

ESTADÍSTICA 2

Aquí tenemos dos agrupamientos

distintos (con 6 y 7 clases) de los mismos datos:

LA MEDIA ES LA SUMA DE TODOS LOS DATOS
ENTRE EL NÚMERO TOTAL DE DATOS: $\bar{X}$

$$\begin{array}{c} \times \quad = \\ \begin{array}{c|c|c} x_i & f_i & x_i \cdot f_i \\ \hline 3 & 2 & 6 \\ 4 & 3 & 12 \\ 5 & 7 & 35 \\ 6 & 9 & 54 \\ 7 & 7 & 49 \\ 8 & 8 & 64 \\ 9 & 4 & 36 \\ \hline & 40 & 256 \end{array} \end{array}$$

PRIMERO PASAREMOS
LOS DATOS A UNA TABLA,
COMO VIMOS EN
EL VIDEO ANTERIOR
Y TRAS EL RECuento:

6 8 5 7 9 6 8 6
4 7 8 7 6 5 8 6
6 5 6 9 9 8 7 3
5 4 8 7 5 6 5 9
7 8 3 6 5 7 8 4

Ahora, con la información organizada,
sumar las 40 notas será muy fácil.
Los dos treses suman 6
Los tres cuatros suman 12
Bastará con multiplicar cada dato
por su frecuencia... y sumar

La media es $\bar{X} = 256 / 40 = 6'4$

1

	x_i	f_i	$x_i \cdot f_i$
$49 \leq x < 56$	52'5	8	420
$56 \leq x < 63$	59'5	4	238
$63 \leq x < 70$	66'5	8	532
$70 \leq x < 77$	73'5	9	661'5
$77 \leq x < 84$	80'5	6	483
$84 \leq x < 91$	87'5	5	437'5
		40	2772

Cuando tenemos datos agrupados
usamos la marca para los cálculos

$$\bar{X} = 2772 / 40 = 69'3$$

Ha sido algo más incómodo
porque las marcas eran
cantidades decimales

PULSAC.	x_i	f_i	$f_i \cdot x_i$
[49,56)	52,5	8	420
[56,63)	59,5	4	238
[63,70)	66,5	8	532
[70,77)	73,5	9	661,5
[77,84)	80,5	6	483
[84,91)	87,5	5	437,5
TOTAL		40	2772

La media es 69'3 (2772/40)

PULSAC.	x_i	f_i	$f_i \cdot x_i$
[49,55)	52	8	416
[55,61)	58	3	174
[61,67)	64	5	320
[67,73)	70	9	630
[73,79)	76	5	380
[79,85)	82	6	492
[85,91)	88	4	352
TOTAL		40	2764

La media es 69'1 (2764/40)

En este caso, los cálculos han resultado
más fáciles porque las marcas son enteras

En realidad, si sumamos
esos datos sin agrupar
y dividimos entre 40,
obtenemos una media
real de 68'925

Al agrupar los datos, sustituimos
todos los valores de cada clase
por su marca. Cometemos un error
minúsculo, pero si no lo hacemos,
no podríamos tratar grupos
grandes de datos

MEDIA
PONDERADA
cuando unos datos
no tienen todos
el mismo PESO.

Por ejemplo, si el profesor de Matemáticas decide que
la nota de actitud en clase tiene un peso de 2,
que la nota de la libreta tiene un peso de 5,
y que la nota de los exámenes tiene un peso de 8,
y Antonio ha obtenido 9, 6 y 4 respectivamente,
su media se calculará usando los pesos como frecuencia:

APARTADO	NOTA	PESO	NOTA x PESO
ACTITUD	9	2	18
LIBRETA	6	5	30
EXÁMENES	4	8	32
TOTAL		15	80

Su media será
 $80 / 15 = 5'33$

La MODA (M_o) es el valor más repetido (mayor frecuencia absoluta)

x_i	f_i
3	2
4	3
5	7
6	9
7	7
8	8
9	4
	40

La moda vale
 $M_o = 6$

En datos agrupados:

Decimos que la CLASE MODAL es el intervalo de datos entre 70 y 77

x_i	f_i
$49 \leq x < 56$	52'5
$56 \leq x < 63$	59'5
$63 \leq x < 70$	66'5
$70 \leq x < 77$	73'5
$77 \leq x < 84$	80'5
$84 \leq x < 91$	87'5
	40

