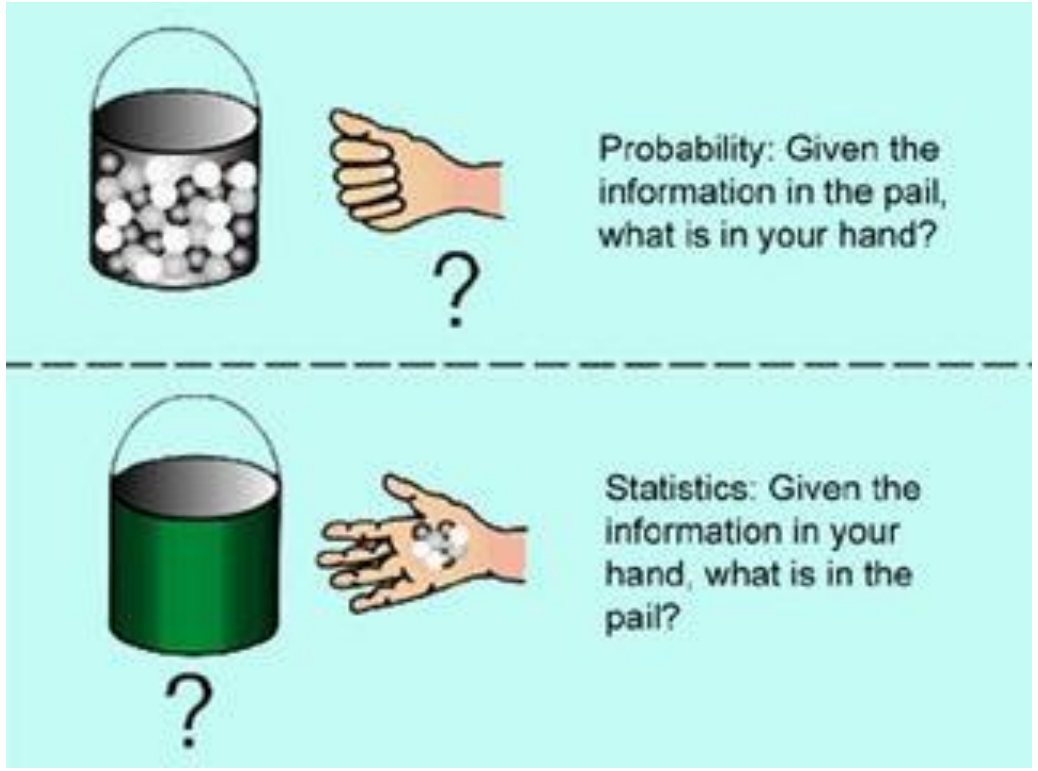
Estadística Descriptiva

Javier Martínez
Xavier Vilasís

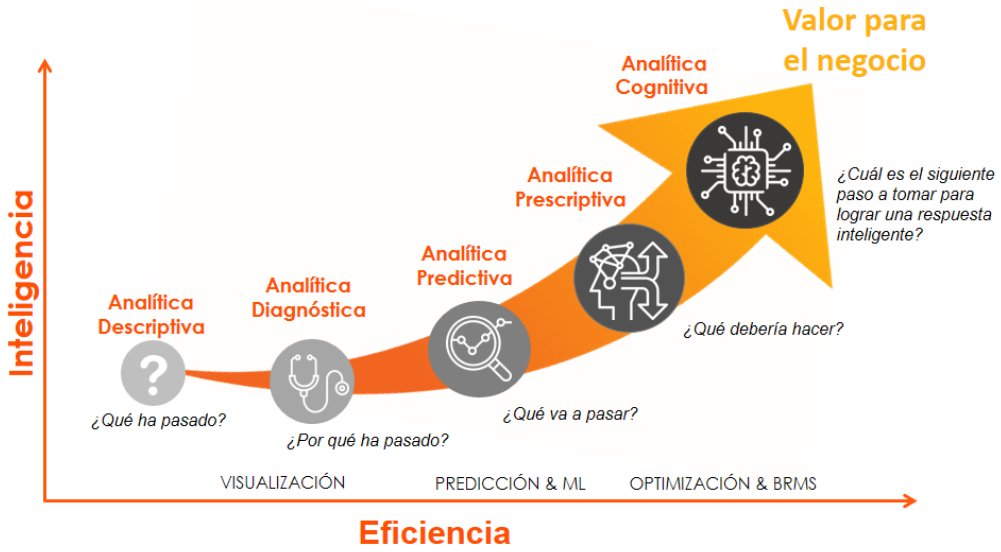
Master Universitario en Ciencia de los Datos/Data Science

Probabilidad vs Estadística



laSalle
UNIVERSITAT RAMON LLULL

Analítica de datos para tomar mejores decisiones



MISIÓN

Nos centraremos en estudiar las principales **técnicas estadísticas** que nos proporcionarán las herramientas necesarias para acabar entiendo **¿qué sucedió?** y poder hacer una aproximación a **¿por qué sucedió?** todo ello con el objetivo de avanzar hacía la inferencia y poder interpretar **¿qué sucederá?**

Obtención de datos

Proceso de recolección de datos necesarios para resolver los objetivos del proyecto.

Analítica descriptiva

¿Qué pasó?
Qué está pasando?

Estadística descriptiva, visualizaciones, clustering…

Analítica predictiva

¿Qué pasará?

Data mining, forecasting, machine learning, deep learning…

Analítica prescriptiva

¿Qué se tendría que hacer?

Optimización matemática, simulación…

laSalle
UNIVERSITAT RAMON LLULL

Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. **Conceptos Clave**
2. Medidas de Posición
3. Medidas de Dispersión
4. Medidas de Asimetría
5. Medidas de Asociación
6. Visualización
7. Consideraciones

laSalle
UNIVERSITAT RAMON LLULL

Estadística

La Estadística es la rama de las matemáticas que se encarga del estudio de una determinada característica en una **población**, **recogiendo datos**, **organizándolos** y **analizándolos** para extraer **conclusiones** de dicha población.

Estadística

La Estadística es la rama de las matemáticas que se encarga del estudio de una determinada característica en una **población**, **recogiendo datos**, **organizándolos** y **analizándolos** para extraer **conclusiones** de dicha población.

Estadística Descriptiva

La Estadística Descriptiva es un conjunto de **técnicas** y **procedimientos** que ayudan a **describir**, **mostrar** y **resumir** la información de un conjunto de datos.

Estadística

La Estadística es la rama de las matemáticas que se encarga del estudio de una determinada característica en una **población**, **recogiendo datos**, **organizándolos** y **analizándolos** para extraer **conclusiones** de dicha población.

Estadística Descriptiva

La Estadística Descriptiva es un conjunto de **técnicas** y **procedimientos** que ayudan a **describir**, **mostrar** y **resumir** la información de un conjunto de datos.

Estadística Inferencial

La Estadística Inferencial realiza un estudio descriptivo de un **subconjunto de la población** (muestra) con el objetivo de **extender los resultados** a toda la población.

laSalle
UNIVERSITAT RAMON LLULL

Estadística

La Estadística es la rama de las matemáticas que se encarga del estudio de una determinada característica en una **población**, **recogiendo datos**, **organizándolos** y **analizándolos** para extraer **conclusiones** de dicha población.

Estadística Descriptiva

La Estadística Descriptiva es un conjunto de **técnicas** y **procedimientos** que ayudan a **describir**, **mostrar** y **resumir** la información de un conjunto de datos.

Estadística Inferencial

La Estadística Inferencial realiza un estudio descriptivo de un **subconjunto de la población** (muestra) con el objetivo de **extender los resultados** a toda la población.

Universo

Conjunto de sujetos u observaciones con una característica en común, **observable** y **medible**.

laSalle
UNIVERSITAT RAMON LLULL

Estadística

La Estadística es la rama de las matemáticas que se encarga del estudio de una determinada característica en una **población**, **recogiendo datos**, **organizándolos** y **analizándolos** para extraer **conclusiones** de dicha población.

Estadística Descriptiva

La Estadística Descriptiva es un conjunto de **técnicas** y **procedimientos** que ayudan a **describir**, **mostrar** y **resumir** la información de un conjunto de datos.

Estadística Inferencial

La Estadística Inferencial realiza un estudio descriptivo de un **subconjunto de la población** (muestra) con el objetivo de **extender los resultados** a toda la población.

Universo

Conjunto de sujetos u observaciones con una característica en común, **observable** y **medible**.

laSalle
UNIVERSITAT RAMON LLULL

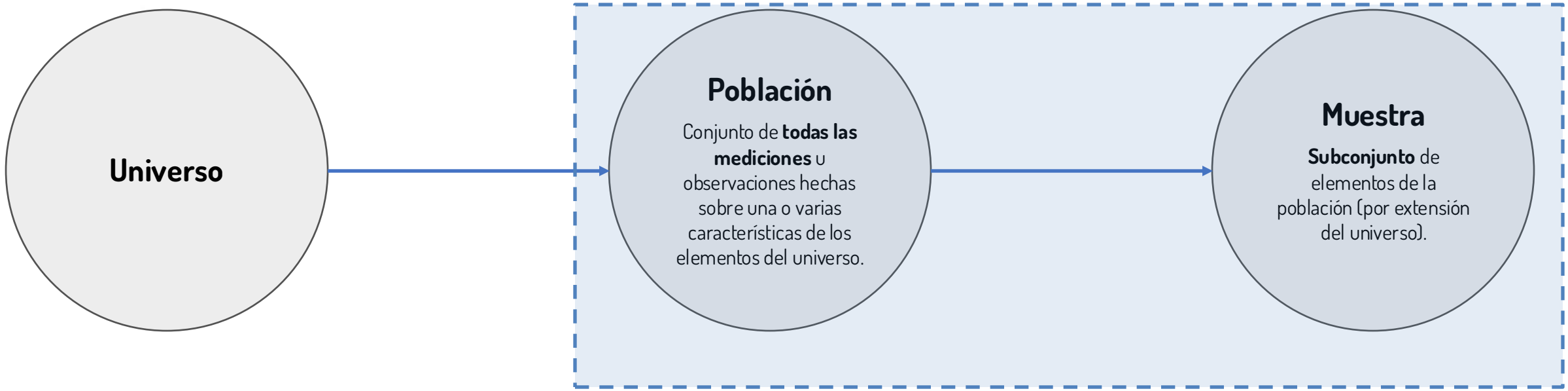
Ejemplo: ***Determinar la altura media de los habitantes de un país.***

Universo: conjunto finito formado por todos los habitantes censados en el país

Población: conjunto finito de todas las alturas de cada habitante de un país

Muestra: selección aleatoria de un número finito de habitantes de las que obtenemos su altura

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Cada habitante será un individuo.



