Tema 2. Medidas descriptivas de los datos

Resumen del tema

2.1. Medidas de posición

Son valores que nos sirven para indicar la posición alrededor de la cual se distribuyen las observaciones.

2.1.1. Mediana

La **mediana** es un valor que deja a su izquierda el 50 % de los datos de la muestra ordenada. La denotaremos por M_e. Su unidad de medida es la misma que la de la variable.

a) Cálculo con datos no agrupados en intervalos:

- n impar: M_e es el valor central de la muestra ordenada.
- n par: M_e es el punto medio de los dos valores centrales de la muestra ordenada.

b) Cálculo con datos agrupados en intervalos:

Intervalo mediano: es el que contiene a la mediana. Es el primer intervalo cuya frecuencia absoluta acumulada es igual o mayor que $\frac{n}{2}$.

$$M_e = \ell_i + \frac{\frac{n}{2} - F_{i-1}}{f_i} \left(\ell_{i+1} - \ell_i\right),$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo mediano, f_i es su frecuencia absoluta y F_{i-1} es la frecuencia absoluta acumulada del intervalo anterior al mediano.

2.1.2. Cuantiles o percentiles

El **cuantil** o **percentil** al r % es un valor que deja por debajo el r % de los datos de la muestra ordenada de menor a mayor. Lo denotaremos por C_r. Su unidad de medida es la misma que la de la variable.

CASOS PARTICULARES:

- Cuartiles:

$$\begin{aligned} 1^{er} \text{ cuartil} &= Q_1 = C_{25} \\ 2^{o} \text{ cuartil} &= Q_2 = C_{50} = M_e \\ 3^{er} \text{ cuartil} &= Q_3 = C_{75} \end{aligned}$$

- Deciles:

$$\begin{array}{rcl} 1^{er} \text{ decil} & = & D_1 = C_{10} \\ 2^{o} \text{ decil} & = & D_2 = C_{20} \\ \vdots & \vdots & \vdots \\ 9^{o} \text{ decil} & = & D_9 = C_{90} \end{array}$$

Si los datos están agrupados en intervalos de clase, el intervalo que contiene a C_r es el primero cuya frecuencia acumulada absoluta es igual o mayor que

$$\frac{n\,r}{100}$$

y el cuantil al $r\,\%$ se determina mediante la fórmula:

$$C_r = \ell_i + \frac{\dfrac{n\,r}{100} - F_{i-1}}{f_i}\,(\ell_{i+1} - \ell_i)\,,$$

donde $(\ell_i, \ell_{i+1}]$ es el intervalo que contiene a C_r, f_i es su frecuencia absoluta y F_{i-1} es la frecuencia absoluta acumulada del intervalo anterior.

2.1.3. Media

Llamaremos **media** a la **media aritmética**. (Hay otras medias, como, por ejemplo, la media geométrica, la media cuadrática y la media armónica.)

Si la variable se denota por X, la media de los datos de una muestra será denotada por $\overline{x}$. (Si tenemos los datos de toda la población, entonces representaremos la media por μ.)

a) **Cálculo con datos no agrupados en intervalos:**
Si $x_1, x_2, \ldots, x_n$ son los n valores de la muestra, entonces:

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}\,.$$

Si los datos son $x_1, x_2, \ldots, x_k$, y aparecen con frecuencias absolutas respectivas $f_1, f_2, \ldots, f_k$, entonces:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i f_i}{n}\,.$$

De las fórmulas anteriores se deduce que la unidad de medida de $\overline{x}$ es la misma que la de la variable.

b) **Cálculo con datos agrupados en intervalos:**
La fórmula es la misma que la anterior, siendo x_i la marca de clase del intervalo $(\ell_i, \ell_{i+1}]$ y f_i su correspondiente frecuencia absoluta.

2.2. Medidas de dispersión

Miden el grado de separación de las observaciones entre sí o con respecto a ciertas medidas de posición, como la media o la mediana.

2.2.1. Recorrido

La fórmula del **recorrido** es:

$$R = x_{max} - x_{min}\,.$$

De la fórmula anterior se deduce que la unidad de medida de R es la misma que la de la variable.

