2.2: Resumen numérico

- Medidas de localización.
- Medidas de dispersión.
- Medidas de forma.

Lecturas recomendadas:

- Capítulos 2 a 6 del libro de Peña y Romo (1997)
- Capítulos 3 a 7 del libro de Portilla (2004)

Medidas Descriptivas

¿Para qué nos sirven?

¿Se pueden calcular todas con todo tipo de variables?

¿Cuáles son las más adecuadas en cada caso?

¿De qué forma podemos sacar partido a nuestra calculadora?

Medidas de localización

Existen tres medidas comunes: la moda, la mediana y la media.

Una muestra del número de años en el ayuntamiento de los últimos 24 alcaldes de Madrid

3	1	1	1	1	1	2	1
7	6	13	8	3	2	1	1
2	1	1	7	3	2	12	6

La moda

Es el valor más frecuente

Clase	*Frecuencia*
1	10
2	4
3	3
4	0
5	0
6	2
7	2
8	1
9	0
10	0
11	0
12	1
13	1
y mayor...	0

¿Podemos calcular la moda con datos cualitativos?

¿Tiene sentido esta definición con datos continuos?

Puede haber más de una moda: bimodal-trimodal-plurimodal

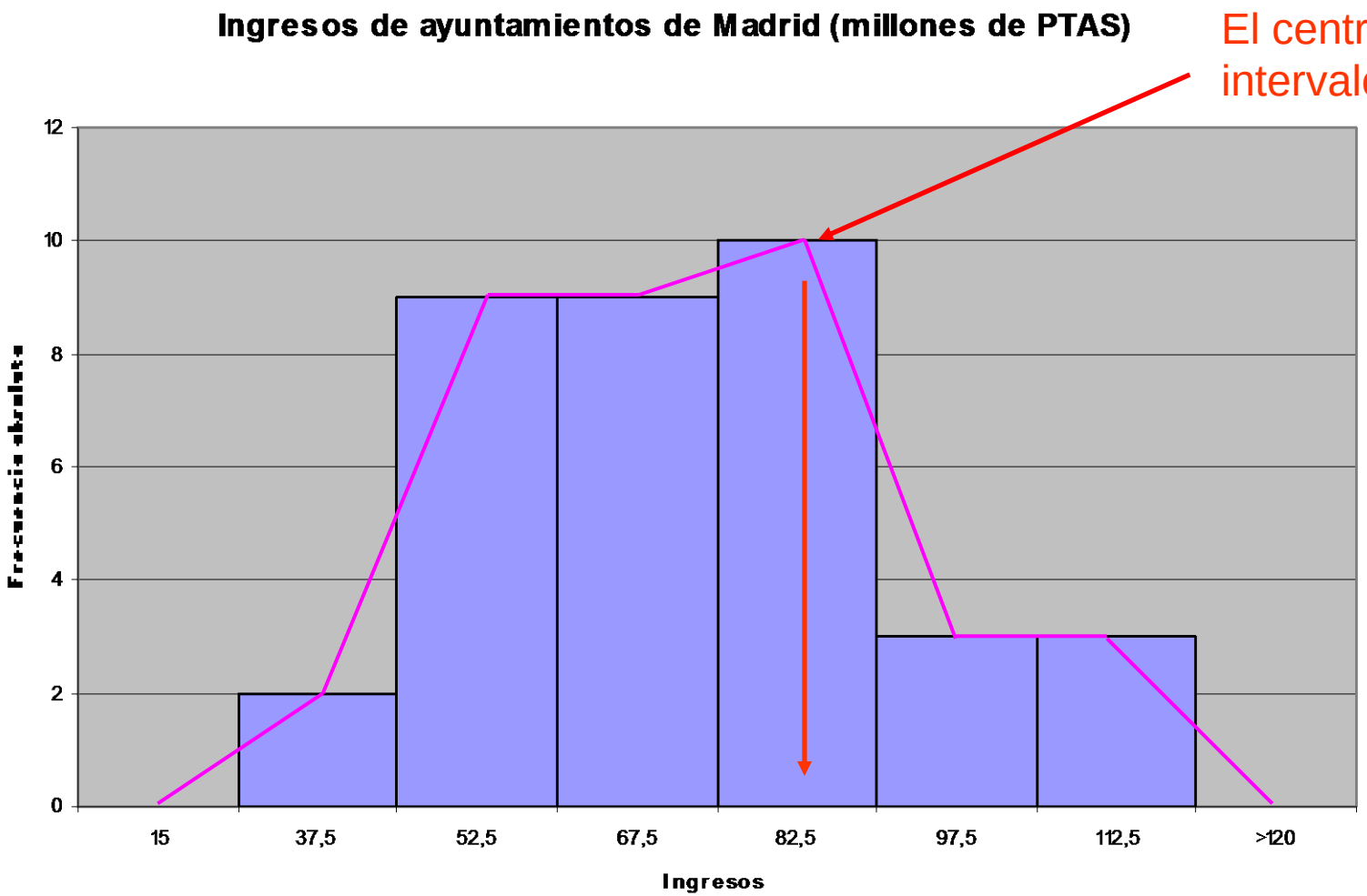
La moda con datos (continuos) agrupados

Ingresos y Derechos liquidados (millones de PTAS)	Frecuencia absoluta
≤ 30	0
(30,45]	2
(45,60]	9
(60,75]	9
(75,90]	10
(90,105]	3
(105,120]	3
> 120	0
Total	60

Tenemos una clase modal

¿Qué hacemos si las clases son de distinto tamaño?

Un valor exacta para la moda con datos agrupados



La mediana

Es la observación que ocupa el lugar central.

5 3 11 21 7 5 2 1 3

¿Qué valor toma la mediana?

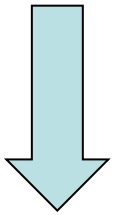
1. Ordenamos los datos de menor a mayor.
2. Tenemos en cuenta también los que se repiten.
