FORMULARIO DE ESTADÍSTICA

Conceptos básicos

Población: conjunto de todos los elementos objeto de nuestro estudio

Muestra: subconjunto, extraído de la población,(mediante técnicas de muestreo) cuyo estudio sirve para inferir características de toda la población

Individuo: cada uno de los elementos que forman la población o la muestra

Variable estadística: característica objeto de estudio

- Discreta: Es la variable que presenta separaciones o interrupciones en la escala de valores que puede tomar
- Continua: Es la variable que puede adquirir cualquier valor dentro de un intervalo especificado de valores

Notaciones y frecuencias:

- *Variables discretas*

  $X : x_1,...,x_k$ con frecuencias $f_1,...,f_k$

  $f_i =$ número de veces que aparece el dato $x_i \equiv$ frecuencia absoluta de x_i

  $N =$ número total de datos

  $F_i = \sum_{j \le i} f_j \equiv$ frecuencia absoluta acumulada de x_i

  $h_i = \frac{f_i}{N} \equiv$ frecuencia relativa de x_i

  $H_i = \sum_{j \le i} h_j \equiv$ frecuencia relativa acumulada de x_i

- *Variables continuas*

  $X : I_1,...,I_k$ (intervalos)

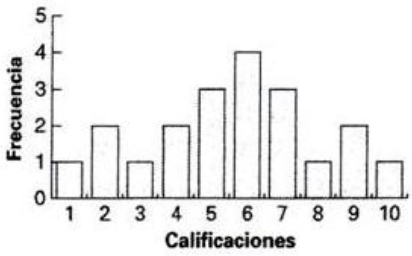
  $x_i =$ punto medio del intervalo $I_i \equiv$ marca de clase de I_i

Tablas de frecuencias:

x_i	f_i	F_i	h_i	H_i	…

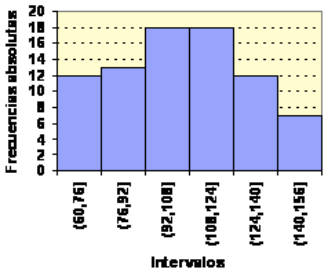
Gráficos estadísticos

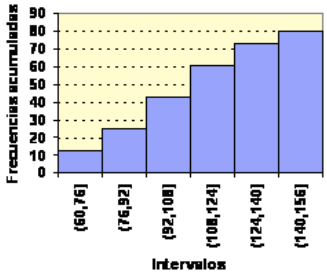
- Diagrama de barras o columnas



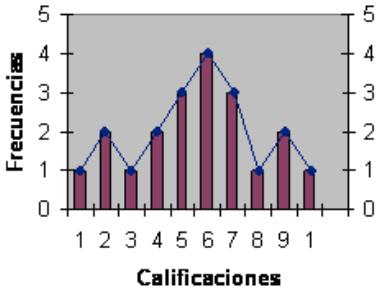
- Diagrama de sectores

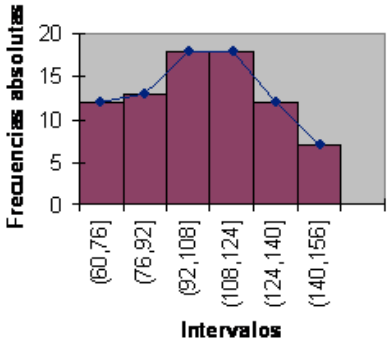
- Histogramas



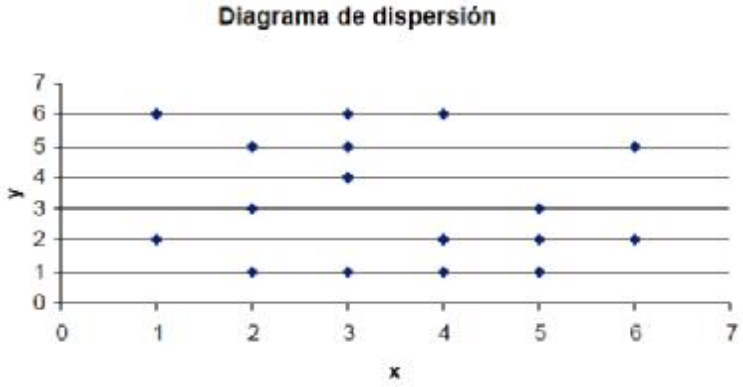


- Polígonos de frecuencias





- Diagrama de dispersión



Medidas de tendencia central:

Media (aritmética):

$$\bar{x} = \frac{\sum f_i x_i}{n}$$

Mediana:

Se ordenan los datos. Si hay un número par de datos la mediana es la media de los dos datos centrales; si el número de datos es impar, la mediana es justamente el dato central.

