

Resumen Tema 1

" La estadística es la ciencia que busca saber todo de todos sin saber nada de nadie".

Conceptos básicos

- **Fenómenos determinísticos:** Dan lugar a un **mismo resultado** bajo **mismas condiciones**.
- **Fenómenos aleatorios:** El **resultado puede variar**.
- **Población:** Conjunto de **unidades con características en común** de los que **se quiere obtener información**.
- **Muestra:** Subconjunto **representativo** de la población.
- **Característica estadística:** Propiedad que se desea estudiar de la población.
 - **Modalidad:** Formas en las que se puede manifestar el carácter (1 por individuo)
 - **Carácter cualitativo:** Modalidades no medibles (noméricamente)
 - **Carácter cuantitativo:** Modalidades medibles (noméricamente)
- **Escala de medida:** Asignación de símbolos o números a las distintas modalidades.
 - **Escala nominal:** $x_A = x_B \text{ o } x_A \neq x_B$
 - **Escala ordinal:** $x_A \leq x_B \text{ o } x_B \leq x_A$
 - **Escala de intervalo:** Tiene sentido que A es $x_A - x_B$ unidades diferentes B.
 - **Escala de razón:** A es $\frac{x_A}{x_B}$ veces superior a B.
- **Variable:** Símbolo que **representa distintos valores numéricos**. Si son el resultado de **observaciones estadísticas** se llama **variable estadística**.
 - **VARIABLES DISCRETAS:** Si el paso de un valor a otro representa un salto (puntos cuestados)
 - **VARIABLES CONTINUAS:** Si, a priori, puede tomar cualquier valor entre dos dadas

Supongamos una población de tamaño n en la que se ha observado una variable estadística x que ha presentado K modalidades.

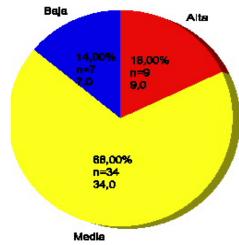
- **Frecuencia absoluta de la modalidad x_i :** Número de individuos que presentan dicha modalidad.
- **Frecuencia relativa de la modalidad x_i :** Proporción de individuos que presentan dicho valor.
- **Frecuencia absoluta acumulada de la modalidad x_i :** Número de individuos que presentan un valor menor o igual a x_i .
- **Frecuencia relativa acumulada de la modalidad x_i :** Proporción de individuos que presentan un valor menor o igual a x_i .
- Se denomina **distribución de frecuencia** de una variable al conjunto formado por cada uno de las modalidades junto con su frecuencias.

- (*) En una población de tamaño n SE HA OBSERVADO una variable estadística X que HA PRESENTADO K modalidades distintas con distribución de frecuencia (absoluta) $\{x_i, n_i\} \quad i = 1, \dots, K$.

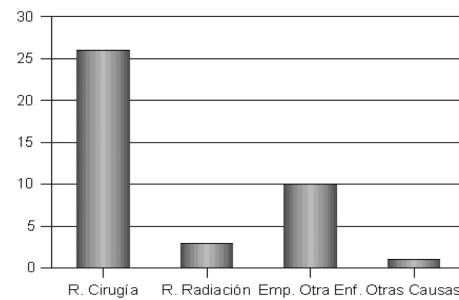
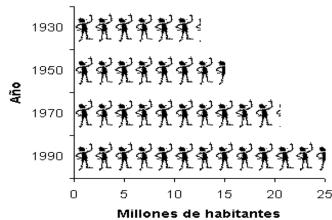
Representación gráfica

- Para atributos (variables cualitativas):

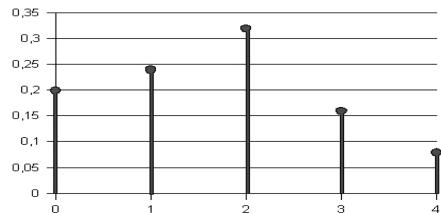
- Diagrama de sectores: Círculo dividido en tantos sectores como modalidades, siendo cada uno proporcional a la frecuencia.



