Módulo 0: Nociones matemáticas básicas

0.3. Estadística Descriptiva

Rafael Zambrano
rafazamb@gmail.com

Introducción

PARTE 3. ESTADÍSTICA DESCRIPTIVA

3.1 Conceptos básicos
3.2 Media, varianza y desviación estándar
3.3 Estadísticos de posición
3.4 Frecuencias e histogramas
3.5 Relación entre variables numéricas
3.6 Análisis exploratorio de datos
3.7 Interpretación y presentación de los datos

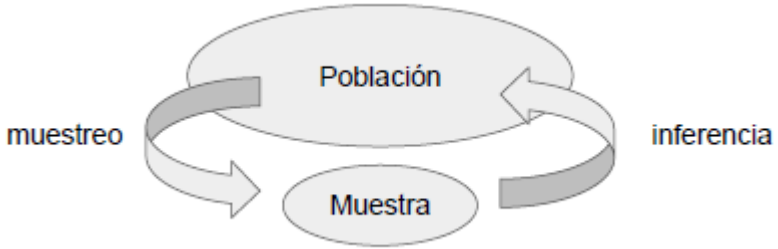
3. Estadística Descriptiva

3.1. Conceptos básicos

Conceptos básicos

- **Población**: conjunto sobre el que estamos interesados en obtener conclusiones (hacer inferencias)
  - ➢ Normalmente es demasiado grande para poder abarcarlo

- **Muestra**: subconjunto de la población al que tenemos acceso y sobre el que realmente hacemos las observaciones (mediciones)
  - ➢ Debe ser representativo



Conceptos básicos

- **Variable**: Característica observable que varía entre los diferentes individuos de una población. La información que disponemos de cada individuo es resumida en variables

- **Dato**: Es un valor particular de la variable

Ejemplo: En los individuos de la **población** de alumnos universitarios en España, de uno a otro es **variable:**

- Su altura: {1.7m, 1.64m, 1.55m...}
- Su color de pelo: {rubio, moreno, pelirrojo...}
- Su edad: {20,18,26....}

Conceptos básicos

- **Parámetro**: Cantidad numérica calculada sobre una población
  - Intenta resumir toda la información que hay en la población en unos pocos números (parámetros)
  - Ejemplo: la media de edad de los universitarios españoles

- **Estadístico**: cantidad numérica calculada sobre una muestra que resume su información sobre algún aspecto
  - Normalmente nos interesa conocer un parámetro, pero por la dificultad que conlleva estudiar a toda la población, calculamos un estimador sobre una muestra y confiamos en que sean próximos
  - Ejemplo: la media de edad de los universitarios de Salamanca

Definición de Estadística

La estadística es la Ciencia de la **sistematización, recogida, ordenación y presentación** de los datos referentes a un fenómeno que presenta variabilidad o incertidumbre para su estudio metódico, con objeto de **deducir las leyes** que rigen estos fenómenos y poder hacer previsiones sobre los mismos, tomar **decisiones** y obtener **conclusiones**

División de la Estadística

- La **estadística descriptiva** busca obtener información sobre la población basándose en el estudio de los datos de una muestra tomada a partir de ella.

- La **estadística inferencial** se preocupa de llegar a conclusiones basadas en la muestra y luego hacerlos válidos para toda la población.

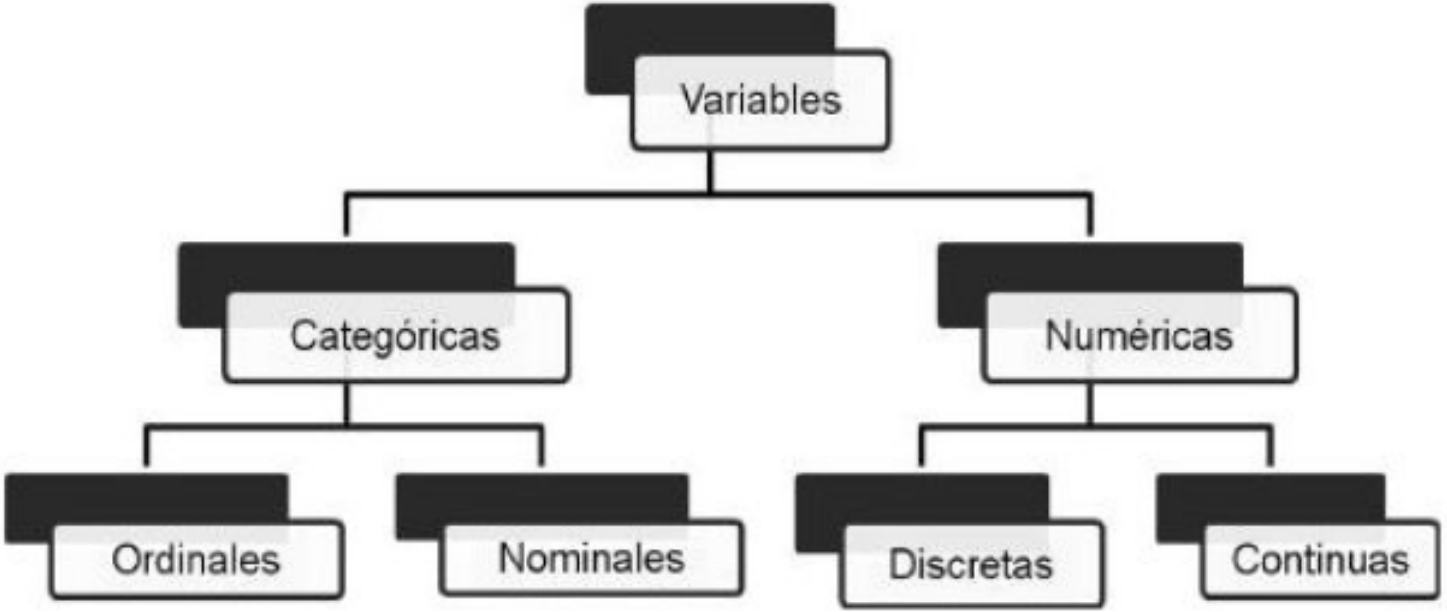
Estadísticos

- **Posición** (basados en el orden): Dividen un conjunto ordenado de datos en grupos con la misma cantidad de individuos
  - Cuantiles, percentiles, cuartiles, deciles...

- **Centralización**: Indican valores con respecto a los cuales los datos parecen agruparse
  - Media, mediana y moda

- **Dispersión**: Indican la mayor o menor concentración de los datos con respecto a las medidas de centralización
  - Desviación estándar, coeficiente de variación, rango,varianza...

Tipos de variables



3. Estadística Descriptiva

3.2. Media, Varianza y Desviación estándar

Estadística descriptiva

Ejemplo: Se han tomado mediciones de la concentración de nitrito en agua. Los datos son los siguientes:

```
0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47 0.23 0.55 3.21 4.02 0.23
```

Cuando se describen los datos las preguntas básicas que hemos de tener en la cabeza pueden ser:

1. ¿De qué orden son?
2. ¿Cómo de dispersos están?
3. ¿Hay datos anormales que estén muy alejados de los demás?

Media aritmética

La medida de localización más utilizada es la **media aritmética** o *media muestral*

```
x=0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47 0.23 0.55 3.21 4.02 0.23
```

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.32 + 0.36 + 0.24 + 0.11 + 0.11 + 0.44 + 2.79 + 2.99 + 3.47 + 0.23 + 0.55 + 3.21 + 4.02 + 0.23}{14} = 1.36$$

En Excel: `PROMEDIO(matriz)`

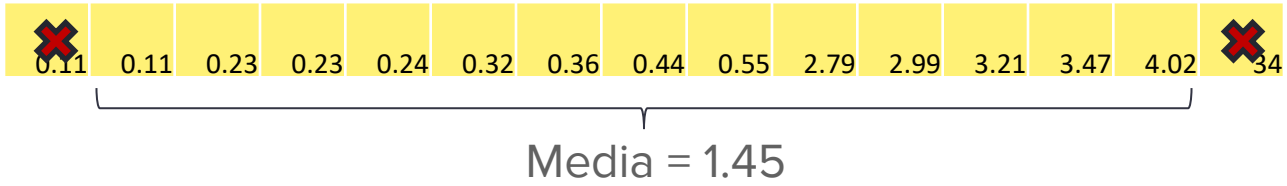
NOTA: La media **poblacional** suele expresarse con la letra $\mu = \frac{\sum_{i=1}^{n} X_i}{N}$

Media aritmética

La media muestral es muy sensible a datos anómalos o extremos. Por ejemplo, a nuestros datos originales les vamos a añadir un dato anómalo: el valor 34

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{0.32 + 0.36 + 0.24 + 0.11 + 0.11 + 0.44 + 2.79 + 2.99 + 3.47 + 0.23 + 0.55 + 3.21 + 4.02 + 0.23 + 34}{15} = 3.54$$

Para evitar estos datos anómalos podemos usar la *media ajustada*, que ordena la muestra y "recorta" una proporción determinada, por ejemplo, el 15%



En Excel: `MEDIA.ACOTADA(matriz, porcentaje)`

Varianza y desviación estándar

Para cuantificar lo dispersos que están nuestros datos se suele utilizar la **varianza** y la **desviación estándar**.