El recorrido nos mide el grado de variabilidad de los datos de la muestra: cuanto más grande sea el resultado del recorrido, más dispersos están los datos.

2.2.2. Recorrido intercuartílico

La fórmula del **recorrido intercuartílico** es:

$$R_I = Q_3 - Q_1 = C_{75} - C_{25}\,.$$

De la fórmula anterior se deduce que la unidad de medida de R_I es la misma que la de la variable.

Cuanto más pequeño sea el resultado del recorrido intercuartílico, menos dispersión respecto de la mediana hay; es decir, los datos están menos alejados de la mediana y, por tanto, la mediana es más representativa. Pero, ¿cuándo podríamos decir que el valor del recorrido intercuartílico es pequeño? …Como entre el primer cuartil, Q_1, y el tercer cuartil, Q_3, hay exactamente la mitad de los datos, podríamos comparar la mitad del recorrido total con el recorrido intercuartílico, y podríamos decir que la mediana es representativa si R_I es menor o igual que $R/2$.

2.2.3. Varianza y desviación típica

I) Varianza

Si la variable se denota por X, la varianza de los datos procedentes de una muestra será denotada por s_x^2. (Si disponemos de los datos de toda la población, entonces representaremos la varianza por σ^2.)

La fórmula de la **varianza** es:

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n}\,.$$

Una fórmula equivalente es:

$$s_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \overline{x}^2 = \frac{\sum_{i=1}^{k} x_i^2 f_i}{n} - \overline{x}^2\,.$$

De las fórmulas anteriores se deduce que unidad de medida de s_x^2 es la unidad de la variable elevada al cuadrado.

II) Desviación típica

Si la variable se denota por X, la desviación típica de los datos procedentes de una muestra será denotada por s_x. (Si disponemos de los datos de toda la población, entonces representaremos la desviación típica por σ.)

La fórmula de la **desviación típica** es:

$$s_x = \sqrt{\text{Varianza}}\,.$$

De la fórmula anterior se deduce que la unidad de medida de s_x es la misma que la de la variable.

Cuanto más pequeño sea el resultado de la desviación típica, menos dispersión respecto de la media hay; es decir, los datos están menos alejados de la media y, por tanto, la media es más representativa. Pero, ¿cuándo podríamos decir que el resultado de la desviación típica es pequeño? ... Como entre $\overline{x} - s$ y $\overline{x} + s$ hay, para la mayoría de las variables, más de las dos terceras partes de los datos, podríamos comparar la amplitud del intervalo $(\overline{x} - s, \overline{x} + s)$ con los dos tercios del recorrido total; es decir, podríamos comparar el resultado de $2\,s$ con el resultado de $2\,R/3$, lo que es lo mismo que comparar s con $R/3$. En consecuencia, podríamos decir que la media es representativa si s es menor o igual que $R/3$.

III) Cuasivarianza o varianza corregida

Se utiliza, sobre todo, en Estadística Inferencial.

Si la variable se denota por X, la cuasivarianza o varianza corregida de los datos procedentes de una muestra será denotada por S_x^2.

La fórmula de la **cuasivarianza** es:

$$S_x^2 = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{\displaystyle\sum_{i=1}^{k}(x_i - \overline{x})^2 f_i}{n-1}\,.$$

Una fórmula equivalente es:

$$S_x^2 = \frac{\left(\displaystyle\sum_{i=1}^{n} x_i^2\right) - n\overline{x}^2}{n-1} = \frac{\left(\displaystyle\sum_{i=1}^{k} x_i^2 f_i\right) - n\overline{x}^2}{n-1}\,.$$

De las fórmulas anteriores se deduce que unidad de medida de S_x^2 es la unidad de la variable elevada al cuadrado.

Relación entre la varianza y la cuasivarianza:

$$n\,s_x^2 = (n-1)\,S_x^2\,.$$

IV) Cuasidesviación típica o desviación típica corregida

Se utiliza, sobre todo, en Estadística Inferencial.

La fórmula de la **cuasidesviación típica** es:

$$S_x = \sqrt{\text{Cuasivarianza}}\,.$$

De la fórmula anterior se deduce que la unidad de medida de S_x es la misma que la de la variable.