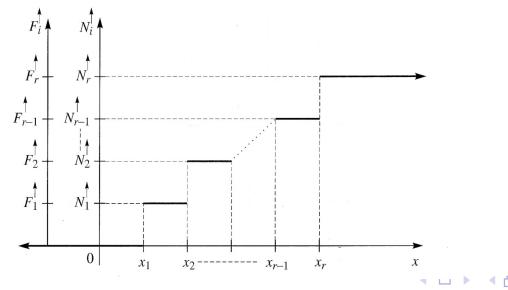
- Diagrama de rectángulos o barras: Rectángulos de base constante y alturas proporcionales a la frecuencia.



- Para variables discretas:



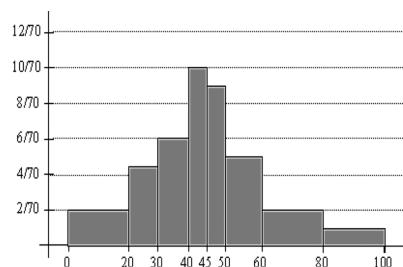
- Diagrama de barras: igual que el diagrama de barras que para atributos, pero trazando líneas rectas de longitud proporcional a la frecuencia.



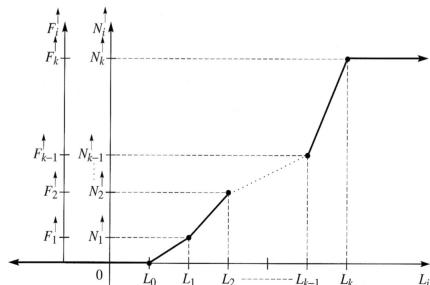
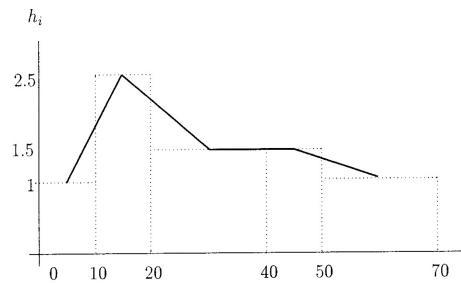
- Curva acumulativa o de distribución: Función definida para cada valor x como la proporción de datos menores o iguales a x .

- Variables continuas:

- Histograma: Rectángulos juxtapuestos, cuyas bases son los diferentes intervalos o clases y cuyas alturas son las frecuencias medias (densidades de frecuencia)



- Polígonal de frecuencias: Línea poligonal resultante de unir los vértices de las marcas de clase del histograma.



- Curva acumulativa o distribución: Análogo a la de atributos, pero continua

Entendemos por medidas estadísticas resúmenes cuantitativos de los datos que reflejan la información de los mismos, haciendo más fácil su interpretación y facilitando la comparación entre distintos conjuntos de datos.

Propiedades deseables de Yule:

- 1) Deben definirse de manera objetiva
- 2) Deben usar todas las observaciones
- 3) Tener un significado concreto y fácilmente interpretable.
- 4) Sencillas de calcular
- 5) Poco sensible a fluctuaciones muestrales

Sensible a cambios en los extremos

| | |
|---------------------------------|-----------------------------------|
| 1, 2, 3 | 1, 2, 30 |
| $\bar{x} = \frac{1+2+3}{3} = 2$ | $\bar{x} = \frac{1+2+30}{3} = 11$ |

Tipos:

- **Medidas de posición:** Permiten localizar una distribución en la recta real:
 - **Promedios:** Proporcionan un valor central representativo, alrededor del que se agrupan datos.
 - **Cuartiles:** Proporcionan valores representativos de partes de la distribución
- **Medidas de dispersión:** Grado de esparcimiento de los datos de una distribución
- **Medidas de forma:** caracterizan la forma de una distribución sin necesidad de hacer su gráfica.

Tipos de media

- **Media aritmética:** Suma de todos los valores de la variable entre el número total de observaciones

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i$$

- Si la variable es continua y los datos están agrupados en intervalos de clase :

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i c_i \rightarrow \text{marca de clase (centro del intervalo)}$$

- Está acotada por los valores extremos de la variable.

$$x_1 < \bar{x} < x_k$$



La desigualdad es estricta porque según el "piñero *". No puede ser todos los datos superiores o todos inferiores a la media y no vale el caso de $k=1$ porque no tiene nada de interés estadístico.

