Estadística

Estadística

- La estadística es la ciencia que se ocupa de la recogida de datos, su organización y análisis.
- Según el método de estudio y el problema a resolver se puede distinguir:
  - **Estadística descriptiva**: Se encarga de recolectar datos, organizarlos y en el cálculo de valores que describan el conjunto objeto de estudio.
  - **Estadística inferencial**: se encarga de elaborar conclusiones para el conjunto estudiado a partir del estudio de una muestra.

Conceptos estadísticos I

- Población: conjunto de elementos a estudiar.
- Individuo: cada uno de los elementos de la población.
- Muestra: Subconjunto de la población que se toma para representar a la población en el estudio.
- Tamaño de la muestra: Número de elementos que componen la muestra.

Conceptos estadísticos II

- Carácter estadístico: propiedad de los individuos de una población que permite clasificarlos.
  - Cualitativo: no pueden expresarse numéricamente.
  - Cuantitativo: puede expresarse numéricamente.
- Variable estadística: conjunto de valores que puede tomar un carácter estadístico cuantitativo.
  - Discreta: la variable puede tomar un número finito de valores
  - Continua: la variable puede tomar un valor que se encuentre dentro de un intervalo de números reales.

Estudio estadístico

Para realizar un estudio estadístico se siguen las siguientes pautas:

- Extracción de una muestra de la población

  Debe ser representativa de la población

- Recuento de datos

  Proporciona la frecuencia absoluta de la tabla

- Elaboración de una tabla estadística

Recuento de datos

- El recuento de datos proporciona la frecuencia absoluta para cada valor de la variable estadística
- En la primera columna se disponen los distintos valores que puede tomar la variable estadística
- En la siguiente, tras contar los casos en los que aparece cada valor de la variable, se calcula la frecuencia absoluta.

Ejemplo I

A las 50 familias que viven en una calle se les ha realizado una encuesta preguntando el número de personas que viven en su vivienda.

3	3	5	8	3
7	2	4	4	3
2	6	1	5	4
4	3	3	4	1
4	6	4	7	8
7	4	2	2	7
3	2	2	2	2
6	3	3	3	4
2	4	6	8	1
3	5	5	3	6

El **carácter estadístico** se corresponde con el número de personas que viven juntas.
El **tamaño de la muestra** es 50.

Ejemplo I (continuación)

Valor de la variable	Frecuencia absoluta
x_i	f_i
1	3
2	9
3	12
4	10
5	4
6	5
7	4
8	3
Total	50

El rango de valores de la variable va de 1 a 8. La variable es **discreta**.

La frecuencia absoluta establece en este caso el número de viviendas que está ocupada por el número de inquilinos que le corresponde al valor de la variable.

Hay 4 viviendas de las 50 ocupadas por 7 personas.

La suma de las frecuencias absolutas es el tamaño de la muestra.

Intervalos o clases

- Cuando una variable estadística cuantitativa puede tomar una gran cantidad de valores distintos, se agruparán en intervalos.
- Normalmente se toman intervalos de igual amplitud.
- Para realizar cálculos se utilizará la marca de clase que no es más que el punto medio del intervalo.

Ejemplo II

A los 100 empleados de una empresa , se les ha realizado una prueba sobre resolución de problemas. En una escala de 0 a 100 se ha obtenido las siguientes puntuaciones:

2	58	69	62	17	53	5	30	76	33
84	76	69	70	63	65	46	99	65	2
5	82	83	80	72	74	39	21	84	48
68	5	15	90	48	99	32	17	27	62
21	54	11	4	7	33	90	76	4	18
27	53	47	11	88	66	64	29	59	99
46	61	20	27	49	44	37	90	72	35
84	7	85	14	28	71	89	76	48	52
70	41	25	85	90	4	81	78	55	36
73	8	2	94	84	68	11	13	37	9

Ejemplo II (continuación)

