UNIDAD 13.- **Distribuciones unidimensionales. Parámetros**

(tema 13 del libro)

1. **PARÁMETROS DE CENTRALIZACIÓN**

Aunque las tablas estadísticas y las representaciones gráficas contienen toda la información relativa a un problema, muchas veces interesa simplificar ese conjunto de datos por uno o varios parámetros que caractericen de la mejor forma posible esa distribución de frecuencias y que, además nos permita comparar unas distribuciones con otras. En este sentido hay unos *parámetros de centralización*, que tienden a situarse en el centro de la distribución, unos *parámetros de dispersión* cuyo valor indica si los datos están concentrados o dispersos alrededor de un valor prefijado; y unos parámetros de posición que tienden a situarse en un determinado lugar de la distribución.

a) Media aritmética

La media aritmética de una variable estadística es el cociente que resulta de dividir la suma de todos los valores por el nº total de éstos. Se representa por $\bar{x}$ ó por μ

Si los datos de la variable no vienen con frecuencias, la media se obtiene así:

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Si los datos de la variable están con sus frecuencias absolutas, entonces:

$$\bar{x} = \frac{x_1 f_1 + x_2 f_2 + \ldots + x_n f_n}{f_1 + f_2 + \ldots + f_n} = \frac{\sum_{i=1}^{n} x_i f_i}{\sum_{i=1}^{n} f_i} = \frac{\sum_{i=1}^{n} x_i f_i}{N}$$

Cuando tenemos datos agrupados en intervalos, consideraremos como valor de variable xi al punto medio de cada intervalo, es decir, la **marca de clase**. El valor calculado, evidentemente no es el valor real de la media, pero compensa con la reducción de operaciones que hay que realizar. Además si los datos dentro del intervalo están distribuidos de un modo más o menos uniforme la media calculada se aproxima mucho a la real.

Ventajas:

- La media es el valor medio o promedio de las observaciones.
- La media es el parámetro de centralización más utilizado
- Es un valor situado entre los valores extremos de la variable.
- Su cálculo sólo tiene sentido cuando la variable es cuantitativa.
- Presenta rigor matemático
- Es sensible a cualquier cambio en los datos

Desventajas:

- No siempre es posible calcular la media e incluso a veces ésta carece de significado. En estos casos se utilizan otras medidas de centralización.
- Es sensible a los valores extremos
- No es recomendable emplearla en distribuciones muy asimétricas
- Si se emplean variables discretas o cuasi-cualitativas, la media aritmética puede no pertenecer al conjunto de valores de la variable

Ejemplo: Preguntamos a 20 alumnos el nº de miembros de su familia y sus respuestas fueron:

3, 5, 4, 3, 5, 6, 8, 3, 3, 5, 7, 5, 6, 5, 4, 4, 7, 4, 5, 3

Realizamos una tabla de frecuencias:

Miembros por familia x_i	Frecuencia f_i	Frecuencia acumulada	$x_i \cdot f_i$
3	5	5	15
4	4	9	16
5	6	15	30
6	2	17	12
7	2	19	14
8	1	20	8
	20		95

Entonces tenemos que $\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{95}{20} = 4{,}75$

Ejemplo: A los 100 empleados de una empresa de piezas de precisión, se les ha realizado una prueba de habilidad manual. En una escala de 0 a 100 se han obtenido las siguientes puntuaciones:

27, 66, 32, 55, 46, 37, 75, 81, 18, 33, 47, 74, 37, 52, 47, 66, 80, 87, 37, 29,
46, 15, 29, 90, 76, 67, 23, 35, 94, 23, 25, 56, 73, 78, 17, 28, 76, 58, 45, 36,
55, 60, 17, 56, 23, 82, 64, 50, 51, 45, 37, 65, 62, 26, 69, 36, 54, 42, 40, 54,
27, 62, 28, 65, 46, 92, 36, 33, 23, 66, 18, 82, 47, 49, 59, 45, 73, 43, 47, 83,
78, 65, 39, 36, 53, 91, 38, 35, 68, 78, 91, 23, 34, 43, 55, 56, 74, 56, 62, 38.

Observamos que los valores extremos son 15 y 94. La amplitud total entre los datos es de 80 puntos, ya que ambas puntuaciones están incluidas.

Agruparemos los datos en 8 intervalos de amplitud 10:

$(14{,}24]$, $(24{,}34]$, …, $(84{,}94]$. Realizando el recuento con atención, se obtiene la tabla que sigue:

Habilidad manual	**Marca de clase** x_i	**Frecuencias** f_i	$x_i \cdot f_i$
$(14{,}24]$	19	10	190
$(24{,}34]$	29	12	348
$(34{,}44]$	39	17	663
$(44{,}54]$	49	18	882
$(54{,}64]$	59	13	767
$(64{,}74]$	69	13	897
$(74{,}84]$	79	11	869
$(84{,}94]$	89	6	534
		100	5150

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N} = \frac{5150}{100} = 51{,}50$$

b) Moda

La moda *Mo* es el dato que más se repite, es decir el valor de la variable con mayor frecuencia absoluta. Es la única medida de centralización que tiene sentido estudiar en una variable cualitativa, pues no precisa la realización de ningún cálculo. La moda no tiene por qué ser única, sino que puede haber distribuciones multimodales.

