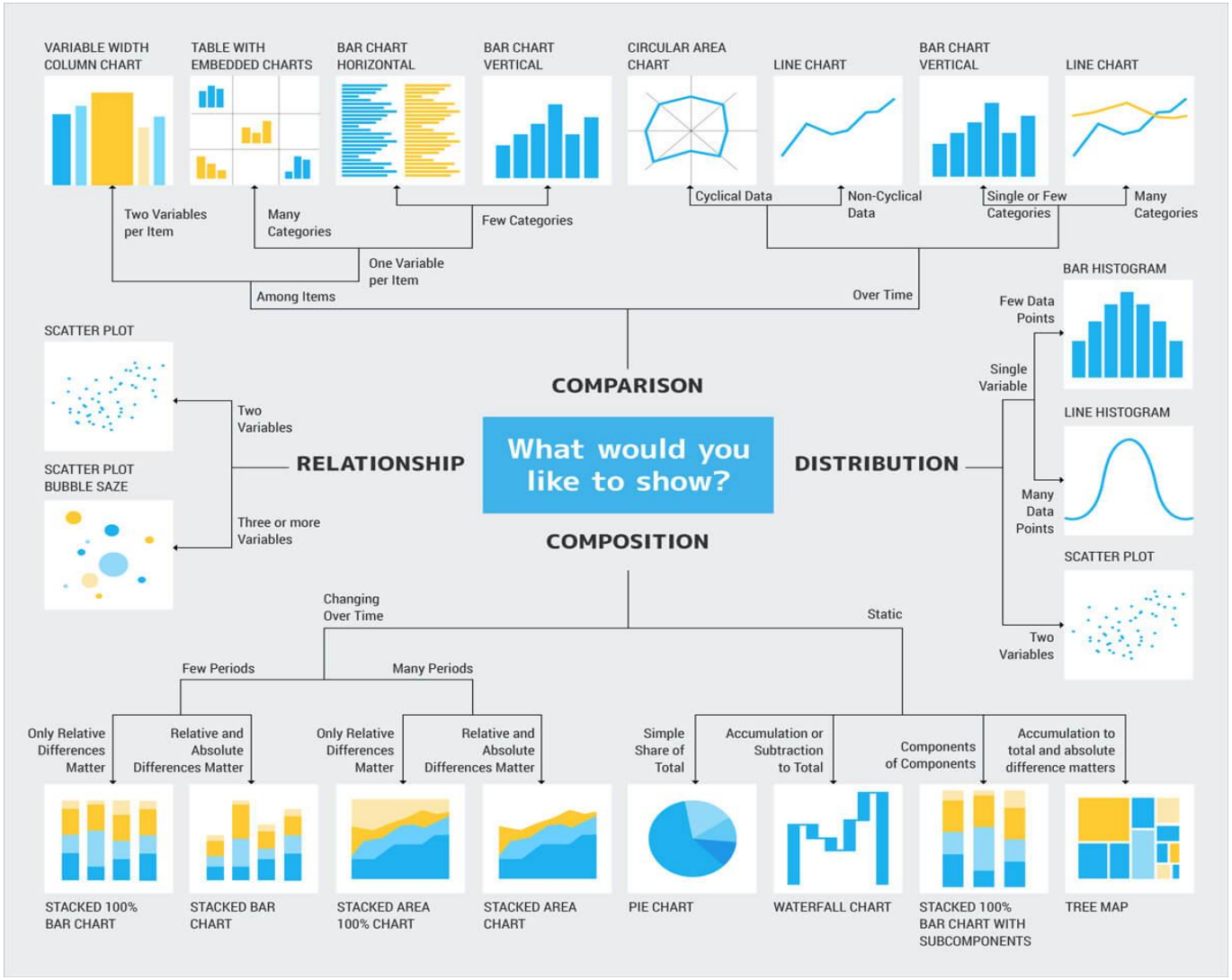
Whisker (Bigotes): Desde la caja hasta el valor mínimo y máximo de las observaciones que no superen 1.5 veces el rango intercuartílico

Valores atípicos: Los que están fuera del intervalo (Q1-1.5RIC,Q3+1.5RIC)