Se ha decidido agrupar los valores en 5 clases, como el menor valor es 2 y el mayor 100, la amplitud del intervalo será:

$$\frac{Máximo\ valor - Mínimo\ valor}{Número\ de\ clases} = \frac{100 - 2}{5} = 19{,}6$$

Tomaremos para la longitud del intervalo 20, pues es más fácil de realizar cálculos:

Intervalo	Marca de clase	Frecuencia Absoluta
$(x_i, x_{i+1}]$	x_i	f_i
(0,20]	10	23
(20,40]	30	17
(40,60]	50	16
(60,80]	70	25
(80,100]	90	19
	Total	100

Frecuencias

Frecuencia absoluta: Número de individuos de la muestra que presenta un valor de la variable estadística. Su suma debe ser el tamaño de la muestra.

Frecuencia relativa: Fracción que representa la frecuencia absoluta respecto del tamaño de la muestra. Se calcula dividiendo la frecuencia absoluta entre el tamaño de la muestra. La suma total debe ser 1.

Frecuencia porcentual: Porcentaje que representa cada frecuencia absoluta respecto del tamaño de la muestra. La suma total debe ser 100.

Frecuencias acumuladas: se corresponde con la suma de las frecuencias anteriores a las de un valor de la variable más la suya propia.

Ejemplo

Valor de la variable	Frecuencia absoluta	Frecuencia relativa	Frecuencia porcentual	Frecuencia acumulada
x_i	f_i	h_i	p_i	F_i
1	3	0,06	6	3
2	9	0,18	18	12
3	12	0,24	24	24
4	10	0,2	20	34
5	4	0,08	8	38
6	5	0,1	10	43
7	4	0,08	8	47
8	3	0,06	6	50
Total	50	1	100	

Distribuciones unidimensionales

PARÁMETROS DE CENTRALIZACIÓN

Parámetros estadísticos

Permiten describir de forma resumida el comportamiento de la variable estadística

Existen dos grandes grupos:

- Parámetros de **centralización**
  - Moda, mediana y media
- Parámetros de **dispersión**
  - Rango, varianza, desviación típica y coeficiente de variación

Parámetros de centralización

La moda (*Mo*): de los valores de la variable estadística el de mayor frecuencia.

La mediana (*Me*): tras haber ordenado los valores de la variable (junto con sus frecuencias), aquel valor de la variable que ocupa el lugar central.

La media ($\overline{x}$): valor que se obtiene al sumar los productos de los valores de la variable estadística por la frecuencia absoluta, dividido por el tamaño de la muestra.

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

Ejemplo

Valor de la variable	Frecuencia absoluta
x_i	f_i
1	3
2	9
3	12
4	10
5	4
6	5
7	4
8	3
Total	50

La moda es 3, pues es el valor de la variable que más frecuencia absoluta tiene

La mediana es 4, pues contiene el elemento 25 y 26 (mitad y mitad mas uno del tamaño de la población)

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i \cdot f_i}{N} = \frac{3 \cdot 1 + 2 \cdot 9 + 3 \cdot 12 + \cdots .. + 8 \cdot 3}{50} = 3,98$$

A tener en cuenta para calcular la mediana

Cuando la variable estadística es discreta y el número de elementos es par, para calcular la mediana se utilizan los dos valores centrales y se realiza la media aritmética de los valores centrales de la variable estadística.

