

GUÍA BÁSICA PARA LA REALIZACIÓN DE TRATAMIENTO ESTADÍSTICO DE DATOS.

Las variables estadísticas pueden estudiarse de forma conjunta cuando se desea ver la relación entre dos o más caracteres del mismo individuo. Por ejemplo, se puede estudiar la relación que hay entre la estatura y el peso de unos jugadores de un equipo de baloncesto, o la relación que hay entre la cantidad de un medicamento y el tiempo que tarda en hacer reacción, etcétera.

ESTADÍSTICA DESCRIPTIVA

Variable estadística: es el conjunto de los distintos valores numéricos que adopta un carácter cuantitativo.

Las variables suelen ser de dos tipos:

- **cualitativas** o categóricas, no se pueden medir numéricamente (ejemplo: nacionalidad)
- **cuantitativas**, tienen un valor numérico, (ejemplo: edad, precio de un producto).

Las variables se pueden clasificar también en:

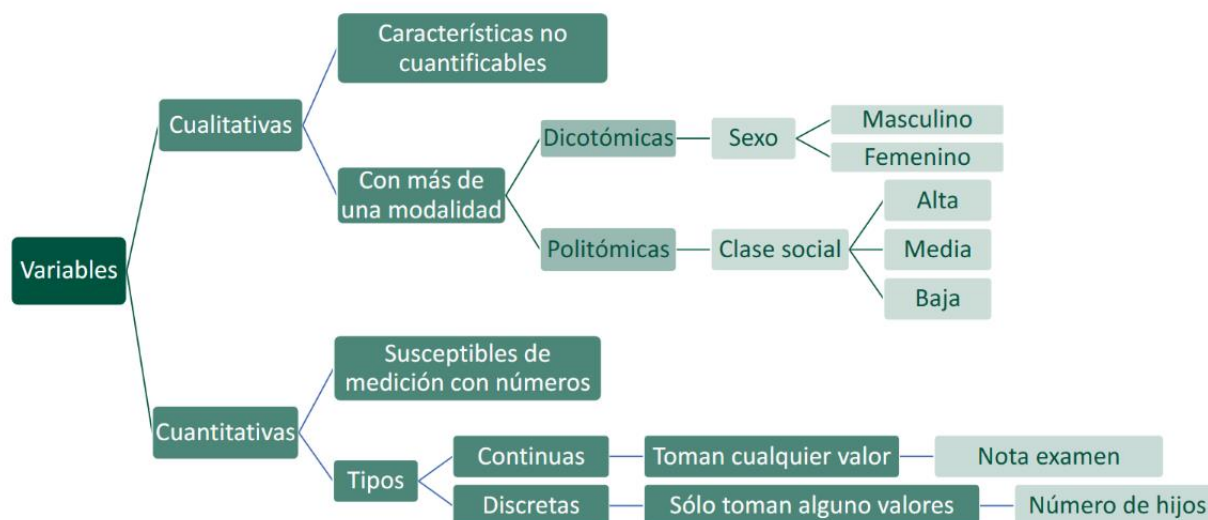
- Variables unidimensionales: sólo recogen información sobre una característica
- Variables bidimensionales: recogen información sobre dos características. Ejemplo: edad y altura de un mismo individuo)
- Variables pluridimensionales: recogen información sobre tres o más características.

Las **variables cuantitativas** se pueden clasificar en discretas y continuas:

- **Discretas:** solo pueden tomar valores enteros. (Ejemplo: número de hermanos)
- **Continuas:** pueden tomar cualquier valor dentro de un intervalo. Ejemplo: velocidad de un vehículo.

Algunas veces, cuando el número de datos recogido es muy grande, las variables discretas se tratan como si fuesen continuas.

Variable: es una característica que admite, al menos, dos valores.



