



Madrid Internet
of Things Institute
Make-build-learn

12 de marzo de 2020

Estadística para Data Science

Sesión 5: Regresión y Correlación

Jesús Hernando Corrochano



Estadística para Data Science

● Programa

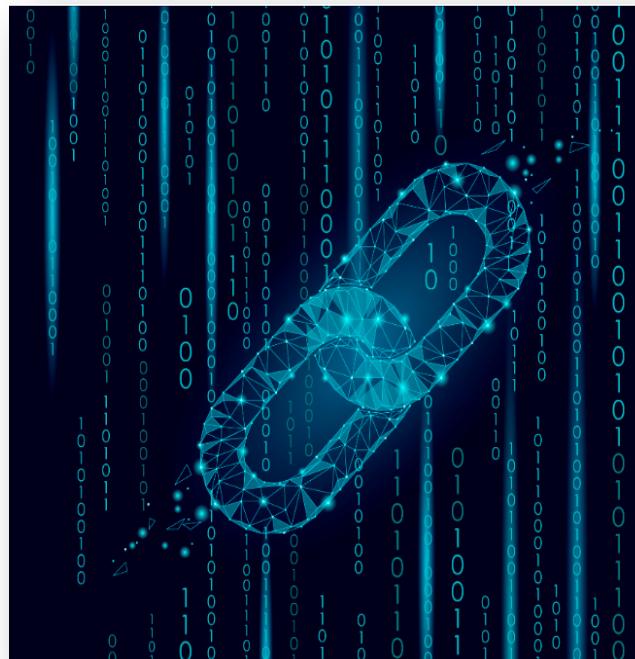


	Sesión 1 13/2	Sesión 2 20/2	Sesión 3 27/2	Sesión 4 5/3	Sesión 5 12/3	Sesión 6 19/3	Sesión 7 26/3	Sesión 8 2/3
Introducción a la estadística								
Introducción a la combinatoria y la probabilidad								
Estadística descriptiva								
Regresión y correlación								
Estadística inferencial								
Probabilidad Total. Teorema de Bayes. Test A/B								

1. Regresión Lineal
2. Correlación
3. Recta de Regresión
4. Intervalos de Confianza
5. A saber...

Diagramas de Caja-Bigotes (boxplots o box and whiskers)

Diagrama "tallo y hojas" (Stem-and-Leaf Diagram)



Regresión



El objetivo de inferencia estadística es estudiar datos para inferir conocimientos que van más allá del alcance inmediato.

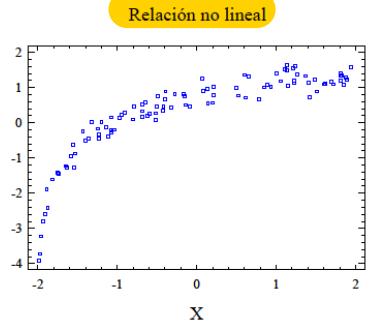
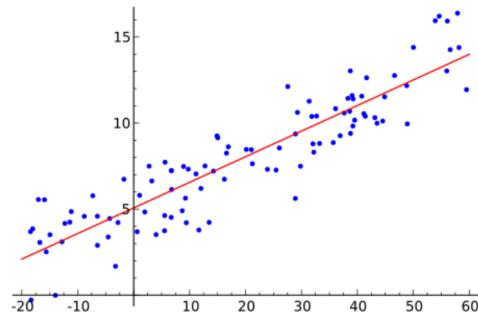
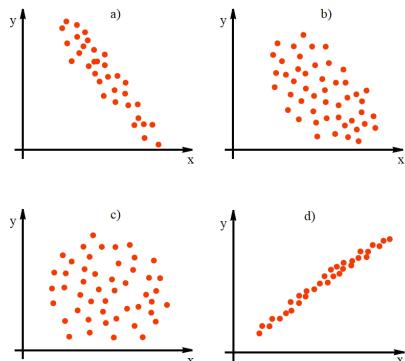
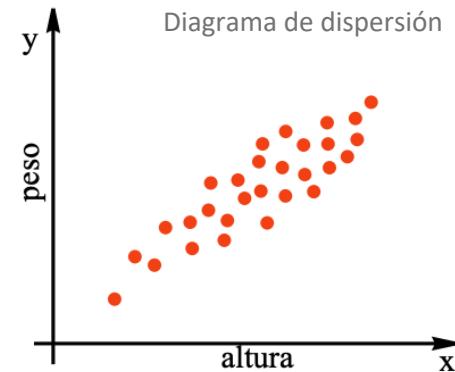
La inferencia estadística es el proceso de deducción de propiedades sobre una muestra donde se supone que la población es mayor que el conjunto de datos observados con los que está trabajando actualmente.

