

## **TEMA 8 - ESTADÍSTICA**

### **8.1 NOCIONES GENERALES DE ESTADÍSTICA**

#### **8.1.1 INTRODUCCIÓN**

**Objetivo:** La estadística tiene por objeto el desarrollo de técnicas para el conocimiento numérico de un conjunto de datos empíricos (recogidos mediante experimentos o encuestas).

**Población:** Es el conjunto de todos los elementos cuyo conocimiento nos interesa y serán objetos de nuestro estudio.

**Muestra:** Es un subconjunto, extraído de la población, cuyo estudio sirve para inferir características de toda la población.

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

**Caracteres y variables:** Caracteres son los aspectos que deseamos estudiar en los individuos de una población. Cada carácter puede tomar distintos valores o modalidades. Una variable estadística recorre todos los valores de un cierto carácter.

#### **Clasificación de las variables estadísticas:**

- *Cualitativas:* No toman valores numéricos
- *Cuantitativas discretas:* Toman valores numéricos aislados
- *Cuantitativas continuas:* Pueden tomar todos los valores de un intervalo.

#### **8.1.2 DOS RAMAS DE LA ESTADÍSTICA**

- **La estadística descriptiva:** Trata de “describir” y analizar algunos caracteres de los individuos de un grupo dado, sin extraer conclusiones para un grupo mayor. Para este estudio, se siguen estos pasos:
  - Selección de caracteres dignos de estudio
  - Análisis de cada carácter, anotando los valores que toman los individuos en ellos.
  - Clasificación y organización en tablas de los resultados obtenidos.
  - Cálculo de ciertos valores numéricos (los parámetros estadísticos) a partir de los datos obtenidos.
- **La estadística inferencial:** Trabaja con muestras y pretende, a partir de ellas, “inferir” características de toda la población. Es decir, se pretende tomar como generales propiedades que solo se han verificado para casos particulares. En ese proceso hay que operar con mucha cautela: ¿Cómo se elige la muestra?, ¿Qué grado de confianza se puede tener en el resultado obtenido? (No la estudiaremos este año)

## 8.2 TABLAS DE FRECUENCIAS

### 8.2.1 DEFINICIÓN

Las **tablas de frecuencias** sirven para ordenar y organizar los datos estadísticos. Con ellas, una masa amorfa de datos pasa a ser una colección ordenada y perfectamente inteligible.

Con los datos se construye la tabla de frecuencias:

- En la primera columna, la variable  $x_i$ , con todos sus posibles valores
- En la segunda columna, la correspondiente frecuencia,  $f_i$ : número de veces que aparece cada valor.

$x_i$	$f_i$

### 8.2.2 FRECUENCIAS RELATIVAS

Cuando se desea comparar varias distribuciones similares con distinto número de elementos, se debe recurrir a las **frecuencias relativas**. Estas vienen dadas en “tanto por uno” ( $f_r$ ) o en “tantos por ciento” (%). Si  $N$  es el número de individuos:  $f_r = \frac{f}{N}$       % =  $100 \cdot f_r = \frac{100 \cdot f}{N}$

### 8.2.3 FRECUENCIAS ACUMULADAS

En una distribución de frecuencias, se llama **frecuencia acumulada**,  $F_i$ , correspondiente al valor  $i$ -ésimo,  $x_i$ , a la suma de la frecuencia de ese valor con todas las anteriores:  $F_i = f_1 + f_2 + \dots + f_i$

Análogamente se puede definir frecuencia relativa acumulada o porcentaje acumulado.

### 8.2.4 TABLAS CON DATOS AGRUPADOS

Cuando en una distribución estadística el número de valores que toma la variable es muy grande, conviene elaborar una tabla de frecuencias agrupándolos en intervalos. Para ello:

- Se localizan los valores extremos,  $a$  y  $b$ , y se halla su diferencia,  $r = b - a$
- Se decide el número de intervalos que se quiere formar, teniendo en cuenta la cantidad de datos que se poseen. El número de intervalos no debe ser inferior a 6 ni superior a 15.
- Se toma un intervalo,  $r'$ , de longitud algo mayor que el recorrido  $r$  y que sea múltiplo del número de intervalos, con objeto de que estos tengan una longitud entera.
- Se forman los intervalos de modo que el extremo inferior del primero sea algo menor que  $a$  y el extremo superior del último sea algo superior a  $b$ . Es deseable que los extremos de los intervalos no coincidan con ninguno de los datos. Para ello, puede convenir que los extremos de los intervalos tengan una cifra decimal más que los datos.

El punto medio de cada intervalo se llama **marca de clase**. Es el valor que representa a todo el intervalo para el cálculo de algunos parámetros.