3. La mediana, es el “CENTRO FÍSICO”

¿Cómo cambia el cálculo si *N* es par o impar?

¿Podemos calcular la mediana para datos cualitativos?

Ejemplo de los alcaldes

3	1	1	1	1	1	2	1
7	6	13	8	3	2	1	1
2	1	1	7	3	2	12	6



1	1	1	1	1	1	1	1
1	1	2	**2**	**2**	2	3	3
3	6	6	7	7	8	12	13

La mediana es ½*(2+2)=2

La mediana a través de la tabla de frecuencias (datos discretos)

Mediana →

x_i	n_i	N_i	f_i	F_i	
1	10	10	0,41666667	0,41666667	<0,5
2	4	14	0,16666667	0,58333333	>0,5
3	3	17	0,125	0,70833333	
4	0	17	0	0,70833333	
5	0	17	0	0,70833333	
6	2	19	0,08333333	0,79166667	
7	2	21	0,08333333	0,875	
8	1	22	0,04166667	0,91666667	
9	0	22	0	0,91666667	
10	0	22	0	0,91666667	
11	0	22	0	0,91666667	
12	1	23	0,04166667	0,95833333	
13	1	24	0,04166667	1	
y mayor...	0	24	0	1	

La mediana con datos agrupados

Intervalo mediano →

Ingresos	n_i	N_i	f_i	F_i
≤ 30	0	0	0	0
(30,45]	2	2	0,05555556	0,05555556
(45,60]	9	11	0,25	0,30555556
(60,75]	9	20	0,25	0,55555556
(75,90]	10	30	0,27777778	0,83333333
(90,105]	3	33	0,08333333	0,91666667
(105,120]	3	36	0,08333333	1
> 120	0	36	0	1
Total	36		1	

La media

La media o media aritmética es el promedio de todos los datos de la muestra.

Para el ejemplo de los alcaldes, la suma de los datos es:

3 + 1 + 1 + 1 + 1 + 1 + 2 + 1
7 + 6 + 13 + 8 + 3 + 2 + 1 + 1
2 + 1 + 1 + 7 + 3 + 2 + 12 + 6
= 86

Luego, la media es 86/24 ≈ 3,583 años.

Para datos $x_1, x_2, \ldots, x_N$, la media es

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

¿Podemos calcular la media para datos cualitativos?