- *Varianza poblacional*

$$\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N}$$

- *Desviación estándar poblacional*

$$\sigma = \sqrt{\sigma^2} = \sqrt{\sum_{i=1}^{n} \frac{(X - \mu)^2}{N}}$$

- *Varianza muestral*

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

- *Desviación estándar muestral*

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

Varianza y desviación estándar

Varianza poblacional

$$\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N}$$

Ejemplo: Tengo los pesos en Kg de pacientes en dos clínicas diferentes, y quiero saber en qué clínica existe mayor dispersión

	Pesos Clínica 1	Pesos Clínica 2		$(X_i - \mu)$	$(X_i - \mu)$		$(X_i - \mu)^2$	$(X_i - \mu)^2$
	80	80		0	0		0^2	0^2
	90	90		10	10		10^2	10^2
	70	70		-10	-10		-10^2	$(-10)^2$
		81			1			1^2
		79			-1			$(-1)^2$
μ	80	80	$\sum$	0	0	$\sum$	200	202

Clínica 1

$$\sigma^2 = \frac{200}{3} = 66.7$$

Clínica 2

$$\sigma^2 = \frac{202}{5} = 40.4$$

Varianza y desviación estándar

Desviación estándar

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

Ejemplo: Tengo los pesos en Kg de pacientes en dos clínicas diferentes, y quiero saber en qué clínica existe mayor dispersión

Clínica 1

$$\sigma^2 = \frac{200}{3} = 66.7 \text{ Kg}^2 \Rightarrow \sigma = \sqrt{66.7} = 8.16 \text{ Kg}$$

Clínica 2

$$\sigma^2 = \frac{202}{5} = 40.4 \text{ Kg}^2 \Rightarrow \sigma = \sqrt{40.4} = 6.35 \text{ Kg}$$

- La desviación estándar nos permite entender mejor la dispersión de los datos

Varianza y desviación estándar

¿Por qué la varianza muestral tiene n-1 en el denominador?

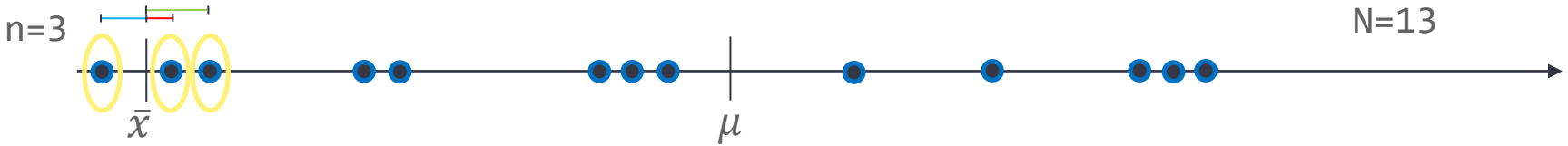
- *Varianza poblacional*

$$\sigma^2 = \sum_{i=1}^{n} \frac{(X_i - \mu)^2}{N}$$

- *Varianza muestral*

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

En general se sub-estima el valor de la varianza poblacional al realizar un muestreo. Al dividir por n-1 se obtiene un valor de la varianza muestral ligeramente mayor, la cual es una mejor estimación de la varianza poblacional.

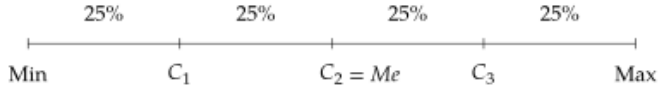


3. Estadística Descriptiva

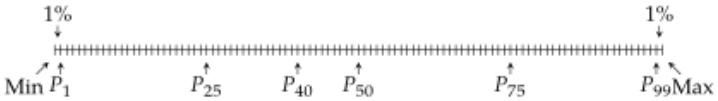
3.3. Estadísticos de posición

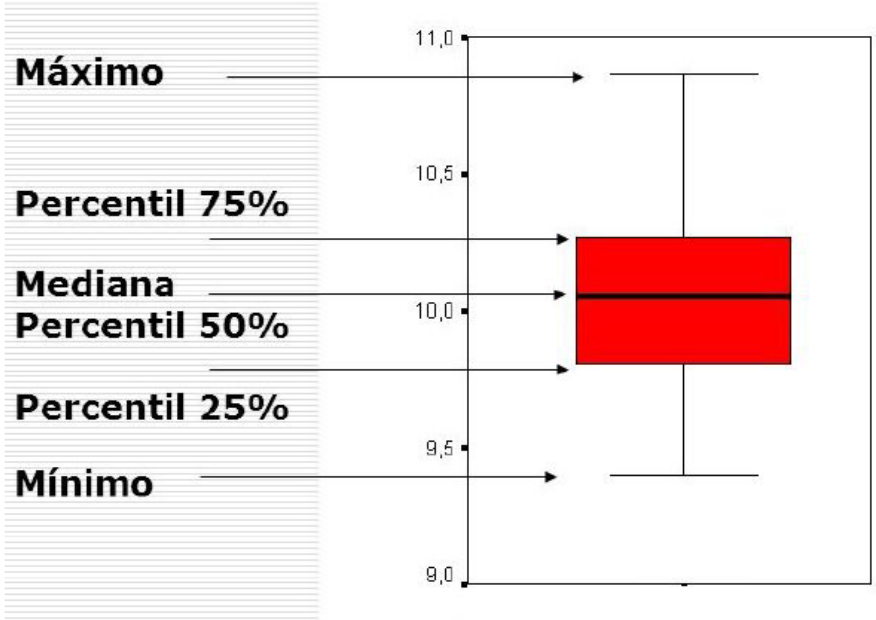
Cuartiles y percentiles

Cuartiles



Percentiles





En Excel, para obtener un percentil de orden *p* hacemos: `PERCENTIL(matriz,p)`

Para obtener un cuartil deteriminado: `CUARTIL(matriz,N)`

Mediana

- La *mediana muestral* es el percentil de orden 50% (o el valor del cuartil 2)

¿Cómo se calcula?

- Para muestras con longitud impar:

Mediana = 8

- Para muestras con longitud par:

Mediana = (8+10)/2 = 9

- La mediana no se ve afectada por valores extremos, por lo que es adecuado usarla en lugar de la media para describir un conjunto de datos

Rango

El **rango** de una muestra es el intervalo comprendido entre el valor máximo y el mínimo

Ejemplo:

```
x = 0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47 0.23 0.55 3.21 4.02 0.23
```

```
MAX(x) = 4.02
MIN(x) = 0.11
```

Una medida más robusta es el rango intercuartil (diferencia entre los cuartiles 3 y 1), el cual nos dice la dispersión del 50% de los valores centrales de una muestra:

```
IQR = CUARTIL(x,3)-CUARTIL(x,1)
```

Descripción inicial de los datos

Antes de comenzar a trabajar con datos, es recomendable realizar una descripción básica, con los estadísticos mínimo, primer cuartil, media, mediana, tercer cuartil y máximo

Ejemplo:

```
x = 0.32 0.36 0.24 0.11 0.11 0.44 2.79 2.99 3.47 0.23 0.55 3.21 4.02 0.23
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.1100  0.2325  0.4000  1.3621  2.9400  4.0200
```