Ejemplo: ***Determinar la altura media de los habitantes de un país.***

Universo: conjunto finito formado por todos los habitantes censados en el país

Población: conjunto finito de todas las alturas de cada habitante de un país

Muestra: selección aleatoria de un número finito de habitantes de las que obtenemos su altura

Individuo: cualquier elemento que porte información sobre el fenómeno que se estudia. Cada habitante será un individuo.

Variable estadística:

- Variables **cualitativas o categóricas**: no se pueden medir numéricamente (por ejemplo: nacionalidad, color de la piel, sexo…).
- Variables **cuantitativas**: tienen valor numérico (edad, precio de un producto, ingresos anuales).

Variables cuantitativas

- **Discretas**: sólo pueden tomar **valores enteros** (1, 2, 8, -4, etc.).
  Por ejemplo: número de hermanos (puede ser 1, 2, 3…., etc., pero, por ejemplo, nunca podrá ser 3.45)
- **Continuas**: pueden tomar **cualquier valor real** dentro de un intervalo.
  Por ejemplo, la velocidad de un vehículo puede ser 90.4 km/h, 94.57 km/h…etc.

Variables cualitativas

- **Nominal**: es una variable cualitativa en la que **no es posible ni está implícito ningún orden en los niveles**.
  Por ejemplo, la variable género es nominal porque no hay orden en los niveles femenino / masculino.
- **Ordinal**: es una variable cualitativa con un orden implícito en los niveles.
  Por ejemplo, si la gravedad de los accidentes de tráfico se ha medido en una escala como accidentes leves, moderados y mortales, esta variable es una variable ordinal cualitativa porque hay un orden claro en los niveles.

Frecuencia – f:

Disponer la información de tal forma que resulte fácil responder a preguntas que se planteen.

x	f
1	6
2	11
3	12
4	30
5	40
6	25
7	14
8	9
9	3
Total:	150

Frecuencia Acumulada – F

La suma de todas las frecuencias comprendidas entre el primer valor de la tabla y el valor que interesa, ambos inclusive.

x	f	F
1	6	6
2	11	17
3	12	29
4	30	59
5	40	99
6	25	124
7	14	138
8	9	147
9	3	150
Total:	150	

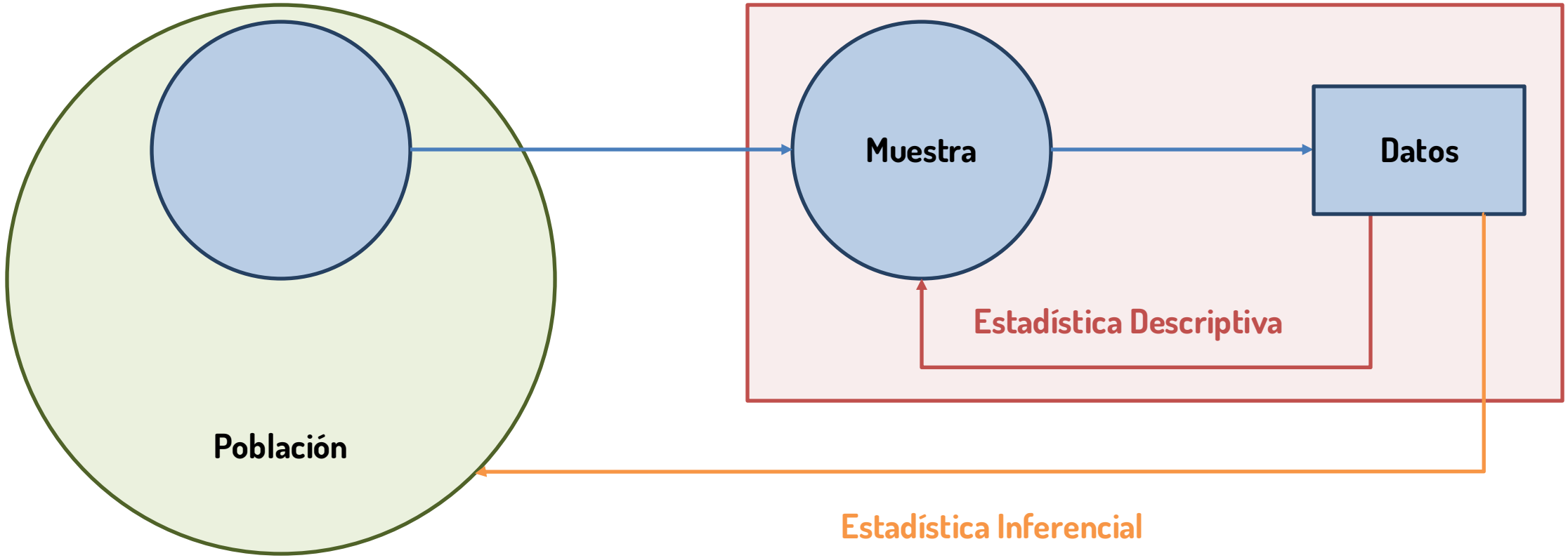
Frecuencia Relativa – f_r:

Expresar las frecuencias en términos relativos en vez de absolutos. Esto es precisamente lo que consiguen las proporciones: expresar una cantidad con respecto al total.

	Nuevos datos		Datos anteriores	
x	f	f_r	f	f_r
1	200	0.1754	6	0.0400
2	170	0.1491	11	0.0733
3	120	0.1053	12	0.0800
4	60	0.0526	30	0.2000
5	40	0.0351	40	0.2667
6	60	0.0526	25	0.1667
7	120	0.1053	14	0.0933
8	170	0.1491	9	0.0600
9	200	0.1754	3	0.0200
Total	1,140	1.0000	150	1.0000

laSalle
UNIVERSITAT RAMON LLULL

Resumen



laSalle
UNIVERSITAT RAMON LLULL

Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. **Medidas de Posición**
3. Medidas de Dispersión
4. Medidas de Asimetría
5. Medidas de Asociación
6. Visualización
7. Consideraciones

laSalle
UNIVERSITAT RAMON LLULL

Media Aritmética o Promedio

$$Media(X) = \overline{x} = \frac{\sum_{i=1}^{N} X_i}{N}$$

siendo $(X_1, X_2, ..., X_N)$ el conjunto de observaciones

1 - Es una medida totalmente numérica, es decir sólo puede calcularse en datos de **características cuantitativas**.

2 - En su cálculo se toman en cuenta **todos los valores** de la variable.

3 - La media aritmética se ve **altamente afectada** por valores extremos (*outliers*).

4 - La media aritmética es única, es decir, un conjunto de datos numéricos tiene **una y solo una** media aritmética.

Mediana o valor central

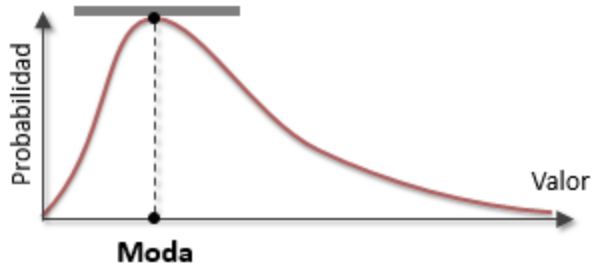
Mediana	$x_{\frac{N+1}{2}}$ si N impar $\frac{1}{2}\cdot\left(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}\right)$ si N par

1 - En su cálculo no se incluyen todos los valores de la variable.

2 - La Mediana **no es afectada** por valores extremos.

Moda o valor más frecuente

Moda	Datos x_i más repetidos.



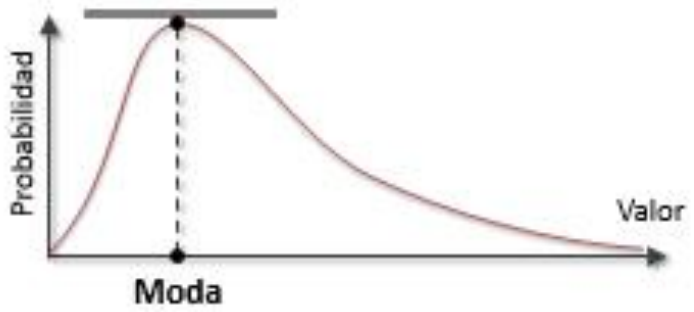
1 - En su cálculo no se incluyen todos los valores de la variable.

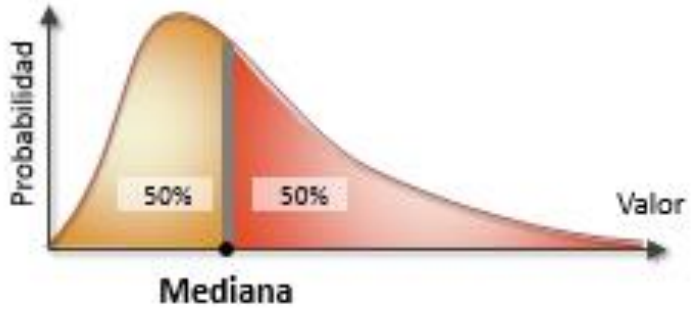
2 – (No está definida algebraicamente).

