1. NATURALEZA Y TIPOS DE DATOS

1.1 Definición de estadística

La **Estadística** es la ciencia que trata de la teoría y aplicación de métodos apropiados para coleccionar, representar, resumir y analizar datos para hacer inferencias a partir de ellos.

1.2 Clasificaciones de estadística

- **Estadística descriptiva o deductiva**

  La **estadística descriptiva o deductiva** se encarga de recoger y resumir las características de una población o muestra deduciendo de esta descripción conclusiones sobre su estructura, además de las relaciones existentes entre otros colectivos distintos con los cuales se compara.

- **Estadística inductiva o inferencial**

  La **estadística inductiva o inferencial** se basa en los resultados del análisis de la muestra de la población induce o estima las leyes generales de comportamiento de la población.

1.3 Objetivo de la estadística

Población y muestra

- **Población** es el conjunto de todos los individuos o elementos que tienen unas características comunes y sobre el cual queremos efectuar nuestro estudio.
- **Muestra** es una parte de la población seleccionada aleatoriamente. Atención: muestra representativa es aquella que mantiene las características del conjunto de la población.

Ejemplo: Imaginemos que queremos realizar un estudio sobre la estatura de los alumnos de la facultad de Informática. En este caso, la población serían todos los alumnos de la facultad. Una muestra sería escoger al azar una parte de estos alumnos ,por ejemplo, una clase de segundo.

El tamaño de la muestra dependerá del grado de exactitud que queramos dar a nuestro estudio. Generalmente, a mayor tamaño de la muestra obtendremos resultados más fiables pero también nos supondrá mayores costes.

El objetivo de la estadística es encontrar un punto óptimo entre costes y fiabilidad.

1.4 Nivel observacional de una variable: datos y errores de medida

Antes de adentrarnos en la estadística económica, es muy importante tener claros conceptos que utilizaremos a lo largo de toda la asignatura:

- **Variable** es la característica que nosotros observamos/estudiamos. Por ejemplo, la altura.
- **Observación** es cada uno de los datos que recojo sobre la variable. Tendremos tantos datos como individuos analicemos.

Los DATOS son el conjunto de observaciones de una o más características obtenidas en una población o en una muestra.

Ejemplo: Variable de estudio: "sueldo" de los habitantes de Barcelona. La respuesta de cada individuo encuestado, sería una observación, y el conjunto de observaciones, serían los datos de que disponemos para nuestro estudio.

En este punto, cabe también señalar que a pesar de que la estadística es un instrumento muy útil, hemos de ser escépticos respecto a ella ya que existen errores de medida de los datos.
A pesar de ello, nosotros **supondremos**, en lo que sigue, **que los datos han sido medidos sin error** y que disponemos del valor de la variable que nos interesa para todos los individuos que queremos estudiar.

1.5 Tipos de datos

1.5.1. Datos cualitativos y datos cuantitativos

- Los **datos cualitativos** corresponden a observaciones de variables que sólo pueden tomar un número finito de valores, que no son numéricas y que no precisan de una ordenación de las diferentes clases.

  Ejemplo: nivel de estudios de una población. A este tipo de datos se les llama atributos, y a cada uno de los valores que toman se les conoce como categorías. (Si la variable tiene únicamente 2 categorías, se conoce como dicotómica. Ej: Sexo; hombre o mujer).

- Los **datos cuantitativos** hacen referencia a observaciones de variables que pueden medirse, que toman un número infinito de valores y que pueden representarse de forma numérica. Los datos cuantitativos, a diferencia de los cualitativos, siempre pueden ordenarse.

  Ejemplo: edad, número de hermanos.

Dos tipos de datos cuantitativos:

- **Discretos**: sólo pueden tomar valores enteros. Ej: nº de hermanos.
- **Continuos**: hacen referencia a variables que pueden tomar valores fraccionarios, y dados 2 valores cualesquiera, siempre pueden tomar infinitos valores entre ellos. Ej: altura.

1.5.2. Datos de corte transversal y datos temporales

- Los **datos de corte transversal** (o cross-section) se miden en un instante concreto del tiempo, y hacen referencia a varios individuos.

  Ejemplo: beneficios de ciertas empresas durante el ejercicio 1998.

- En los **datos temporales**, analizamos a un único individuo a lo largo del tiempo.

  Ejemplo: beneficios de una empresa desde el ejercicio 1995 hasta el de 1998.

1.5.3. Datos unidimensionales y datos multidimensionales

- Hablamos de datos **unidimensionales** cuando nos referimos a una única variable.

- Los datos **multidimensionales** se obtienen observando simultáneamente varias variables. Estos datos dan lugar a las matrices de datos (que estudiaremos en el siguiente tema).

DESCRIPCIÓN DE DATOS CUANTITATIVOS

1.6 Datos unidimensionales

1.6.1. Métodos de tabulación

La tabulación es un proceso que consiste en la construcción de tablas que sintetizan la información recogida. Estas tablas se conocen como **tablas de frecuencias** o **tablas de distribución de frecuencias**.

Ejemplo: Variable de estudio la estatura.
Consideraremos una población de 150 alumnos de estadística. Extraemos de la misma una muestra de 6 alumnos que llamaremos "n" (siendo n el tamaño muestral). Como n=6, tenemos 6 observaciones (o 6 datos) de nuestra variable. Que son los siguientes: (1'77, 1'65, 1'65, 1'84, 1'82, 1'75).

A partir de aquí construiremos la tabla, colocando en la misma la variable "X" con sus respectivos valores "x1, x2,…". ¡OJO!: en la tabla representamos los valores de la variable, no los datos u observaciones. En nuestro caso, tenemos 5 valores diferentes de la variable.

Definiciones:

- **Frecuencia absoluta** es el número de veces que se repite cada valor de la variable. Se representa por n_i. Siendo $\sum n_i = n$.

- **Frecuencia relativa** es la frecuencia absoluta dividida por el tamaño muestral. Se representa por f_i, siendo $\sum f_i = 1$.
  $f_i = n_i / n$

- **Porcentaje** (de la modalidad xi). Viene dada por el producto $p_i = f_i \cdot 100$.

Tipos de frecuencias para variables cuantitativas.

- **Frecuencia acumulada**
  - **Frecuencia absoluta acumulada**: indica el nº de observaciones que hemos obtenido igual al considerado o inferiores a él. Se representa por N_i.
  - **Frecuencia relativa acumulada**: acumula frecuencias relativas. Se representa por F_i.

La tabla del ejemplo anterior sería:

X = altura.	***ni***	***fi***	***Ni***	***Fi***
x1 = 1'65	n1 = 2	f1 = 2/6	2	2/6
x2 = 1'75	n2 = 1	f2 = 1/6	3	3/6
x3 = 1'77	n3 = 1	f3 = 1/6	4	4/6
x4 = 1'82	n4 = 1	f4 = 1/6	5	5/6
x5 = 1'84	n5 = 1	f5 = 1/6	6	6/6 = 1
	∑ni = n = 6	∑fi = 1		

Hemos de tener en cuenta que en algunas ocasiones será necesario construir la tabla de frecuencias agrupando los posibles valores de la variable en **intervalos**. Esto sucederá cuando nos encontremos con variables continuas.

La información presentada de esta manera se conoce como **información agrupada** o **intervalos de clase**, y presenta diferentes problemas: elegir el nº de intervalos, decidir y calcular la amplitud de cada intervalo…

Definiciones:

- **Límite inferior de clase** es el valor numérico más pequeño que puede formar parte de aquella clase definida por un intervalo. Se representa como $\mathbf{L_{i-1}}$. Análogamente se define el **límite superior de clase** por el valor más grande. Se representa por $\mathbf{L_i}$.