Es el centro de gravedad, no tiene que ser un extremo

- Si se somete una variable X a una transformación lineal :

$$Y = aX + b \Rightarrow \bar{y} = a\bar{x} + b$$

- La media aritmética de los cuadrados de las desviaciones respecto a la media aritmética es mínima:

$$\sum_{i=1}^k f_i (x_i - \bar{x})^2 \leq \sum_{i=1}^k f_i (x_i - a)^2 \quad \forall a \neq \bar{x}$$

- **Media geométrica:** Se usa cuando se desea promediar datos con efectos multiplicativos y se calcula como la raíz n -ésima del producto de los n valores de la distribución.

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

Se puede apreciar que si una modalidad es 0 no tiene sentido calcularla

- **Media armónica:** Para promediar datos que son cocientes de dos magnitudes. (por ejemplo la velocidad $v = \frac{\Delta x}{\Delta t}$)

$$H = \frac{n}{\frac{n_1}{x_1} + \frac{n_2}{x_2} + \cdots + \frac{n_k}{x_k}}$$

- **Media cuadrática:** Su uso se reduce al promedio de superficies.

$$Q = \sqrt{\sum_{i=1}^k f_i x_i^2}$$

Cada variable solo tiene una media, pero algunos libros plantean la siguiente desigualdad : $G \leq H \leq \bar{x} \leq Q$ (Solo se puede usar para acotar)

Características unidimensionales

- Mediana de una distribución: Valor que divide a los individuos de la población en dos efectivos iguales (Suponiendo el conjunto de datos ordenados, M_e lo divide en dos partes iguales)
 - Para variables discretas: Se busca el primer valor cuya frecuencia acumulada sea mayor o igual a $\frac{n}{2}$.
 - $N_i > \frac{n}{2} \Rightarrow M_e = x_i$
o'
 $N_i' > \frac{n}{2} \text{ en } [x_i, x_{i+1}) \Rightarrow M_e = \bar{x} = \frac{x_i + x_{i+1}}{2}$
- Para variables continuas: Equivale a calcular el percentil 50 como viene en el dibujo del final de la página.

Propiedad: La desviación absoluta media respecto a la mediana es mínima.

- Moda: Valor de mayor frecuencia "Está de moda" ☺
 - Para variables discretas: Valor de mayor frecuencia
 - Para variables continuas: Está en el denominado intervalo modal, el de mayor densidad de frecuencia (Mayor altura del histograma)
- Percentiles: El percentil de orden r ($r=1, \dots, 100$) es un valor, P_r , que divide al conjunto ordenada de datos en dos partes, tales que el $r\%$ del total son inferiores o iguales a P_r .
 - Variables discretas: $\frac{x_i}{N_{i-1}} < \frac{nr}{100} \leq \frac{x_i}{N_i} \Rightarrow P_r = x_i$ (si es un intervalo de clase se toma la marca de clase)
 - Para variables continuas: Usamos el esquema de abajo.

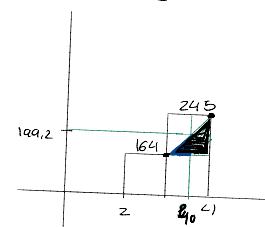
Los cuartiles Q_1, Q_2, Q_3 equivalen a los percentiles de orden 25, 50, 75 y los deciles a los de orden 10, 20, ..., 90

La mejor forma de calcular los percentiles es razonando

$$\frac{\text{base } 1}{P_{40} - 3} = \frac{\text{altura } 1}{199,2 - 164}$$

$$\frac{4-3}{P_{40} - 3} = \frac{245 - 164}{245 - 164}$$

$$\text{base } 2 = \text{altura } 2$$



Características unidimensionales

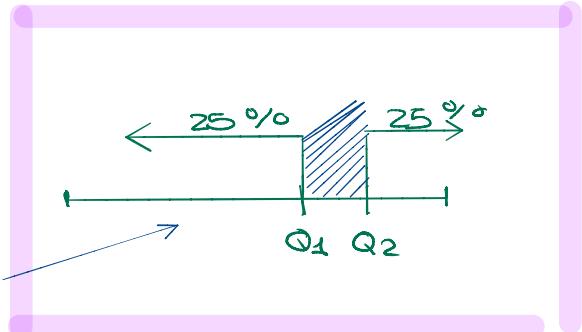
- Recorrido o rango: Distancia entre la primera y última modalidad del carácter (previamente las modalidades han sido ordenadas)