En otras palabras, la inferencia estadística ayuda a estimar los parámetros de una población mayor cuando los datos observados con los que está trabajando son un subconjunto de esa población

● Regresión Lineal

Supongamos que tenemos la altura, peso y edad, al respecto de los alumnos de Mioti ¿existe alguna relación entre estos valores?

El conjunto de datos dados por dos variables X e Y , se denominan distribuciones bidimensionales o bivariadas.



Covarianza

La covarianza es el valor que refleja en qué cuantía dos variables aleatorias varían de forma conjunta respecto a sus medias

$$Cov(X, Y) = S_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$S_{xy} = \frac{\sum_{i=1}^n x_i \cdot y_i}{n} - \bar{x} \cdot \bar{y}$$

La covarianza de una variable bidimensional es la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas

● Coeficiente de Correlación de Pearson

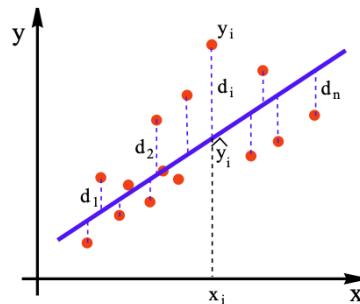
El inconveniente de la covarianza como medida de la asociación lineal entre dos variables es que depende de las unidades de X e Y , por ello se define el **coeficiente de correlación** entre dos variables r_{xy} , por:

$$r_{xy} = \frac{S_{xy}}{S_x \cdot S_y}$$

- $-1 \leq r_{xy} \leq 1$
- Los valores extremos 1 y -1 se alcanzan solamente si todos los datos se sitúan exactamente sobre una recta.
- Si la relación lineal es muy pequeña, el valor de r_{xy} es próximo a cero

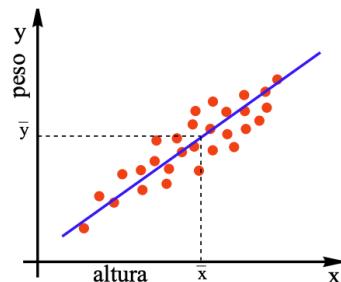
● Recta de Regresión

La recta de regresión corresponde a la recta que mejor se aproxima a los puntos del diagrama de dispersión para la variable X y la Y .



Recta de regresión de y sobre x:

$$r_{y/x} \equiv y - \bar{y} = \frac{S_{xy}}{S_x^2}(x - \bar{x})$$



Recta de regresión de x sobre y:

$$r_{x/y} \equiv x - \bar{x} = \frac{S_{xy}}{S_y^2}(y - \bar{y})$$

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

b) El producto de las pendientes de las rectas es el cuadrado del coeficiente de correlación.

$$\frac{S_{xy}}{S_x^2} \cdot \frac{S_{xy}}{S_y^2} = \left(\frac{S_{xy}}{S_x \cdot S_y} \right)^2 = r_{xy}^2$$

c) Las rectas de regresión se usan para predecir el valor de una variable cuando se conoce la otra, y se debe cumplir que el coeficiente de correlación sea próximo a -1 o a 1.

Recta de Regresión (*Modelo de Regresión Lineal*)

El **modelo de regresión lineal simple** supone que,

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

donde:

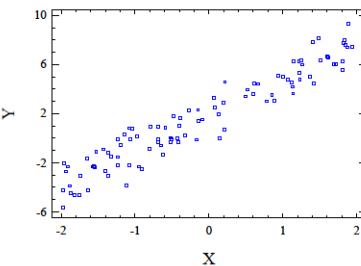
- ▶ y_i representa el valor de la variable respuesta para la observación i-ésima.
- ▶ x_i representa el valor de la variable explicativa para la observación i-ésima.
- ▶ u_i representa el error para la observación i-ésima que se asume normal,

$$u_i \sim N(0, \sigma)$$

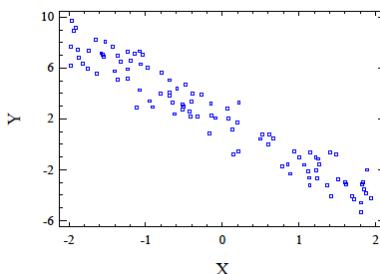
- ▶ β_0 y β_1 son los **coeficientes de regresión**:
 - ▶ β_0 : **intercepto**
 - ▶ β_1 : **pendiente**

Los parámetros que hay que estimar son: β_0 , β_1 y σ .