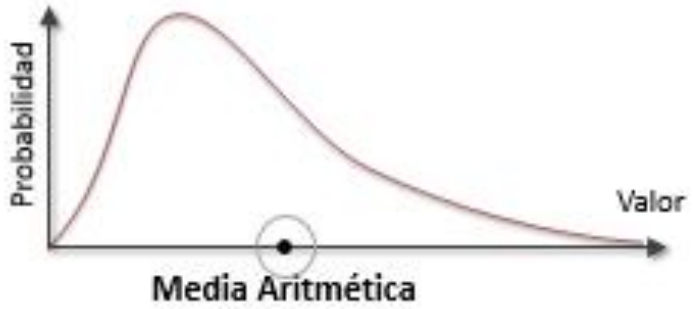
3 - No es afectada por valores extremos.

laSalle
UNIVERSITAT RAMON LLULL

Resumen







laSalle
UNIVERSITAT RAMON LLULL

Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. Medidas de Posición
3. **Medidas de Dispersión**
4. Medidas de Asimetría
5. Medidas de Asociación
6. Visualización
7. Consideraciones

Varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística.

$$S_X^2 = \frac{\Sigma_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$ la media

1 - **Var(X) ≥ 0** *La varianza es un valor siempre positivo*

2 - **Var(X + b) = Var (X)** *Si a todos los datos se les suma una constante, la varianza de esos datos sigue siendo la misma*

3 - **Var(a · X) = a^2 · Var (X)** *Si todos los datos se multiplican por una constante, la varianza queda multiplicada por el cuadrado de la constante.*

4 - **Var(a · X + b) = a^2 · Var(X)** *En base a las dos anteriores, para todo par de números reales a i b.*

5 - **Var(X + Y) = Var(X) + Var(Y)** únicamente en el caso que X e Y sean independientes.

Varianza es la media aritmética del cuadrado de las desviaciones respecto a la media de una distribución estadística.

$$S_X^2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^2}{N-1}$$

siendo $(X_1, X_2, ..., X_N)$ un conjunto de datos y $\overline{x}$

Se ve altamente afectada por los valores extremos (outliers)

'a es un valor siempre positivo

Si a todos los datos se les suma una esos datos sigue siendo la misma

X) *Si todos los datos se multiplican por 'a queda multiplicada por el cuadrado de la constante.*

4 - **Var(a · X + b) = a² · Var(X)** *En base a las dos anteriores, para todo par de números reales a i b.*

5 - **Var(X + Y) = Var(X) + Var(Y)** únicamente en el caso que X e Y sean independientes.

Desviación estándar (o típica) muestra la variación sobre la media aritmética

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

La desviación estándar o típica es la raíz cuadrada de la varianza.

Las interpretaciones que se deducen de la desviación típica son, por lo tanto, parecidas a las que se deducían de la varianza

Desviación estándar (o típica) muestra la variación sobre la media aritmética

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

La desviación estándar o típica es la raíz cuadrada de la varianza.

Las interpretaciones que se deducen de la desviación típica son, por lo tanto, parecidas a las que se deducían de la varianza

¿Qué diferencia existe entre la varianza y la desviación típica?

Desviación estándar (o típica) muestra la variación sobre la media aritmética

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

La desviación estándar o típica es la raíz cuadrada de la varianza.

Las interpretaciones que se deducen de la desviación típica son, por lo tanto, parecidas a las que se deducían de la varianza

¿Qué diferencia existe entre la varianza y la desviación típica?

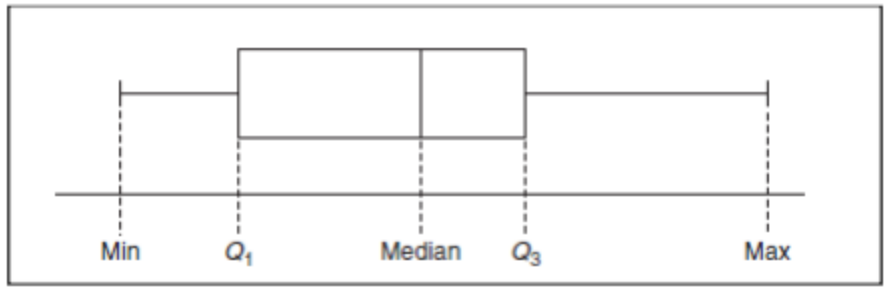
*Su cálculo es necesario para obtener el valor de **otros parámetros estadísticos**.*

Para calcular la covarianza necesitamos la varianza y no la desviación típica, para calcular algunas matrices econométricas se utiliza la varianza y no la desviación típica.

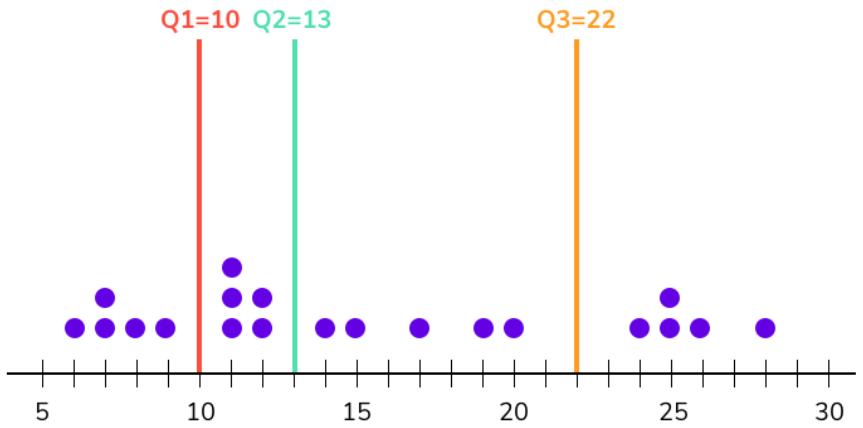
Es una cuestión de comodidad a la hora de trabajar con los datos en según qué cálculos.

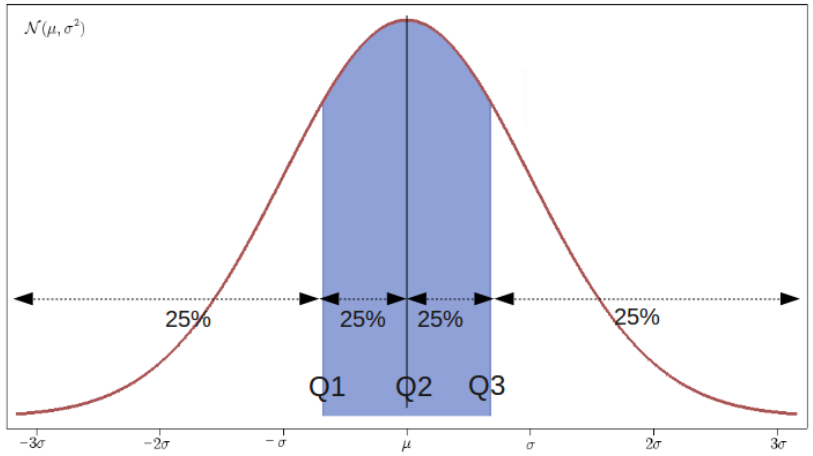
laSalle
UNIVERSITAT RAMON LLULL

Cuartiles



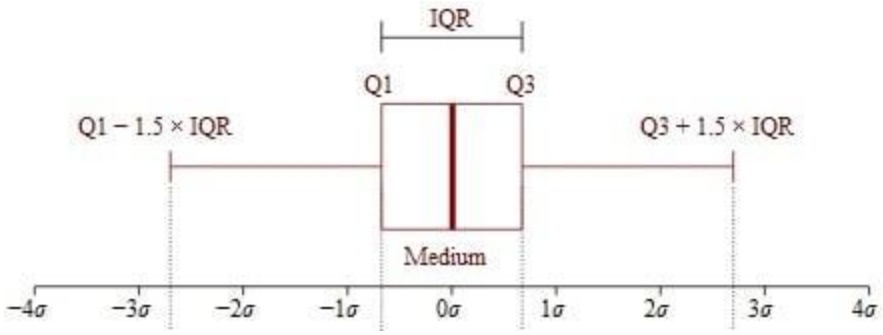
Divide los datos clasificados en 4 grupos iguales.

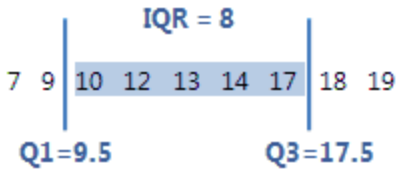




Rango intercuartil (IQR)

$$IQR = Q_3 - Q_1$$





1 - Es la distancia entre el primer cuartil (Q1) y el tercer cuartil (Q3). El 50% de los datos está dentro de este rango.

2 – Sirve para describir la dispersión de los datos. A medida que aumenta la dispersión de los datos, el IQR se hace más grande.

3 - Mediante esta medida se eliminan los valores extremos (outliers).

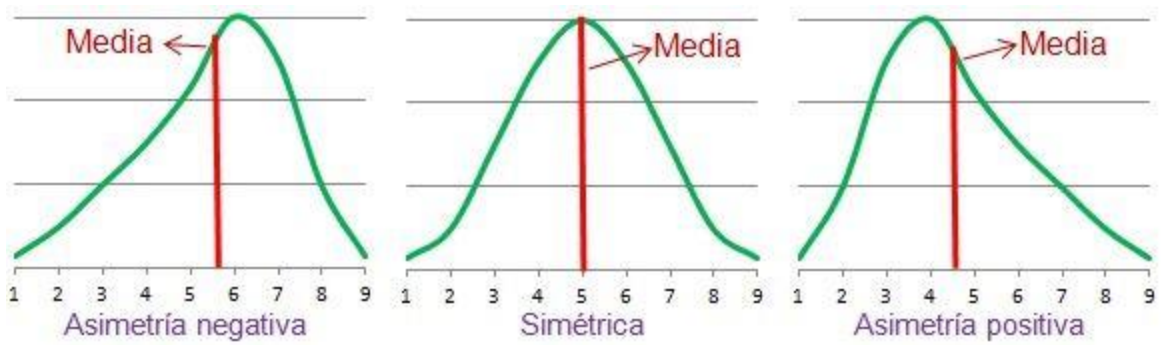
laSalle
UNIVERSITAT RAMON LLULL

Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. Medidas de Posición
3. Medidas de Dispersión
4. **Medidas de Asimetría**
5. Medidas de Asociación
6. Visualización
7. Consideraciones

La asimetría es la medida que indica la simetría de la distribución de una variable respecto a la media aritmética, sin necesidad de hacer la representación gráfica.



Existen tres tipos de curva de distribución según su asimetría:

Asimetría negativa: la cola de la distribución se alarga para valores inferiores a la media.

Simétrica: hay el mismo número de elementos a izquierda y derecha de la media. En este caso, coinciden la media, la mediana y la moda. La distribución se adapta a la forma de la campana de Gauss, o distribución normal.