3. Estadística Descriptiva

3.4. Frecuencias e histogramas

Tabla de Frecuencias

Muestra la distribución de los datos mediante sus **frecuencias**

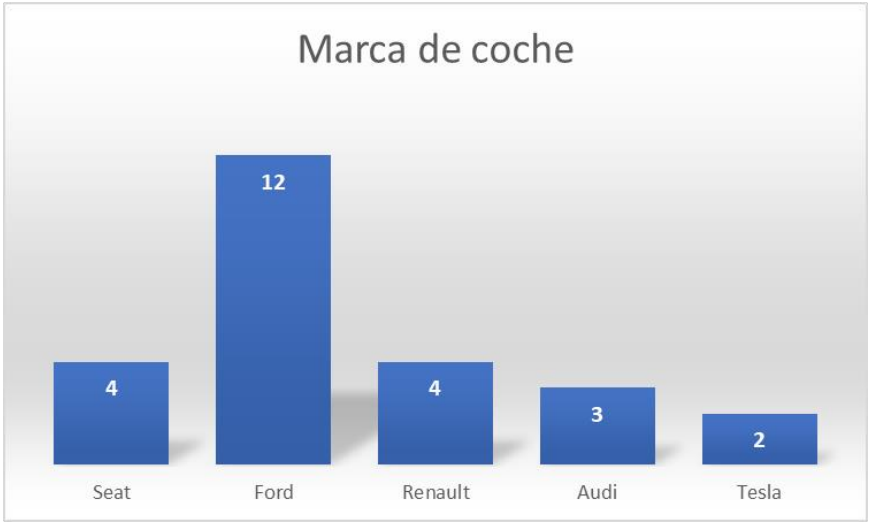
Ejemplo: En una empresa, 25 empleados tienen las siguientes marcas de coches:

Tesla Seat Ford Renault Audi Ford Seat Ford Ford Ford Tesla Seat Ford Audi Renault Renault Ford Ford Seat Ford Renault Audi Ford Ford Ford

X_i	Frecuencia absoluta f_i	Frecuencia abs. acumulada F_i	Frecuencia relativa $h_i=f_i/N$	%	Frecuencia acumulada H_i
Seat	4	4	4/25 = 0.16	16	0.16
Ford	12	16	12/25 = 0.48	48	0.64
Renault	4	20	4/25 = 0.16	16	0.8
Audi	3	23	3/25 = 0.12	12	0.92
Tesla	2	25	2/25 = 0.08	8	1

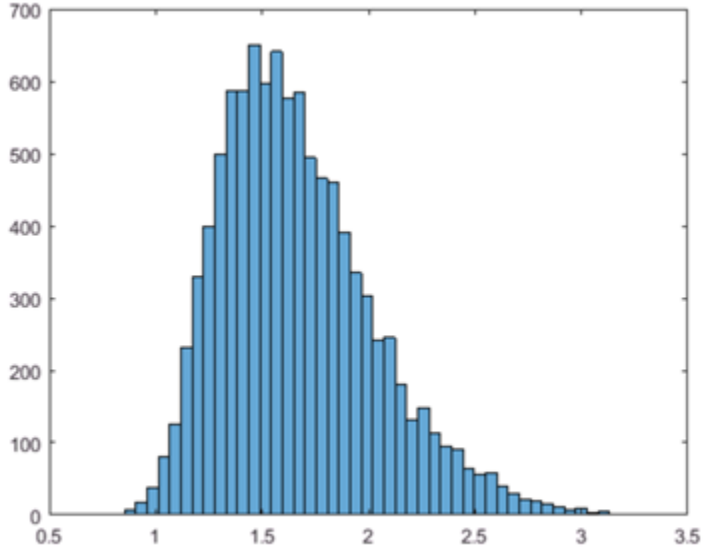
Tabla de Frecuencias

Para representar las frecuencias gráficamente podemos utilizar un diagrama de barras o tarta

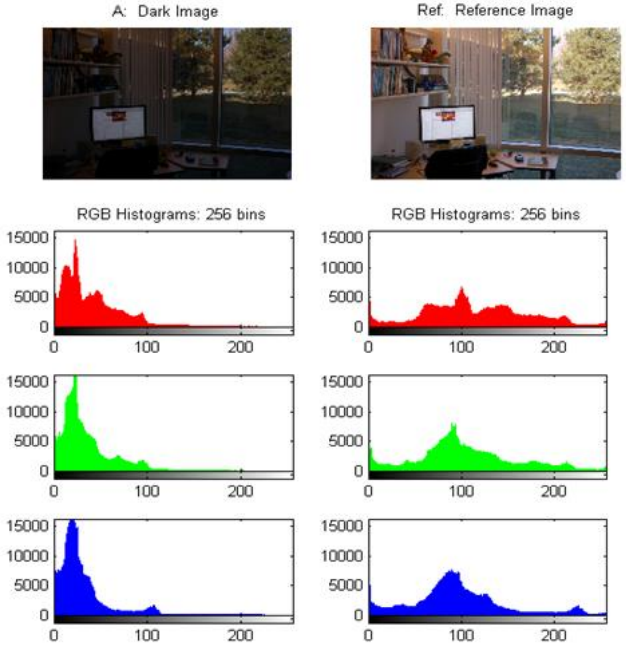


Histogramas

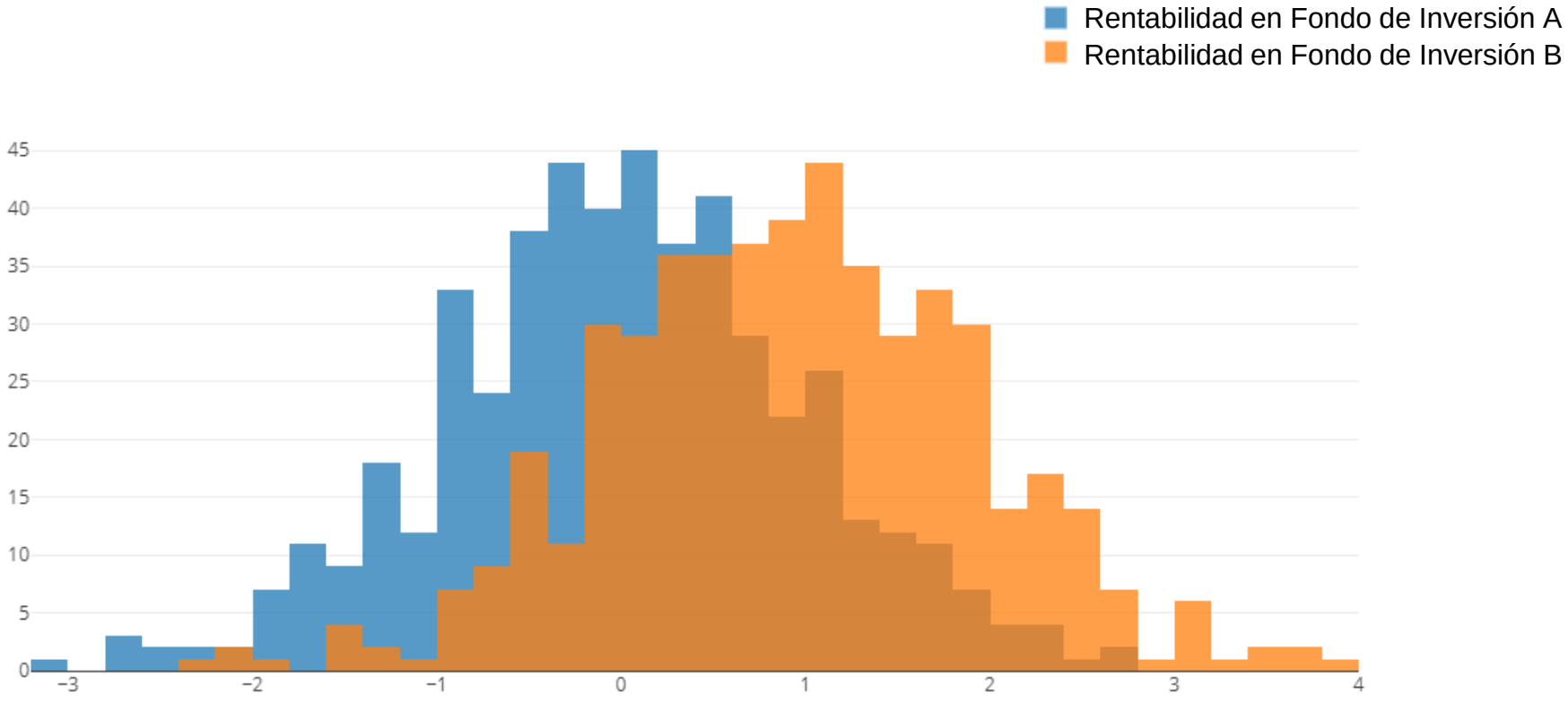
Un histograma es una representación gráfica de datos agrupados mediante intervalos. Los datos provienen de una variables numéricas continuas.