Relación lineal positiva



Relación lineal negativa



● Recta de Regresión (*Modelo de Regresión Lineal*)

El modelo de regresión lineal simple

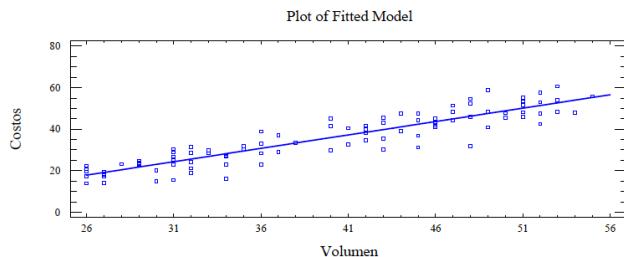
El objetivo es obtener estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ de β_0 y β_1 para calcular la recta de regresión:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

que se ajuste lo mejor posible a los datos.

Ejemplo: Supongamos que la recta de regresión

$$\text{Costo} = -15,65 + 1,29 \text{ Volumen}$$



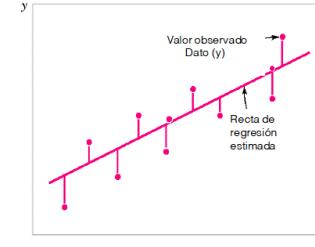
Se estima que una empresa que produce 25 mil unidades tendrá un costo:

$$\text{costo} = -15,65 + 1,29 \times 25 = 16,6 \text{ mil euros}$$

El modelo de regresión lineal simple

La diferencia entre cada valor y_i de la variable respuesta y su estimación \hat{y}_i se llama **residuo**:

$$e_i = y_i - \hat{y}_i$$



Ejemplo (cont.): Indudablemente, una empresa determinada que haya producido exactamente 25 mil unidades no va a tener un gasto de exactamente 16,6 mil euros. La diferencia entre el costo estimado y el real es el residuo. Si por ejemplo el costo real de la empresa es de 18 mil euros, el residuo es:

$$e_i = 18 - 16,6 = 1,4 \text{ mil euros}$$

● Modelo de Regresión Lineal Simple

Hipótesis del modelo de regresión lineal simple

- ▶ **Linealidad:** La relación existente entre X e Y es lineal,

$$f(x) = \beta_0 + \beta_1 x$$

- ▶ **Homogeneidad:** El valor promedio del error es cero,

$$E[u_i] = 0$$

- ▶ **Homocedasticidad:** La varianza de los errores es constante,

$$\text{Var}(u_i) = \sigma^2$$

- ▶ **Independencia:** Las observaciones son independientes,

$$E[u_i u_j] = 0$$

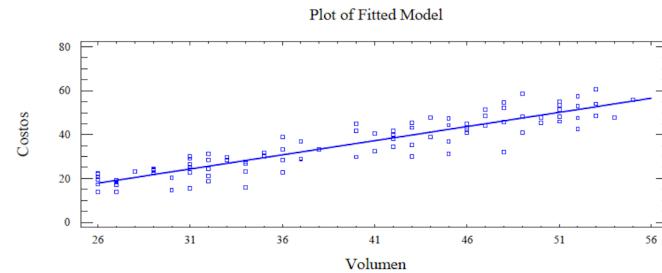
- ▶ **Normalidad:** Los errores siguen una distribución normal,

$$u_i \sim N(0, \sigma)$$

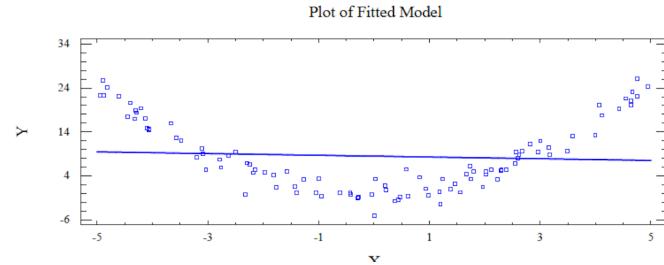
Hipótesis del modelo de regresión lineal simple

Linealidad

- ✓ Los datos deben ser razonablemente rectos.