Cuando se elabora una tabla con datos agrupados, se pierde algo de información (pues en ella se ignora cada valor concreto, que se difumina dentro de un intervalo). A cambio, se gana en claridad y eficacia.

## 8.3 GRÁFICOS ESTADÍSTICOS

### 8.3.1 GRÁFICOS PARA VARIABLES CUALITATIVAS O CUANTITATIVAS DISCRETAS

#### Diagrama de barras:

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia:  $f$ ,  $f_r$  ó %
- Se levanta para cada valor de la X una barra que representa la frecuencia de dicho valor.

Si unimos mediante una poligonal los puntos más altos de cada barra obtenemos **el polígono de frecuencias**.

#### Diagrama de barras acumuladas:

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia acumulada:  $F$ ,  $F_r$  ó %a
- Se levanta para cada valor de la X una barra que representa la frecuencia acumulada de dicho valor.

Si unimos mediante una poligonal los puntos más altos de cada barra obtenemos **el polígono de frecuencias acumuladas**.

### 8.3.2 GRÁFICOS PARA VARIABLES CUANTITATIVAS CONTINUAS

#### SI TODOS LOS INTERVALOS TIENEN LA MISMA AMPLITUD

##### Histograma :

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia:  $f$ ,  $f_r$  ó %
- Se levanta para cada valor del intervalo de la X un rectángulo de altura la frecuencia de dicho intervalo.

Si unimos mediante una poligonal los puntos medios de cada uno de dichos rectángulos **el polígono de frecuencias**.

#### Diagrama de barras acumuladas:

- En el eje de las X : Se representan los valores de la variable
- En el eje de las Y : Se representan los valores de la frecuencia acumulada:  $F$ ,  $F_r$  ó %a
- Se levanta para cada valor del intervalo de la X un rectángulo de altura la frecuencia acumulada de dicho valor.

Si unimos mediante una poligonal las diagonales de dichos rectángulos obtenemos **el polígono de frecuencias acumuladas**.

#### SI LOS INTERVALOS NO SON TODOS DE LA MISMA AMPLITUD

En el eje de las Y: En vez de representar la frecuencia se representa la densidad de frecuencia :  $d_i = f_i/a_i$  siendo  $a_i$  la amplitud de dicho intervalo, para que así la frecuencia coincida con el área del rectángulo.

### 8.3.3 DIAGRAMAS DE SECTORES

Se dibuja un círculo y los porcentajes correspondientes a cada valor.

## 8.4 PARÁMETROS ESTADÍSTICOS

Los parámetros estadísticos sirven para sintetizar la información dada por una tabla. Los hay de dos tipos: de **centralización** y de **dispersión**.

Los **parámetros de centralización** nos indican en torno a qué valor (centro) se distribuyen los datos. Son de este tipo la media, la mediana y la moda.

Los **parámetros de dispersión** informan sobre cuánto se alejan del centro, los valores de la distribución. La mayoría de los parámetros de dispersión están asociados con un parámetro central al cual complementa:

- Mediana: Me  $\Rightarrow$  Cuartiles ( $Q_1$  y  $Q_3$ ) y recorrido intercuartílico ( $Q_3 - Q_1$ )
- Media:  $\bar{x}$   $\Rightarrow$  Desviación media (D.M.), varianza ( $\sigma^2$ ), desviación típica ( $\sigma$ ), coeficiente de variación (C.V.)

El recorrido (diferencia entre los valores extremos) es una medida de dispersión asociable tanto a la mediana como a la media.

Para que la información sea eficaz, hay que escoger parámetros que se complementen, por ejemplo, la media y la desviación típica.

Las definiciones siguientes sirven tanto para datos aislados como para datos agrupados en intervalos:

- Si los datos son aislados: los  $x_i$  son los valores que toma la variable
- Si los datos están agrupados en intervalos: los  $x_i$  son las marcas de clase.

### 8.4.1 MEDIA

$$\bar{x} = \frac{\sum f_i \cdot x_i}{\sum f_i} = \frac{\sum f_i \cdot x_i}{N}$$

Es el centro de gravedad de la distribución. Es decir, si las barras tuvieran peso, la media es el punto donde habría que sostener la tablilla en que se sitúan. Análogamente ocurre con las distribuciones dadas mediante histogramas.

### 8.4.2 VARIANZA

$$\text{Var} = \sigma^2 = \frac{\sum f_i \cdot (x_i - \bar{x})^2}{N} = \frac{\sum f_i \cdot x_i^2}{N} - \bar{x}^2$$

### 8.4.3 DESVIACIÓN TÍPICA

$\sigma = \sqrt{\text{Varianza}}$  Nos dice cómo de alejados de la media, cómo de dispersos, se encuentran los datos.