Cuando se estudia el comportamiento de una variable hay que distinguir los siguientes conceptos:

- **Individuo:** cualquier elemento que nos de información sobre el carácter en estudio
- **Población:** conjunto de todos los individuos que nos dan información sobre el fenómeno en estudio.
- **Muestra:** subconjunto seleccionado de una población.

ORGANIZACIÓN DE DATOS

➤ Tabla de frecuencias

Los datos se recogen en tablas con más o menos columnas, según el estudio a realizar.

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	

X valores de la variable.

f frecuencia absoluta (número de veces que aparece un dato).

F frecuencia absoluta acumulada (suma de las frecuencias).

N tamaño de la muestra o población (número de datos recogidos)

x	f	f_r	%
1	6	0.0400	4.00
2	11	0.0733	7.33
3	12	0.0800	8.00
4	30	0.2000	20.00
5	40	0.2667	26.67
6	25	0.1667	16.67
7	14	0.0933	9.33
8	9	0.0600	6.00
9	3	0.0200	2.00
Total	150	1.0000	100.00

f_r frecuencias relativas (f/N).

Suman 1

% frecuencias porcentuales
(frecuencias relativas por 100).

Suman 100

Una **distribución de frecuencias** de datos cualitativos es una lista de los diferentes valores posibles y sus frecuencias

Deportes favoritos en una clase

Fútbol	Fútbol	Atletismo
Baloncesto		Baloncesto
	Fútbol	
Fútbol	Natación	
Baloncesto		Fútbol

Distribución de frecuencias

Deporte	Frecuencia	Frecuencia relativa	Frecuencia acumulada
Atletismo	1	0.1	0.1
Baloncesto	3	0.3	0.4
Fútbol	5	0.5	0.9
Natación	1	0.1	1.0
TOTAL (N)	10		

$$f_{relativa} = \frac{f_{absoluta}}{N}$$

Número de casos observados puede ser alto ➡ Los agrupamos en **clases**

Notas de un grupo

9	7	3	8	5
	5	7	6	7
6		6	8	4
	8	7		5
4		5	6	10

Nota	f_a
3	1
4	2
5	4
6	5
7	4
8	3
9	1
10	1

Agrupamiento
en **clases**

Distribución de frecuencias

Clase	f	f_r	f_{acum}
3-4	3	0.14	0.14
5-6	9	0.43	0.57
7-8	7	0.33	0.90
9-10	2	0.10	1.0
TOTAL	21		

Los agrupamientos son arbitrarios
Tres directrices a tener en cuenta

1. El número de clases del agrupamiento no debe ser muy grande ni muy pequeño
2. Cada observación debe pertenecer a una única clase
3. Siempre que sea posible, las clases deben tener la misma amplitud

GRÁFICAS

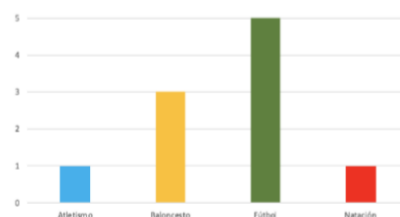
Los **gráficos** también nos ayudan a organizar los datos facilitando su visualización

Deporte	Frecuencia	Frecuencia relativa	Porcentaje
Atletismo	1	0.1	10%
Baloncesto	3	0.3	30%
Fútbol	5	0.5	50%
Natación	1	0.1	10%
TOTAL (N)	10		