- ✓ Si no, la recta de regresión no representa la estructura de los datos.

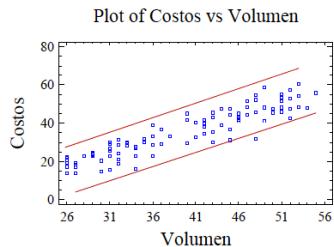


● Modelo de Regresión Lineal Simple

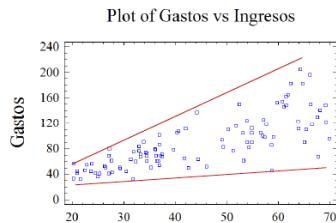
Hipótesis del modelo de regresión lineal simple

Homocedasticidad

La dispersión de los datos debe ser constante para que los datos sean **homocedásticos**.



Si no se cumple, los datos son **heterocedásticos**.



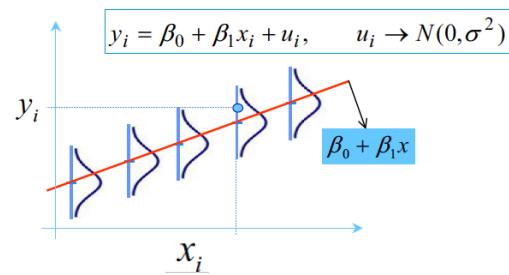
Hipótesis del modelo de regresión lineal simple

Independencia

- ▶ Los datos deben ser independientes.
- ▶ Una observación no debe dar información sobre las demás.
- ▶ Habitualmente, se sabe por el tipo de datos si son adecuados o no para el análisis.
- ▶ En general, las series temporales no cumplen la hipótesis de independencia.

Normalidad

- ▶ Se asume que los datos son normales a priori.



● Modelo de Regresión Lineal Simple

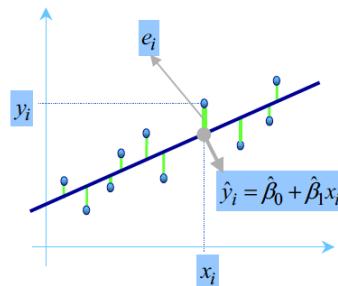
Estimadores de mínimos cuadrados

Gauss propuso en 1809 el **método de mínimos cuadrados** para obtener los valores $\hat{\beta}_0$ y $\hat{\beta}_1$ que mejor se ajustan a los datos:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

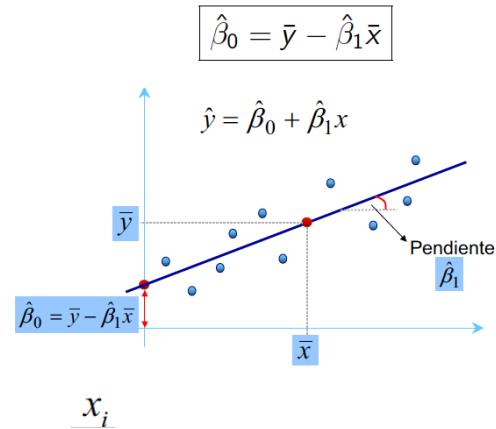
El método consiste en minimizar la suma de los cuadrados de las distancias verticales entre los datos y las estimaciones, es decir, **minimizar la suma de los residuos al cuadrado**,

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right)^2$$



El resultado que se obtiene es:

$$\hat{\beta}_1 = \frac{\text{cov}(x, y)}{s_x^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$



● Modelo de Regresión Lineal Simple. Ejemplo

- ✓ Los datos de la producción de trigo en toneladas (X) y el precio del kilo de harina en pesetas (Y) en la década de los 80 en España fueron:

Producción de trigo	30	28	32	25	25	25	22	24	35	40
Precio de la harina	25	30	27	40	42	40	50	45	30	25

Ajusta la recta de regresión por el método de mínimos cuadrados

Resultados

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{10} x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^{10} x_i^2 - n \bar{x}^2} = \frac{9734 - 10 \times 28,6 \times 35,4}{8468 - 10 \times 28,6^2} = -1,3537$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 35,4 + 1,3537 \times 28,6 = 74,116$$

La recta de regresión es:

$$\hat{y} = 74,116 - 1,3537x$$

Intervalos de Confianza

Intervalos de confianza para la media poblacional

- **Intervalo de confianza** es el intervalo que contiene al parámetro que se está estimando con un cierto nivel de confianza.
- **Nivel de confianza ($1 - \alpha$)**, significa que el $(1 - \alpha) \cdot 100\%$ de los intervalos de confianza contienen el parámetro poblacional que se está estimando.