### 8.4.4 COEFICIENTE DE VARIACIÓN

$\text{C.V.} = \frac{\sigma}{\bar{x}}$  Sirve para comparar las dispersiones de poblaciones heterogéneas, pues indica la variación relativa.

## 8.5 PARÁMETROS DE POSICIÓN PARA DATOS AISLADOS

### 8.5.1 MEDIANA

Si los individuos de una población están colocados en orden creciente según la variable que se estudia y partimos la población en cuatro trozos con el mismo número de individuos, los puntos de separación son los cuarteles y la mediana:

**1º Cuartil o cuartil inferior:**  $Q_1$ , es el valor de la variable que supera al 25% de la población y, por tanto, queda por debajo del otro 75%.

La **mediana**,  $Me$ , es el valor de la variable que supera al 50% de los individuos.

**3º Cuartil o cuartil superior:** supera al 75% de los individuos. Se llama tercer cuartil porque la mediana es el segundo cuartil.

### 8.5.2 CENTILES O PERCENTILES

Si partimos la población en 100 partes y señalamos el lugar que deja debajo  $k$  de ellas, el valor de la variable correspondiente a ese lugar se designa por  $p_k$  y se denomina centil  $k$  o percentil  $k$ .

La mediana es  $Me = p_{50}$  y los cuartiles  $Q_1 = p_{25}$ ,  $Q_3 = p_{75}$ .

Si en vez de dividir en 100 partes dividimos sólo en 10, obtenemos los deciles:  $D_2 = p_{20}$

### 8.5.3 OBTENCIÓN PERCENTILES (mediana y cuarteles) EN TABLAS DE FRECUENCIAS

Para hallar el percentil  $p_k$  en una tabla de frecuencias, se obtienen los porcentajes acumulados. El percentil  $p_k$  es el valor para el cual la frecuencia acumulada correspondiente supera el  $k\%$ .

En el caso de que una de ellas coincida con  $k\%$ , se toma como  $p_k$  el valor intermedio entre ese valor de  $x$  y el siguiente.

## 8.6 MEDIDAS DE POSICIÓN PARA DATOS AGRUPADOS EN INTERVALOS

### 8.6.1 INTRODUCCIÓN

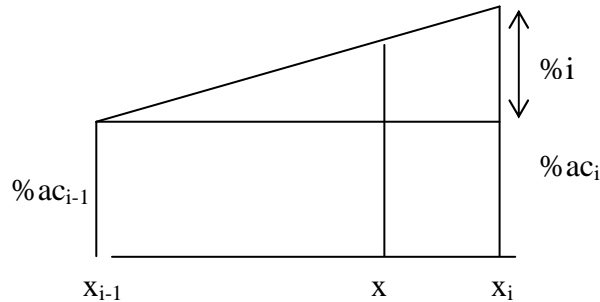
En las tablas de frecuencias con datos agrupados en intervalos se ha perdido el valor concreto de cada individuo. ¿Cómo saber, pues, dónde está la mediana o el percentil 20?

Teniendo en cuenta el convenio: En una tabla de frecuencias con datos agrupados en intervalos, suponemos que los datos de cada intervalo se reparten uniformemente en él.

Según esto, los valores de las frecuencias acumuladas deben asignarse a los extremos superiores de los intervalos, pues es al final de cada intervalo cuando se han contabilizado todos los individuos.

### 8.6.2 CÁLCULO

Procedemos como si los datos no estuviesen agrupados para hallar el intervalo correspondiente, y teniendo en cuenta el **polígono de porcentajes acumuladas** en dicho intervalo:



Aplicando semejanza de triángulos, obtendremos “x”.

## 8.7 INTERPRETACIÓN DE LAS MEDIDAS DE POSICIÓN. DIAGRAMA DE CAJA

El significado de cada una de las medidas de posición es muy claro: separa a los individuos de la población en dos fracciones cuyas proporciones respecto del total (25%-75%, 50%-50%, 10%-90%) quedan explícitas.

Para la interpretación conjunta utilizaremos los diagramas de caja y bigotes

### 8.7.1 DIAGRAMA DE CAJA Y BIGOTES

Se construyen del siguiente modo:

- La caja abarca el intervalo  $Q_1, Q_3$  (llamado recorrido intercuartílico) y en ella se señala expresamente el valor de la Mediana,  $Me$ .
- Los bigotes se trazan hasta abarcar la totalidad de los individuos, con la condición de que cada lado no se alargue más de una vez y media la longitud de la caja.
- Si uno (o más) de los individuos quedara por debajo o por encima de esta longitud, el correspondiente bigote se dibujará con esa limitación y se añadiría, mediante asterisco, el individuo en el lugar que le corresponde.