La media a través de la tabla de frecuencias (datos discretos)

x_i	n_i	$n_i * x_i$	
1	10	10	
2	4	8	
3	3	9	
4	0	0	
5	0	0	
6	2	12	
7	2	14	
8	1	8	
9	0	0	
10	0	0	
11	0	0	
12	1	12	
13	1	13	
y mayor …	0	0	
Total	24	86	3,58333333

La fórmula

Para datos $x_1, x_2, \ldots, x_k$ con frecuencias absolutas $n_1, n_2, \ldots, n_k$, la media es

$$\begin{aligned}
\bar{x} &= \frac{1}{N}(n_1 \times x_1 + n_2 \times x_2 + \cdots + n_k \times x_k) \\
&= \frac{1}{N}\sum_{i=1}^{k} n_i \times x_i \\
&= \sum_{i=1}^{k} \frac{n_i}{N} \times x_i \\
&= \sum_{i=1}^{k} f_i \times x_i
\end{aligned}$$

La media con datos agrupados

Ingresos	x_i	n_i	x_i*n_i	
<= 30	22,5	0	0	
(30,45]	37,5	2	75	
(45,60]	52,5	9	472,5	
(60,75]	67,5	9	607,5	
(75,90]	82,5	10	825	
(90,105]	97,5	3	292,5	
(105,120]	112,5	3	337,5	
> 120	127,5	0	0	
Total		36	2610	72,5

Es la misma fórmula pero usando la marca de clase.

Otros puntos de la distribución: mínimo, máximo y cuartiles

Ordenando los datos, el mínimo y máximo son fáciles de calcular.

1	1	1	1	1	1	1	1
1	1	2	2	2	2	3	3
3	6	6	7	7	8	12	**13**

¿Y los cuartiles?

1er cuartil = (1+1)/2

1	1	1	1	**1**	**1**	1	1
1	1	2	**2**	**2**	2	3	3
3	**6**	**6**	7	7	8	12	13

3er cuartil = (6+6)/2

2º cuartil = mediana = (2+2)/2

Cálculo de cuartiles

Tenemos el siguiente conjunto de datos:

47	52	52	57	63	64	69	71
72	72	78	81	81	86	91	

1. Ordenamos los datos de menor a mayor.

2. Calculamos c_2 , que ocupa la posición correspondiente a la “mitad”, ¿con qué parámetro visto ya coincide este *segundo cuartil*?

3. Ahora calculamos, la “mitad” de la primera parte: c_1 .

4. Y la “mitad” de la segunda parte: c_3 .

	47	47		
	52	52		
	52	52		
	57	57	c1 = 60	
	63	63		
	64	64		
	69	69		
c2 = 71	71	71	71	
	72		72	
	72		72	
	78		78	c3 = 79,5
	81		81	
	81		81	
	86		86	
	91		91	

Medidas de dispersión

Existen varias medidas:

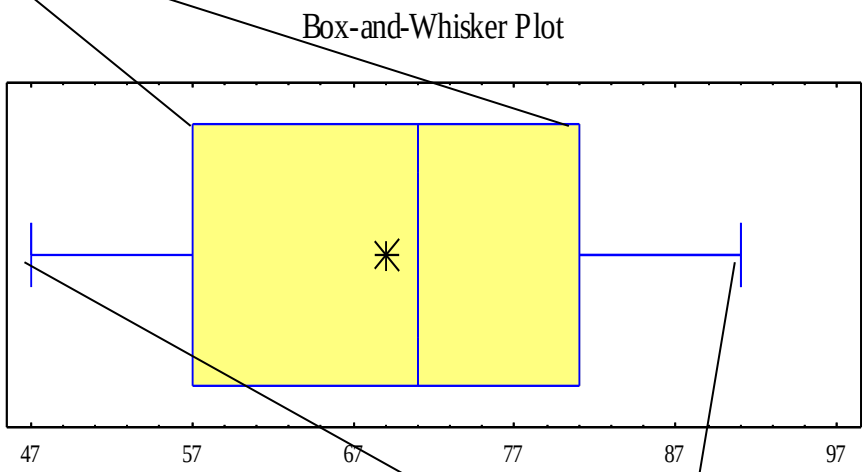
- El rango
- El rango intercuartilico
- La desviación típica
- El coeficiente de variación

Lecturas recomendadas:

Capítulos 4 y 5 del libro de Peña y Romo (1997)
Capítulos 6 y 7 del libro de Portilla (2004)

El rango y el rango intercuartílico

El rango intercuartílico



Calculamos el rango y rango intercuartílico en los ejemplos anteriores.

¿Cuál de los dos es más sensible a datos atípicos?

El rango $

$ Notar que en este ejemplo no hay valores atípicos.