- **Marca de clase es** el punto medio del intervalo de clase, es decir, su centro. Se representa por $\mathbf{X_i}$. $\mathbf{X_i = (L_{i-1} + L_i)/2}$.

- **Amplitud de clase (o amplitud de intervalo)** es la diferencia entre el límite superior y el inferior. Se representa por $\mathbf{C_i}$. $\mathbf{C_i = L_i - L_{i-1}}$.

  Para cada intervalo la **frecuencia absoluta acumulada** consiste en contar las unidades que hay en la categoría y sumar las que hay en categorías inferiores. La definición es análoga para **frecuencia relativa**.

- **Densidad de frecuencia del intervalo** es la frecuencia absoluta dividida por la amplitud. Se representa por $\mathbf{d_i}$. $\mathbf{d_i = n_i / C_i}$.

Ejemplo:

Li-1 -Li	Ci	Xi	ni	fi	Ni	Fi	di
145-150	150-145 = 5	147'5	***20***	0.2	20	0.2	4
150-160	160-150 = 10	155	***40***	0.4	60	0.6	4
160-180	180-160 = 20	170	***40***	0.4	100	1	2
			100	1			

¡OJO! En el anterior ejemplo observamos que la frecuencia absoluta puede llevarnos a interpretaciones erróneas, esto es debido a que en el intervalo tenemos muchas más observaciones. La densidad de frecuencia (d_i) se encarga de recoger los posibles errores derivados de la existencia de intervalos de diferente amplitud.

EJERCICIO.

Se ha observado el precio de una habitación en 50 hoteles diferentes, y se han obtenido los siguientes datos (en pesetas):

700 700 500 1500 300 400 750 500 500 750 300 750 400
800 700 1200 500 500 10000 800 400 500 300 500 1000
300 400 500 700 500 300 400 700 400 700 500 400 700
1000 750 700 800 750 700 750 800 700 700 1200 800.

Tabular las observaciones calculando la distribución de frecuencias de los precios:

a) de manera no agrupada.

b) agrupados en 5 intervalos de igual amplitud. Valor inferior del primer intervalo=250.

Calcular frecuencias absolutas, relativas, frecuencias acumuladas, marca de clase, densidad de frecuencia…

1.6.2. Métodos gráficos para variables unidimensionales

Consiste en crear gráficos que sintetizan el comportamiento de la variable. En una representación gráfica siempre debe haber constancia en el gráfico de las variables que estamos representando. Un resumen gráfico es un método complementario al resumen numérico. Por sí solo no nos da información. Los gráficos se basarán en el sistema cartesiano. Van acompañados siempre de resúmenes numéricos. Se dividen en dos grupos:

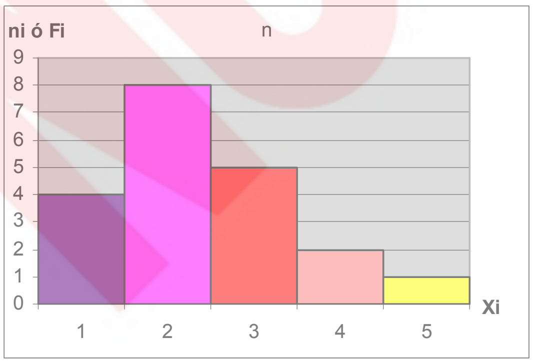
1. Cualitativos.
2. Cuantitativos.

1.6.2.1. Métodos gráficos para datos cualitativos

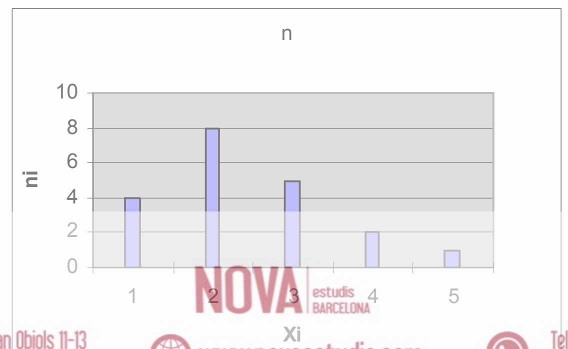
- **Diagramas de barras** (no histogramas)

Se construye colocando las distintas modalidades de la variable cualitativa sobre el eje de abscisas y sobre cada una de ellas se levanta un rectángulo de igual base y altura igual a su frecuencia (absoluta o relativa).

- **Caso vertical:**

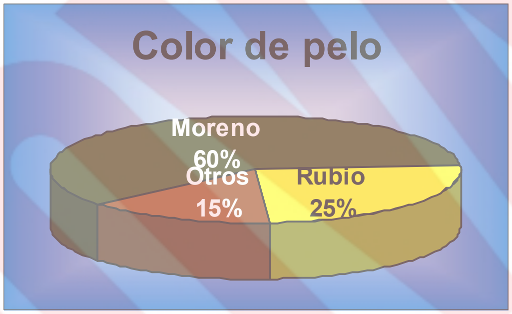


- **Caso horizontal:**



- **Diagramas de sectores**

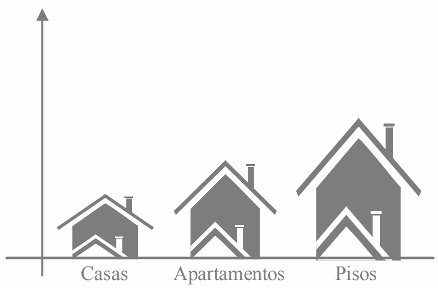
Sirve para variables cualitativas no agrupadas. Se construye repartiendo el área del círculo en sectores de tamaño proporcional a la frecuencia de cada modalidad. Hay tantos sectores como valores de la variable y los ángulos se calculan de forma proporcional a las frecuencias relativas de cada sector: $\alpha = fi \cdot 360^{o}$



- **Pictogramas**

Consiste en que se emplean figuras relacionadas con el fenómeno que se está estudiando de forma que su tamaño (tipo 1) o número (tipo 2) nos indique la frecuencia asociada a cada modalidad. Las figuras tienen un tamaño proporcional a la frecuencia en el tipo 1.

Tipo de viviendas	ni
Casas	200
Apartamentos	400
Pisos	600



El segundo tipo de pictograma sería el siguiente:

- **Cartogramas**

Cada variante identifica una variable de la modalidad que queramos representar.

1.6.2.2. Métodos gráficos para datos cuantitativos

<u>**Datos cuantitativos sin agrupar**</u>

- **Diagrama de dispersión o nube de puntos**

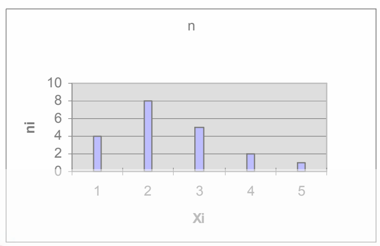
Se colocan los valores de la variable en el eje de las ordenadas y en el eje de las abcisas, la frecuencia (absoluta o relativa) asociada a cada valor.

- **Diagrama de barras** (para tablas de frecuencias no agrupadas)

Se construye igual que el anterior pero en lugar de puntos dibujamos barras. Se anotan sobre el eje de abscisas los distintos valores del carácter y sobre cada uno de ellos se levanta una barra con altura igual a su frecuencia absoluta o relativa.

Ejemplo:

x_i	n_i	f_i	N_i	F_i
1	4	0’20	4	0’2
2	8	0’40	12	0’6
3	5	0’25	17	0’85
4	2	0’10	19	0’95
5	1	0’05	20	1



Observaciones: la altura es proporcional a las frecuencias, y la base es constante (ya que corresponde a un valor puntual). Al ser la base constante, el área será proporcional a la frecuencia.