Si los datos están agrupados en intervalos elegimos el intervalo modal, que es aquel con mayor frecuencia absoluta. Si se desea mayor precisión en el cálculo de la moda a partir del intervalo modal se aplica la siguiente fórmula:

$$M_0 = L_i + \frac{f_{M_0} - f_{M_0-1}}{\left(f_{M_0} - f_{M_0-1}\right)+\left(f_{M_0} - f_{M_0+1}\right)} \cdot c \text{ , donde:}$$

- L_i es el límite inferior de la clase modal
- c es la amplitud del intervalo modal
- f_{M_0}, f_{M_0-1} y f_{M_0+1} son, respectivamente, las frecuencias absolutas de la clase modal, la clase anterior y la posterior.

Ejemplo: Se ha aplicado un test a los empleados de una fábrica, obteniéndose la siguiente tabla:

x	$(38,44]$	$(44,50]$	$(50,56]$	$(56,62]$	$(62,68]$	$(68,74]$	$(74,80]$
Nº trabajadores	7	8	15	25	18	9	6

Calcular la clase modal y su moda.

Como vemos la clase modal es $(56,62]$, y para la moda usamos la fórmula dada anteriormente:

$L_i = 56$ $\quad c = 6$ $\quad f_{M_0} = 25$ $\quad f_{M_0-1} = 15$ $\quad f_{M_0+1} = 18$

Luego, $M_0 = L_i + \frac{f_{M_0} - f_{M_0-1}}{\left(f_{M_0} - f_{M_0-1}\right)+\left(f_{M_0} - f_{M_0+1}\right)} \cdot c = 56 + \frac{25-15}{(25-15)+(25-18)} \cdot 6 = 56 + \frac{10}{17} \cdot 6 = 59{,}53$

c) Mediana

La mediana de una distribución es un valor M_e que divide a la distribución en dos partes iguales; es decir, deja tantas observaciones a la izquierda como a la derecha.

- Para calcular la mediana en caso de pocos datos y sin agrupar se colocan estos en orden creciente de magnitud. Si el número de datos es impar la mediana coincide con el valor central. Si el número de datos es par, cualquier valor comprendido entre los dos valores centrales es una mediana, pero se suele tomar el valor medio de los dos valores centrales.

  Ejemplo: Calcula la mediana de los siguientes datos: 6, 4, 3, 2, 8, 6, 5, 6, 7, 3.
  Tenemos que ordenarlos de menor a mayor: 2, 3, 3, 4, 5, 6, 6, 6, 7, 8
  El nº de datos es N = 10, por tanto, la mitad de los datos es 5. Tomamos los datos quinto y sexto y hacemos la media aritmética: $M_e = \frac{5+6}{2} = 5{,}5$

- Si tenemos muchos datos y sin agrupar, se construye la tabla de frecuencias acumuladas F_i, y se toma la mediana como aquel valor de la variable x_i para el cual F_i sea igual o supere $\frac{N}{2}$

  Ejemplo: Preguntamos a 20 alumnos el número de miembros de su familia, y sus respuestas fueron:
  3, 5, 4, 3, 5, 6, 8, 3, 3, 5, 7, 5, 6, 5, 4, 4, 7, 4, 5, 3
  Construimos la tabla de frecuencias absolutas y acumuladas

x_i	f_i	F_i
3	5	5
4	4	9
5	6	15
6	2	17
7	2	19
8	1	**20**

Tenemos que $\frac{N}{2}=\frac{20}{2}=10$ Mirando en la columna de las frecuencias absolutas acumuladas, la que primero supera a 10 es la correspondiente a $x_i=5$, por tanto $M_e=5$

- En caso de datos agrupados en intervalos primero buscaremos el intervalo mediano, que es el primer intervalo de clase cuya frecuencia acumulada es igual o superior a la mitad del número de observaciones, $\frac{N}{2}$.