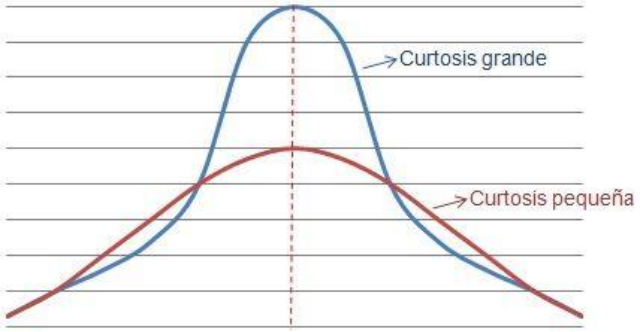
Asimetría positiva: la cola de la distribución se alarga para valores superiores a la media.

CURTOSIS

$$g_2 = \frac{\sum_{i=1}^{N}(X_i - \overline{x})^4 \cdot n_i}{N S_x^4}$$

n_i la frecuencia absoluta de x_i

Este coeficiente indica la cantidad de datos que hay cercanos a la media, de manera que a mayor grado de curtosis, más escarpada (o apuntada) será la forma de la curva.

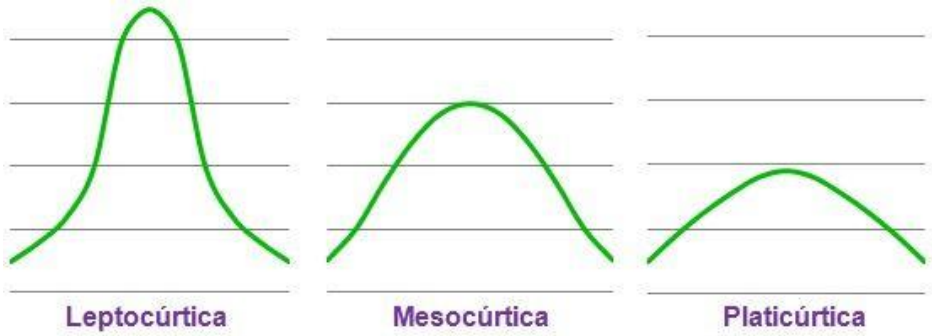


Las curvas se pueden clasificar en tres grupos según el signo de su curtosis, es decir, según la forma de la distribución:

Leptocúrtica: la Curtosis>0. Los datos están muy concentrados en la media, siendo una curva muy apuntada.

Mesocúrtica: la Curtosis=0. Distribución normal.

Platicúrtica: la Curtosis<0. Muy poca concentración de datos en la media, presentando una forma muy achatada.



Coeficiente de asimetría de Fisher

$$g_1 = \frac{m_3}{S^3}$$

$$CA_F = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3}{N \cdot S_x^3}$$

siendo $\overline{x}$ la media y S_x la desviación típica

Según sea el valor de g1, diremos que la distribución es asimétrica a derechas o positiva, a izquierdas o negativa, o simétrica, o sea:

- Si g1 > 0 × la distribución será asimétrica positiva o a derechas (desplazada hacia la derecha).
- Si g1 < 0 × la distribución será asimétrica negativa o a izquierdas (desplazada hacia la izquierda).
- Si g1 = 0 × la distribución puede ser simétrica; si la distribución es simétrica, entonces si podremos afirmar que g1 = 0.

laSalle
UNIVERSITAT RAMON LLULL

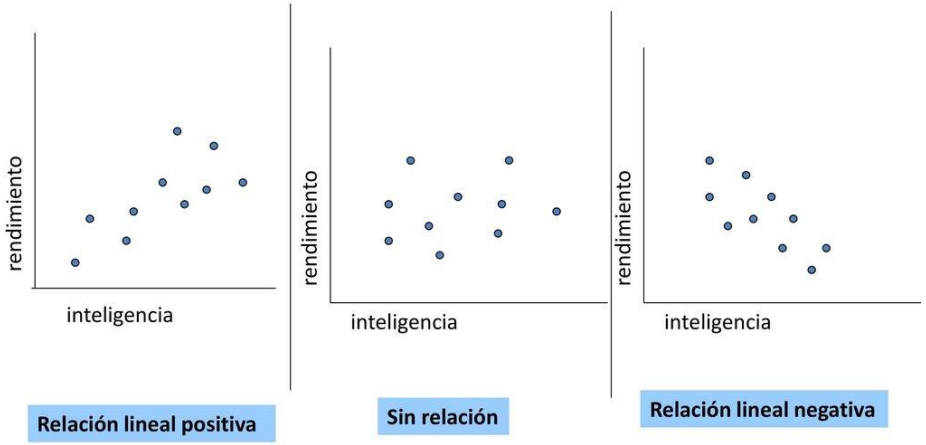
Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. Medidas de Posición
3. Medidas de Dispersión
4. Medidas de Asimetría
5. **Medidas de Asociación**
6. Visualización
7. Consideraciones

Correlación

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

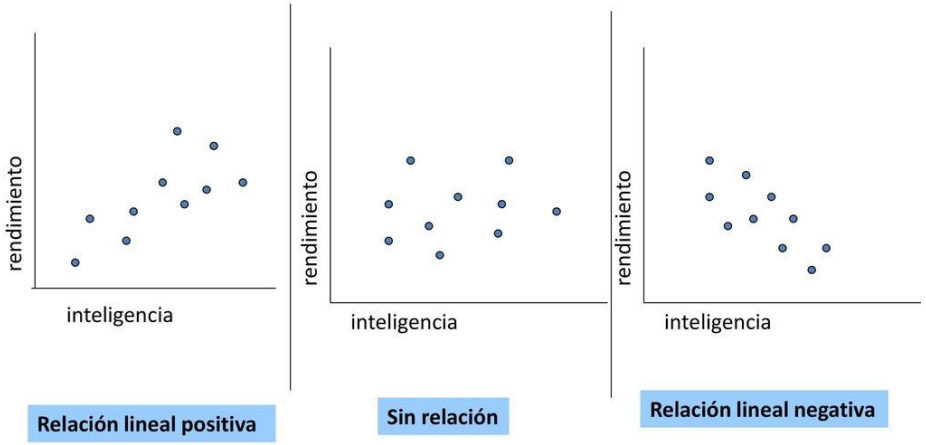


Mide la fuerza de una relación lineal entre dos variables.

- La correlación es entre 1 y 1
- Signo = dirección de la relación
- Valor absoluto: fuerza de la relación (0,6 es una relación más fuerte que +0,4)
- La correlación de una variable consigo misma es 1
- La correlación no implica causalidad

Correlación

$$\rho_{xy} = \frac{Cov_{xy}}{\sigma_x \sigma_y}$$

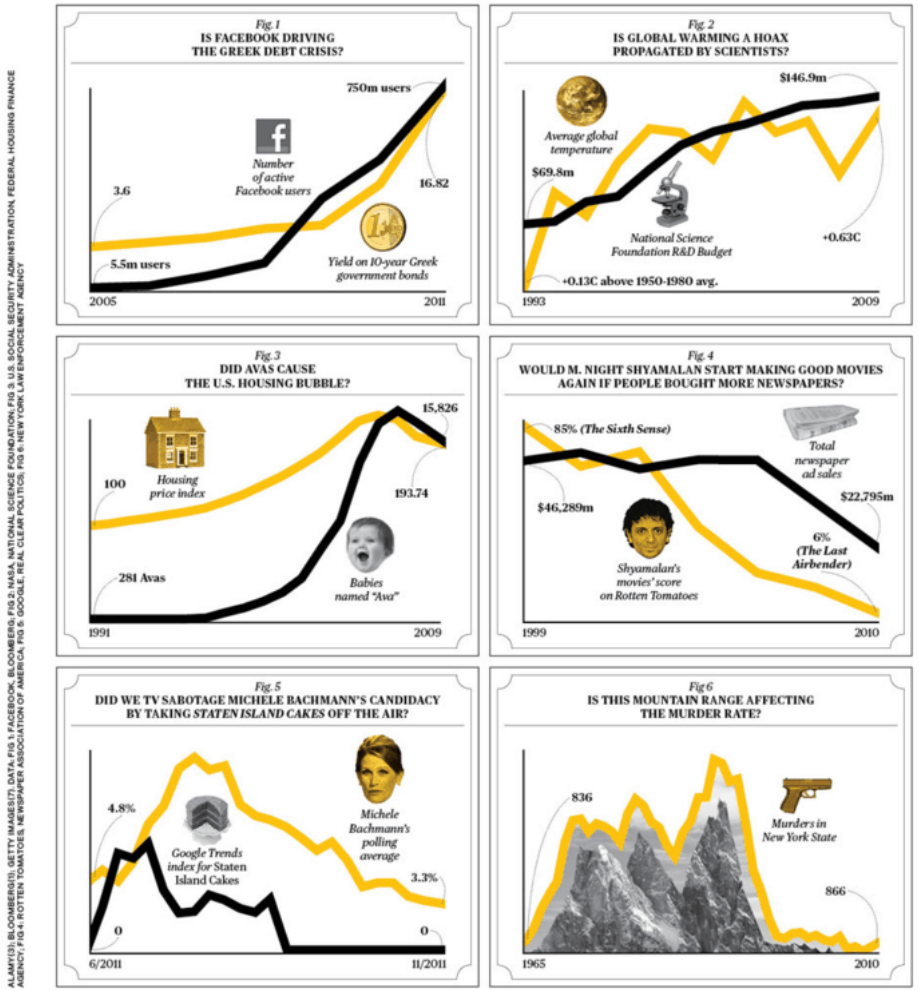


Mide la fuerza de una relación lineal entre dos variables.