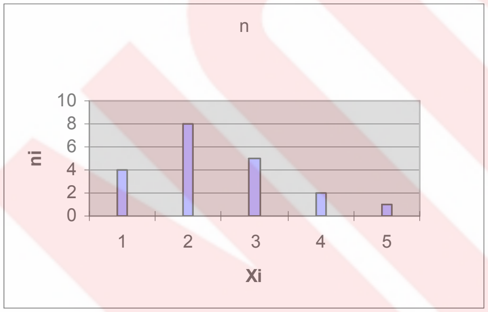
Si unimos los puntos medios de cada barra obtenemos el **polígono de frecuencias**

- **Polígono de frecuencias:**

  - **Caso no agrupado en intervalos y frecuencias no acumuladas**

En este caso se construye sobre el diagrama de barras uniendo los extremos superiores de barras consecutivas mediante una línea.

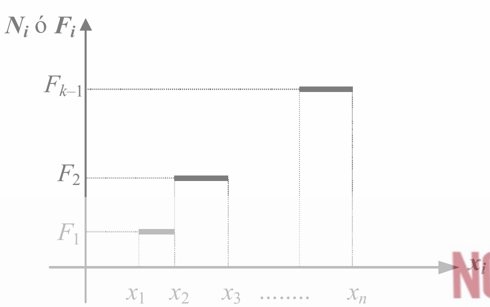
Para variables discretas



  - **Caso no agrupado en intervalos y frecuencias acumuladas**

La curva acumulativa, una vez colocados en el eje de abscisas los distintos valores de la variable y en el eje de ordenadas las frecuencias acumuladas correspondientes (absolutas y correlativas) se trazan segmentos desde el punto.

Para variables discretas



- **Box-Plot (Diagrama de caja y bigotes)**

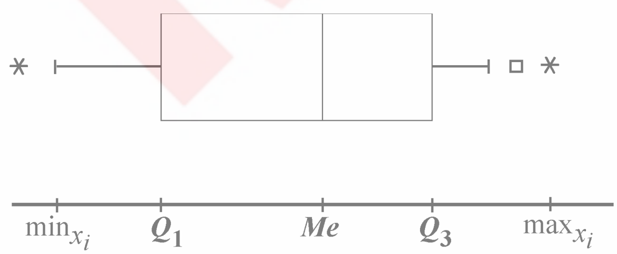
Nos permite hacer representaciones semigráficas de una distribución que nos dejan ver sus características principales y señalar los posibles datos atípicos ("outliers"). El gráfico no se ve afectado por esos datos atípicos.

Proceso de construcción del boxplot:

- se ordenan los datos de menor a mayor y se calculan los cuartiles y la mediana. La caja central queda limitada por Q1 y Q3
- se calculan los límites inferior y superior (LI y LS)
- se consideran atípicos los datos fuera del intervalo [LI, LS]
- se dibuja una línea desde cada extremo de la caja central hasta el valor más lejano no atípico
- identificamos los datos fuera de [LI, LS], señalándolos como atípicos.

Limite Superior: $LS = Q_3 + 1{,}5 \cdot IQR$

Limite Inferior: $LI = Q_1 - 1{,}5 \cdot IQR$



Los boxplots son útiles para comparar la distribución de una variable entre diferentes poblaciones.

Datos cuantitativos agrupados

- **Histograma** (para tablas de frecuencia sí agrupadas)

Se construye representando sobre las abscisas cada uno de los intervalos asociados a la variable y sobre cada uno de esos intervalos se levanta con área igual (proporcional) a la frecuencia de dicho intervalo.

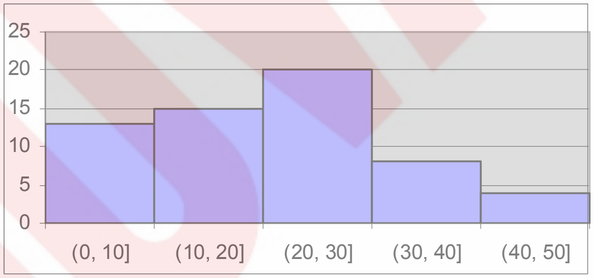
hi = altura de li en el histograma

Ar = b×alt.

Ni = ai × hi

$$h_i = \frac{n_i}{a_i} \; ; \; h_i = \frac{f_i}{a_i}$$

Salarios	n_i
(0, 10]	13
(10,20]	15
(20,30]	20
(30,40]	8
(40,50]	4



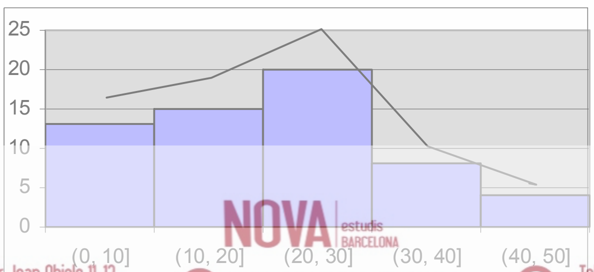
- **Polígono de frecuencias:**

  - <u>**Caso agrupado en intervalos y frecuencias no acumuladas**</u>

1º Se construye el histograma.

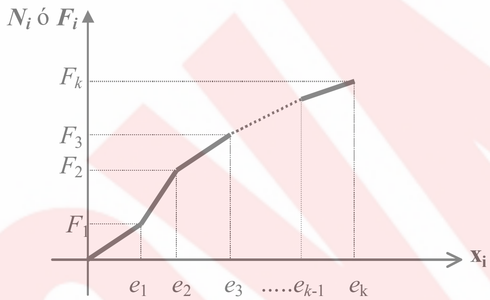
2º Sobre este se unen los puntos medios de la base superior del rectángulo y uniéndolos mediante una poligonal

Para variables contínuas:



- <u>**Caso agrupado en intervalos y frecuencias acumuladas**</u>

Se construye marcando los extremos de los intervalos en el eje de abscisas y las frecuencias acumuladas (relativas o absolutas) con la marca en el eje y.



1.6.3. Medidas de síntesis para variables unidimensionales

Son números que, al igual que las tablas o las gráficas vistas anteriormente, pretenden resumir la información recogida a la vez que ponen de manifiesto las principales características de una distribución.

Aunque la clasificación no es definitiva ni excluyente veremos:

- Medidas de posición
- Medidas de dispersión
- Momentos
- Medidas de forma
- Medidas de concentración

1.6.3.1. Medidas de síntesis para datos cuantitativos no agrupados

Llamaremos X a la variable de estudio, y n al tamaño de la muestra. Entonces, $x_1, x_2, x_3, \ldots, x_n$, serán los valores observados de la variable, es decir, los datos.