A cada nivel de confianza (N_c) le corresponde un **valor crítico $z_{\alpha/2}$** correspondiente a la distribución normal $N(0, 1)$ y que cumple:

$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$$

A los extremos del intervalo de confianza se les llama **límites de confianza**.

El **valor crítico $z_{\alpha/2}$** correspondiente a un nivel de confianza N_c en tanto por ciento, se calcula mediante la expresión:

$$P(Z \leq z_{\alpha/2}) = \frac{1 + \frac{N_c}{100}}{2}$$

Y después buscando el resultado **dentro** de las tablas de la distribución normal.

Si fijamos el nivel de confianza $N_c = 95\%$ hallar el valor crítico $z_{\alpha/2}$.

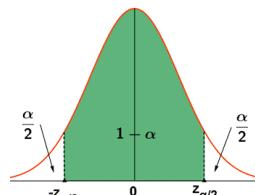
$$P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 0,95$$

$$P(Z \leq z_{\alpha/2}) = \frac{1 + \frac{N_c}{100}}{2} = \frac{1 + \frac{95}{100}}{2} = \frac{1,95}{2} = 0,975$$

En las tablas de la distribución normal, el número 0,975 corresponde a: **1,96**

En una distribución $N(\mu, \sigma)$ el intervalo correspondiente a una probabilidad $p = 1 - \alpha$ es:

$$(\mu - z_{\alpha/2} \cdot \sigma, \quad \mu + z_{\alpha/2} \cdot \sigma)$$



En una distribución normal $N(70, 6)$, obtener los intervalos característicos para el 90%, 95% y 99%.

$$90\% \Rightarrow 1 - \alpha = 0,9 \Rightarrow z_{\alpha/2} = 1,645 \Rightarrow (70 - 1,645 \cdot 6, \quad 70 + 1,645 \cdot 6) = (60,13, \quad 79,87)$$

$$95\% \Rightarrow 1 - \alpha = 0,95 \Rightarrow z_{\alpha/2} = 1,96 \Rightarrow (70 - 1,96 \cdot 6, \quad 70 + 1,96 \cdot 6) = (58,24, \quad 81,76)$$

$$99\% \Rightarrow 1 - \alpha = 0,99 \Rightarrow z_{\alpha/2} = 2,575 \Rightarrow (70 - 2,575 \cdot 6, \quad 70 + 2,575 \cdot 6) = (54,55, \quad 85,45)$$

● Intervalos de confianza para la media poblacional

Un intervalo de confianza para la media poblacional de una distribución normal con desviación típica σ conocida, con un nivel de confianza $1 - \alpha$ construido a partir de una muestra de tamaño n , es:

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right)$$

Si σ es desconocida y n es grande $n \geq 30$, el intervalo de confianza viene dado por:

$$\left(\bar{x} - z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{\hat{\sigma}}{\sqrt{n}} \right)$$

Donde $\hat{\sigma}^2$ es la cuasivarianza muestral.

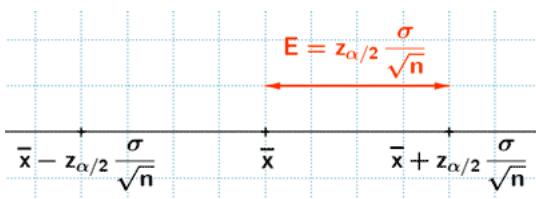
$$\hat{\sigma}^2 = \sigma^2 \cdot \frac{n}{n-1}$$

● Intervalos de confianza para la media poblacional

El error máximo admisible en la estimación de la media poblacional utilizando el intervalo de confianza para la media con un nivel de confianza $1 - \alpha$ es :

$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

El error será igual o menor que la mitad de la amplitud del intervalo, es decir, el radio del intervalo.



- Cuanto mayor sea n, menor será el error cometido.
- Cuanto mayor sea $1-\alpha$, mayor será $z_{\alpha/2}$ y , por tanto, también el error.