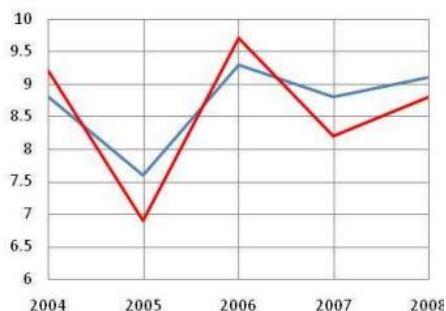
Diagrama de sectores



Diagrama de barras

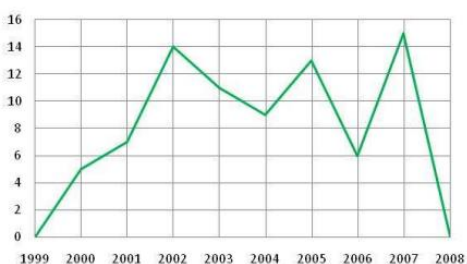


Gráfica simple de líneas



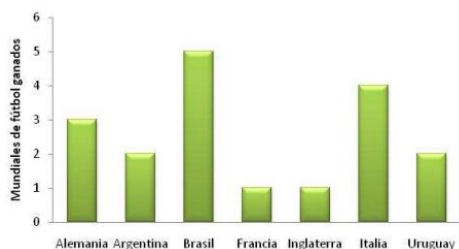
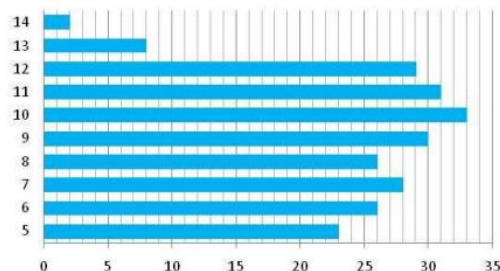
Permite comparar el comportamiento de dos caracteres

Polígono de frecuencias



Para hacer la representación basta con tener las frecuencias absolutas.

Diagrama de barras

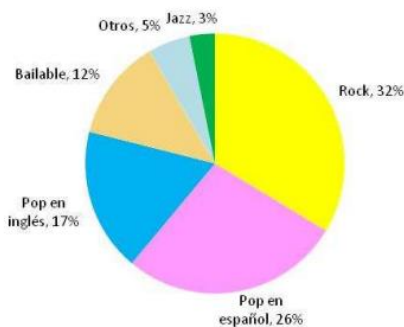


No muestran frecuencias acumuladas, la barra con mayor altura representa la mayor frecuencia. La suma de las alturas de las columnas equivale al 100% de los datos.

El papel de los ejes se intercambia. Se suele usar para poblaciones divididas por tramos de edad.

Diagrama de sectores

Este tipo de gráficas es muy útil cuando lo que se desea es resaltar las proporciones que representan algunos subconjuntos con respecto al total.

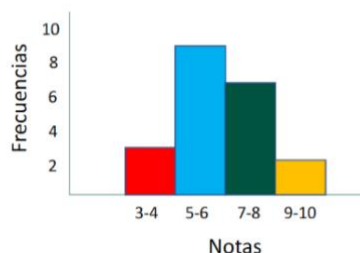


Histogramas

Distribución de frecuencias

Clase	f	f _r	f _{acum}
3-4	3	0.14	0.14
5-6	9	0.43	0.57
7-8	7	0.33	0.90
9-10	2	0.10	1.0
TOTAL	21		

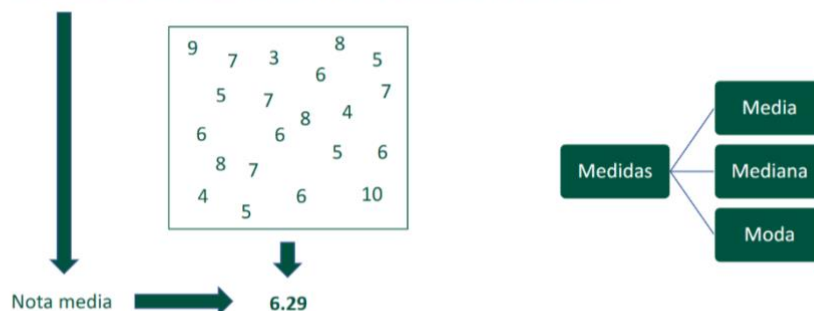
Histograma



MEDIDAS DE TENDENCIA CENTRAL

Objetivo de las **medidas de tendencia central**: describir los datos con un número que proporcione una idea global de los mismos

Ejemplo ➡ Representación de la nota de un grupo de estudiantes en un examen



MEDIA.