$$R = x_k - x_1$$

- Recorrido intercuartílico:

$$R_I = Q_3 - Q_1$$

Localiza el 50 % de la distribución



- Desviación absoluta respecto a la media \bar{x} :

$$D_{\bar{x}} = \frac{\sum_{i=1}^k |x_i - \bar{x}| n_i}{n}$$

Es muy útil para estudiar la representatividad del promedio

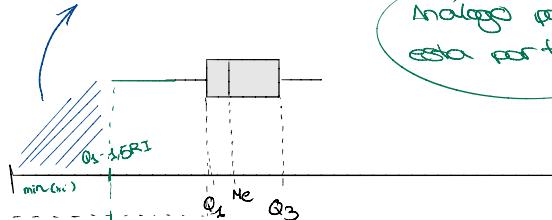
- Desviación absoluta respecto a la mediana M_e :

$$D_{M_e} = \frac{\sum_{i=1}^k |x_i - M_e| n_i}{n}$$

Grafica box & Whisker

Outlier - Datos atípicos → datos que habría que ver si hay que distribuir la distribución

Análogo para esta parte



$$\max \left\{ \min(x_i), Q_1 - 1.5 R_I \right\}$$

- Varianza:

$$\text{Var}(x) = \sigma^2 = \frac{\sum_{i=1}^k n_i (x_i - \bar{x})^2}{n}$$

No puede aparecer $\text{Var}(x) = 0$, ya que eso supone que solo hay una modalidad ' x_1 ', lo cual, con \oplus no tiene sentido

$$\begin{aligned} \text{Var}(x) &= \frac{1}{n} \sum_{i=1}^k n_i x_i^2 + \frac{1}{n} \sum_{i=1}^k n_i \bar{x}^2 - \frac{2}{n} \sum_{i=1}^k n_i x_i \bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i^2 + \bar{x}^2 \sum_{i=1}^k \frac{n_i}{n} + 2\bar{x} \sum_{i=1}^k \frac{n_i x_i}{n} = \\ &= \frac{1}{n} \sum_{i=1}^k (n_i x_i^2) - \bar{x} \end{aligned}$$

- Como la varianza es una suma de cuadrados solo puede ser positiva.

- La varianza es la medida cuadrática de la dispersión óptima

$$\frac{\sum_{i=1}^k n_i (x_i - \bar{x})^2}{n} < \frac{\sum_{i=1}^k n_i (x_i - b)^2}{n} \quad \forall b \neq \bar{x}$$

- La varianza está acotada superior e inferiormente

La varianza es una medida de cuadrados. Pensemos en la medida aritmética

$$x_1 < \bar{x} < x_k \Rightarrow \min(x_i) < \bar{x} < \max(x_k). \text{ Sabiendo esto, } \min(x_i - \bar{x})^2 < \bar{v}_x^2 < \max(x_i - \bar{x})^2$$

$$\text{Además, } +\sqrt{\min(x_i - \bar{x})^2} < \bar{v} < +\sqrt{\max(x_i - \bar{x})^2}$$

- Si se somete a transformaciones afines tenemos que:

$$y_i = a x_i + b \quad i = 1, \dots, k \Rightarrow \bar{v}_y^2 = a^2 \bar{v}_x^2$$

Medidas de dispersión absolutas:

• Desviación típica: La raíz cuadrada positiva de la varianza

$$\bar{v} = +\sqrt{\bar{v}^2}$$

- No puede ser negativa

- Acotada inferior y superiormente.