  Como primera aproximación puede tomarse la mediana como la marca de clase de dicho intervalo; sin embargo podemos calcularla de forma más exacta con la siguiente fórmula:

$$\mathrm{M_e}=\mathrm{L_i}+\frac{\frac{\mathrm{N}}{2}-\mathrm{F}_{\mathrm{M_e}-1}}{\mathrm{f}_{\mathrm{M_e}}}\cdot \mathrm{c}$$

Ejemplo: Se ha realizado un test de habilidad numérica a los alumnos de una clase. Los resultados obtenidos son:

Puntos	$[10,15)$	$[15,20)$	$[20,25)$	$[25,30)$	$[30,35)$	$[35,40)$	$[40,45)$	$[45,50)$
Nºde alumnos	4	6	6	10	8	10	3	3

Hacemos la tabla de frecuencias absolutas y acumuladas

Puntos	Marca de clase x_i	f_i	F_i
$[10,15)$	12,5	4	4
$[15,20)$	17,5	6	10
$[20,25)$	22,5	6	16
$[25,30)$	27,5	10	26
$[30,35)$	32,5	8	34
$[35,40)$	37,5	10	44
$[40,45)$	42,5	3	47
$[45,50)$	47,5	3	50

Tenemos $\frac{N}{2}=25$, así que la clase mediana es $[25,30)$, aplicando la fórmula:

$$\mathrm{M_e}=\mathrm{L_i}+\frac{\frac{\mathrm{N}}{2}-\mathrm{F}_{\mathrm{M_e}-1}}{\mathrm{f}_{\mathrm{M_e}}}\cdot \mathrm{c}=25+\frac{\frac{50}{2}-16}{10}\cdot 5=29{,}5$$

d) Percentiles

La mediana de los valores de una variable estadística divide a la distribución en dos partes iguales. Es decir, la mediana parte la distribución en dos mitades, cada una correspondiente al 50 %. Si generalizamos esta idea, se puede pensar en obtener valores que dividan a los datos en diversas partes iguales. Estos son los percentiles.

- *Cuartiles*

Hay tres cuartiles que dividen a los datos en 4 partes:

Q_1: (cuartil primero) Deja al 25 % de los datos a su izquierda, es decir, $\frac{N}{4}$. Se obtiene con la fórmula $Q_1 = L_i + \frac{\frac{N}{4} - F_{Q_1-1}}{f_{Q_1}} \cdot c$

$Q_2 = M_e$: (cuartil segundo) Deja al 50 % de los datos a su izquierda, es decir, $\frac{N}{2}$. Obviamente se trata de la mediana

Q_1: (cuartil tercero) Deja al 75 % de los datos a su izquierda, es decir, $\frac{3 \cdot N}{4}$. Se obtiene con la fórmula $Q_3 = L_i + \frac{\frac{3 \cdot N}{4} - F_{Q_3-1}}{f_{Q_3}} \cdot c$

- *Deciles*

Análogamente a los cuartiles, los deciles son nueve valores que dividen a los datos en diez partes iguales. Se notan por $D_1, D_2, ..., D_9$

Se calculan con la siguiente fórmula: $D_j = L_i + \frac{\frac{j \cdot N}{10} - F_{D_j-1}}{f_{D_j}} \cdot c$

- *Percentiles*

Lo mismo que los anteriores pero en cien partes, luego hay 99 percentiles. Se notan por $P_1, P_2, ..., P_{99}$

Se calculan con la siguiente fórmula: $P_j = L_i + \frac{\frac{j \cdot N}{100} - F_{P_j-1}}{f_{P_j}} \cdot c$

Ejemplo: Los varones que entre 20 y 60 años contrajeron matrimonio durante el año 1961 en España presentan la distribución por edades que muestra la tabla adjunta. Calcula la mediana, el cuartil tercero el percentil 37 y el orden del percentil que corresponde a un valor de 26,6

Edad	Nº de varones f_i	F_i
$[20,25)$	41000	41000
$[25,30)$	123000	164000
$[30,35)$	44000	208000
$[35,40)$	13000	221000
$[40,50)$	7000	228000
$[50,60)$	3000	231000

1) Tenemos que $\frac{N}{2}=\frac{231000}{2}=115500$ luego el intervalo de la mediana es $[25,30)$

La mediana es $M_e = L_i + \frac{\frac{N}{2} - F_{M_e-1}}{f_{M_e}} \cdot c = 25 + \frac{115500 - 41000}{123000} \cdot 5 = 28{,}03$

2) Para el cuartil tercero $\frac{3\cdot N}{4}=173250$, que le corresponde el intervalo $[30,35)$

$$Q_3 = L_i + \frac{\frac{3\cdot N}{4} - F_{Q_3-1}}{f_{Q_3}} \cdot c = 30 + \frac{173250 - 164000}{44000} \cdot 5 = 31{,}05$$

3) Para el percentil 37, tenemos $\frac{37\cdot N}{100}=85470$, que le corresponde el intervalo $[25,30)$

$$P_{37} = L_i + \frac{\frac{37\cdot N}{100} - F_{P_{37}-1}}{f_{P_{37}}} \cdot c = 25 + \frac{85470 - 41000}{123000} \cdot 5 = 26{,}81$$