La varianza y la desviación típica

Podemos mirar las distancias de las observaciones de la media

Empresa A	$x_i - \overline{X}$	Empresa B	$x_i - \overline{X}$
30700	**-2800**	27500	**-6000**
32500	**-1000**	31600	**-1900**
32900	**-600**	31700	**-1800**
33800	**300**	33800	**300**
34100	**600**	34000	**500**
34500	**1000**	35300	**1800**
36000	**2500**	40600	**7100**

¿Cuánto suman estas dos nuevas columnas?

$$
\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\
&= x_1 + x_2 + \cdots + x_n - \underbrace{\bar{x} - \bar{x} - \cdots - \bar{x}}_{n \text{ veces}} \\
&= x_1 + x_2 + \cdots + x_n - n\bar{x} \\
&= n\bar{x} - n\bar{x} \\
&= 0 \qquad \Rightarrow \\
\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}) &= 0
\end{aligned}
$$

Entonces, la distancia media no nos vale como medida de dispersión.

¿Cómo podemos resolver el problema?

La varianza

Es la distancia cuadrada media

Empresa A		Empresa B	
30700	**7840000**	27500	**36000000**
32500	**1000000**	31600	**3610000**
32900	**360000**	31700	**3240000**
33800	**90000**	33800	**90000**
34100	**360000**	34000	**3240000**
34500	**1000000**	35300	**250000**
36000	**6250000**	40600	**50410000**
	16900000		**96840000**
	2414285,7		**13834285,7**

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

¿Qué unidades tiene este nuevo estadístico?

¿Podemos cambiarlas?

La desviación típica

Es la raíz de la varianza.

Empresa A s = 1553,8
Empresa B s = 3719,4

¿Cuál es más sensible a atípicos: la desviación típica o el rango intercuartílico?

¿Y si queremos una medida sin unidades?

El coeficiente de variación

Cuando la media sea distinta de 0, podemos calcular una medida de dispersión normalizada.

$$CV = s/|\bar{x}|$$

Nos permite comparar, porque no tiene unidades.

¿Para qué nos sirve con una única base de datos?

EJERCICIO:

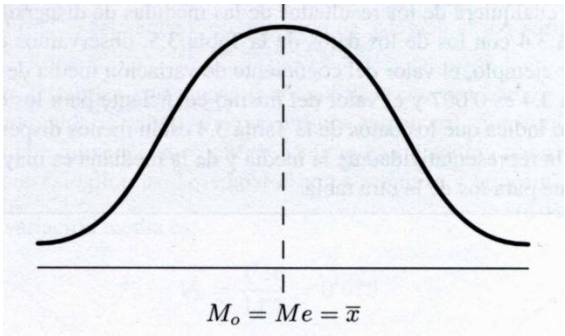
Analizamos el volumen de consultas durante el período de exámenes en 10 bibliotecas universitarias, y se comparan con las anotadas el año anterior. El % de incremento de consultas fue:

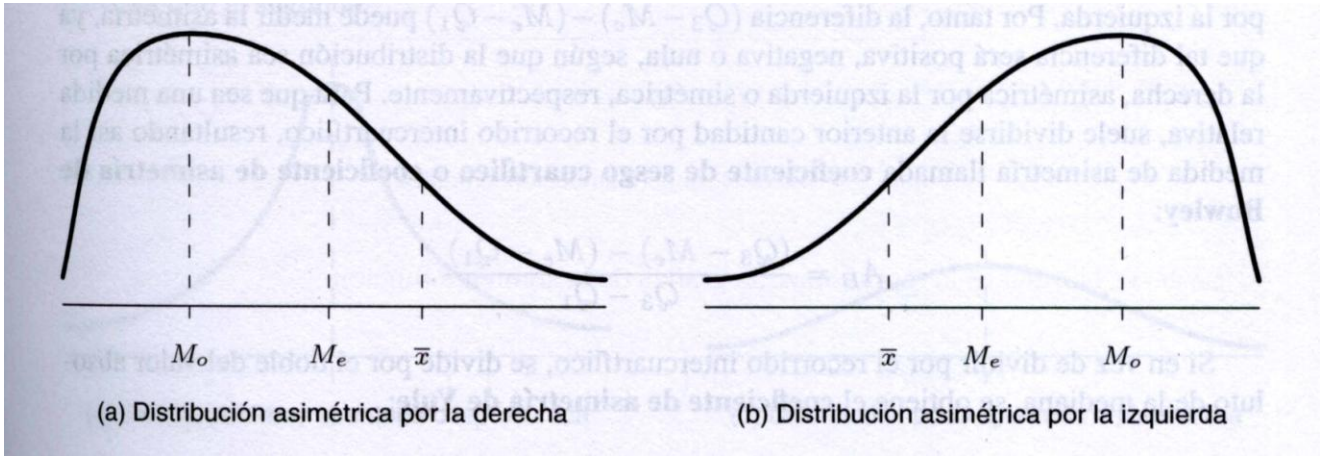
10.2	2.9	3.1	6.8	5.9
7.3	7.0	8.2	3.7	4.3