Si la variable estadística se encuentra agrupada en intervalos el que contiene el valor medio es el intervalo mediano pero la mediana se calculará mediante la siguiente fórmula:

$$M_e = L_i + a\frac{\frac{N}{2} - F_{i-1}}{f_i}$$

Donde:

- L_i Límite inferior del intervalo mediano
- a Amplitud del intervalo
- f_i Frecuencia absoluta de la clase mediana
- N Tamaño de la muestra
- F_{i-1} Frecuencia absoluta acumulada de la clase anterior

Ejemplo

Intervalo	Marca de clase	Frecuencia Absoluta
$(x_i,x_{i+1}]$	x_i	f_i
(0,20]	10	23
(20,40]	30	17
(40,60]	50	16
(60,80]	70	25
(80,100]	90	19
	Total	100

El intervalo mediano es [40,60] que acumula los datos 50 y 51

$$M_e = L_i + a\frac{\frac{N}{2} - F_{i-1}}{f_i} = 40 + 20\frac{\frac{100}{2} - 40}{16} = 52{,}5$$

Parámetros estadísticos de dispersión I

El rango (amplitud o recorrido): diferencia entre el mayor y menor valor de la variable estadística.

La varianza: mide la distancia entre la media y los valores de la variable estadística.

Las siguientes fórmulas permiten el cálculo de la varianza.

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2 f_i}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{n} x_i^2 f_i}{N} - \overline{x}^2$$

Parámetros estadísticos de dispersión II

La desviación típica: mide si los datos se encuentran agrupados en las proximidades de la media. Se mide en las mismas unidades que los datos

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

La desviación típica es el parámetro de dispersión más utilizado

Si se suma una constante a cada uno de los valores de la variable, la desviación típica no varía

Si se multiplica una constante a cada uno de los valores de la variable, la desviación típica queda multiplicada por él.

Coeficiente de variación

Cuando se desea comparar dos variables aleatorias se utiliza el **coeficiente de variación,** que no es más que la razón entre la desviación típica y la media de la variable

$$CV = \frac{\sigma}{\overline{x}}$$

El coeficiente de variación siempre es positivo y no tiene unidades.

Cuanto más pequeño es, los valores de la variable se encuentran más concentrados alrededor de la media

Permite comparar dos variables aleatorias heterogéneas.

VARIABLES ESTADÍSTICAS BIDIMENSIONALES

Variable estadística bidimensional

Cuando en una población se observan dos caracteres de los individuos, se recurre a una variable estadística bidimensional

Cada individuo tendrá asociado dos valores, uno para cada uno de los caracteres estadísticos estudiados.

El par ordenado (X,Y) se compone de dos variables estadísticas unidimensionales. Si n es el tamaño de la muestra o población, el par (x_i,y_i) representa los valores de cada una de las variables estadísticas para el individuo *i.*

Relaciones funcionales y estadísticas

La relación entre dos variables podría ser funcional, si conocida una de ellas se conoce la otra de forma unívoca. Este tipo de relaciones no son estudiadas por la estadística.

Cuando en una relación puede obtenerse un valor aproximado si se conoce el valor de la otra variable, se dice que hay una relación estadística (correlación).

Ejemplos

La longitud de un niño conocidos los meses desde su nacimiento

El peso y la altura de una persona

Tabla simple

Para organizar los datos de las variables bidimensionales se pueden utilizar las tablas simples, donde para cada individuo de la muestra aparece los valores de los dos caracteres estadísticos a estudiar.

Ejemplo:
Supongamos que se estudia la puntuación de 10 estudiantes en un test de expresión oral y escrita.

Expresión oral (X)	8	5	4	9	3	7	8	5	9	9
Expresión escrita (Y)	9	6	6	7	6	7	4	6	8	6

Cada columna representa la puntuación obtenida por cada uno de los estudiantes (la frecuencia de cada par es 1).

Tabla simple con frecuencias

Para cada par de valores de las variables estadísticas se les asocia la frecuencia.