- Es una medida de dispersión óptima

$$\left(\bar{v} < \sqrt{\sum_{i=1}^k f_i (x_i - b)^2} \quad \forall b \neq \bar{x} \right)$$

Medidas de dispersión relativas:

• Coeficiente de apertura: Cociente entre los extremos de una distribución ordenada.

$$CA = \frac{x_k}{x_1}$$

• Recorrido relativo: Cociente entre el recorrido y la medida aritmética

$$R_R = \frac{R}{\bar{x}} = \frac{x_k - x_1}{\bar{x}}$$

• Recorrido semi-intercuantílico: Cociente entre el recorrido intercuantílico y la suma del primer y tercer cuartil

$$RSI = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

- Coeficiente de variación de Pearson: Relación entre la desviación típica y la media

$$CV(x) = \frac{\sigma_x}{\bar{x}}$$

- Índice de dispersión respecto a la mediana: Cociente entre la desviación absoluta media respecto la mediana:

$$V_{Me} = \frac{DMe}{Me}$$

Momentos

Sea r un número entero y positivo. Se llama momento de orden r respecto al valor " a " a la siguiente cantidad:

$$amr = \sum_{i=1}^n f_i (x_i - a)^r = \frac{1}{n} \sum_{i=1}^n n_i (x_i - a)^r$$

Según los valores de " a " hay dos clases de momentos:

- Momentos no centrales: $a = 0$ (se denota m_r)
- Momentos centrales: $a = \bar{x}$ (se denota μ_r)

$$m_0 = 1$$

$$m_1 = \bar{x}$$

$$\mu_0 = 1$$

$$\mu_1 = 0$$

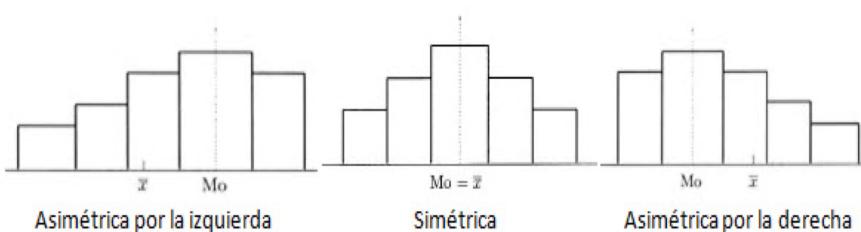
$$\mu_2 = \sigma_x^2$$

$$\sigma_x = \sqrt{m_2 - m_1^2}$$

Medidas de asimetría

Dada una variable estadística X , se entiende por asimetría de X a la falta de simetría respecto del eje vertical $x = \bar{x}$.

Diremos, pues, que una distribución es simétrica si la perpendicular que pasa por la media aritmética divide al diagrama diferencial (o histograma) en dos partes iguales. En caso contrario, la distribución es asimétrica.



• Coeficiente de asimetría de Fisher:

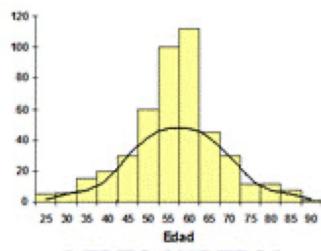
$$\gamma_1(x) = \frac{\mu_3}{\sqrt{3} s_x}$$

- Si $\gamma_1(x) > 0 \Rightarrow$ distribución asimétrica por la derecha o positiva
 - Si $\gamma_1(x) < 0 \Rightarrow$ distribución asimétrica por la izquierda o negativa.
 - Si $\gamma_1(x) = 0 \Rightarrow$ distribución simétrica
- Coeficiente de asimetría de Pearson:

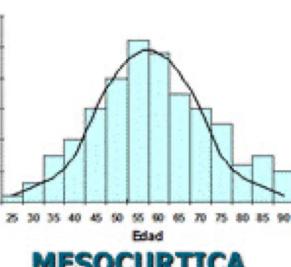
$$A_p = \frac{\bar{x} - \mu_0}{s_x} \quad (\text{misma interpretación})$$

Medidas de apuntamiento o curtosis

Miden la concentración de frecuencias de una distribución respecto a la que presenta una distribución "Normal" de su misma media y desviación típica.



LEPTOCURTICA



MESOCURTICA



PLATICURTICA

• Coeficiente de curtosis de Fisher:

$$\gamma_2(x) = \frac{\mu_4}{s^4} - 3$$

- $\gamma_2(x) < 0 \Rightarrow$ Platicúrtica
- $\gamma_2(x) = 0 \Rightarrow$ Mesocúrtica
- $\gamma_2(x) > 0 \Rightarrow$ Leptocúrtica

• Coeficientes de curtosis de Kelley:

$$K = \frac{1}{2} \frac{Q_3 - Q_1}{D_x - D_x} - 0,263$$

(misma interpretación)