Tiempo de espera(s)



Histogramas



3. Estadística Descriptiva

3.5. Relación entre variables numéricas

Relación entre variables numéricas

Ejemplo: Se dispone de los siguientes datos acerca de las horas de sueño y el peso de una serie de personas

Horas de sueño	Peso (kg)
7	74
4	50
12	89
11	84
8	65
6	60
11	70
5	52

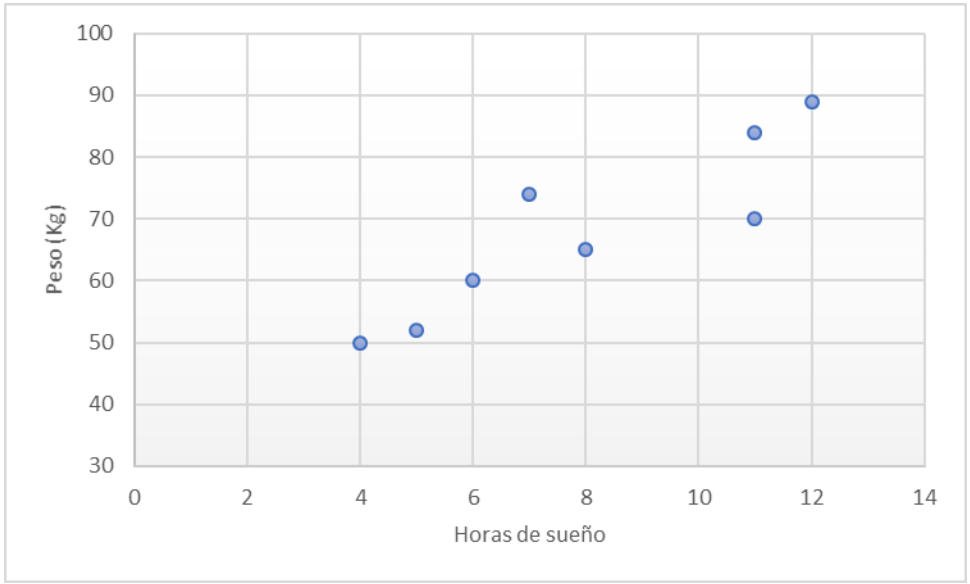


Gráfico de dispersión

Coeficiente de correlación

- Un coeficiente de correlación mide el grado en que dos variables tienden a cambiar al mismo tiempo. El coeficiente describe tanto la fuerza como la dirección de la relación
- La correlación de Pearson evalúa la relación lineal entre dos variables continuas. Una relación es lineal cuando un cambio en una variable se asocia con un cambio proporcional en la otra variable

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y - \bar{y})^2}}$$

- El valor está comprendido entre -1 y +1

Coeficiente de correlación

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y - \bar{y})^2}}$$

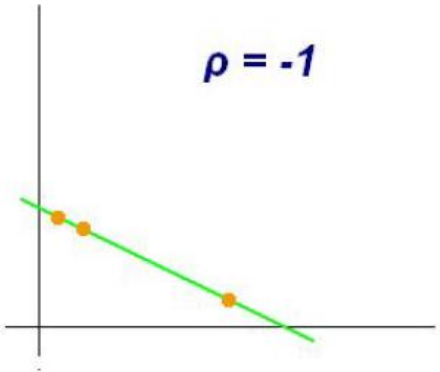
Horas de sueño	Peso (kg)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\,(y_i - \bar{y})$
7	74	-1	6	-6
4	50	-4	-18	72
12	89	4	21	84
11	84	3	16	48
8	65	0	-3	0
6	60	-2	-8	16
11	70	3	2	6
5	52	-3	-16	48
$\bar{x} = 8$	$\bar{y} = 68$			**268**

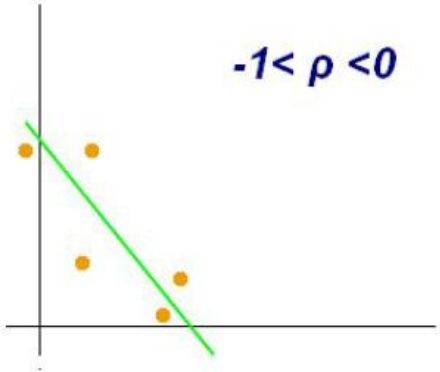
$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = 64$$

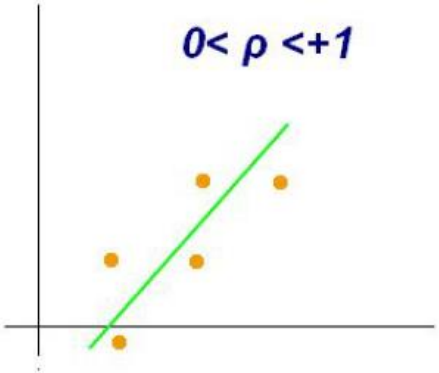
$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = 1390$$

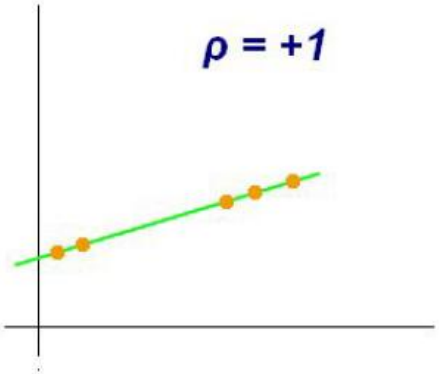
$$\rho(x,y) = \frac{268}{\sqrt{64}\cdot\sqrt{1390}} \approx 0.9$$

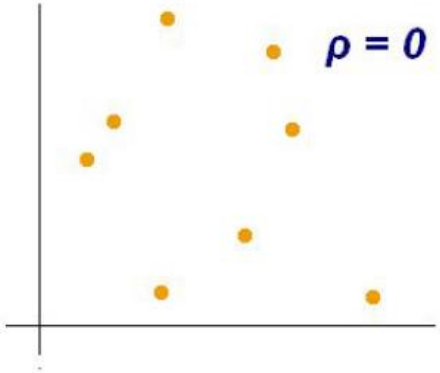
Coeficiente de correlación





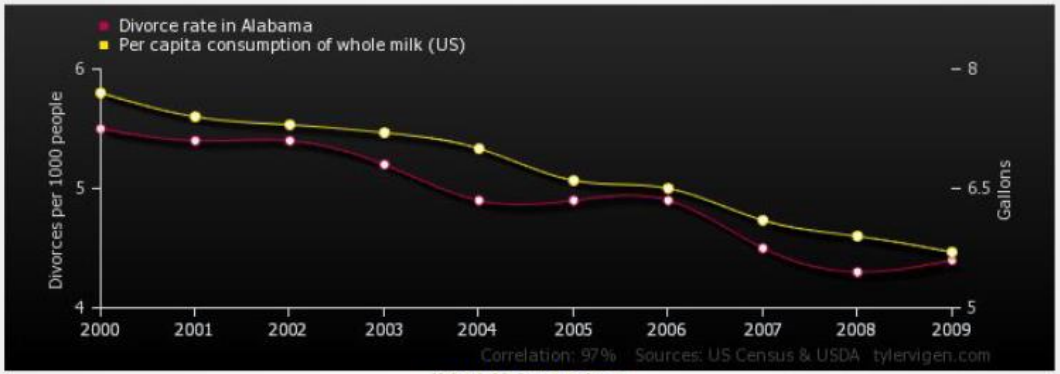






La correlación no implica causalidad

Divorce rate in Alabama
correlates with
Per capita consumption of whole milk (US)



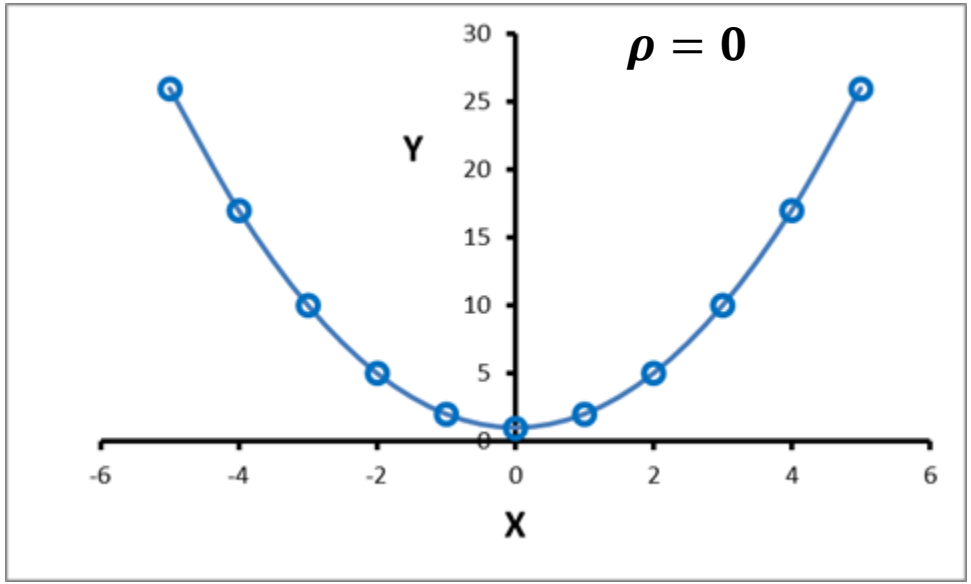
Upload this image to imgur

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Alabama *Divorces per 1000 people (US Census)*	5.5	5.4	5.4	5.2	4.9	4.9	4.9	4.5	4.3	4.4
Per capita consumption of whole milk (US) *Gallons (USDA)*	7.7	7.4	7.3	7.2	7	6.6	6.5	6.1	5.9	5.7