Ejemplos que se van a resolver en clase

Ejemplo 2.1. Observamos la edad de 8 alumnos de clase y calculamos la mediana.

Ejemplo 2.2. Observamos la edad de 9 alumnos de clase y calculamos la mediana.

Ejemplo 2.3. La distribución de frecuencias de las calificaciones de 13 alumnos en un determinado examen viene dada por la tabla siguiente. Calcular la mediana.

Tabla 2.1		
x_i	f_i	F_i
2	2	2
4	3	5
6	5	10
8	3	13

Ejemplo 2.4. La distribución de frecuencias de las calificaciones de 12 alumnos en un determinado examen viene dada por la tabla siguiente. Calcular la mediana.

Tabla 2.2		
x_i	f_i	F_i
2	1	1
4	5	6
6	4	10
8	2	12

Ejemplo 2.5. En la tabla siguiente aparece el número de palabras por resumen de los artículos científicos de autores españoles que han publicado en una determinada revista de investigación durante un año concreto (datos del Problema 1.9). Calcular la mediana.

Tabla 2.3									
10	15	16	20	17	19	21	14	13	19
11	14	17	19	20	20	22	15	13	12
12	15	17	19	18	23	22	17	21	20
15	18	16	18	12	17	14	15	17	15

Ejemplo 2.6. En las columnas 1ª y 3ª de la siguiente tabla aparece la distribución de frecuencias de la altura, en metros, de una muestra de 15 alumnos. Los datos están agrupados en intervalos de la misma amplitud.

Tabla 2.4					
$(\ell_i, \ell_{i+1}]$	x_i	f_i	F_i	$x_i f_i$	$x_i^2 f_i$
$[1'60, 1'64]$	$1'62$	6	6	$9'72$	$15'7464$
$(1'64, 1'68]$	$1'66$	3	9	$4'98$	$8'2668$
$(1'68, 1'72]$	$1'70$	3	12	$5'10$	$8'6700$
$(1'72, 1'76]$	$1'74$	1	13	$1'74$	$3'0276$
$(1'76, 1'80]$	$1'78$	2	15	$3'56$	$6'3368$
suma		15		$25'1$	$42'0476$

a) Calcular el valor aproximado de la mediana a partir del gráfico de frecuencias acumuladas absolutas.

b) Calcular la mediana mediante la fórmula.

Ejemplo 2.7. Con los datos de la Tabla 2.3 calcular: el primer decil, el primer cuartil, el tercer cuartil y el noveno decil.

Ejemplo 2.8. Con los datos de la Tabla 2.4 calcular el primer y el tercer cuartil.

Ejemplo 2.9. Calcular la media de los datos de la Tabla 2.3.

Ejemplo 2.10. Calcular la media de los datos de la Tabla 2.4.

Ejemplo 2.11. ¿Cuál es el grado de dispersión de los datos de la Tabla 2.3? Razonar la respuesta.

Ejemplo 2.12. ¿Cuál es el grado de dispersión de los datos de la Tabla 2.4? Razonar la respuesta.

Ejemplo 2.13. Con los datos de la Tabla 2.3 ¿cuál es el grado de representatividad de la mediana: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.14. Con los datos de la Tabla 2.4 ¿cuál es el grado de representatividad de la mediana: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.15. Con los datos de la Tabla 2.3 ¿cuál es el grado de representatividad de la media: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Ejemplo 2.16. Con los datos de la Tabla 2.4 ¿cuál es el grado de representatividad de la media: muy fuerte, fuerte, regular, débil o muy débil? Razonar la respuesta.

Problemas propuestos

Problema 2.1. Se preguntó a varias personas, elegidas al azar, el número de periódicos distintos que leían trimestralmente, y se obtuvo las siguientes respuestas:

Nº de periódicos	0	1	2	3	4	5	6	7
Nº de lectores	7	13	18	15	11	6	4	2

a) Dibujar el gráfico de frecuencias acumuladas absolutas. Calcular la mediana.

b) ¿Cuál es el grado de representatividad de la mediana: muy poco representativa, poco, regular, bastante o muy representativa?