Ejemplo:
Supongamos que se estudia la puntuación de 63 estudiantes en una prueba de matemáticas y economía, recogiendo para cada par distinto de valores obtenidos el número de alumnos que lo obtienen:

Matemáticas (X)	3	4	5	5	6	7	7	8	8	9
Economía (Y)	4	4	4	5	7	7	8	8	9	9
Número de alumnos	4	5	6	7	9	9	8	6	6	3

Tablas de doble entrada

Para organizar los datos de las variables bidimensionales se pueden utilizar las tablas de doble entrada:

Nota de Física y Química X

Nota de Matemáticas Y

	[2.5,5)	[5,6)	[6,7)	[7,9)	[9,10)	Total
[2.5,5)	2	3	2	2	1	10
[5,6)	1	5	6	6	2	20
[6,7)	7	7	14	5	5	38
[7,9)	1	2	9	3	6	21
Total	12	19	35	20	14	100

Hay 19 alumnos que han obtenido una nota de física entre 5 y 6.

Número total de alumnos

Distribuciones marginales

A partir de las tablas de doble entrada, es posible obtener las distribuciones marginales de cada una de las variables:

Distribuciones marginales ⇨

[2.5,5)	[5,6)	[6,7)	[7,9)	[9,10)
12	19	35	20	14

⇩

[2.5,5)	10
[5,6)	20
[6,7)	38
[7,9)	21

Nota de Física y Química

Nota de Matemáticas

	[2.5,5)	[5,6)	[6,7)	[7,9)	[9,10)	Total
[2.5,5)	2	3	2	2	1	10
[5,6)	1	5	6	6	2	20
[6,7)	7	7	14	5	5	38
[7,9)	1	2	9	3	6	21
Total	12	19	35	20	14	100

Distribuciones condicionadas

En una distribución bidimensional podemos estudiar la distribución que resulta de fijar un valor en una de las variables y estudiar las frecuencias correspondientes a la otra.
A la distribución así formada se denomina distribución condicionada.

Ejemplo:
Si fijamos la nota de matemáticas al intervalo [2.5,5), la distribución condicionada para la nota de física es:

	[2.5,5)	[5,6)	[6,7)	[7,9)	[9,10)
[2.5,5)	2	3	2	2	1

Esta distribución se notará por: $X|_{Y=[2.5,5)}$ siendo X la variable asociada a la puntuación de física e Y la variable asociada a la puntuación de matemáticas.

Diagrama de dispersión I

El diagrama de dispersión es la representación gráfica de los pares de puntos (x_i,y_i) que forman la distribución bidimensional.

Ejemplo:
Supongamos que se estudia la nota de 5 alumnos en bachillerato (X) y en la PAU (Y).