Los CUARTILES separan los datos ordenados en 4 grupos de igual tamaño

El primer cuartil de ja la cuarta parte de los datos por debajo, y el tercero de ja la cuarta parte por encima

x_i	f_i	F_i
3	2	2
4	3	5
5	7	12
6	9	21
7	7	28
8	8	36
9	4	40

El segundo cuartil coincide con la mediana: $Me = Q_2$

Hallamos Q_3

1. Calculo $3/4$ de N : $(3/4) \times 40 = 30$ (entero)

2. Me quedo con el 30º y el 31º

3. Los 28 primeros valen 7 o menos

Hasta el 36º valen todos 8 $\Rightarrow Q_3 = 8$

Las tres cuartas partes de los datos son menores o iguales a 8, y la cuarta parte restante es mayor o igual a 8

La MEDIANA (Me) es el valor central de los datos

Es decir, si ordenamos los datos, es el valor que queda en medio

Sin embargo, si tenemos un número par de datos, ninguno quedará en el centro

2 4 5 7 9

$Me = (4 + 5) / 2 = 4'5$

Podemos observar que la mitad de los datos está por debajo de 4'5, y la otra mitad, por encima

Está claro que este método no sirve si tenemos muchos datos

En estos casos, tomaremos los dos datos centrales y hallaremos la media:

Vamos a ayudarnos de la frecuencia absoluta acumulada:

Los doce primeros datos son menores o iguales a 5

Aquí tenemos todos los datos desde el décimo tercero al vigésimo primero

Ambos son 6

x_i	f_i	F_i
3	2	2
4	3	5
5	7	12
6	9	21
7	7	28
8	8	36
9	4	40

1. Primero calcularemos la mitad de los datos. En este caso es

$$N/2 = 40/2 = 20$$

2. Si el resultado es decimal, nos quedaremos el número entero siguiente. Si el resultado es entero nos quedaremos con ese número y el siguiente.

En nuestro caso, como es entero, nos quedamos con el 20 y el 21

3. Buscamos los datos que correspondan a los ORDINALES de los números que hemos obtenido, y si son dos datos diferentes, hallamos su media:

Tenemos que buscar los datos 20º y 21º (vigésimo y vigésimo primero) $\Rightarrow Me = 6$

Y cada PERCENTIL (P_i) separa los datos en porcentajes

Halla un valor que de ja por debajo al $i\%$ de los datos

x_i	f_i	F_i
3	2	2
4	3	5
5	7	12
6	9	21
7	7	28
8	8	36
9	4	40

Hallamos P_{12}

1. Calculo el 12% de N : $12 \times 40 / 100 = 4'8$ (decimal)

2. Me quedo el dato 5º (el entero siguiente)

3. El dato quinto es un 4

$P_{12} = 4$

El 12% de los datos es menor o igual a 4, y el 88% restante es mayor o igual a 4

PARA HALLAR CON DATOS AGRUPADOS LA MEDIANA, LOS CUARTILES, O LOS PERCENTILES, POR AHORA, PROCEDEMOS DE LA MISMA MANERA QUE CON LA MEDIA: USAREMOS LA MARCA DE CADA INTERVALO.

EL RANGO o RECORRIDO es la diferencia entre el dato mayor y el dato menor

x_i	f_i
3	2
4	3
5	7
6	9
7	7
8	8
9	4
40	

El rango o recorrido es
 $R = 9 - 3 = 6$

Con datos agrupados, si hemos perdido los datos originales, restamos al final de la última clase el inicio de la primera clase:

$$R = 91 - 49 = 42$$

	x_i	f_i
$49 \leq x < 56$	52'5	8
$56 \leq x < 63$	59'5	4
$63 \leq x < 70$	66'5	8
$70 \leq x < 77$	73'5	9
$77 \leq x < 84$	80'5	6
$84 \leq x < 91$	87'5	5
	40	

La DESVIACIÓN MEDIA (D_m) es la media de las desviaciones respecto a la MEDIA ($\bar{X}$)

x_i	f_i	$ x_i - \bar{X} $	DESVIACIÓN TOTAL
3	2	3'4	6'8
4	3	2'4	7'2
5	7	1'4	9'8
6	9	0'4	3'6
7	7	0'6	4'2
8	8	1'6	12'8
9	4	2'6	10'4
	40		54'8

Ya sabemos que la media de estos datos es 6'4
El dato 3 se está desviando 3'4 respecto a la media
Los dos treses se están desviando 6'8 en total