¿Son los datos homogéneos?

Medidas de forma

Las medidas comunes son de asimetría y curtosis.





Datos simétricos, asimétricos a la derecha y a la izquierda

El coeficiente de asimetría de Pearson

CA=0 Simétrica
CA>0 Asimétrica derecha
CA<0 Asimétrica izquierda

$$CA = \frac{\bar{x} - M_o}{s}$$

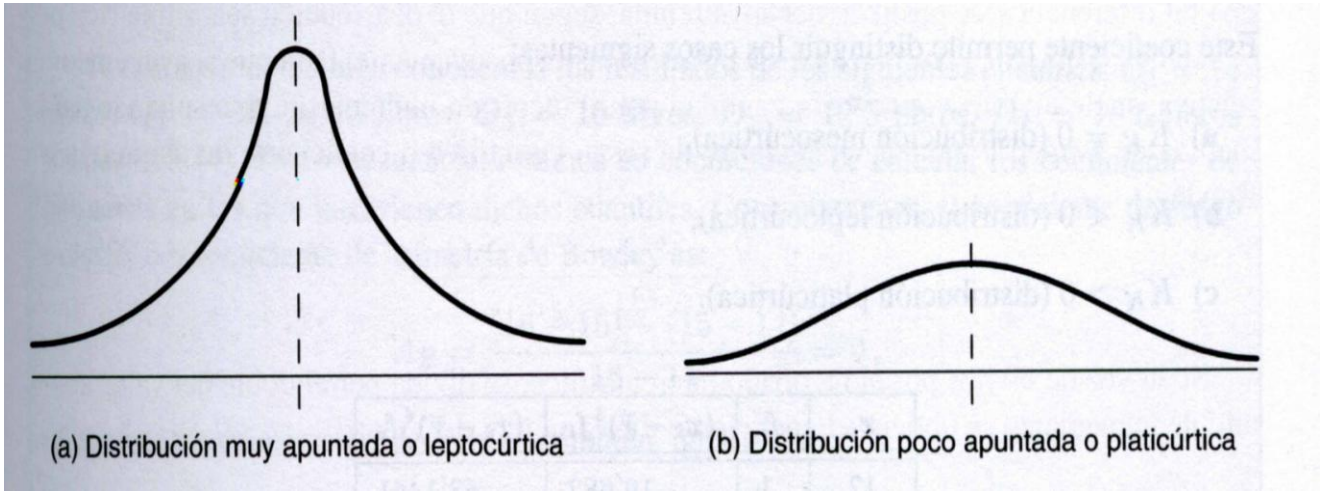
El coeficiente de asimetría de Fisher

(cuando existe más de una moda):

$$CA = \frac{\sum_{i=1}^{N} (x_i - \bar{x})^3}{N s^3}$$

Curtosis

Podemos verlo gráficamente, comparando con la curva normal:



El coeficiente de curtosis de Fisher

$$CC = \frac{\sum_{i=1}^{N}(x_i - \bar{x})^4}{N s^4} - 3$$

CC = 0 (mesocúrtica)

CC > 0 (leptocúrtica)

CC < 0 (platicúrtica)

EJERCICIO: Cálculo de las medidas forma estudiadas.