1.6.3.1.1. Medidas de posición

Hay de dos tipos, centradas y no centradas

- **Centradas**

Determinan el valor de X que ocupa la posición central, es decir, buscan el “centro” alrededor del cual se agrupan todos los datos. Intentan representar los valores de una muestra o población indicando dónde se localizan pero no cómo se localizan. Las más importantes son :

1.- MEDIA ARITMÉTICA

Se calcula sumando todos los valores observados, y dividiéndolos por el tamaño muestral.

$$\overline{X} = \frac{\sum_{i=1}^{n} Xi}{n} \qquad \text{ó} \qquad \overline{X} = \frac{\sum_{i=1}^{n} Xi \cdot n_i}{n}$$

Es la suma ponderada de todas las modalidades de la variable por sus respectuvas frecuencias relativas.

$$\overline{x} = \sum_{i=1}^{k} x_i \cdot f_i = \frac{1}{n} \sum_{i=1}^{k} x_i \cdot n_i$$

Si la tabla de frecuencias está agrupada en intervalos, se calcula con las marcas de clase.

$$\overline{x} = \frac{\sum_{i=1}^{m} n_i c_i}{n}$$

Propiedades:

- Si a todos los valores observados les sumamos una constante K (cambio de origen), la media de los nuevos valores se obtiene sumando a la media de los valores originales esta constante K.

$$X_A = \{x_1, x_2, \ldots, x_n\} \Rightarrow \overline{X}_A$$

$$X_B = \{x_1 + k, x_2 + k, \ldots, x_n + k\} \Rightarrow \overline{X}_B = \overline{X}_A + k$$

- Si todas las observaciones se multiplican por una constante K (cambio de escala), la nueva media es el producto de la anterior por la constante K.

- La media de una constante es la misma constante.
- Es centro de gravedad de la distribución $\sum_{i=1}^{n}\left(Xi - \overline{X}\right) = 0.$
- La media es la cantidad que hace mínima la suma de los cuadrados de las desviaciones respecto a un valor.
- $\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2$: esta diferencia es mínima cuando la constante es la media aritmética.

Ventajas e inconvenientes de la media aritmética sobre otras medidas de síntesis

- **Ventajas**: siempre es calculable, es única (para una distribución dada, siempre toma el mismo valor), y tiene en cuenta a todos los valores de la distribución.

- **Inconvenientes**: al tener en cuenta todos los valores, es sensible a valores extremos. Estos valores pueden distorsionar el sentido de la media aritmética, es poco robusta a errores.

2.- MEDIANA

Es el valor que quedaría en medio de la distribución si ordenásemos de menor a mayor todas las observaciones. La mediana deja un 50% de las observaciones a cada lado.

- **Si n es par** la mediana se calcula como la media aritmética de los dos valores centrales. La media aritmética de los valores n/2 y (n/2)+1

$$Me = \frac{x\left(\frac{n}{2}\right) + x\left(\frac{n}{2}+1\right)}{2}$$

Ejemplo: $X = \{1,2,5,7,9,10,13,14\}$

n=8 (par) $\Rightarrow$ n/2=4 (número que ocupa la 4ª posición, en nuestro ej. es el 7).
(n/2) + 1=5 (número que ocupa la 5ª posición, en nuestro ej. es el 9).

Me = (7+9)/2 = 8.

- **Si n es impar** la mediana es la posición (n+1)/2.

$$Me = x\left(\frac{n+1}{2}\right)$$

Ejemplo: $X = \{1,2,5,7,9,10,13\}$
n = 7 (impar) $\Rightarrow$ (n+1)/2 = 4 (valor que ocupa la 4ª posición, en nuestro ej. es el 7).
Me = 7.

Ventajas e inconvenientes de la mediana

- **Ventajas**: divide la distribución en dos partes iguales y no es sensible a valores extremos (esto es una ventaja respecto la media aritmética).

- **Inconvenientes**: el cálculo de la mediana es más complejo que el de la media ya que previamente hemos de hacer una ordenación de los datos, y no tiene en cuenta todos los datos.

3.- MODA

La moda es el valor que más veces se repite en la distribución. La moda de una variable estadística es el valor/es que tiene/n asociada la frecuencia máxima. Podría no considerarse una medida de posición ya que no nos da información sobre el centro de la distribución.

Puede suceder que en los datos no se repita ningún valor, en este caso podemos decir que no hay moda. Por otra parte, si hay dos valores que se repiten el mismo número de veces, hablamos de datos bimodales (o datos multimodales si se repiten un mayor número de veces).

¡¡ ATENCIÓN !!

Media y mediana coincidirán únicamente en el caso de que la distribución sea simétrica. (La moda no tiene por qué coincidir con ellas)
Si la distribución es simétrica y unimodal, media, mediana y moda coincidirán

4.- RANGO MEDIO

Es la media entre el mayor valor y el menor valor de la distribución.

$$RM = \frac{Xn + Xo}{2}$$

EJEMPLO.
Dada la siguiente distribución de frecuencias:

Xi	ni (frec. Absoluta)	Ni (frec. Acumulada)
1	**3**	**3**
2	**4**	**7**
5	**9**	**16**
7	**17**	**33**
	$\sum$ = 33	

- Media: $\overline{X} = \frac{1 \cdot 3 + 2 \cdot 4 + 5 \cdot 9 + 7 \cdot 17}{33} = 5.30.$
- Mediana: n = 33 (impar) $\Rightarrow$ (n+1)/2 = 17. Vemos en la columna de frecuencia acumulada que el valor que ocupa la posición 17 es el 7. Me = 7.
- Moda: Mo = 7, ya que el 7 es el valor que más veces se repite.
- Rango Medio: RM = (7+1)/2 = 4.

5.- RELACIÓN ENTRE MEDIA, MEDIANA Y MODA

Bajo distribuciones unimodales:

$$\bar{x} - Mo \approx 3(\bar{x} - Me)$$

El caso es que cumplen una simetría. También se puede dar el caso en que las 3 sean iguales.

6.- OTRAS MEDIDAS DE TENDENCIA CENTRAL

- **Media geométrica**

Dadas x1,…, xk (>0)

$$G = \sqrt[n]{x_1^{n_1} x_2^{n_2} \ldots x_k^{n_k}}$$

- **Media armónica**

Dadas x1,…, xk (>0)

$$H = \frac{n}{\sum_{i=1}^{k} \frac{1}{x_i} n_i} \quad ; \quad \frac{1}{H} = \frac{\sum_{i=1}^{n} \frac{1}{x_i} n_i}{n}$$

- **Media cuadrática**

$$Q = \sqrt{\frac{\sum_{i=1}^{k} x_i^2 n_i}{n}}$$

<u>Propiedad</u>:

Siempre se cumple que $H \leq G \leq \bar{x} \leq Q$

- **No centradas**

1.- CUANTILES, CUARTILES, DECILES Y PERCENTILES

- **Cuantil de nivel** α ($0 \leq \alpha \leq 1$)

$$X_\alpha / F_{x_\alpha} = \alpha$$

- **Cuartiles:** Q1 = X0'25 , Q2 = X0'5 = Me , Q3 = X0'75

- **Deciles:** D1 = X0'1 , D2 = X0'2 , … , D9 = X0'9

- **Percentiles:** P1 = X0'01 , … , P99 = X0'99

- **Octiles:** O1 = X0'125 , … , O7 = X0'875

Estas medidas se pueden definir en dos sentidos diferentes:

1) Supongamos que se quiere calcular cual es el percentil del valor X. Esto quiere decir calcular qué tanto por ciento de observaciones de nuestros datos tienen un valor inferior a X respecto al total de las observaciones. Indicará, por tanto, una posición relativa de los datos en una ordenación ascendente de las observaciones. El percentil de una observación X se calcula mediante la siguiente expresión:

$$\frac{(\text{número de observaciones inferiores a x}).\ 100}{n}$$

Ejemplo:

Xi	ni	Ni
1	10	10
2	15	25
5	7	32
8	10	42
9	20	62
20	40	102
	n = 102	

P(5) = ?

Nº de observaciones inferiores a 5 = 32.

P (5) = (32 . 100)/102= 31.37

P (9)=?

Nº de observaciones inferiores a 9 = 62 $\Rightarrow$ P(9)= (62·100)/102= 60.78.

2) Podemos buscar, por ejemplo, el percentil 30 si queremos buscar aquel valor X (puede ser que no haya sido observado), tal que el 40% de las observaciones serían inferiores a él.