x_i	y_i
5,4	4,8
6,2	7,0
5,5	6,2
6,4	4,3
8,2	8,5

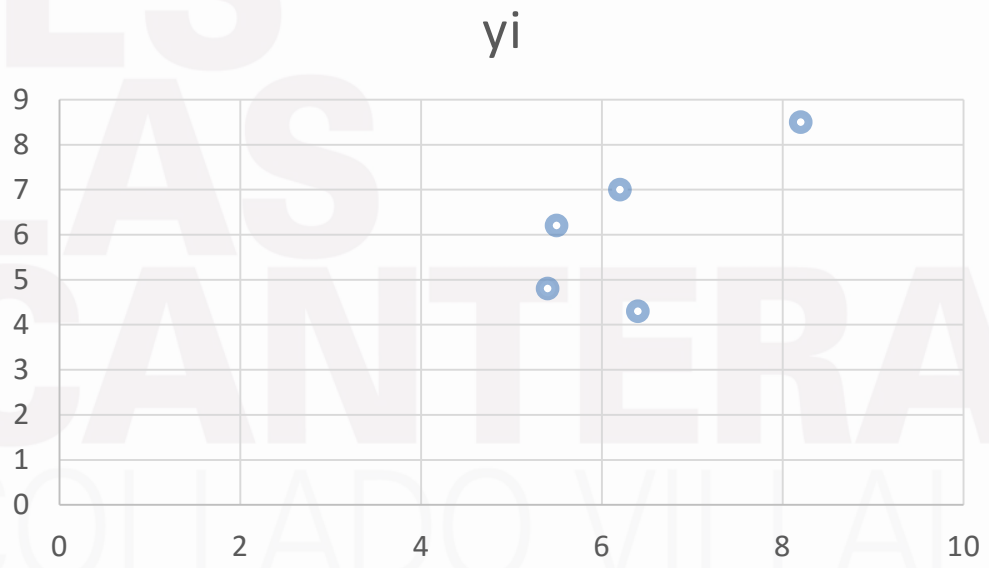
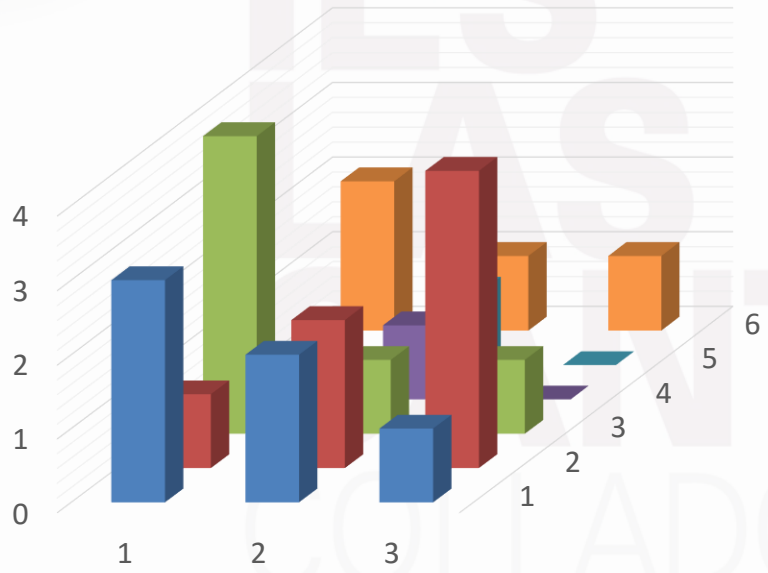


Diagrama de dispersión II

Cuando se dispone de una tabla de doble entrada, la frecuencia de cada punto no es uno, por lo que puede representarse en las inmediaciones del punto representado tantos puntos como frecuencia exista en la tabla, o bien, el tamaño del "punto" representado sea proporcional, o utilizando un gráfico de tridimensional.



Covarianza

Se define la **covarianza** de una variable bidimensional ***(X,Y)*** a la media aritmética de los productos de las desviaciones de cada una de las variables respecto a sus medias respectivas. Para realizar el cálculo de la covarianza se utilizan las siguientes fórmulas (son equivalentes):

$$S_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{N} \qquad S_{xy} = \frac{\sum_{i=1}^{n} f_i x_i y_i}{N} - \bar{x}\bar{y}$$

Interpretación de la covarianza

Dependiendo del signo de la covarianza:

- **Positivo**: Indica que los productos de $\mathbf{x_i}$ e $\mathbf{y_i}$ se alejan en el mismo sentido de sus respectivas medias
- **Negativo** (comportamiento recíproco al anterior)
- **Cero o próximo a cero**: la covarianza indica que no hay relación entre las variables.

Correlación

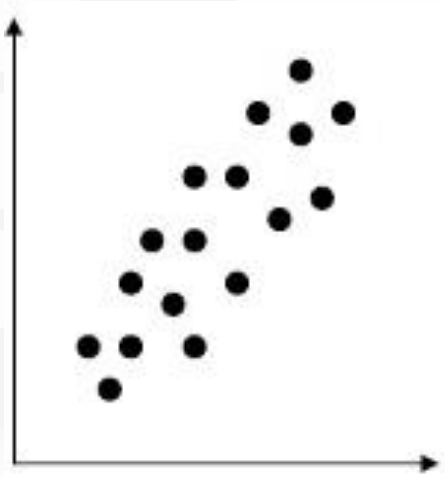
La correlación es la relación que existe entre las dos variables que participan en una distribución bidimensional.