Correlation: 0.970609

Fuente: http://tylervigen.com/discover

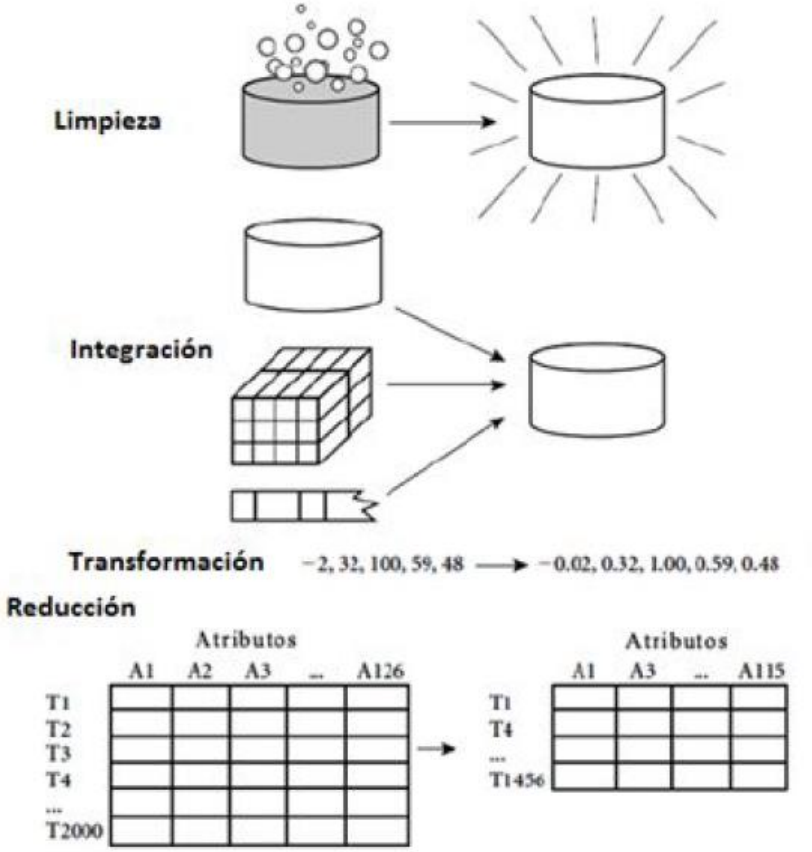
La correlación no implica causalidad



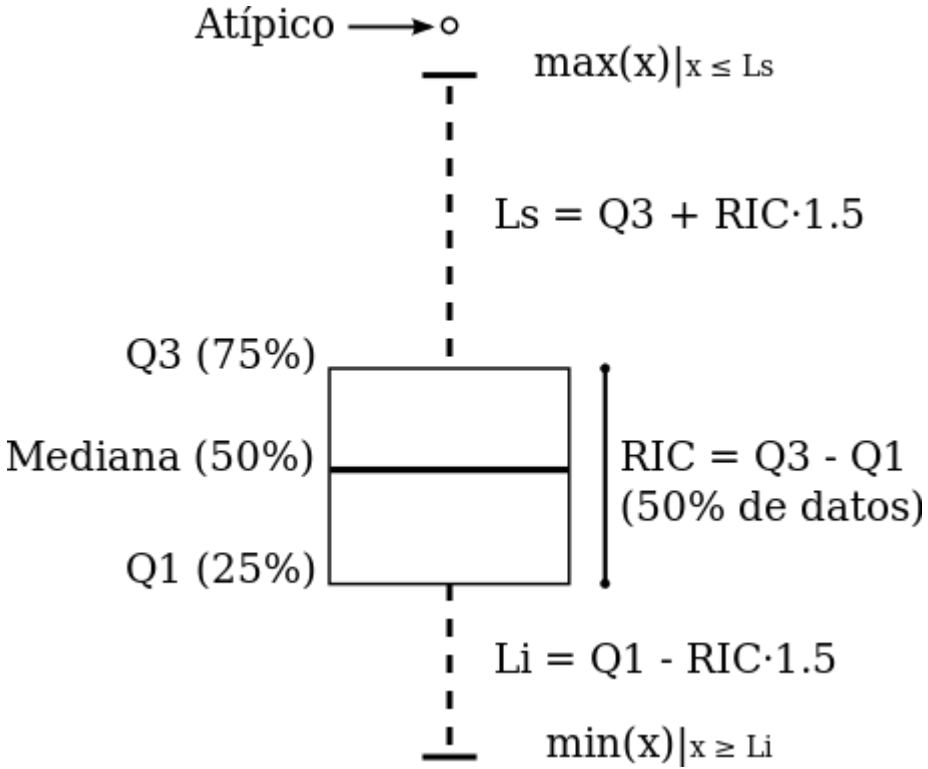
3. Estadística Descriptiva

3.6. Análisis exploratorio de datos

Los Datos


Limpieza
Integración
Transformación
−2, 32, 100, 59, 48 → −0.02, 0.32, 1.00, 0.59, 0.48
Reducción
Atributos
A1 A2 A3 ... A126
T1
T2
T3
T4
...
T2000
Atributos
A1 A3 ... A115
T1
T4
...
T1456