Problema 2.2. El número de personas que visitan diariamente una biblioteca fue observado durante 74 días elegidos al azar, y los resultados fueron:

Nº de personas	47	59	62	64	71	76	78	80
Nº de días	4	6	10	17	16	10	7	4

a) Hallar la media y la mediana.

b) Calcular la medida de dispersión adecuada para medir el grado de representatividad de la media. Interpretar su resultado.

c) Calcular la medida de dispersión adecuada para medir el grado de representatividad de la mediana. Interpretar su resultado.

Problema 2.3. La edad de las personas que aprobaron la oposición de auxiliar de biblioteca en España en un determinado año tiene la siguiente distribución:

Edad	[20,25]	(25,30]	(30,35]	(35,40]	(40,50]	(50,60]
Nº de personas	41	123	44	13	7	3

a) Dibujar el gráfico de frecuencias acumuladas absolutas. A partir de este gráfico, determinar el valor aproximado de la mediana. Determinar, después, el valor de la mediana con la fórmula estudiada.

b) ¿Cuál es el grado de representatividad de la mediana? Justificar la respuesta.

Problema 2.4. Los siguientes datos corresponden al número mensual de nuevos socios de una determinada biblioteca:

27	40	12	3	30	16	20	21	30	12
45	18	25	22	35	24	37	12	21	7
35	17	21	27	14	15	25	45	12	24

a) Determinar la distribución de frecuencias y dibujar el polígono de frecuencias absolutas.

b) Calcular la media y la mediana.

Problema 2.5. El número de veces que fueron consultados 60 artículos de investigación archivados en una hemeroteca, durante un determinado año, viene dado por la siguiente tabla:

8	25	20	4	19	3	21	2	20	22
23	9	1	24	21	22	20	2	22	21
2	24	21	9	3	21	22	3	22	3
12	6	20	2	26	46	2	4	10	37
14	9	7	25	50	26	38	46	36	1
7	1	35	23	45	36	5	65	46	37

Agrupar los datos en intervalos de la misma amplitud, y calcular, a partir de esta clasificación, el valor de la medida de posición que resulte más representativa del conjunto total de los datos.

Problema 2.6. A continuación se ofrecen los datos correspondientes al tiempo de espera (en minutos) de 50 usuarios de una biblioteca hasta que son atendidos por algún miembro del personal de ésta.

1	3	5	20	21	4	7	9	10	12
20	18	6	4	13	11	10	13	15	9
4	20	2	22	8	6	11	4	8	6
5	18	19	20	7	15	16	13	12	14
7	10	5	24	11	8	9	10	11	7

a) Determinar la distribución de frecuencias. Calcular la media y la mediana.

b) Agrupar los datos en intervalos de distinta amplitud, y calcular, a partir de esta nueva clasificación, las mismas medidas descriptivas del apartado anterior. Comparar los resultados.

Soluciones de los problemas propuestos

Solución del problema 2.1. La distribución de frecuencias es:

x_i	f_i	F_i
0	7	7
1	13	20
2	18	38
3	15	53
4	11	64
5	6	70
6	4	74
7	2	76

a)
- *Gráfico de frecuencias acumuladas absolutas*: es la representación gráfica de las frecuencias acumuladas absolutas, F, para todo valor numérico, x. Es una gráfica en forma de "escalera".
- Mediana=$M_e = 2'5$ periódicos.

b) Como el recorrido intercuartílico es $R_I = 3$ periódicos y la mitad del recorrido es $R/2 = 3'5$ periódicos, entonces se cumple que R_I es un poco menor que $R/2$ y, como consecuencia, la mediana es bastante representativa.

Solución del problema 2.2.

a)
- Media=$\overline{x} = 67'7297$ personas.
- Mediana=$M_e = 67'5$ personas.

b) La desviación típica es $s_x = 8'1677$ personas. Como $R/3 = 11$, entonces se cumple que s_x es bastante menor que $R/3$ y, como consecuencia, la media es bastante representativa.

c) El recorrido intercuartílico es $R_I = 14$ personas. Como $R/2 = 16'5$, entonces R_I es bastante menor que $R/2$ y, como consecuencia, la mediana es bastante representativa.