Notas de un grupo

9	7	3	8	5
5	7	6	4	7
6	6	8	5	6
8	7	5	6	
4	5	6	10	

Número de alumnos: **21**

¿Cómo calculamos la media aritmética?



Sumamos todas las notas y dividimos entre el número de las mismas

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \dots + X_N}{N} =$$



$$= \frac{3 + 4 + 4 + 5 + 5 + 5 + 5 + 6 + 6 + 6 + 6 + 6 + 7 + 7 + 7 + 7 + 8 + 8 + 8 + 9 + 10}{21} = \frac{132}{21} = \mathbf{6.29}$$

MEDIANA

Notas de un grupo

9				8		5
	7	3		6		7
	5		7		8	4
6			6		5	6
	8	7				
4		5		6		10

Número de alumnos: **21**

¿Qué es la mediana?

Es el valor de los datos que, una vez ordenados, los divide en dos mitades de igual tamaño

¿Cómo calculamos la mediana?

1. Ordenamos las notas de menor a mayor
2. La mediana será el dato que divide a las notas en dos mitades con el mismo número de notas

NOTA: Si el número de datos es par



Mediana = 6



$$\text{Mediana} = \frac{3 + 4}{2} = 3.5$$

MODA.

Notas de un grupo

9				8		5
	7	3		6		7
	5		7		8	4
6			6		5	6
	8	7				
4		5		6		10

Número de alumnos: **21**

¿Qué es la moda?

Es el valor de los datos que es más frecuente

¿Cómo calculamos la moda?

Simplemente será el dato con mayor frecuencia absoluta

3 4 4 5 5 5 5 6 6 6 6 6 7 7 7 7 8 8 8 9 10



El 6 es el dato que más se repite



Moda = 6

Algunas puntualizaciones

Si ningún dato se repite más de una vez → No hay moda

Si hay dos puntuaciones con la misma frecuencia → Distribución **bimodal**

Si hay más de dos puntuaciones con la misma frecuencia → Distribución **plurimodal**

MEDIA- MODA Y MEDIANA

¿Cuál de las tres utilizar? ➡ En la medida de lo posible las tres: ofrecen información complementaria

Ejemplo: estadísticos descriptivos del rendimiento de una clase en una prueba escrita de lengua

Información estadística	Interpretación
En la clase la media en lengua es 5.2	El grupo tiene un rendimiento medio-bajo
En la clase la media en lengua es 5.2 y la moda 8	El grupo tiene un rendimiento medio-bajo pero la nota más frecuente es de 8. Debe haber un grupo de estudiantes con notas bajas.

La mediana y la moda no son sensibles a los valores extremos, la media sí

2 3 4 4 5 6	<div>Media = 4.0</div> <div>Mediana = 4.0</div> <div>Moda = 4.0</div>	2 3 4 4 5 10	<div>Media = 4.6</div> <div>Mediana = 4.0</div> <div>Moda = 4.0</div>
-----------------------	---	------------------------	---

Aunque las distribuciones son diferentes, la **mediana** y la **moda** nos proporcionan la misma información

La **media** nos permite diferenciarlas

MEDIDAS DE DISPERSIÓN

RANGO- DESVIACIÓN TÍPICA O ESTANDAR Y VARIANZA

Rango: diferencia entre el valor máximo y el mínimo

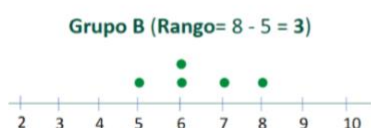
Desviación estándar (o típica): es una medida de la desviación con respecto a la media $\rightarrow s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}}$

Varianza: es una medida de la desviación con respecto a la media. Matemáticamente es el cuadrado de la desviación típica.