- La correlación es entre 1 y 1
- Signo = dirección de la relación
- Valor absoluto: fuerza de la relación (0,6 es una relación más fuerte que +0,4)
- La correlación de una variable consigo misma es 1
- **La correlación no implica causalidad**

laSalle
UNIVERSITAT RAMON LLULL

La correlación no implica causalidad



https://www.jotdown.es/2016/06/correlacion-no-implica-causalidad/

Covarianza

$$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x - \bar{x})(y - \bar{y})$$

La covarianza indica el sentido de la correlación entre las variables

- Si $S_{xy} > 0$ la correlación es directa (cuando uno variable crece la otra variable también).
- Si $S_{xy} < 0$ la correlación es inversa (cuando una variable crece la otra variable decrece).

laSalle
UNIVERSITAT RAMON LLULL

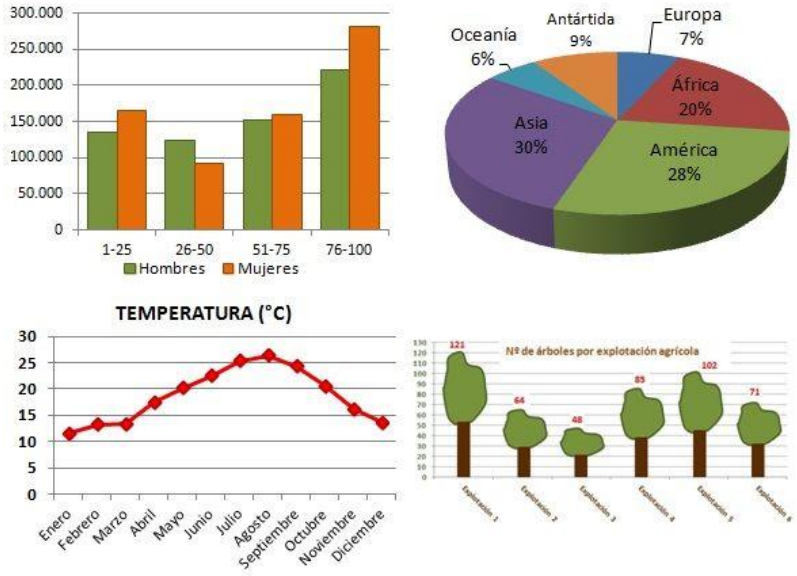
Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. Medidas de Posición
3. Medidas de Dispersión
4. Medidas de Asimetría
5. Medidas de Asociación
6. **Visualización**
7. Consideraciones

laSalle
UNIVERSITAT RAMON LLULL

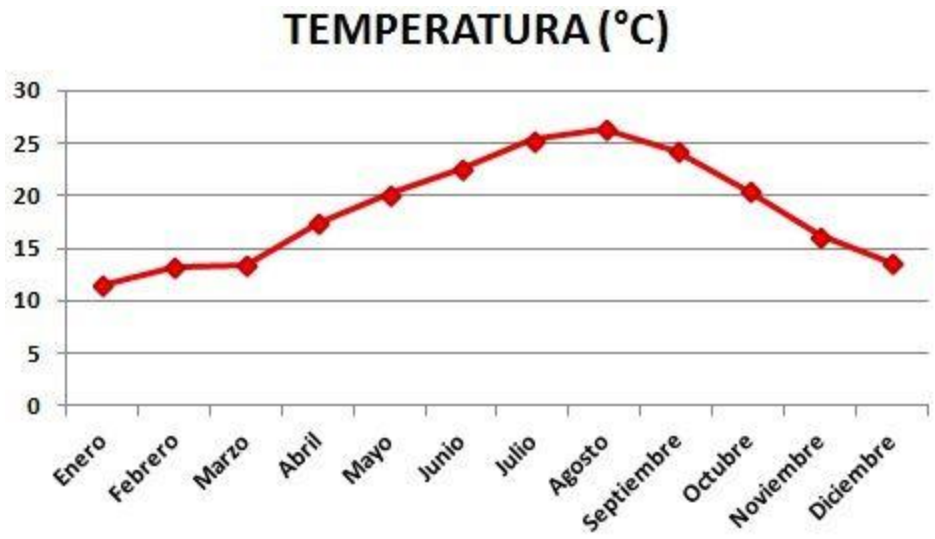
GRÁFICO: es el recurso de representar los datos numéricos por medio de líneas, diagramas, dibujos, etc. La representación gráfica es un importante suplemento al análisis y estudio estadístico



Un buen gráfico puede captar al lector para que a continuación lea todo el estudio.

Si un estudio se compone únicamente de texto y tablas, posiblemente no todos los lectores lean el estudio.

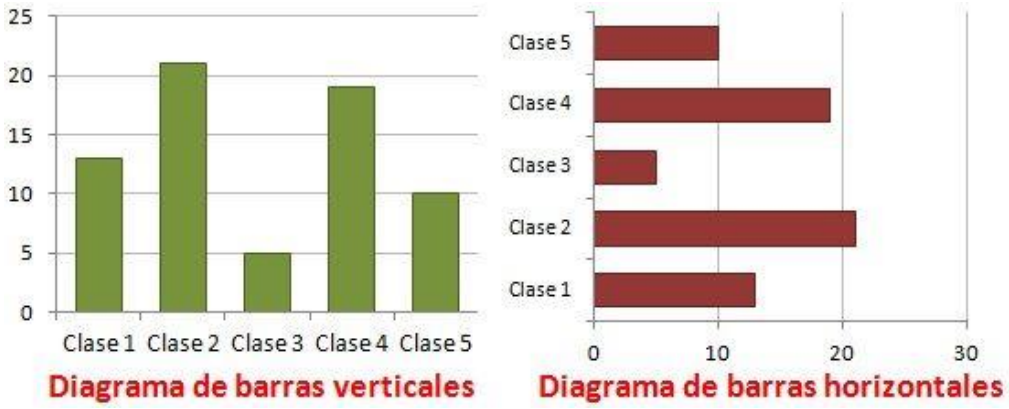
Gráfico lineal



Se suele utilizar con variables cuantitativas, para ver su **comportamiento en el transcurso del tiempo**.

Por ejemplo, en las series temporales mensuales, anuales, trimestrales, etc.

Diagrama de barras

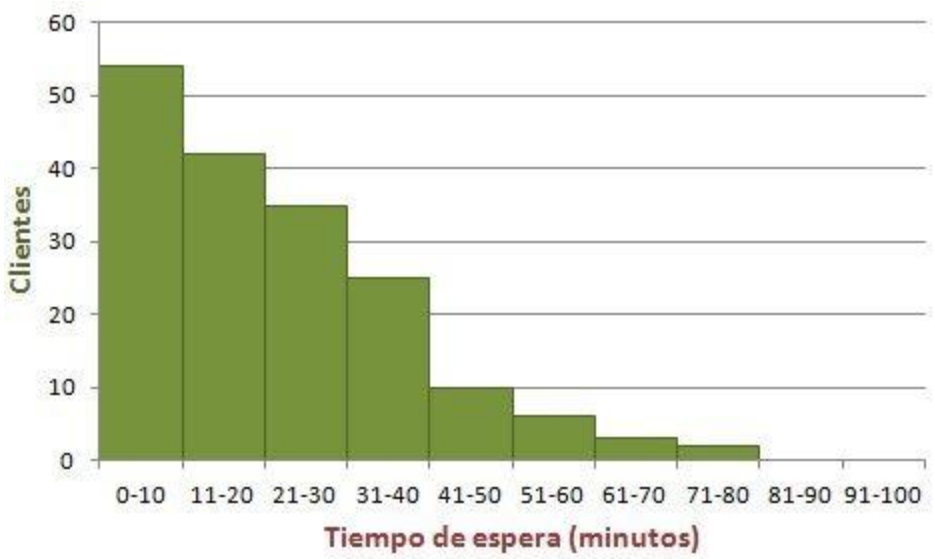


Se utiliza para representar datos de **variables cualitativas o discretas**.

Está formado por barras rectangulares cuya **altura es proporcional** a la **frecuencia** de cada uno de los **valores de la variable**.

- En el **eje de abscisas** se colocan las **cualidades de la variable**, si la variable es cualitativa, o los valores de dicha variable, si es discreta.
- En el **eje de ordenadas** se colocan las barras proporcionales a la **frecuencia relativa o absoluta** del dato.
- Las barras pueden ser horizontales o verticales, según donde se reflejen los valores de la variable.
- Todas las barras deben tener el mismo ancho y no deben superponerse las unas con las otras.

Histograma



Es una representación gráfica de **datos agrupados mediante intervalos**.

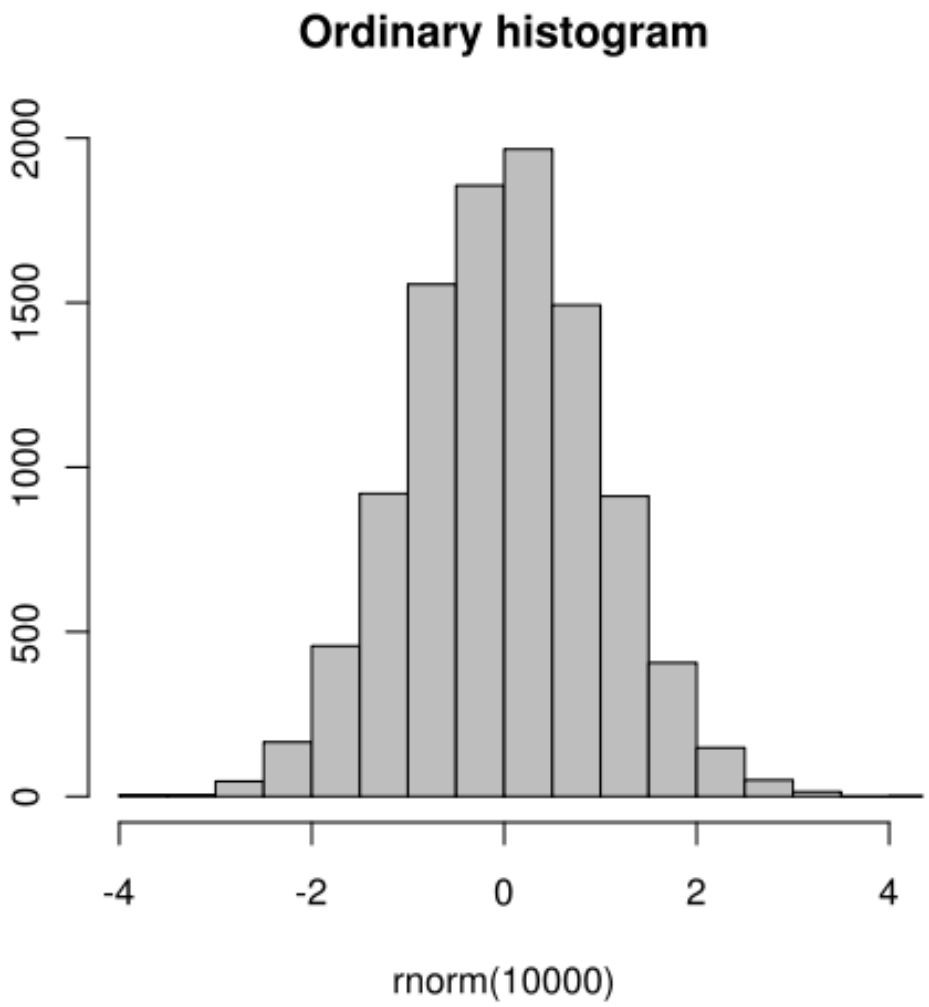
Los datos provienen de una **variables cuantitativas continuas**.