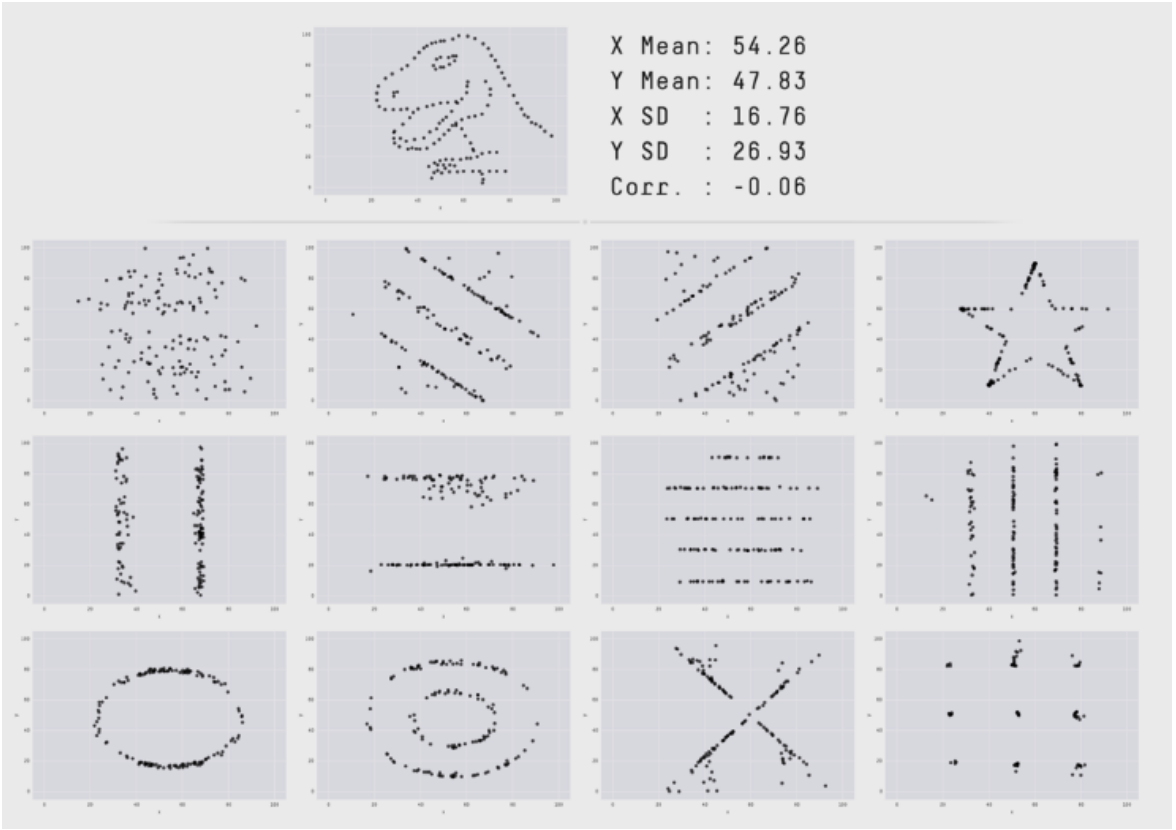


4. Otras representaciones usuales de variables



4. Ejemplos

Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics
https://www.autodeskresearch.com/publications/samestats



“Never trust summary statistics alone; always visualize your data"

Referencias

VISUAL VOCABULARY
https://gramener.github.io/visual-vocabulary-vega/#

From data to Viz | Find the graphic you need https://www.data-to-viz.com

EXAMPLES
https://cran.r-project.org/web/packages/googleVis/vignettes/googleVis_examples.html

https://www.r-graph-gallery.com/

CHEATSHEETS
https://github.com/ShivamPanchal/Complete-CheatSheets

http://www.cheat-sheets.org/saved-copy/427513-Statistics-Reference-Cheatsheet.pdf

https://static1.squarespace.com/static/54bf3241e4b0f0d81bf7ff36/t/55e9494fe4b011aed10e48e5/1441352015658/probability_cheatsheet.pdf

https://media.sumo.com/3d2bc8f118adc00c8fb6de785b9868b6d428cbfb06b9faa36313bbac4a329da3

MATHEMATICS FOR MACHINE LEARNING https://mml-book.github.io/book/mml-book.pdf

Afi Escuela de Finanzas