Detección de *outliers*



3. Estadística Descriptiva

3.7. Interpretación y presentación de los datos

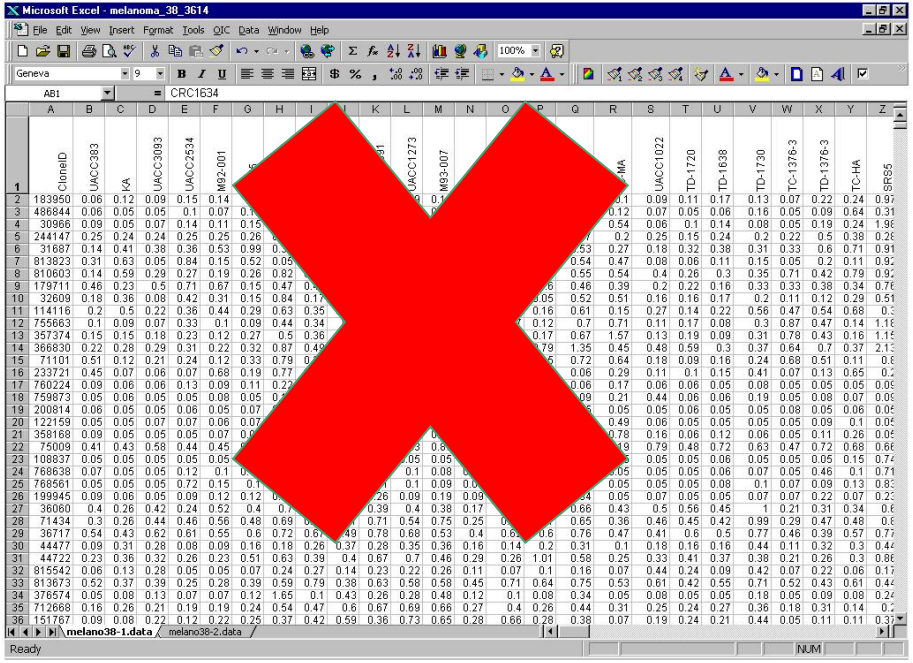
Interpretación y presentación de datos

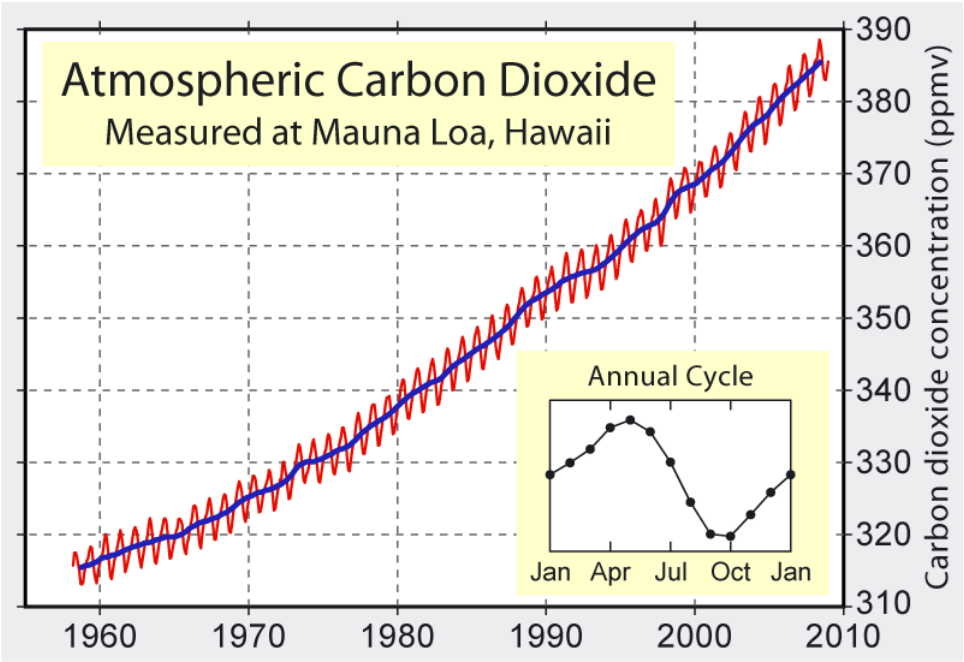
El propósito de analizar datos es obtener información útil y utilizable.

- Filtrar los datos relevantes
- Describir y resumir los datos
- Identificar las relaciones y diferencias entre las variables
- Comparar las variables
- Pronosticar y presentar resultados con la ayuda de gráficas

Interpretación y presentación de datos

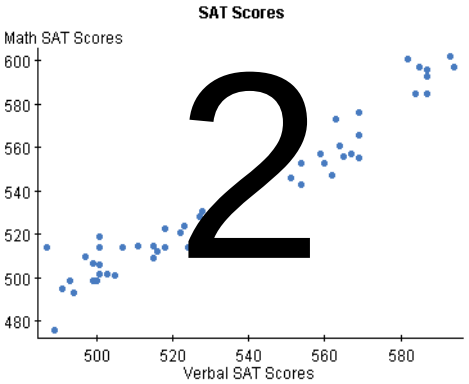
La presentación de los datos debe ser clara y concisa

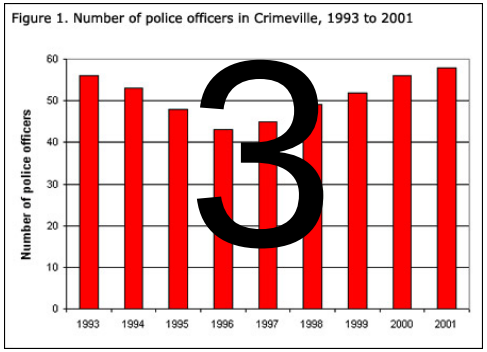


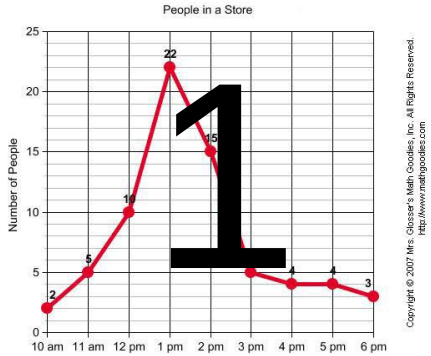


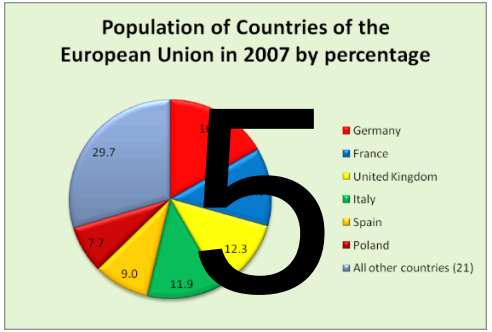
Tipos de Gráficos

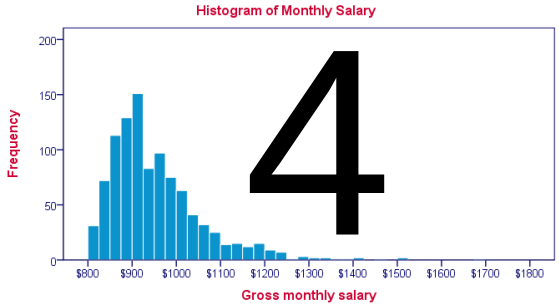
1. Líneas
2. Dispersión (*scatter*)
3. Barras
4. Histograma
5. Circular (tarta)
6. De Caja