El percentil k-ésimo se calcula de la siguiente manera:

$$P_{k\%} = \frac{n}{100} \cdot k$$

Ej. El percentil 40% del ejemplo anterior sería:

$$P_{40\%} = \frac{102}{100} \cdot 40 = 40.8$$

Aproximamos a 41. El 41 nos indica la posición que ocupa el valor que buscamos. Si observamos en la tabla la columna de frecuencias acumuladas, vemos que el valor que ocupa esta posición es el 8. <u>El 8 deja un 40% de las observaciones por debajo</u>.

Los cuartiles Q_1, Q_2, Q_3 son los percentiles 25, 50 y 75 respectivamente. Los deciles $D_1, D_2, \ldots\ldots, D_9$ se corresponden con los percentiles 10, 20, 30, 40, 50, 60, 70, 80 y 90.

<u>Observaciones</u>:

P (Me) = 50%.

$$P_{50\%} = Me = Q_2 = D_5$$

1.6.3.1.2. Medidas de dispersión o variabilidad

Las medidas de tendencia central no siempre son fiables, a veces son un poco engañosas, por eso recurrimos a las medidas de dispersión. Estas intentan medir hasta las medidas de tendencia central para ver cómo son de representativas las medidas.

Las hay de dos tipos, absolutas o relativas.

Supongamos que tenemos los dos conjuntos de datos siguiente:

DATOS A: 65,66,67,68,71,73.74.77.77.77.

DATOS B: 42,54,58,62,67,77,77,85,93,100.

Si se calculan las medidas anteriormente estudiadas de tendencia central, el resultado es:

	Datos A	Datos B
Media:	71.5	71.5
Mediana:	72	72
Moda:	77	77
Rango medio:	71	71

Según las medidas de tendencia central, las distribuciones anteriores son prácticamente iguales, sin embargo, si dibujamos un histograma, observamos que tienen un comportamiento muy diferente: el conjunto A está mucho más próximo a su centro, mientras que el conjunto B está mucho más disperso.

Las medidas de dispersión sirven para conocer si la distribución está próxima a su centro, o si los valores se hallan dispersos, es decir, alejados de su centro.

Medidas de dispersión absoluta

1.- RECORRIDO

Es la diferencia entre el valor más grande y el más pequeño de x.

$$RE = X_{máx} - X_{min}$$

2.- VARIANZA

Es la media aritmética de la suma de los cuadrados de las desviaciones respecto a la media.

$$V(x) = S^2 = \frac{\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2}{n}$$

Fórmula alternativa: $V(x) = S^2 = \frac{\sum_{i=1}^{n} Xi^2}{n} - \overline{X}^2$

Propiedades:

- La varianza no puede ser negativa **V(x) ≥ 0**.
- Es la medida de dispersión óptima, ya que según el teorema de Köning, la media es el valor que minimiza las desviaciones de valores al cuadrado.
- Si sumamos una constante a todas las observaciones (cambio de origen), la varianza no varía.
- Si multiplicamos todos los valores por una constante (cambio de escala), la varianza queda multiplicada por el cuadrado de dicha constante.
- La varianza de una constante es 0.
- El principal inconveniente de la varianza es que al ser una medida de dispersión elevada al cuadrado, hace que perdamos parte de su significado económico. Por ello utilizaremos otra medida de dispersión: la desviación estándar.

3.- VARIANZA MUESTRAL O CORREGIDA

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

4.- DESVIACIÓN ESTÁNDAR

Es la raíz cuadrada positiva de la varianza.

$$S = +\sqrt{\frac{\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^2}{n}}$$

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

Propiedades:

- Siempre es positiva.
- Es una medida de dispersión óptima.
- La desviación estándar de una constante es 0.
- Si sumamos a todos los valores de la variable una constante, la desviación estándar no varía.
- Si multiplicamos todos los valores de la variable por una constante, la desviación estándar queda multiplicada también por la constante.

5.- DESVIACIÓN ESTÁNDAR

$$\sigma_{n-1} = \sqrt{\sigma_{n-1}^2} = \sqrt{\frac{n}{n-1}}\sigma_n$$

6.- RANGO INTERCUARTÍLICO

Es la diferencia entre el tercer cuartil y el primero.

RI = Q_3-Q_1

Medidas de dispersión relativas

Éstas relacionan las medidas de tendencia central con las medidas de dispersión. Ponderan la dispersión respecto la magnitud global de la variable, y de esta manera **permiten hacer comparaciones**.

1.- COEFICIENTE DE APERTURA

Es el cociente entre el valor máximo y el mínimo de la variable.

$$A = \frac{Max\ X_i}{Min\ X_i}$$

2.- COEFICIENTE DE VARIACIÓN DE PEARSON

Expresa en porcentaje la magnitud de la desviación estándar en función de la media aritmética.

$$CV = \frac{S}{\overline{X}}$$

Cuanto más grande sea el coeficiente de variación, más dispersados estarán los datos. Cuanto más pequeño sea, más concentrados estarán los datos alrededor de la media.

El coeficiente de variación no siempre es positivo, ya que la media puede ser negativa.

Cuando la media es un valor cercano a 0 no debe emplearse Pearson.

1.6.3.1.3. Momentos

Los hay de dos tipos, respecto al origen y respecto a la media.

Momentos respecto al origen

El momento respecto al origen de orden r es:

$$a_r = \frac{\sum x_i^r}{n}$$

- Momentos ordinarios o centrados en el origen

$$a_r = E(x^r) = \frac{\sum x_i^r}{n}$$

Distribución discreta $a_r = \sum x_i^r \, p_i$

Distribución continua $a_r = \int_{-\infty}^{\infty} x^r \, f(x) \, dx$

Casos concretos:

$a_0 = 1 \qquad a_1 = \mu = E(x)$

- Momentos centrales o centrados en la media

$$m_r = E[(x-\mu)^r] = \frac{\sum (x_i - \bar{x})^r}{n}$$

Distribución discreta $m_r = \sum (x_i - \mu)^r \, p_i$

Distribución continua $m_r = \int_{-\infty}^{\infty} (x - \mu)^r \, f(x) \, dx$

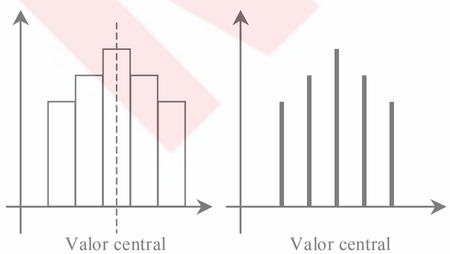
Casos concretos:

$m_0 = 1 \qquad m_2 = V(x)$

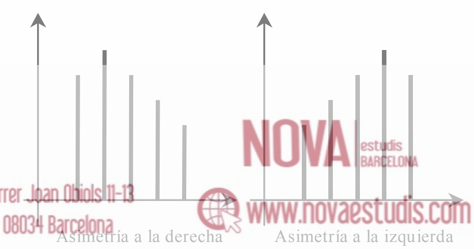
1.6.3.1.4. Medidas de forma: asimetría y curtosis

Medidas de asimetría

Diremos que una distribución de frecuencias es simétrica cuando los valores de la variable equidistantes de un valor central tienen las mismas frecuencias



Se dirá asimétrica a la izquierda o a la derecha según presenten la cola (las frecuencias descienden más por la derecha o por la izquierda) por la derecha o por la izquierda.



1.-ÍNDICE DE ASIMETRÍA DE PEARSON (Ap)

Sirve para calcular si los datos están distribuidos de forma simétrica respecto su "centro" , o por el contrario, se concentran en un lado.