La desviación media es
 $D_m = 54'8 / 40 = 1'37$

De media, los datos se desvían 1'37 respecto a la media

Pero los parámetros que generalmente se usan para medir la homogeneidad de unos datos son la VARIANZA y la DESVIACIÓN TÍPICA

La VARIANZA (V) es la media de las desviaciones al cuadrado respecto a la media

La DESVIACIÓN TÍPICA (S) es la raíz cuadrada de la VARIANZA
Veamos cómo se hallan, recuperando algunas columnas anteriormente usadas

DATOS	FRECUENCIA	SUMA TOTAL	DESVIACIÓN	DESVIACIÓN TOTAL	DESVIACIÓN AL CUADRADO	DESVIACIÓN AL CUADRADO TOTAL	DATOS AL CUADRADO	SUMA TOTAL DE CUADRADOS
x_i	f_i	$x_i \cdot f_i$	$x_i - \bar{X}$	$(x_i - \bar{X}) \cdot f_i$	$(x_i - \bar{X})^2$	$(x_i - \bar{X})^2 \cdot f_i$	x_i^2	$x_i^2 \cdot f_i$
3	2	6	3'4	6'8	11'56	23'12	9	18
4	3	12	2'4	7'2	5'76	17'28	16	48
5	7	35	1'4	9'8	1'96	13'72	25	175
6	9	54	0'4	3'6	0'16	1'44	36	324
7	7	49	0'6	4'2	0'36	2'52	49	343
8	8	64	1'6	12'8	2'56	20'48	64	512
9	4	36	2'6	10'4	6'76	27'04	81	324
	40	256		54'8		105'6		1744

La varianza es $V = 105'6 / 40 = 2'64$

Pero también podemos hallar la varianza calculando la media de los cuadrados menos el cuadrado de la media !!!

MEDIA DE LOS CUADRADOS - CUADRADO DE LA MEDIA

$$V = \frac{1744}{40} - 6'4^2 = 2'64$$

Podemos usar cualquiera de los dos métodos

Con datos pequeños y enteros, el segundo suele ser más sencillo

La DESVIACIÓN TÍPICA es $S = \sqrt{2'64} = 1'625$

La desviación típica es el parámetro más usado
Tiene la misma unidad de medida que los datos

Para realizar los cálculos con datos agrupados, en la desviación media, la varianza, o la desviación típica, usamos la marca de cada clase

Finalmente, en nuestros datos tenemos una nota media de 6'4 y una desviación típica de 1'625

Estos dos parámetros, generalmente, son los que mejor describen los datos de un estudio. Aproximadamente, el 70% de los datos no se desvían de la media más de lo que indica la desviación típica.



En nuestro caso, entre 4'775 ($6'4 - 1'625$) y 8'025 ($6'4 + 1'625$) nos encontramos el 77'5% de los datos

A los valores que se desvían demasiado los denominamos ATÍPICOS

4

Sin embargo, en algunas ocasiones tenemos que recurrir a otros parámetros.

Por ejemplo, si en una empresa la gran mayoría de los trabajadores cobran muy poquito, y unos pocos cobran una barbaridad, puede ser que esos sueldazos (atípicos) provoquen un salario medio decente que enmascare la situación real de la empresa.

¿Qué parámetros deberíamos usar en este caso?

En este caso, para demostrar las desigualdades de la empresa, la moda, la mediana, los cuartiles y los percentiles, describirán mucho mejor los datos del estudio

Con un percentil alto podremos observar que la gran mayoría no alcanza un sueldo digno

Otro caso interesante:

En primero la estatura media es 120cm, con una desviación típica de 10cm

En segundo la estatura media es 150cm, con una desviación típica de 12cm

Está claro que en segundo son más altos, pero... ¿qué grupo es más homogéneo?

$$CV = 10 / 120 = 0'08333...$$

$$CV_1 = 8'33\%$$

$$CV = 12 / 150 = 0'08$$

$$CV_2 = 8\% \downarrow$$

Por un lado, en segundo se aprecia una desviación mayor. Pero, por otro lado, entre personas bajitas, una diferencia de 10cm se aprecia más

Segundo es un grupo más homogéneo

Para resolver este dilema, tenemos el COEFICIENTE DE VARIACIÓN (CV):

$$CV = S / \bar{X}$$

Si lo multiplicamos por 100, se nos queda el porcentaje