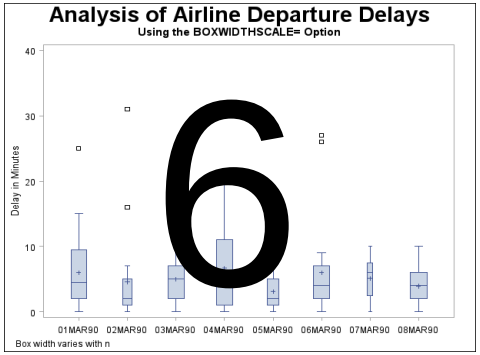






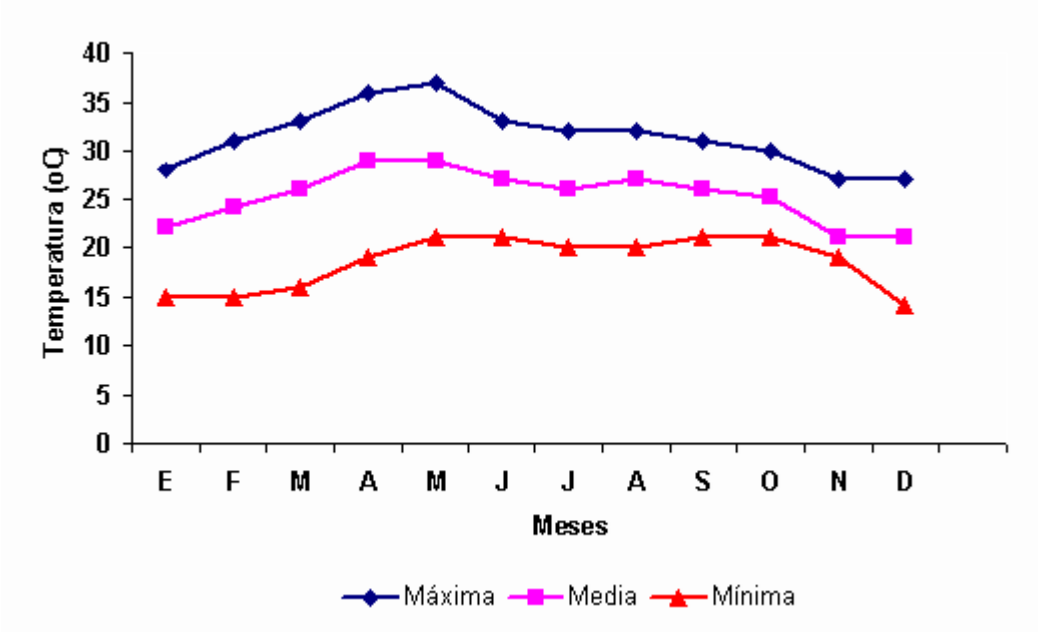


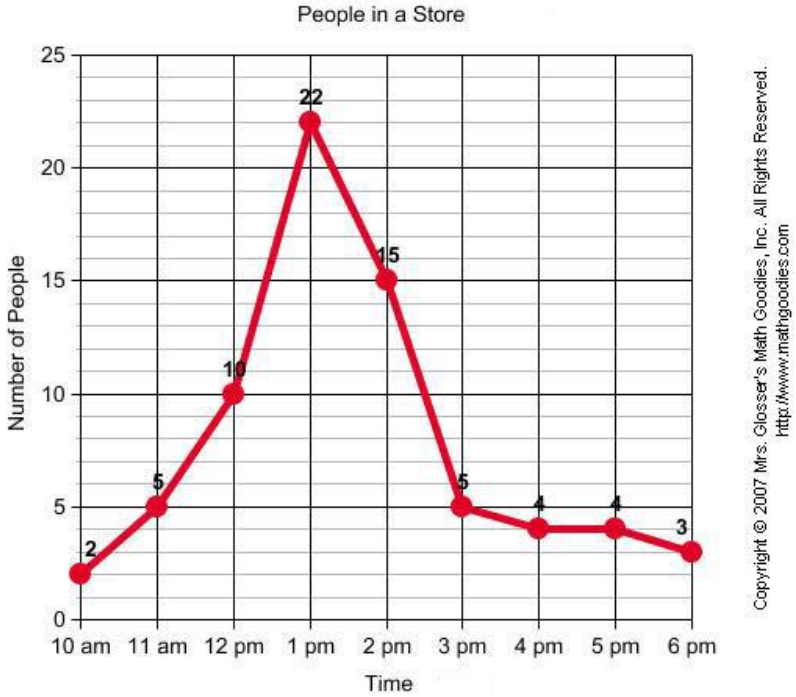




Gráficos de línea

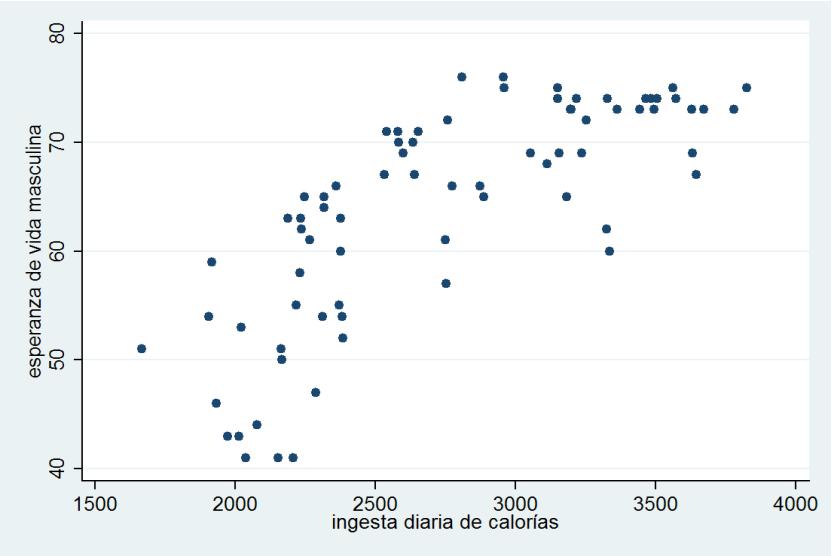
- Cambios en el tiempo

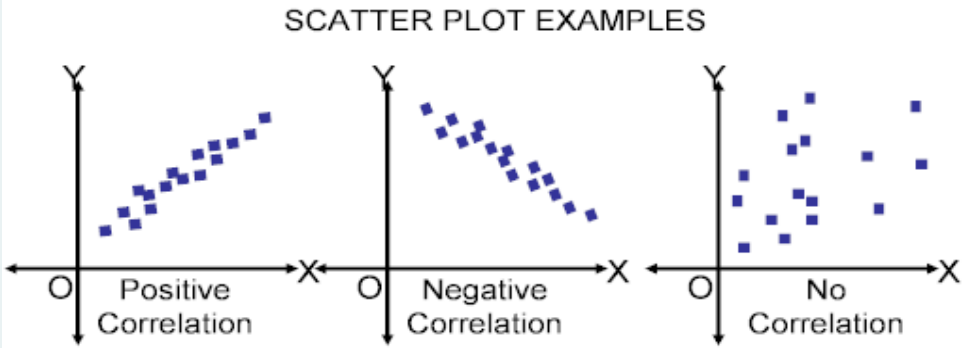




Gráficos de dispersión

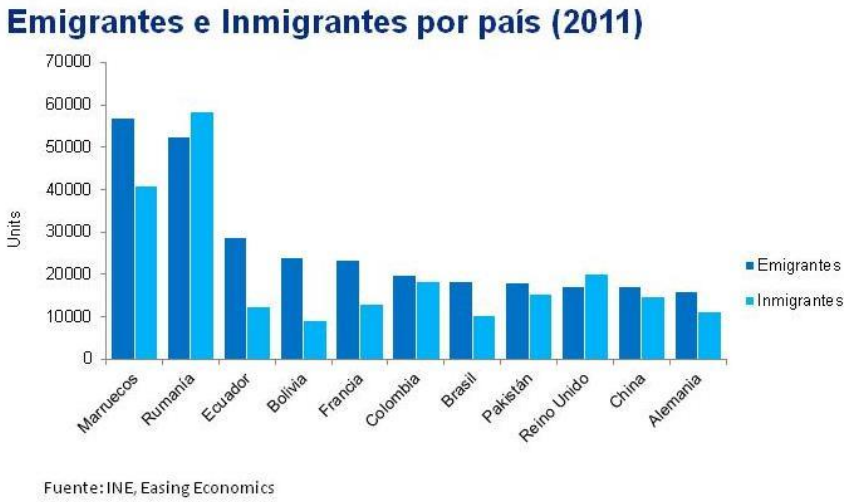
- Correlación entre variables



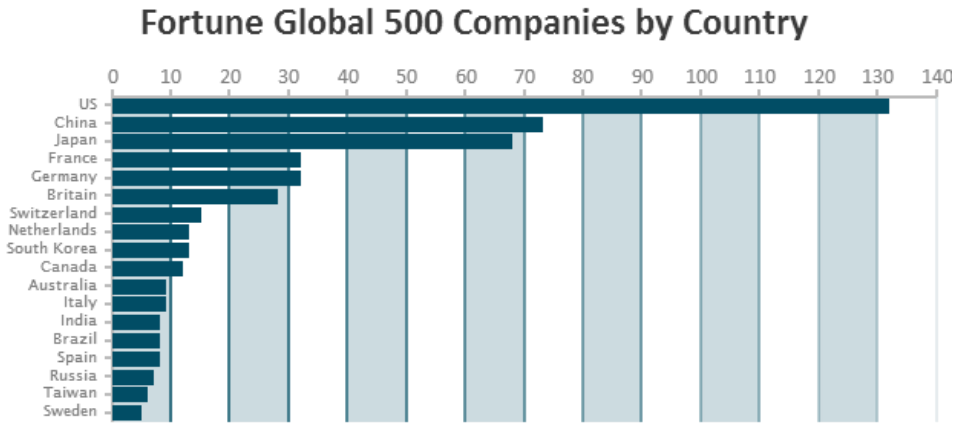


Gráficos de barra

- Comparar grupos

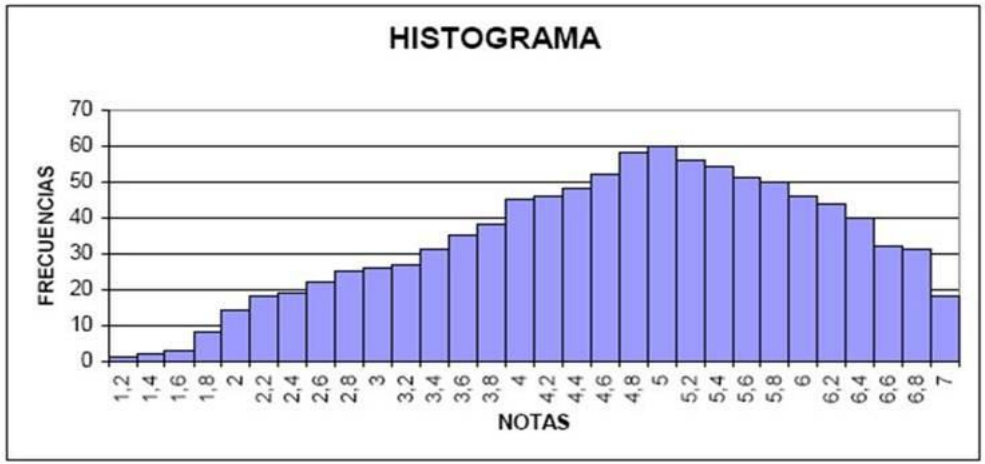


Fuente: INE, Easing Economics



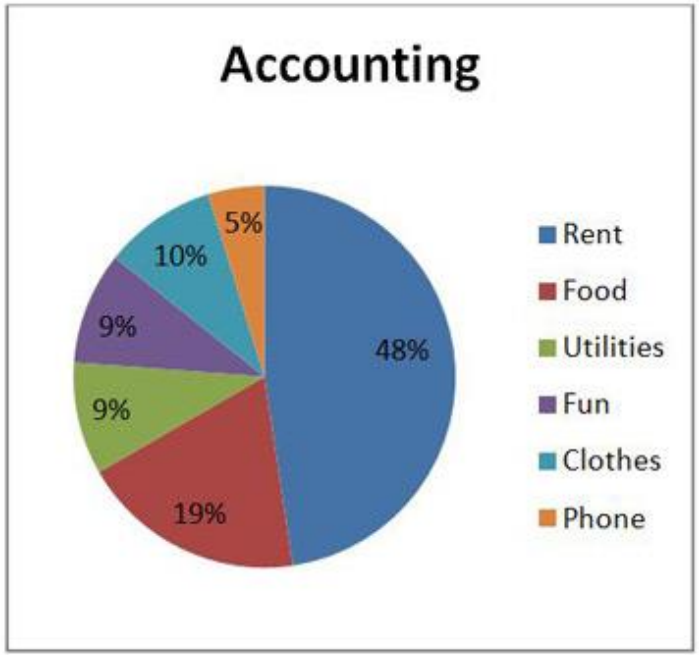
Histogramas

- Distribución de los datos



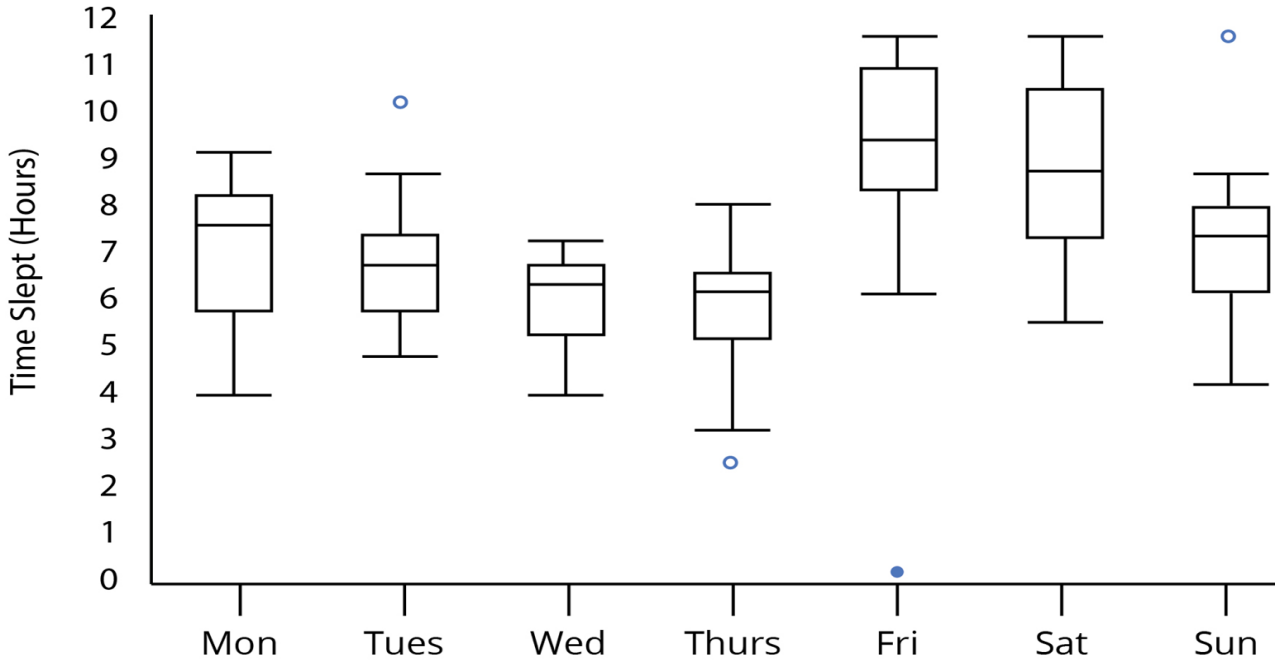
Gráficos de tarta

- Frecuencias o partes del total



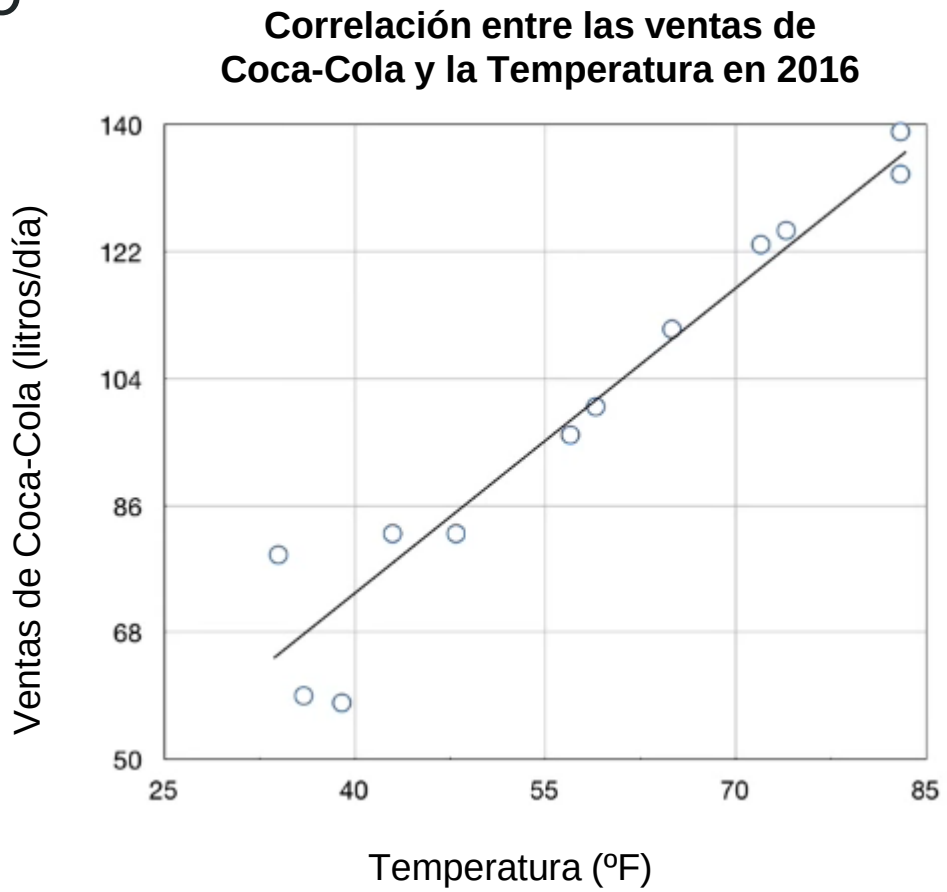