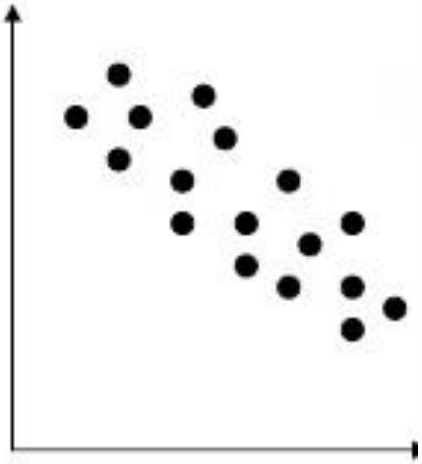
La representación que proporciona la nube de puntos puede indicar el tipo de correlación:

- **Lineal**: la nube de puntos se dispone alrededor de una recta.
  - **Positiva (directa):** la pendiente de la recta es positiva
  - **Negativa (inversa):** la pendiente de la recta es negativa
- **Curvilínea**: la nube de puntos se distribuye alrededor de una curva.
- **Nula**: cuando no existe ninguna relación entre las variables.

Ejemplos

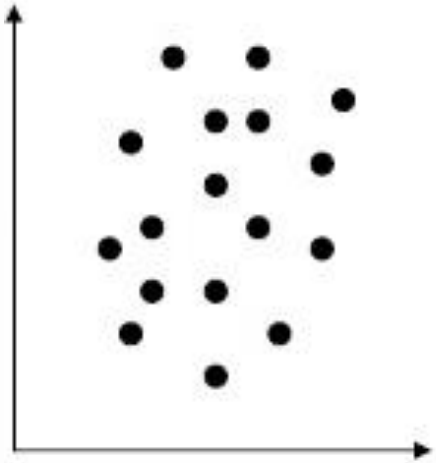
Correlación lineal positiva

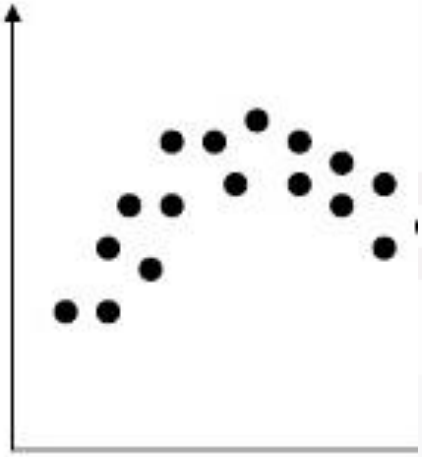




Correlación lineal negativa

Correlación nula





Correlación cuvilínea

Regresión lineal. Coeficiente de Pearson

Para poder cuantificar la correlación lineal entre dos variables se utiliza el coeficiente de correlación lineal de Pearson (r).

$$r = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Este coeficiente tiene el mismo signo que la covarianza, y se encuentre comprendido entre 0 y 1, y proporciona la siguiente información:

- **-1 < r < 0**, la correlación es negativa y es mas fuerte cuanto más se acerque a -1.
- **r = 0**, no existe correlación
- **0 < r < 1**, la correlación es positiva y es mas fuerte cuanto mas se acerque a 1.
- **r = 1 ó r = -1**, existe dependencia funcional.

Rectas de regresión

Tras comprobar que existe correlación entre dos variables estadísticas, es posible realizar el cálculo de las rectas que mejor se ajustan a la nube de puntos:

Recta de regresión de Y sobre X.
Si deseamos calcular el valor estimado de Y, dependiendo del valor de X.

$$y - \overline{y} = \frac{\sigma_{XY}}{{\sigma_X}^2}(x - \overline{x})$$

Recta de regresión de X sobre Y.
Si deseamos calcular el valor estimado de X, dependiendo del valor de Y.

$$x - \overline{x} = \frac{\sigma_{XY}}{{\sigma_Y}^2}(y - \overline{y})$$