Solución del problema 2.3.

a)
- *Gráfico de frecuencias acumuladas absolutas*: se sitúan los puntos que resultan de tomar en el eje horizontal los extremos superiores de los intervalos de clase, y en el eje vertical sus correspondientes frecuencias acumuladas absolutas, uniendo después dichos puntos mediante segmentos rectilíneos.
- A partir del gráfico anterior se deduce que la mediana es aproximadamente igual a 28 años.
- Con la fórmula se obtiene que la mediana es $M_e = 28'0285$ años.

b) El recorrido intercuartílico es $R_I = 5'37$ años. Como $R/2 = 20$ entonces R_I es mucho menor que $R/2$ y, como consecuencia, la mediana es muy representativa.

Solución del problema 2.4.

a)
- La distribución de frecuencias (conteniendo las columnas que posteriormente necesitaremos) es:

x_i	f_i	F_i	$x_i f_i$	$(x_i - \overline{x})^2 f_i$
3	1	1	3	$402'6711$
7	1	2	7	$258'1378$
12	4	6	48	$489'8844$
14	1	7	14	$82'2044$
15	1	8	15	$65'0711$
16	1	9	16	$49'9378$
17	1	10	17	$36'8044$
18	1	11	18	$25'6711$
20	1	12	20	$9'4044$
21	3	15	63	$12'8133$
22	1	16	22	$1'1378$
24	2	18	48	$1'7422$
25	2	20	50	$7'4756$
27	2	22	54	$30'9422$
30	2	24	60	$96'1422$
35	2	26	70	$284'8089$
37	1	27	37	$194'1378$
40	1	28	40	$286'7378$
45	2	30	90	$962'1422$
suma			692	$3297'8\widehat{6}$

- *Polígono de frecuencias absolutas*: se sitúan los puntos que resultan de tomar en el eje horizontal los distintos valores de la variable, x_i, y en el eje vertical sus correspondientes frecuencias absolutas, f_i, uniendo después los puntos mediante segmentos rectilíneos.

b)
- Media=$\overline{x} = 23'0\,\widehat{6}$ socios.
- Mediana=$M_e = 21'5$ socios.

Solución del problema 2.5. La distribución de frecuencias con datos agrupados en intervalos de la misma amplitud es:

$(\ell_i, \ell_{i+1}]$	x_i	f_i	F_i
$(0'8, 10]$	$5'4$	23	23
$(10, 19'2]$	$14'6$	3	26
$(19'2, 28'4]$	$23'8$	22	48
$(28'4, 37'6]$	$33'0$	5	53
$(37'6, 46'8]$	$42'2$	5	58
$(46'8, 56]$	$51'4$	1	59
$(56, 65'2]$	$60'6$	1	60

Como la dispersión es grande, la medida de posición más adecuada es la mediana. Con los datos agrupados en estos intervalos de clase, el valor de la mediana es $M_e = 20'8\,\widehat{72}$ veces.

Solución del problema 2.6.

a) La distribución de frecuencias es:

x_i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	18	19	20	21	22	24
f_i	1	1	1	4	3	3	4	3	3	4	4	2	3	1	2	1	2	1	4	1	1	1
F_i	1	2	3	7	10	13	17	20	23	27	31	33	36	37	39	40	42	43	47	48	49	50
$x_i f_i$	1	2	3	16	15	18	28	24	27	40	44	24	39	14	30	16	36	19	80	21	22	24

- Media=$\overline{x} = 10'86$ minutos.
- Mediana=$M_e = 10$ minutos.

b) Una posible agrupación de los datos en intervalos de distinta amplitud es:

$(\ell_i, \ell_{i+1}]$	f_i	x_i	$x_i f_i$	F_i
(0,4]	7	2	14	7
(4,6]	6	5	30	13
(6,8]	7	7	49	20
(8,10]	7	9	63	27
(10,12]	6	11	66	33
(12,15]	6	$13'5$	81	39
(15,19]	4	17	68	43
(19,24]	7	$21'5$	$150'5$	50
suma			$521'5$	

Con esta clasificación en intervalos, los resultados de las medidas descriptivas anteriores son:

- Media=$\overline{x} = 10'43$ minutos.
- Mediana=$M_e = 9'4286$ minutos.

Los verdaderos resultados de estas medidas descriptivas son los calculados en el apartado anterior.