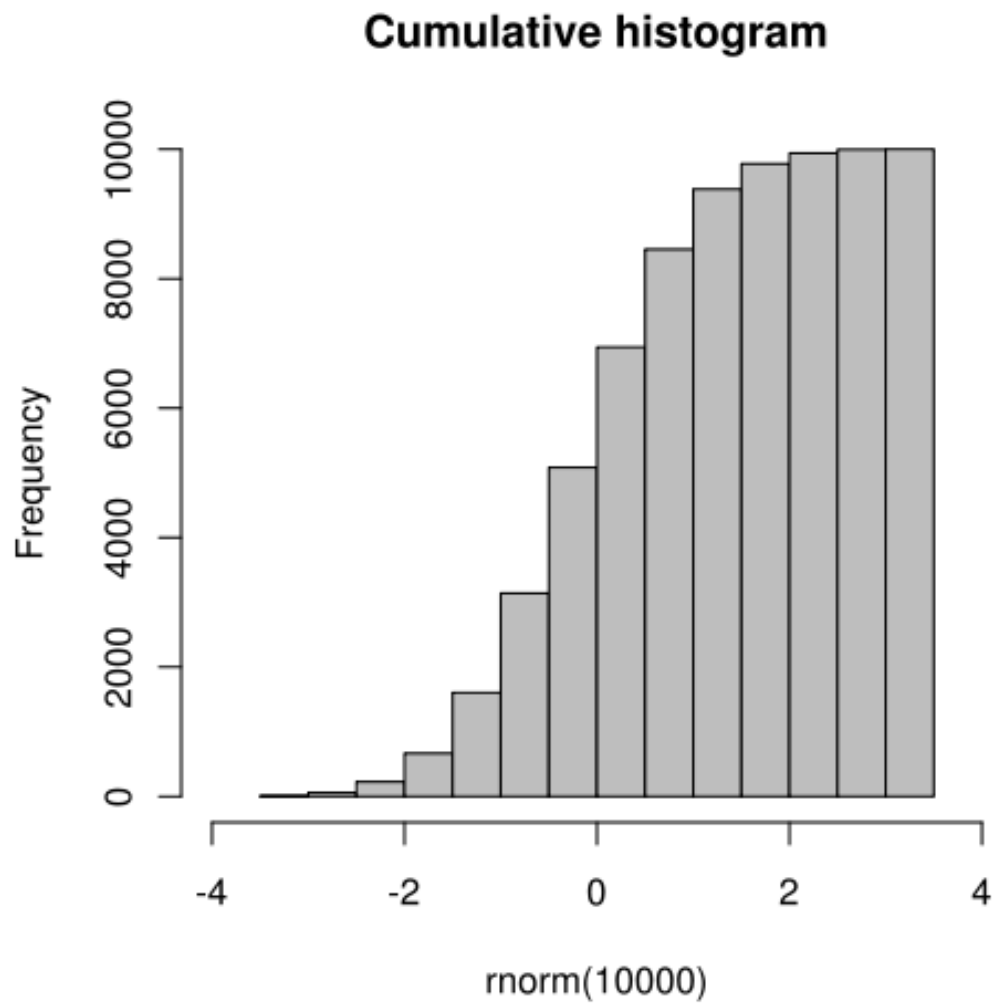
Gracias a él puedes hacerte rápidamente una idea de la distribución de los datos o muestra.

Es un conjunto de rectángulos (bin) que **representan las frecuencias absolutas de cada uno de los intervalos**.

Los intervalos abarcan todo el conjunto sin cortarse, de manera que **un elemento está solo en un intervalo**.

laSalle
UNIVERSITAT RAMON LLULL





laSalle
UNIVERSITAT RAMON LLULL

Assume a discrete random variable or a categorical variable. The probability for each value n_k is given by $f(n_k)$

$f(n_k)=p_k=P[N=n_k]$

We keep repeating the experiment independently while we count how many times we obtain each of the possible values. We will have,

N Total number of experiments i.e. number of times we repeat our exper- iment.
N_1 number of occurences of the first value of N, n_1,
N_2 number of occurences of the second value of N, n_2, ...,
N_n number of occurences of the last value of N, n_n,

p_k is the probability to obtain the kth value of N, $p_k = f(n_k)$. We have
$N_1 + N_2 + \ldots + N_n = N$
The set of numbers $(N_1, N_2, \ldots, N_n)$ follows a multinomial distribution

$$f(N_1, N_2, \ldots, N_n) = \frac{\mathcal{N}!}{N_1!N_2!\ldots N_n!} p_1^{N_1} \ldots p_n^{N_n}$$

If the total number of data elements $\mathcal{N}$ is not fixed but data appear in an independent way, $\mathcal{N}$ should follow a Poisson distribution. We call its average ν. If this is so, each category or value of the histogram shall also follow a Poisson distribution. This Poisson distribution shall have average ν_k,

$$f_k(N_k) = \frac{1}{N_k!}\nu_k^{N_k} e^{-N_k}$$

The expected probability for the corresponding category is

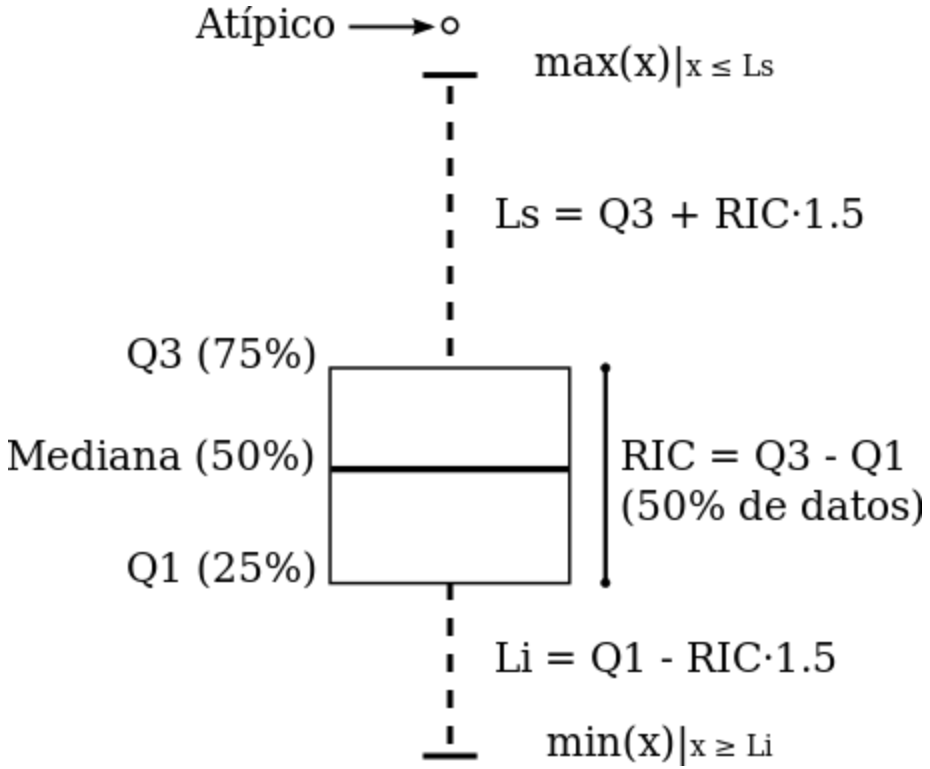
$$p_k = \frac{\nu_k}{\nu}$$

while the probability we evaluate is

$$\frac{N_k}{\mathcal{N}}$$

For large values of ν (say $\nu >> 5$, the Poisson variable also can be approximated by a Gaussian.

BoxPlot

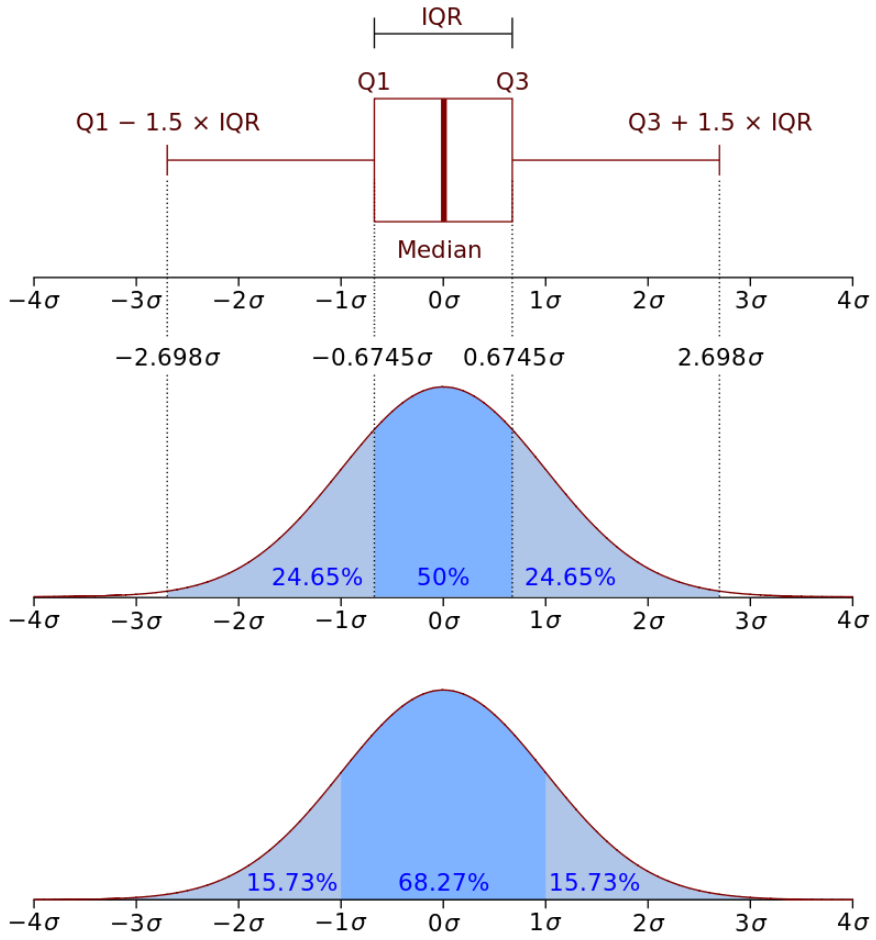


El box-plot es una forma de representación destinada, fundamentalmente, a **resaltar aspectos de la distribución de las observaciones** en una o más series de datos cuantitativos.