Cálculo:

Datos sin agrupar:

$$F_{j-1} = \frac{n}{2} < F_j \Rightarrow Me = \frac{x_{j-1} + x_j}{2}$$

$$F_{j-1} < \frac{n}{2} < F_j \Rightarrow Me = x_j$$

Datos agrupados:

$$F_j = \frac{n}{2} \Rightarrow Me = x_j$$

$$F_{j-1} < \frac{n}{2} < F_j \Rightarrow Me = x_{j-1} + \frac{\frac{n}{2} - F_{j-1}}{f_j}(x_j - x_{j-1})$$

Moda:

Valor más frecuente de la variable.

Interpretación: análisis de los datos

Supongamos que estamos estudiando el número de vuelos semanales que realizan 10 pilotos. Los datos obtenidos son los siguientes:

Nº de vuelos	0	1	2	3
Frecuencia absoluta	2	4	3	1

La media es 1,3, y nos indica, que por término medio, el número de vuelos es de 1,3, es decir, que *por término medio estos pilotos vuelan entre 1 y 2 veces por semana*.

La moda es 1, lo que nos indica que lo más frecuente es que vuelen 2 veces por semana.

Y por último, la mediana es 1, lo que nos dice que *hay tantos pilotos que vuelan 1 o más veces, como pilotos que lo hacen 1 vez o menos*.

Medidas de posición no central:

Cuantiles:

El cuantil $p_{r/k}$, $r=1,...,k-1$, se define como aquel valor de la variable que divide la distribución de frecuencias, previamente ordenada de forma creciente, en dos partes, estando el $100\frac{r}{k}\%$ de ésta formado por valores menores que $p_{r/k}$.

Si $k=4$ los (tres) cuantiles reciben el nombre de **cuartiles**. Si $k=10$ los (nueve) cuantiles reciben, en este caso, el nombre de **deciles**. Por último, si $k=100$ los (noventa y nueve) cuantiles reciben el nombre de **centiles**.

Cálculo:

Datos sin agrupar:

$$F_{j-1}=\frac{r}{k}n<F_j \Rightarrow p_{r/k}=\frac{x_{j-1}+x_j}{2}$$

$$F_{j-1}<\frac{r}{k}n<F_j \Rightarrow p_{r/k}=x_j$$

Datos agrupados:

$$F_j=\frac{r}{k}n \Rightarrow p_{r/k}=x_j$$

$$F_{j-1}<\frac{r}{k}n<F_j \Rightarrow p_{r/k}=x_{j-1}+\frac{\frac{r}{k}n-F_{j-1}}{f_j}(x_j-x_{j-1})$$

Interpretación: análisis de los datos

Para comprar zapatillas a los miembros de una peña de bolos, se les he preguntado por la talla de calzado que usan y los resultados son los siguientes:

Nº de calzado	35	36	37	38	39	40	41	42
Frecuencia absoluta	7	13	20	37	42	50	23	8

El primer cuartil es $Q_1=38$ y lo que nos dice es que el 25 % de los miembros de la peña utilizan una talla de calzado menor o igual que 38.

El segundo cuartil es $Q_2=39$ (que coincide con la mediana) y lo que nos dice es que el 50 % de miembros usa una talla de calzado menor o igual que 39 y el otro 50 % mayor o igual.

El tercer cuartil es $Q_3=40$ que nos dice que el 75 % de los miembros del club de bolos usa una talla de calzado menor o igual que 40.

Medidas de dispersión:

Varianza:

$$\sigma^2 = \frac{\sum f_i \left(x_i - \bar{x}\right)^2}{n} \longrightarrow \sigma^2 = \frac{\sum f_i x_i^2}{n} - \bar{x}^2$$

Desviación típica:

$\sigma = +\sqrt{\sigma^2}$ (Raíz cuadrada positiva de la varianza)

Interpretación: análisis de los datos

Supongamos que estamos estudiando el número de aciertos de 100 alumnos en una prueba de 30 preguntas. Los resultados obtenidos se recogen en la siguiente tabla:

Aciertos	x_i	f_i
$[0,5)$	2,5	3
$[5,10)$	7,5	10
$[10,15)$	12,5	25
$[15,20)$	17,5	38
$[20,25)$	22,5	16
$[25,30]$	27,5	8
Total		100

En este caso el rango es 30, y por tanto, no nos proporciona ninguna información.

La varianza es $\sigma^2 = 33,79$ y la desviación típica es $\sigma = 5,81$, que son relativamente grandes, lo que nos dice que los datos presentan una agrupación relativamente pequeña respecto de la media.

Coeficiente de variación: (Se utiliza para comparar distribuciones)

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

Si $CV_X < CV_Y$ entonces la distribución de X es más homogénea que la de Y

Si $CV = 0 \Rightarrow \sigma = 0 \Rightarrow \bar{x}$ tiene máxima representatividad

Si $\bar{x} < \sigma \Rightarrow \bar{x}$ no tiene representatividad alguna

Interpretación: análisis de los datos

Vamos a comparar las siguientes distribuciones de datos:

7	3	2	4	5
3	2	4	9	8
2	5	6	5	4
8	6	3	4	0
0	2	1	5	6

1	8	6	1	5
1	0	2	4	1
7	1	3	0	5
9	2	5	7	4
4	3	5	2	3

Al calcular los coeficientes de variación obtenemos:

$$CV_1 = 0,57 \quad \text{y} \quad CV_2 = 0,70$$

Esto lo que nos dice es que la primera distribución de datos está menos dispersa que la segunda.