$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} = \sqrt{\frac{(3-6.4)^2 + (5-6.4)^2 + (6-6.4)^2 + (8-6.4)^2 + (10-6.4)^2}{5}}$$

$$= \sqrt{\frac{11.56 + 1.96 + 0.16 + 2.56 + 12.96}{5}} = 2.42 \quad s^2 = 2.72^2 = 5.84$$



$$s = \sqrt{\frac{\sum_{i=1}^N (X_i - \bar{X})^2}{N}} = \sqrt{\frac{(5-6.4)^2 + (6-6.4)^2 + (6-6.4)^2 + (7-6.4)^2 + (8-6.4)^2}{5}}$$

$$= \sqrt{\frac{1.96 + 0.16 + 0.16 + 0.36 + 2.56}{5}} = 1.02 \quad s^2 = 1.02^2 = 1.04$$

DISTRIBUCIÓN BIDIMENSIONAL

Una **distribución bidimensional** es la que se obtiene al estudiar un fenómeno respecto de dos variables estadísticas unidimensionales X e Y

Ejemplo

Dada la siguiente distribución bidimensional:

Variable X	1	1	1	1	1	1	1	1	1	2	2	2	2	2	3	3	4	4	4	4
Variable Y	1	1	3	3	3	3	4	4	4	1	1	2	4	4	2	2	1	3	3	3

se obtienen las tablas de frecuencias.

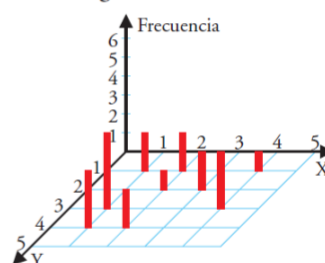
Tabla simple

X	Y	Frecuencia
1	1	2
1	3	4
1	4	3
2	1	2
2	2	1
2	4	2
3	2	2
4	1	1
4	3	3

Tabla de doble entrada

Y \ X	1	2	3	4	
1	2	2	0	1	5
2	0	1	2	0	3
3	4	0	0	3	7
4	3	2	0	0	5
	9	5	2	4	20

Diagrama de barras



Gráficas de dispersión

Una gráfica de dispersión tiene dos ejes de valores, mostrando un conjunto de datos numéricos en el eje x y otro en el eje y.

Combina estos valores en puntos de datos únicos y los muestra en intervalos uniformes o agrupaciones.

Los gráficos de dispersión se utilizan normalmente para mostrar y comparar valores numéricos, como datos científicos, estadísticos y de ingeniería.

RELACIÓN ENTRE LAS VARIABLES: CORRELACIÓN

La **covarianza** de una variable bidimensional (X, Y) es:

$$s_{xy} = \frac{\sum n_i x_i y_i}{N} - \bar{x} \cdot \bar{y}$$

Las **desviaciones típicas marginales** son las desviaciones típicas de las variables X e Y:

$$s_x = \sqrt{\frac{\sum n_i \cdot x_i^2}{N} - \bar{x}^2}$$

$$s_y = \sqrt{\frac{\sum n_i \cdot y_i^2}{N} - \bar{y}^2}$$

Interpretación de la covarianza

Según sea el signo de la covarianza, se interpreta:

- Covarianza positiva:** al aumentar los valores de la variable X, aumentan los valores de la variable Y. La nube de puntos se orienta a la derecha y hacia arriba.
- Covarianza negativa:** al aumentar los valores de la variable X, disminuyen los valores de la variable Y. La nube de puntos se orienta a la derecha y hacia abajo.

Una **nube de puntos** o **diagrama de dispersión** es la representación en unos ejes cartesianos de los datos (xi, yi) de una distribución bidimensional.