Reemplaza, en consecuencia, **al histograma** y a la curva de **distribución de frecuencias** sobre los que tiene ventajas en cuanto a la información que brinda y a la **apreciación global que surge de la lectura**.

Este gráfico **utiliza una sola escala**: la correspondiente a la variable de los datos que se presentan. Es decir, no utiliza escala de frecuencias. Por lo tanto, no corresponde asociarlo a los que utilizan el sistema de coordenadas cartesianas.

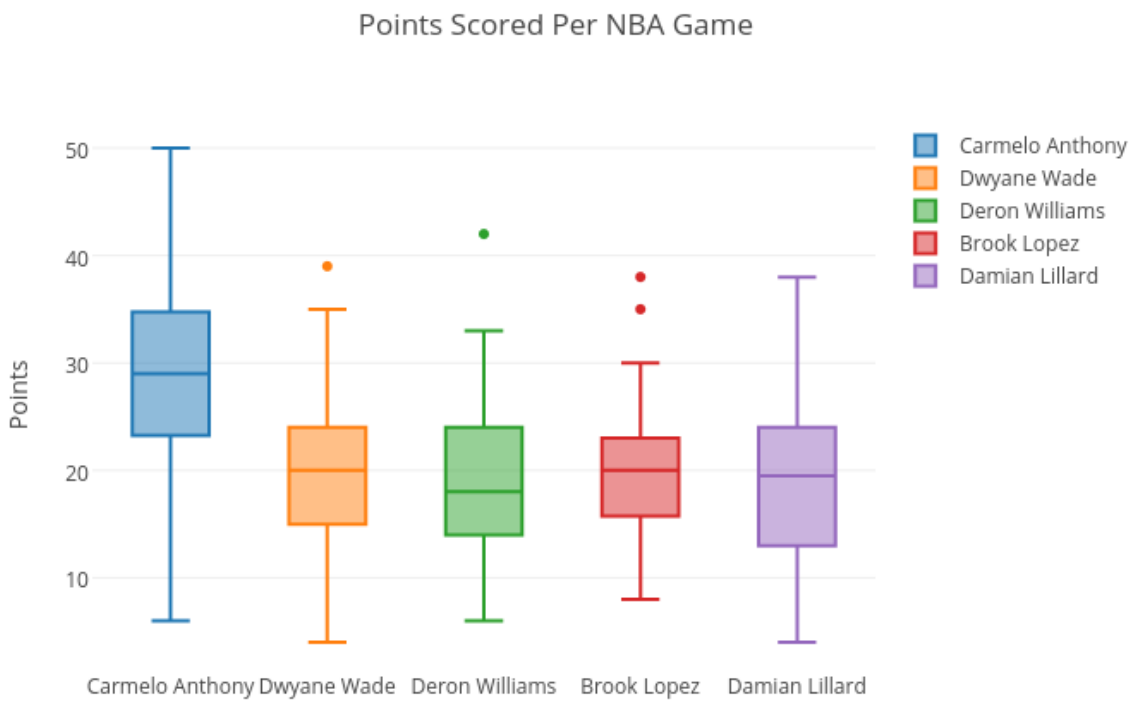
BoxPlot - Interpretación

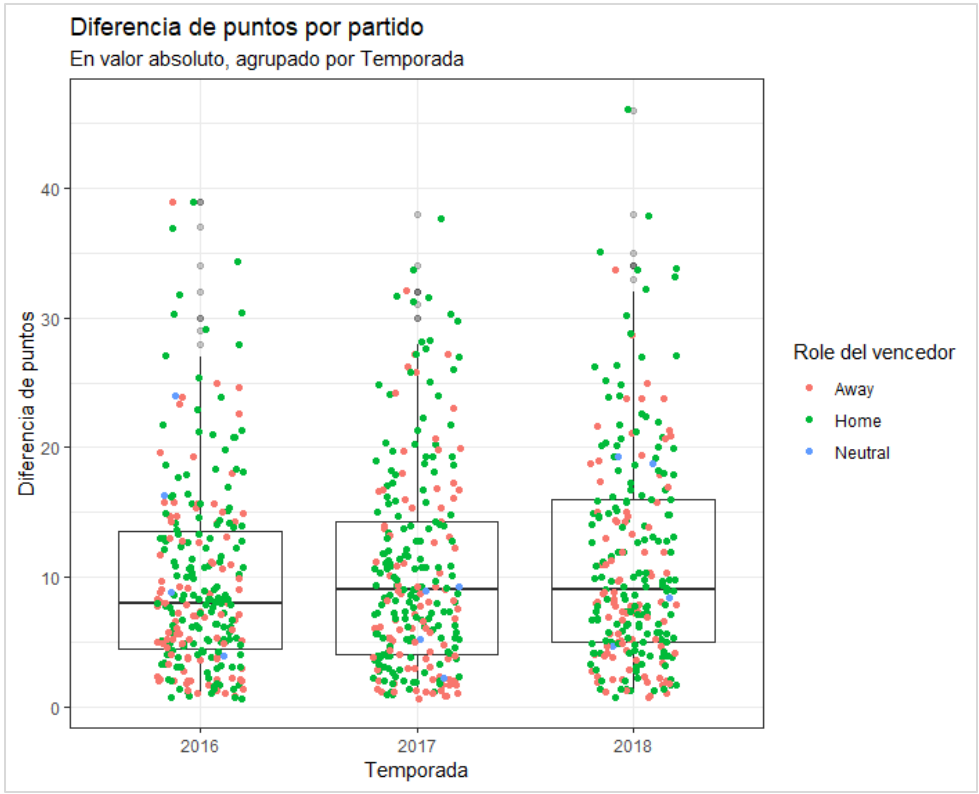


Este gráfico brinda información sobre la forma general de la curva: **simetría**, **curtosis** (curvas más "afinadas" o más "aplanadas"), el punto de la **mediana**, la **distribución de las observaciones** a ambos lados de los valores centrales y la presencia (y el/los valor/es) de **valores atípicos**.