Cuanto más grande es este índice mayor es la asimetría.

$$Ap = \frac{\left(\overline{X} - Mo\right)}{S} \qquad Ap = \frac{3 \cdot \left(\overline{X} - Me\right)}{S}$$

- Si **Ap >0, la <u>distribución</u> es <u>asimétrica por la derecha o asimétrica positiva</u>** (hay más datos a la izquierda de la media).

- Si **Ap <0 la <u>distribución</u> es <u>asimétrica por la izquierda o asimétrica negativa</u>,** hay más datos a la derecha de la media.

- Si **Ap = 0, la <u>distribución</u> es <u>simétrica</u>,** en este caso media y mediana coincidirán.

La asimetría se considera significativa si el índice en valor absoluto es superior a 1

2.- MOMENTOS.

El <u>momento de tercer orden</u>, mantiene el signo (a diferencia del momento de segundo orden) y a la vez nos permite saber si la asimetría es por la derecha o por la izquierda.

$$m_3 = \frac{\sum_{i=1}^{n} \left(Xi - \overline{X}\right)^3}{n}$$

- $m_3 > 0 \Rightarrow$ distribución asimétrica por la derecha o asimétrica positiva
- $m_3 < 0 \Rightarrow$ la distribución asimétrica por la izquierda o asimétrica negativa,
- $m_3 = 0 \Rightarrow$ la distribución es simétrica.

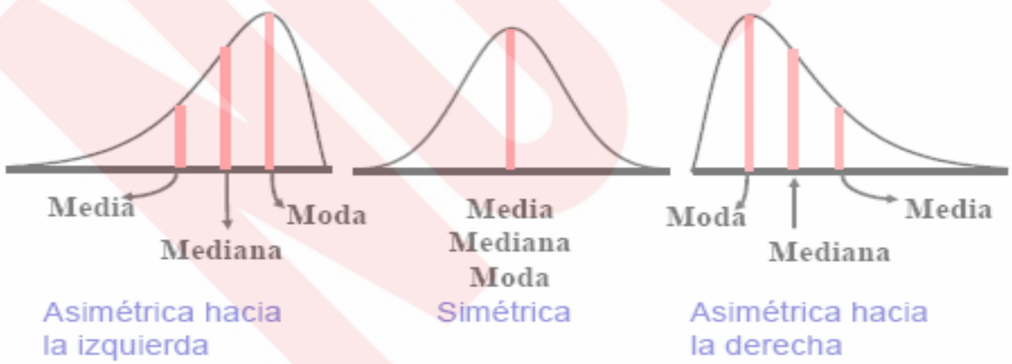
3.- ÍNDICE DE FISHER (g_1)

Convierte el momento de tercer orden en un índice sin medidas.

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

Se interpreta de igual forma que el momento de tercer orden, es decir, de forma inversa al índice de Pearson.

ASIMETRICA NEGATIVA ASIMETRICA POR LA IZQUIERDA	SIMETRICA	ASIMETRICA POSITIVA ASIMETRICA POR LA DERECHA

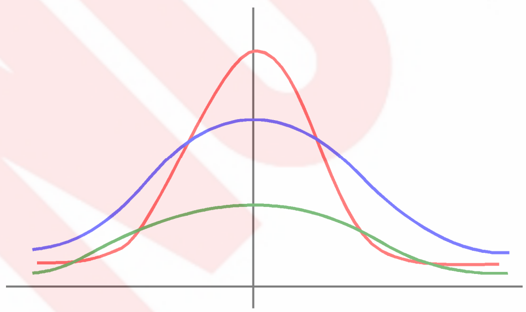


Mo > x   x = Mo = Me   Mo<x

4.- COEFICIENTE DE CURTOSIS g_2

$$g_2 = \frac{m_4}{S^4} - 3 = \frac{\dfrac{\sum_{i=1}^{n}\left(Xi - \overline{X}\right)^4}{n}}{S^4} - 3$$

- $g_2 > 0 \Rightarrow$ distribución leptocúrtica (mucha curtosis).
- $g_2 < 0 \Rightarrow$ distribución platicúrtica
- $g_2 = 0 \Rightarrow$ distribución mesocúrtica.



- En el caso de que sea menos puntiaguda la vamos a denominar **platicúrtica**.
- Si es igual que la campana de Gauss la denominaremos **mesocúrtica**.
- Si es más puntiaguda la llamaremos **leptocúrtica**.

EJEMPLO:

Xi	*ni*	*Ni*
0	2	2
10	4	6
20	7	13
30	5	18
40	2	20
	n=20	

1.- Media: $\overline{X} = \frac{0 \cdot 2 + 10 \cdot 4 + 20 \cdot 7 + 30 \cdot 5 + 40 \cdot 2}{20} = \frac{410}{20} = 20.5$

2.- Mediana: n=20 (par) $\Rightarrow$ N/2 = 10. El nº que ocupa la posición 10 es el 20.

$\Rightarrow$ (N/2)+1 = 11. El valor que ocupa la posición 11 es el 20.

Me= (20+20)/2 = 20.

3.- Moda: el valor que más veces se repite en la distribución es el 20. Mo=20.

4.- Rango medio: RM= (40+0)/2= 20.

5.- Recorrido: Re= 40 -0 =40.

6.- Varianza: $S^2 = \frac{0^2 \cdot 2 + 10^2 \cdot 4 + 20^2 \cdot 7 + 30^2 \cdot 5 + 40^2 \cdot 2}{20} - (20.5)^2 = 124.75$

7.- Desviación estándar: $S = +\sqrt{S^2} = +\sqrt{124.75} = 1.169.$

8.- Coeficiente de variación: $CV = \frac{S}{\overline{X}} = \frac{11.169}{20.5} = 05448$

9.- Índice de Pearson: $A_p = \frac{3 \cdot (20.5 - 20)}{11.169} = 0.13$

La distribución presenta cierta asimetría a la derecha, aunque ésta es prácticamente despreciable, ya que es inferior a la unidad.

10.- Momento de 3er orden:

$$m_3 = \frac{(0-20.5)^3 \cdot 2 + (10-20.5)^3 \cdot 4 + (20-20.5)^3 \cdot 7 + (30-20.5)^3 \cdot 5 + (40-20.5)^3 \cdot 2}{20} = -137.25$$

11.- Índice de Fisher: $g_1 = \frac{m_3}{S^3} = -0.098$: asimetría por la derecha prácticamente despreciable.

12.- Coeficiente de curtosis: $g_2 = \frac{37587.3}{(11.64)^4} - 3 = -0.95$. Distribución platicúrtica.

13.- Percentiles, deciles y cuartiles.

Cuartiles. Dividen la distribución en 4 partes iguales: 20/4 = 5 observaciones en cada cuartil.

$$Q_1 = 10$$
$$Q_2 = 20$$
$$Q_3 = 30$$

Calculamos los cuartiles mirando la columna de frecuencias acumuladas. El valor que acumula las 5 primeras observaciones es el 10, el valor que acumula las 10 primeras observaciones es el 20…

$P_{50} = \frac{20}{100} \cdot k = 0.2 \cdot 50 = 10$, el valor que ocupa la posición 10 es el 20, por tanto, $P_{50} = 20$, vemos que coincide con el segundo cuartil.

El cálculo de los deciles será análogo.

1.7 Datos bidimensionales

2.2.1.Tabulación

En este caso observamos simultáneamente dos variables, datos bidimensionales. Ej. peso-altura

$$(x,y) = \{ (175, 70), (175, 55) \dots \}$$

La distribución de frecuencias es la correspondencia de cada par de valores con su frecuencia conjunta. La frecuencia conjunta es la frecuencia asociada al par de valores, es decir, la frecuencia en la que el valor i-ésimo de x y el valor j-ésimo de y se presentan simultáneamente.