- Muchas variables mantienen una relación entre sí
- El hecho de que dos variables estén relacionadas implica que, cuando una cambia, la otra también lo hace de una forma esperada



- Los objetivos de un análisis correlacional son dos:
 - Explorar si existe relación entre las variables y su naturaleza (positiva o negativa)
 - Estudiar la fortaleza de la relación entre las variables

CORRELACIÓN

La **covarianza** indica cómo es la relación entre dos variables; es decir, cómo se orienta la nube de puntos, pero este parámetro no indica de una forma precisa la medida de esa relación. Para resolver este problema, se definen los conceptos de **correlación** y **coeficiente de correlación**.

Correlación es la relación que existe entre las dos variables que intervienen en una distribución bidimensional.

Posibles situaciones que se pueden dar



Correlación lineal
positiva perfecta



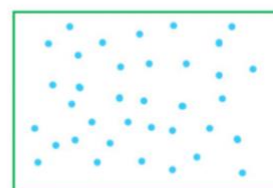
Correlación lineal
positiva



Correlación lineal
negativa perfecta



Correlación lineal
negativa



No hay correlación

COEFICIENTE DE CORRELACIÓN DE PEARSON (ESTÁ COMPRENDIDO ENTRE - 1 Y 1)

Para medir la fortaleza correlación entre variables se utilizan los coeficientes de correlación

Fortaleza	Correlación positiva	Correlación negativa
Perfecta	+1	-1
Fuerte	Entre +0.9 y +0.7	Entre -0.9 y -0.7
Moderada	Entre +0.6 y +0.4	Entre -0.6 y -0.4
Débil	Entre +0.3 y +0.1	Entre -0.3 y -0.1
Nula	0	0

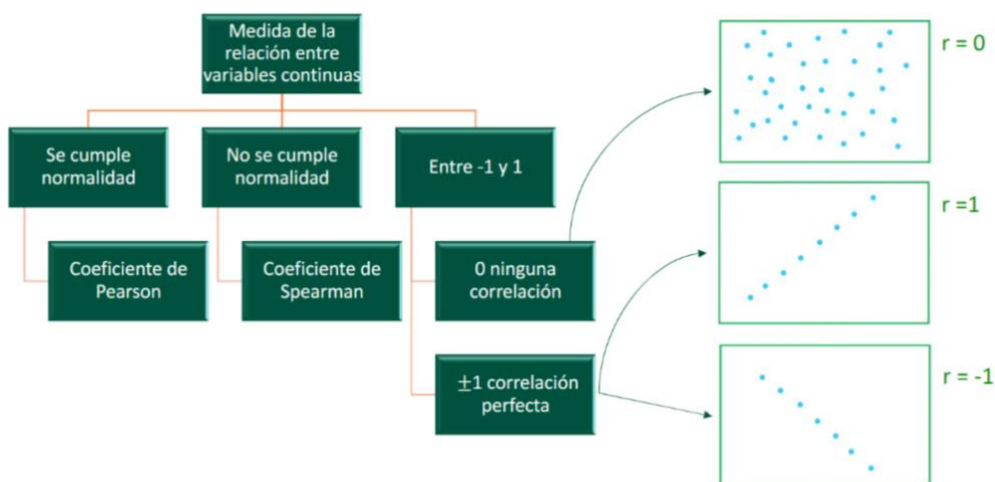
El **coeficiente de correlación de Pearson** es:

$$r = \frac{s_{xy}}{s_x s_y}$$

- Si $r = -1$ o $r = 1$, la **correlación es perfecta o funcional**.
- Si r está próximo a -1 o a 1 , la **correlación es fuerte**.
- Si r está próximo a cero, la **correlación es débil**.

El signo, $r > 0$ o $r < 0$, indica si la correlación es directa o inversa, respectivamente

Correlación fuerte: Se considera que la correlación es fuerte si $|r| > 0,85$



INFORMACIÓN COMPLEMENTARIA:

¿Cómo hacer un análisis estadístico de datos en Excel?

<https://conceptosclaros.com/como-hacer-analisis-estadisticos-de-datos-en-excel/#> Como hacer un analisis estadistico de datos en Excel