laSalle
UNIVERSITAT RAMON LLULL

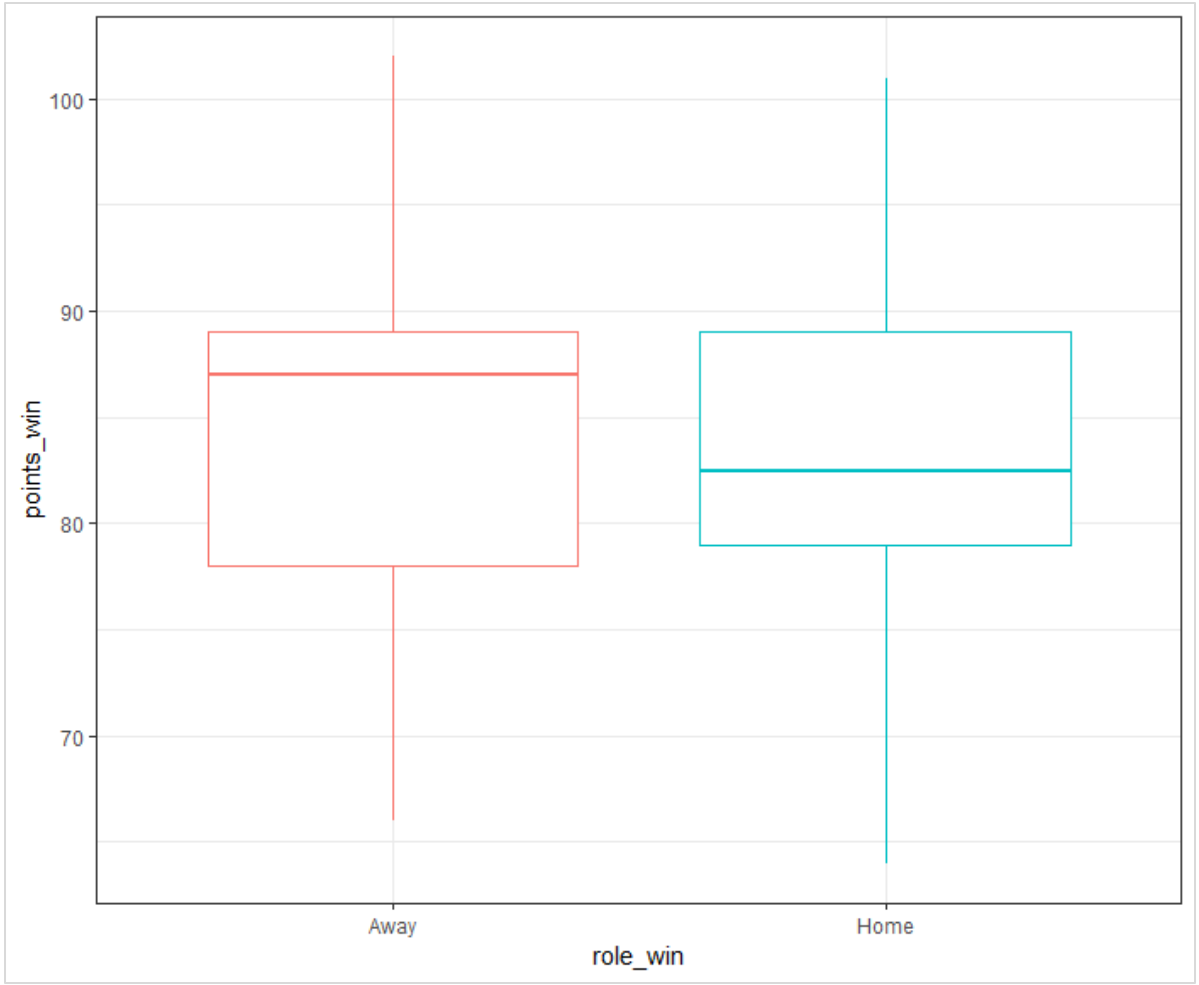
BoxPlot - Ejemplos



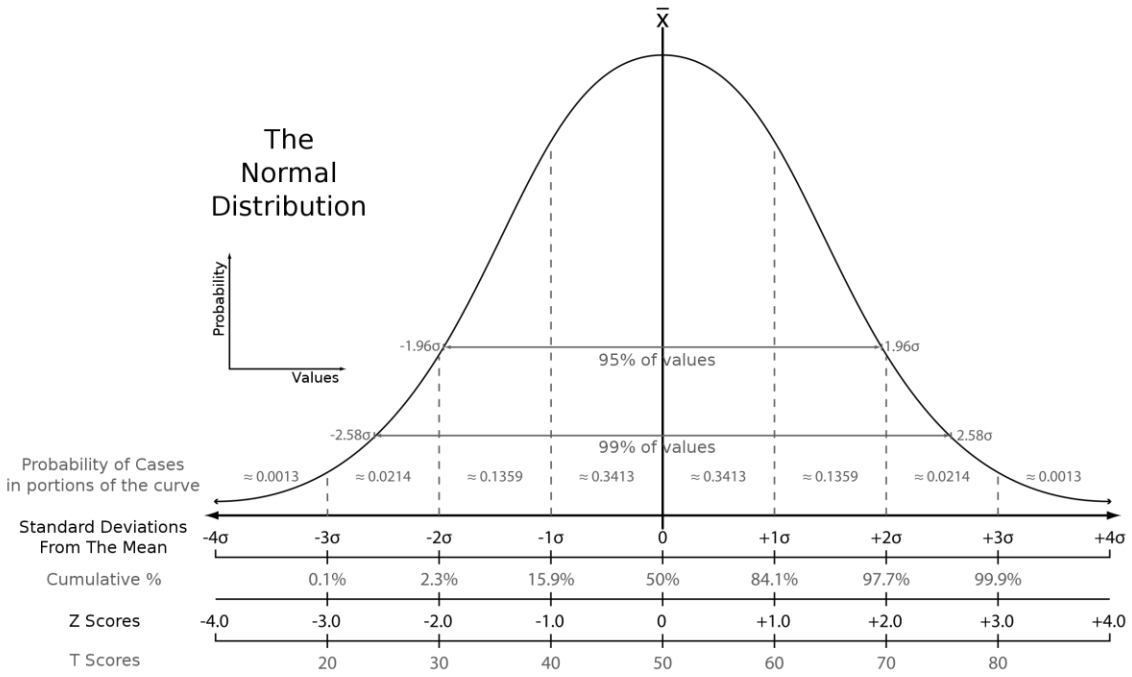


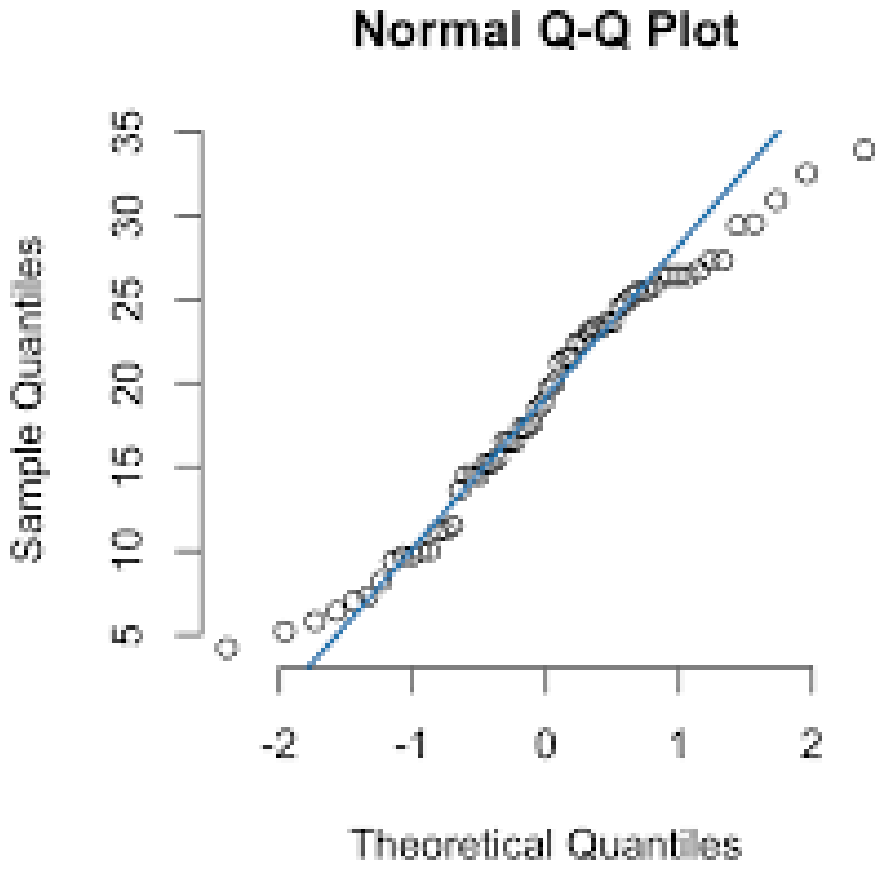
laSalle
UNIVERSITAT RAMON LLULL

BoxPlot - Ejemplos



laSalle
UNIVERSITAT RAMON LLULL





laSalle
UNIVERSITAT RAMON LLULL

Descripción de sesión

ESTADÍSTICA DESCRIPTIVA

1. Conceptos Clave
2. Medidas de Posición
3. Medidas de Dispersión
4. Medidas de Asimetría
5. Medidas de Asociación
6. Visualización
7. **Consideraciones**

laSalle
UNIVERSITAT RAMON LLULL

Muestra vs Población

Promedio Población

$$\mu = \frac{\sum X_i f_i}{n}$$

Varianza Población

$$\sigma^2 = \frac{\sum (X_i - \mu)^2 f_i}{N}$$

Desviación Estándar Población

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

Promedio Muestra

$$\bar{X} = \frac{\sum X_i f_i}{n}$$

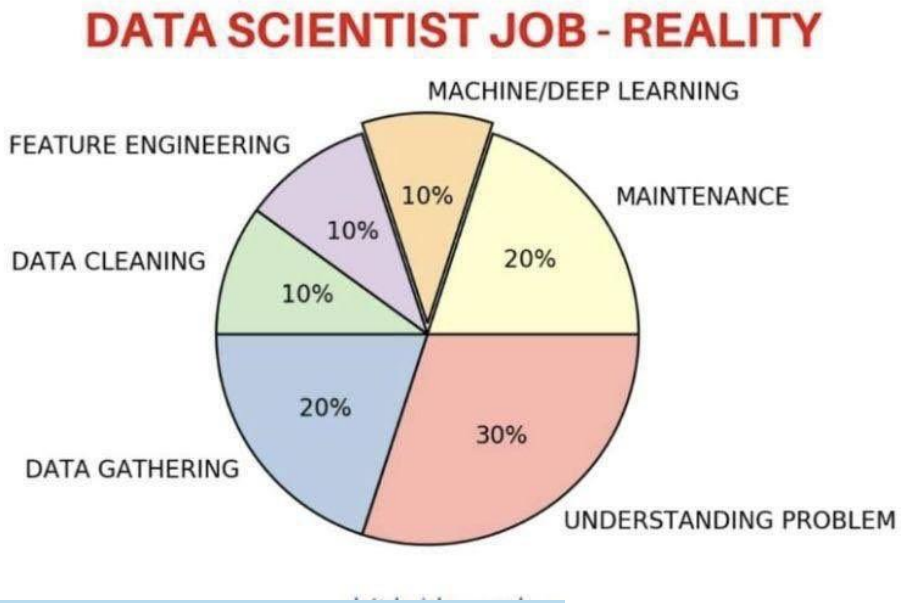
Varianza Muestra

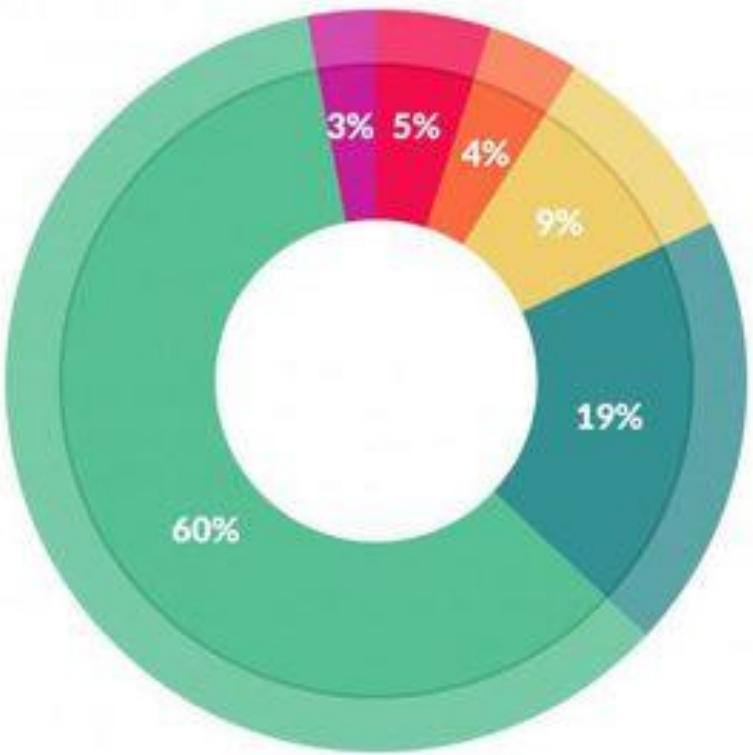
$$S^2 = \frac{\sum (X_i - \bar{X})^2 f_i}{n-1}$$

Desviación Estándar Muestra

$$S = \sqrt{S^2}$$


DATA SCIENTIST JOB - EXPECTATION
@drangshu
DEEP LEARNING
20%
80%
MACHINE LEARNING
Follow: Dr. Angshuman Ghosh


DATA SCIENTIST JOB - REALITY
MACHINE/DEEP LEARNING
10%
FEATURE ENGINEERING
10%
MAINTENANCE
20%
DATA CLEANING
10%
20%
30%
DATA GATHERING
UNDERSTANDING PROBLEM



What data scientists spend the most time doing

- *Building training sets: 3%*
- *Cleaning and organizing data: 60%*
- *Collecting data sets; 19%*
- *Mining data for patterns: 9%*
- *Refining algorithms: 4%*
- *Other: 5%*