Para tabular los datos utilizaremos las **tablas de doble entrada**, en un lado está el valor de **x** y en el otro el valor de **y**, dentro de la tabla están las frecuencias conjuntas.

1.7.1.1. Tabla de frecuencias absolutas

x\y	Y_1	…	Y_j	…	Y_k
x_1	n_{11}		n_{1j}		n_{1k}
…					
x_i	n_{i1}		n_{ij}		n_{ik}
…					
x_h	n_{h1}		n_{hj}		n_{hk}

Alguna combinación puede tener valor cero.

El sumatorio de todas las $\Sigma\Sigma\ n_{ii} = N$. Para calcular la tabla de frecuencias relativas, sólo tenemos que dividir cada n_{ii} por N.

Si nos interesa estudiar aislada alguna de las características de la variable bidimensional, nos remitiremos a las **DISTRIBUCIONES MARGINALES.** Estas son distribuciones unidimensionales que se originan al analizar una de las características de una población bivariante aislada, ignorando la otra.

1.7.1.2. Distribuciones marginales de frecuencias absolutas

Por ejemplo supongamos la siguiente tabla de frecuencias absolutas.

X\Y	60	65	70	75	X
160	3	4	1	0	8
162	4	5	4	2	15
163	6	5	3	3	17
Y	13	14	8	5	N=40

1.7.1.2.1. Distribución marginal de y

Y	60	65	70	75	
$n_{.j}$	13	14	8	5	N=40

El símbolo n.j quiere decir que sólo tenemos en cuenta la segunda componente de los datos (x,y), que están en las columnas j y por eso en vez de la i ponemos un punto, porque no tenemos en cuenta las filas i.

1.7.1.2.2. Distribución marginal de x

X	160	162	163		
$n_{i.}$	8	15	17	N=40	

En este caso sólo tenemos en cuenta la primera parte de los datos, las x, o sea las filas i y no las columnas j.

Para hacer la tabla de frecuencias relativas sólo tenemos que dividir nij por N.

1.7.1.3. Distribuciones marginales de frecuencias relativas

X\Y	60	65	70	75	X
160	0.075	0.100	0.025	0.000	0.200
162	0.100	0.125	0.100	0.050	0.375
163	0.150	0.125	0.075	0.075	0.425
Y	0.325	0.350	0.200	0.125	Σ = 1

1.7.1.3.1. Distribución marginal de frecuencias relativas de y

Y	60	65	70	75	
$f_{.j}$	0.325	0.350	0.200	0.125	$\Sigma = 1$

1.7.1.3.2. Distribución marginal de frecuencias relativas de x

X	160	162	163		
$f_{i.}$	0.200	0.375	0.425	$\Sigma = 1$	

1.7.1.4. Distribuciones condicionadas

Son distribuciones unidimensionales, se observa una única característica. Se fija una condición previa para el valor de la otra variable. Condicionamos el valor de una de las dimensiones a un valor concreto de la otra dimensión.

Por ejemplo, el valor de la variable x cuando y=70.

X|Y=70	$n_i\|y=70$	$f_i\|y=70$
160	1	0.025
162	4	0.100
163	3	0.075
Σ	8	0.200

Cuando condicionamos el valor de x, estamos restringiendo la población a los datos que cumplen la condición, y por eso el sumatorio de las frecuencias ha de ser uno. Por eso multiplicamos por $1/\Sigma$ las frecuencias. En nuestro caso 1/0.2 = 5.

X\Y=70	$f_i\|y=70$
160	0.125
162	0.5
163	0.375
Σ	1

En general,la frecuencia de x condicionada a y = yj se calcula:

$$(f_i|y=y_j)=\frac{n_i \mid y_j}{n_{\cdot j}} \qquad (f_j|y=y_i)=\frac{n_j \mid x_i}{n_{i\cdot}}$$

Dividimos cada frecuencia por la suma de toda la fila o columna.

1.7.2. Independencia estadística de dos variables

Dos variables son independientes estadísticamente cuando la frecuencia relativa conjunta es igual al producto de las frecuencias relativas marginales.

$$\frac{n_{ij}}{N}=\frac{n_{i\cdot}}{N}\cdot\frac{n_{\cdot j}}{N} \quad \forall \text{ i,j.}$$

$$f_{ij}=f_{i\cdot}\, f_{\cdot j} \qquad \forall \text{ i,j.}$$

La ventaja de la independencia estadística es que la frecuencia condicionada es igual a la frecuencia marginal relativa.

$$f_i|y=y_j= f_{\cdot i}$$

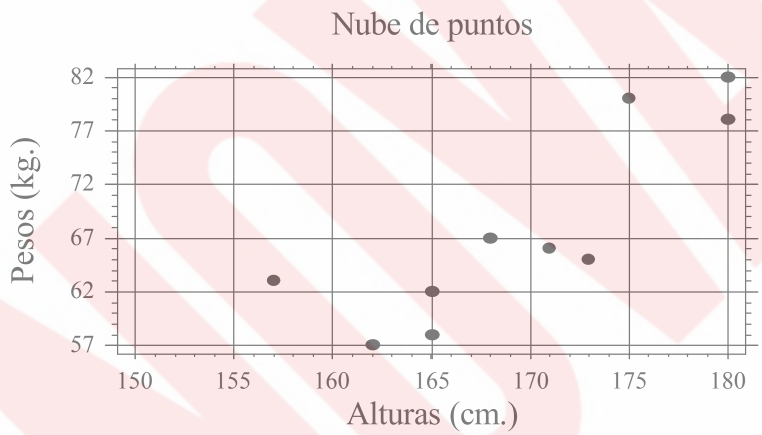
Cuando las variables sean independientes estadísticamente, podremos pasar de datos unidimensionales, de las frecuencias marginales, a las tablas de datos bidimensionales.

X/Y	Y_1	...	y_i	...	y_n	X
x_1	$f_{11}= f_{\cdot 1} f_{1\cdot}$					$f_{1\cdot}$
...						...
x_i			$f_{ii} = f_{\cdot i}\, f_{i\cdot}$		$f_{in}= f_{\cdot n}\, f_{i\cdot}$	$f_{i\cdot}$
...						...
x_n					$f_{nn} = f_{\cdot n}\, f_{n\cdot}$	$f_{n\cdot}$
Y	$f_{\cdot 1}$	...	$f_{\cdot i}$	...	$f_{\cdot n}$	1

1.7.3. Representación gráfica

Las representaciones gráficas se hacen en diagramas de puntos de dos dimensiones, llamados nubes de puntos o en diagramas de dos dimensiones de manera análoga a las variables de una dimensión.

Representamos en ejes coordenados, una de las dos variables en el eje X, y la otra en el eje Y. Para indicar el número de coincidencias, o bien ponemos símbolos diferentes, o bien indicamos entre paréntesis, el número n_{ij}.



1.7.4. Medidas de Síntesis. Covarianza

Son muy parecidas a las ya vistas para variables unidimensionales.

- <u>Media o centro de gravedad</u>

Es la media de cada una de las variables, (x,y).