4) Para calcular el percentil correspondiente a un valor de 26,6 , en la fórmula sustituimos teniendo en cuenta que de estar en el intervalo $[25,30)$

$25 + \frac{\frac{j\cdot 231000}{100} - 41000}{123000} \cdot 5 = 26{,}6$. De donde resulta que $j = 34{,}78 \approx 35$. Le corresponde el percentil 35

2. PARÁMETOS DE DISPERSIÓN

Las medidas de centralización representan bien a un conjunto de datos cuando están agrupados en torno a ellas, pero no cuando hay bastantes observaciones alejadas de ellas. Las medidas de dispersión miden, por tanto, el grado de alejamiento de los datos respecto a las medidas de centralización, fundamentalmente respecto de la media.

a) Recorrido

El recorrido de una distribución es la diferencia entre el dato mayor y el dato menor obtenidos al observar los valores de la variable. Se nota por $R = x_{máx} - x_{mín}$

A veces se usa el recorrido intercuartílico que es la diferencia entre el cuartil tercero y el primero

$$R_I = Q_3 - Q_1$$

b) Desviación media

Se llama desviación media de una serie de datos x_1, x_2, x_3, …, x_n, que tienen frecuencias f_1, f_2, f_3, …, f_n respectivamente, y se representa por ***DM***, a la media aritmética de los valores absolutos de las desviaciones respecto de la media, esto es:

$$DM = \frac{\sum_{i=1}^{n} f_i \left| x_i - \bar{x} \right|}{N}$$

c) Varianza

Se llama varianza de una serie de datos x_1, x_2, x_3, …, x_n, que tienen frecuencias f_1, f_2, f_3, …, f_n respectivamente, y se representa por σ^2 (o S^2), a la media aritmética de los cuadrados de las desviaciones respecto de la media, esto es:

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

También se puede usar esta otra fórmula que es equivalente y es más usada:

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

d) Desviación típica

Es la raíz cuadrada positiva de la varianza y se denota por σ (o S).

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2 \cdot f_i}{N}} \text{ o bien } \sigma = \sqrt{\frac{\sum_{i=1}^{n} {x_i}^2 \cdot f_i}{N} - \bar{x}^2}$$

La desviación típica es el parámetro más utilizado.
Si se suma una constante a todos los valores de la variable, la desviación típica no varía.
Si se multiplican todos los valores de la variable por un mismo número, la desviación típica queda multiplicada por el mismo número.

e) Coeficiente de variación

Se llama coeficiente de variación y se representa por CV al cociente entre la desviación típica y el valor absoluto de la media.

$$CV = \frac{\sigma}{\left| \bar{x} \right|}$$

El valor del coeficiente de variación suele expresarse en tanto por ciento.
Cuanto más pequeño sea este valor, los datos estarán más concentrados alrededor de la media, y esta por tanto será más representativa.

Propiedad: Si x e y son dos variables estadísticas cuyas medias son $\bar{x}$ y $\bar{y}$, y sus desviaciones típicas σ_x y σ_y, se tiene que:

- Si $\bar{x} = \bar{y}$ y $\sigma_x < \sigma_y$, entonces $\bar{x}$ es más representativa.
- Si $\bar{x} \neq \bar{y}$ y $\frac{\sigma_x}{|\bar{x}|} < \frac{\sigma_y}{|\bar{y}|}$, entonces $\bar{x}$ es más representativa.

Ejemplo: Tenemos la siguiente tabla de distribución de frecuencias:

X	(60, 76]	(76, 92]	(92, 108]	(108, 124]	(124, 140]	(140, 156]
Frecuencia	12	13	18	19	11	7

Calcula el recorrido, la desviación media, la varianza, la desviación típica y el coeficiente de variación

Construimos una tabla de frecuencias completa

X	Marca de clase x_i	f_i	F_i	$x_i \cdot f_i$	$x_i^2 \cdot f_i$	$\lvert x_i - \bar{x} \rvert \cdot f_i$
(60, 76]	68	12	12	816	55488	444
(76, 92]	84	13	25	1092	91728	273
(92, 108]	100	18	43	1800	180000	90
(108, 124]	116	19	62	2204	255664	209
(124, 140]	132	11	73	1452	191664	297
(140, 156]	148	7	80	1036	153328	301
Sumas		80		8400	927872	1614

La media aritmética es: $\bar{x} = \frac{8400}{80} = 105$

Recorrido: $R = 156 - 60 = 96$

Desviación media: $DM = \frac{1614}{80} = 20{,}18$

Varianza: $\sigma^2 = \frac{927872}{80} - (105)^2 = 573{,}4$

Desviación típica: $\sigma = \sqrt{573{,}4} = 23{,}95$

Coeficiente de variación: $CV = \frac{23{,}95}{105} = 0{,}228$