Trabaja con la siguiente base de datos (calificaciones de un grupo de alumnos):

100	112	88	105	100	102	98	113
102	87	93	93	117	100	98	92
100	117	97	100	83	67	76	100
106	117	89	83	100	109	109	93
105	108	104	63	81	109	100	98

Ejercicio (Pregunta de Examen)

Se ha tomado una muestra aleatoria de 10 madrileños que están trabajando y se les ha preguntado cuántas horas de trabajo realizan habitualmente cada semana. Los resultados son:

40	40	35	50	50	40	40	60	50	35

Selecciona la respuesta correcta entre las siguientes:

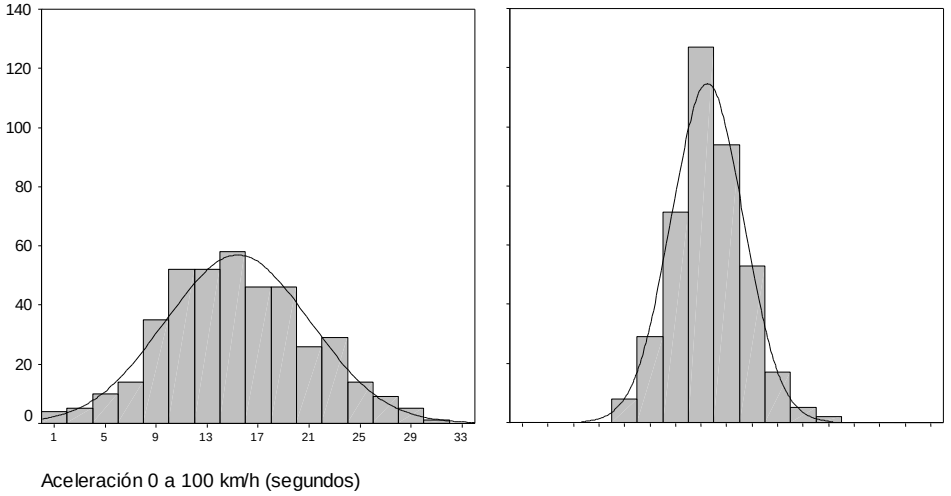
a) La moda y la mediana son iguales a 40, y la media es 44.
b) La media y la moda son iguales a 40, y la mediana es 44.
c) La media y la mediana son iguales a 40, y la moda es 44.
d) Ninguna de las anteriores es correcta.

Ejercicio (Pregunta de examen)

En el histograma de la izquierda se muestra el tiempo (en segundos) que 406 modelos de coches tardan en pasar de 0 a 100 km/h (aceleración), medido por un trabajador A. Para confirmar los resultados un segundo trabajador B realiza de nuevo la medición de la aceleración en los mismos coches (histograma de la derecha).

Analizando estos resultados, ¿cuál de las siguientes afirmaciones es correcta?

a) La aceleración media medida por A es claramente distinta de la aceleración media medida por B.
b) La desviación típica de la aceleración medida por A es menor que la desviación típica de la aceleración medida por B.
c) La varianza de la aceleración medida por B es menor que la varianza de la aceleración medida por A.



Ejercicio (Pregunta de examen)

Se desea estudiar el tiempo de duración de las ponencias de un portavoz político durante el año 2007/2008. Para ello se contabiliza el tiempo (en minutos) en cada una de las 34 veces que habló:

¿Cuál es aproximadamente el tiempo medio de las ponencias?

a) 32.50 minutos
b) 27.52 minutos
c) 30.96 minutos
d) 33.88 minutos

Tiempo	**n_i**	**f_i**
(0-25]	6	0.18
(25-35]	13	0.38
(35-40]	13	0.38
(40-60]	2	0.06
Total	**34**	**1**

Ejercicio (Pregunta de examen)

La tabla muestra las edades y el sexo de un grupo de ministros del gobierno.

Nombre	Sexo	Ministerio	Edad
Bibiana Aído	M	Igualdad	33
Carme Chacón	M	Defensa	38
Ángeles González-Sinde	M	Cultura	44
Cristina Garmendia	M	Ciencia e innovación	47
Trinidad Jiménez	M	Sanidad y Política Social	47
José Blanco	V	Fomento	48
Ángel Gabilondo	V	Educación	60
Elena Salgado	M	Economía y Hacienda	60

¿Cuál de las siguientes afirmaciones es la correcta?

a) El primer cuartil de las edades es 41 y el tercer cuartil es 54.
b) El primer cuartil de las edades es 47 y el porcentaje de ministros varones es de 25%.
c) El rango de edades es 33 y la frecuencia absoluta de mujeres es 6.
d) La moda de las edades es de 60 y la media es 47.