- <u>Varianzas y covarianzas</u>

Además de S_x y S_y que son las mismas que para variables unidimensionales definiremos Sxy que es la covarianza de los puntos (x,y) y la definimos como:

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

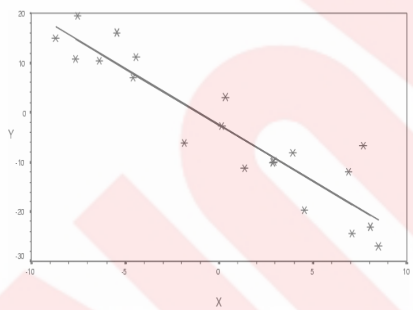
$S_{xy} = S_{yx}$ y puede ser cero , negativo o positivo .

Si la covarianza es diferente de cero existe alguna correlación lineal entre las dos variables.

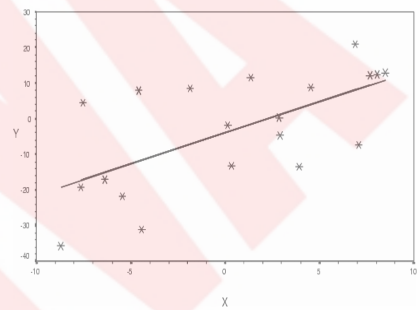
Observaciones:

- Si $S_{xy}>0$ hay dependencia directa (positiva), es decir a grandes valores de x corresponden grandes valores de y.
- Si $S_{xy} = 0$ las variables están incorreladas, es decir no hay relación lineal.
- Si $S_{xy}<0$ hay dependencia inversa o negativa, es decir a grandes valores de x corresponden grandes valores de y.

$S_{xy} < 0$



$S_{xy} > 0$



Cuando las variables x e y son independientes, $S_{xy}=0$, y por tanto $r_{xy}=0$. Es decir, si dos variables son independientes su covarianza vale cero. No podemos asegurar lo mismo en sentido contrario. Si dos variables tienen covarianza cero, no podemos decir que son independientes. Sabemos que linealmente no tienen relación, pero podrían tener otro tipo de relación y no ser independientes.

<u>Propiedades de la covarianza:</u>

- Si a todos los valores de la variable x, les sumamos una constante k y a todos los valores de la variable y les sumamos una constante k', la covarianza no varía.
- Si a todos los valores de una variable x los multiplicamos por una constante k y a todos los valores de la variable y los multiplicamos por una constante k', su covarianza queda multiplicada por el producto de las constantes.
- A partir de las anteriores: si tenemos dos variables x, y con la covarianza S_{xy}, y transformaciones lineales de las variables de la forma z=ax+b, y t=cy+d, la nueva covarianza se relaciona con la anterior de la forma: $S_{zt}=acSxy$.

La covarianza no es directamente interpretable, en el sentido de que no es directamente comparable su tamaño, depende del tamaño de la variable, para poder comparar diferentes variables se emplea la correlación que viene definida por el coeficiente de correlación lineal de Pearson.

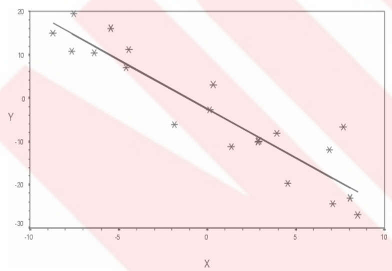
1.7.5. Coeficiente de correlación lineal de Pearson

Con este parámetro podemos comparar la dependencia lineal de diferentes pares de variables.

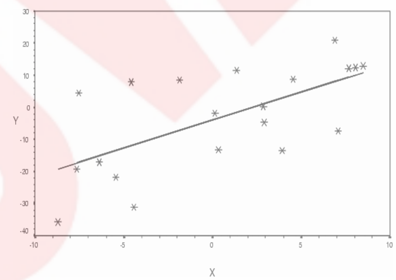
$$r_{xy} = \frac{S_{xy}}{S_x S_y}$$

- Los valores de r_{xy} varían entre –1 y 1.
- Cuando el valor absoluto es 1 la correlación lineal es perfecta.
- Cuando vale cero no hay correlación lineal.
- El signo determina si la correlación es positiva o negativa.

$-1 < r_{xy} < 0$



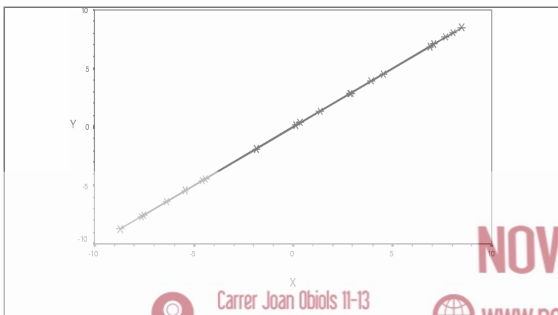
$0 < r_{xy} < 1$



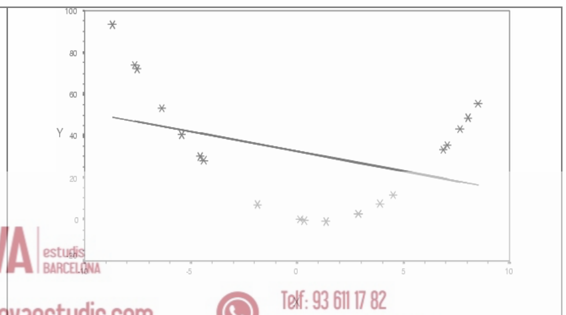
Interpretación:

- $r_{xy} = 1$: dependencia funcional lineal directa o positiva, ya que tiene el mismo signo que b.
- $r_{xy} = –1$: dependencia funcional lineal inversa o negativa.
- $r_{xy} = 0$: no existe dependencia lineal ninguna.

$r_{xy} = 1$



$r_{xy} = 0$



1.7.6. Recta de regresión

Si en el diagrama bivariante hemos visto una relación lineal calcularemos la recta de regresión. La recta de regresión es una recta que pasa entre las observaciones colocadas en los ejes de coordenadas X e Y. Es la recta que se puede construir de modo que los puntos estén a distancia mínima de esta. Sólo existe una y se construye de la siguiente forma:

- Recta de regresión de Y sobre X

  $y = bx + a$ con $b = \dfrac{S_{xy}}{S_x^2}$ y $a = \overline{y} - b\overline{x}$

- Recta de regresión de X sobre Y

  $x = by + a$ con $b = \dfrac{S_{xy}}{S_y^2}$ y $a = \overline{x} - b\overline{y}$

La recta de regresión es la recta que mejor aproxima la nube de puntos y a b se le denomina coeficiente de regresión (es la pendiente de la recta de regresión).

La recta de regresión pasa por el punto
$$\begin{cases} (\overline{x}, \overline{y}), \text{ recta de regresión de Y sobre X} \\ (\overline{y}, \overline{x}), \text{ recta de regresión de X sobre Y} \end{cases}$$

1.7.7. Bondad de ajuste. Coeficiente de determinación

Se trata de medir lo bien que se ajusta la nube de puntos a la recta. El coeficiente de determinación explica la cantidad de variación de un dato que se puede explicar con el otro dato, muy cercano a uno tenemos una relación muy fuerte si esta cerca de 0 tenemos que no hay prácticamente relación. El coeficiente de determinación explica la variación total de los valores de Y que está explicada por una relación lineal con los valores de X.

$$R^2 = \frac{\sum(\hat{y} - \overline{y})^2}{\sum(y_i - \overline{y})^2} = 1 - \frac{\sum e_i^2}{\sum y_i^2 - n\overline{y}^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} \qquad 0 \leq R^2 \leq 1$$

La proporción de variabilidad que explica los valores ajustados de la variable Y.

Interpretación:

- $R^2 = 1$ dependencia funcional lineal.
- $R^2 = 0$ no existe dependencia funcional.
- $R^2 \to 1$ más cerca de la dependencia lineal estamos ($\geq 0'75$)