Covarianza: (Es una medida de dispersión conjunta de las variables X e Y)

$$\sigma_{(X,Y)} = \frac{\sum f_{ij} x_i y_j}{n} - \bar{x} \cdot \bar{y}$$

Rectas de regresión:

Determina la estructura de dependencia (en nuestro caso una recta) que mejor expresa el tipo de relación entre las variables.

1) **de** Y/X : $y - \bar{y} = \frac{\sigma_{(X,Y)}}{\sigma^2_X}(x - \bar{x})$

2) **de** X/Y : $x - \bar{x} = \frac{\sigma_{(X,Y)}}{\sigma^2_Y}(y - \bar{y})$

Índices de correlación:

Es frecuente que estudiemos sobre una misma población los valores de dos variables estadísticas distintas, con el fin de ver si existe alguna relación entre ellas, es decir, si los cambios en una de ellas influyen en los valores de la otra. Si ocurre esto decimos que las variables están correlacionadas o bien que hay correlación entre ellas.

1) **Razón de correlación**: $r^2 = \frac{\sigma^2_{(X,Y)}}{\sigma^2_X \cdot \sigma^2_Y}$

2) **Coeficiente de correlación lineal de Pearson**: $r = \frac{\sigma_{(X,Y)}}{\sigma_X \cdot \sigma_Y}$

El coeficiente de correlación lineal es un número real comprendido entre −1 y 1: $-1 \le r \le 1$

Si el coeficiente de correlación lineal toma valores cercanos a −1 la correlación es fuerte e inversa, y será tanto más fuerte cuanto más se aproxime r a −1.

Si el coeficiente de correlación lineal toma valores cercanos a 1 la correlación es fuerte y directa, y será tanto más fuerte cuanto más se aproxime r a 1.

Si el coeficiente de correlación lineal toma valores cercanos a 0, la correlación es débil.

Si r = 1 ó −1, los puntos de la nube están sobre la recta creciente o decreciente. Entre ambas variables hay dependencia funcional.

Ejemplo:

Una compañía de seguros considera que el número de vehículos (Y) que circulan por una determinada autopista a más de 120 km/h, puede ponerse en función del número de accidentes (X) que ocurren en ella.

Durante 5 días obtuvo los siguientes resultados:

X	5	7	2	1	9
Y	15	18	10	8	20

a) Calcula el coeficiente de correlación lineal.
b) Si ayer se produjeron 6 accidentes, ¿cuántos vehículos podemos suponer que circulaban por la autopista a más de 120 kms/h?
c) ¿Es buena la predicción?

Solución:

Disponemos los cálculos de la siguiente forma:

(Accidentes) x_i	Vehículos y_i	x_i^2	y_i^2	$x_i y_i$
5	15	25	225	75
7	18	49	324	126
2	10	4	100	20
1	8	1	64	8
9	20	81	400	180
24	71	160	1113	409

$$\bar{x} = \frac{\sum x_i}{N} = \frac{24}{5} = 4{,}8\,; \quad \bar{y} = \frac{\sum y_i}{N} = \frac{71}{5} = 14{,}2\,; \quad \sigma_x^2 = \frac{\sum x_i^2}{N} - \bar{x}^2 = \frac{160}{5} - 4{,}8^2 = 8{,}96$$

$$\sigma_y^2 = \frac{\sum y_i^2}{N} - \bar{y}^2 = \frac{1113}{5} - 14{,}2^2 = 20{,}96\,; \; \sigma_{xy} = \frac{\sum x_i y_i}{N} - \bar{x}.\bar{y} = \frac{409}{5} - 4{,}8.14{,}2 = 13{,}64$$

a) Coeficiente de correlación lineal de Pearson: $r = \dfrac{\sigma_{xy}}{\sigma_x \cdot \sigma_y} = \dfrac{13{,}64}{\sqrt{8{,}96}.\sqrt{20{,}96}} = 0{,}996$

b) Recta de regresión de ***y*** sobre ***x***: $y - \bar{y} = \dfrac{\sigma_{xy}}{\sigma_x^2}(x - \bar{x})$

$$y - 14{,}2 = \frac{13{,}64}{8{,}96}(x - 4{,}8)\,; \quad y - 14{,}2 = 1{,}53(x - 4{,}8)$$

Para $x = 6$, $y - 14{,}2 = 1{,}53(6 - 4{,}8)$, es decir, y = 16,04. Podemos suponer que ayer circulaban 16 vehículos por la autopista a más de 120 kms/h.

c) La predicción hecha es buena ya que el coeficiente de correlación está muy próximo a 1.