● Intervalos de confianza para la diferencia de medias

Un intervalo de confianza para la diferencia de medias poblacionales de dos distribuciones normales con desviaciones típicas σ_1 y σ_2 conocidas, con un nivel de confianza $1-\alpha$ construido a partir de dos muestras de tamaño n_1 y n_2 es:

$$\left((\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}, (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \right)$$

El error máximo admisible en la estimación de la diferencia de medias utilizando el intervalo de confianza para la diferencia de medias con un nivel de confianza $1 - \alpha$ es su radio:

$$E = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

- Cuanto mayores sean los tamaños de las muestras, menor será el error cometido.
- Cuanto mayor sea el nivel de confianza, mayor será $z_{\alpha/2}$ y, por tanto, también el error.



Diagramas de Caja-Bigotes (boxplots o box and whiskers)

Son una presentación visual que describe varias características importantes, al mismo tiempo, tales como la dispersión y simetría (visualizan, la tendencia central, la dispersión y la presencia posible de datos atípicos)

Para su realización se representan los tres cuartiles y los valores mínimo y máximo de los datos, sobre un rectángulo, alineado horizontal o verticalmente.

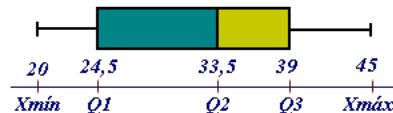
Ejemplos:

1. Dada la distribución: 20 23 24 24 24 25 29 31 31 33 34 36 36 37 39 39 40 40 41 45

$$Q_1 = (24 + 25) / 2 = 24,5$$

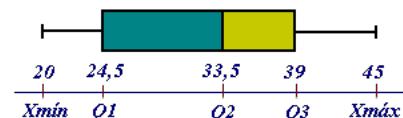
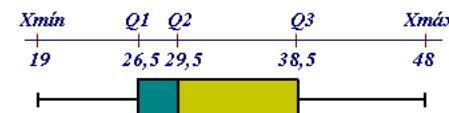
$$m_e = Q_2 = (33 + 34) / 2 = 33,5$$

$$Q_3 = (39 + 39) / 2 = 39$$



2. Dadas dos distribuciones:

35 38 32 28 30 29 27 19 48 40
39 24 24 34 26 41 29 48 28 22



La mayor utilidad de los diagramas caja-bigotes es para comparar dos o más conjuntos de datos

Diagramas de Caja-Bigotes (boxplots o box and whiskers)



Diagrama "tallo y hojas" (Stem-and-Leaf Diagram)

Permite obtener simultáneamente una distribución de frecuencias de la variable y su representación gráfica. Para construirlo basta separar en cada dato el último dígito de la derecha (que constituye la **hoja**) del bloque de cifras restantes (que formará el **tallo**).

Ejemplos:

1. Dada la distribución:

- 36 25 37 24 39 20 36 45 31 31 39 24 29 23 41 40 33 24 34 40
- 35 38 32 28 30 29 27 19 48 40 39 24 24 34 26 41 29 48 28 22

2. Dadas dos distribuciones:

5.03 7.32 9.02 11.07 13.32 15.07 16.50 18.32 20.07 22.38
6.02 7.37 9.07 11.32 13.37 15.20 17.02 18.37 20.20
6.18 7.50 9.24 11.37 13.50 15.32 17.07 18.50 20.32
6.37 8.02 9.32 12.02 14.02 15.37 17.20 19.02 20.37
6.48 8.05 9.37 12.07 14.07 15.50 17.32 19.07 20.50
6.55 8.20 10.02 12.32 14.20 16.02 17.37 19.20 21.02
7.02 8.24 10.07 12.37 14.32 16.07 17.50 19.32 21.07
7.07 8.32 10.32 13.02 14.37 16.20 18.02 19.37 21.20
7.20 8.37 10.37 13.07 14.50 16.32 18.07 19.50 21.32
7.25 8.51 11.02 13.20 15.02 16.37 18.20 20.02 21.37

(N = 20) Hojas										Tallos	Hojas (N = 20)									
9	9	8	8	7	6	4	4	2	9	1	0	3	4	4	4	5	9			
9	8	5	4	2	0				2	2	1	1	3	4	6	6	7	9	9	9
8	8	1	0						3	3	0	0	1	5						

05 | 03
06 | 02 18 37 48 55
07 | 02 07 20 25 32 37 50
08 | 02 05 20 24 32 37 51
09 | 02 07 24 32 37
10 11 12 | 02 07 32 37
13 14 15 16 17 18 19 20 | 02 07 20 32 37 50
21 | 02 07 20 32 37
22 | 38



Calle Rufino González 25
28037 Madrid
+34810527241
www.mioti.es