Diagramas de caja

- Informa sobre el máximo, mínimo, mediana, cuartiles y posibles valores atípicos

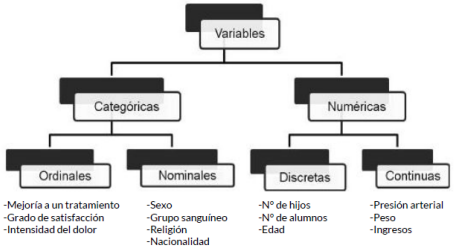


Elementos de un gráfico

- Independencia
- Buen título
- Etiquetar los ejes
- Especificar unidades
- Escalas lineales

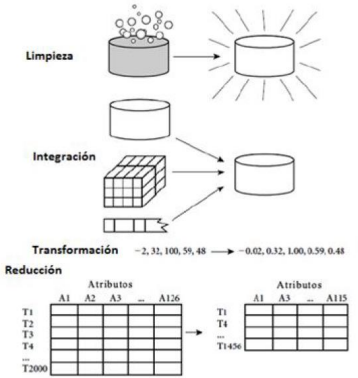


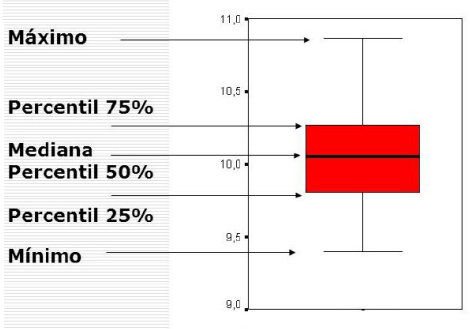
Conclusiones

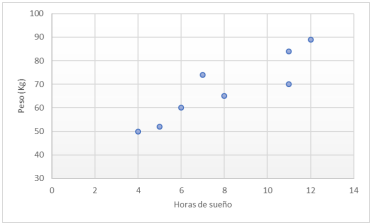


3. ESTADÍSTICA DESCRIPTIVA

3.1 Conceptos básicos
3.2 Media, varianza y desviación estándar
3.3 Estadísticos de posición
3.4 Frecuencias e histogramas
3.5 Relación entre variables numéricas
3.6 Análisis exploratorio de datos
3.7 Interpretación y presentación de los datos







¡Gracias!

Contacto: Rafael Zambrano

rafazamb@gmail.com