

Curso de Estadística Aplicada a las Ciencias Sociales

Tema 5. Descripción numérica (1)

Capítulo 4 del manual

1

Tema 5 – Descripción numérica (1)

Introducción

1. La media
2. La desviación típica
3. El coeficiente de variación
4. El coeficiente de asimetría
5. Descripción numérica para distribuciones de frecuencias

Resumen

Tema 5- Descripción numérica (1)

2

Introducción

- Hasta ahora: descripción de variables con tablas y gráficos
 - ◆ Información que se da: la distribución de los casos entre los diferentes valores
 - ◆ Para todo tipo de variables
- Con **variables cuantitativas**, podemos resumir información de otra forma: valores numéricos sobre
 - ◆ “Centro” de los datos (medidas de posición)
 - ◆ Concentración de los datos en torno al centro (medidas de dispersión)
 - ◆ Otros rasgos de la distribución

Tema 5- Descripción numérica (1)

3

1. La media

- Descripción de un conjunto de datos más elemental: su “centro”
- Media o promedio: el “centro de gravedad”
- Ejemplos: la nota media en un examen, ingreso medio por familia, número de hijos medio por pareja
- **MUY IMPORTANTE:** la media no tiene por qué ser “representativa”

Tema 5- Descripción numérica (1)

4

1. La media: cálculo

- ❑ Tenemos una variable, que llamamos X
- ❑ Llamamos N al número de casos u observaciones de la variable
- ❑ Los valores que toman cada una de las observaciones, los llamamos

$$x_1, x_2, \dots, x_{n-1}, x_n$$

- ❑ La media se obtiene dividiendo la suma de los valores de todo los sujetos por el número de sujetos

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_N}{N} = \frac{\sum x_i}{N}$$

1. La media: cálculo

- Si tenemos varios casos con el mismo valor, podemos hacerlo de una forma más rápida
- Llamamos c_i a cada uno de los valores de la escala
- Así, dada una escala con los valores

$$c_1, c_2, \dots, c_{k-1}, c_k$$

- Cuyas frecuencias absolutas son:

$$n_1, n_2, \dots, n_{k-1}, n_k$$

- La fórmula de la media se reelaboraría así:

$$\bar{x} = \frac{c_1 n_1 + c_2 n_2 + \dots + c_n n_n}{N} = \frac{\sum x_i}{N}$$

1. La media: cálculo

- Esa fórmula se puede simplificar aún más:

ATENCIÓN:
CUADROS COMO
ESTE CON
LÍNEAS
PUNTEADAS SON
EXPLICACIONES
MATEMÁTICAS
PARA
"AFICIONADOS".
NO SON
NECESARIOS
PARA SEGUIR LA
ASIGNATURA

$$\bar{x} = \frac{(c_1 n_1) + (c_2 n_2) + \dots + (c_k n_k)}{N}$$
$$\bar{x} = \frac{c_1 n_1}{N} + \frac{c_2 n_2}{N} + \dots + \frac{c_k n_k}{N}$$
$$\bar{x} = (c_1 f_1) + (c_2 f_2) + \dots + (c_k f_k)$$

$$\bar{x} = \sum c_i f_i$$

1. La media: Cálculo con EXCEL

- Cálculo con EXCEL
- Ejemplos de variables GTINE y AHORRO
- **=promedio(rango)**
- GTINE: 261.277 (* según fichero HOGARES, con datos tabla 2.1; pero cálculos libro con datos apéndice, que son un poco distintos)
- AHRR: 14.763
- Comprobación
=suma(rango)/contar(rango)

1. La media: propiedades-1

- Suma de las **desviaciones** de un conjunto de observaciones respecto a su media, es igual a cero (se compensan unas con otras) (Ver con GTINE o AHORRO)

$$\sum (x_i - \bar{x}) = 0$$

$$\begin{aligned} \sum (x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \dots + (x_n - \bar{x}) = \\ (x_1 + x_2 + \dots + x_n) - N\bar{x} &= \sum x_i - N \frac{\sum x_i}{N} = \\ \sum x_i - \sum x_i &= 0 \end{aligned}$$

9

1. La media: propiedades -2

- El valor de la media puede verse muy afectado por unas pocas observaciones cuyo valor sea muy diferente de los demás
- Ejemplo 7 sueldos en empresa: 10.200€, 10.400€, 10.700€, 11.200€, 11.300€, 11.500€ y 200.000€
- Sueldo medio es 37.900€
- Un solo **valor atípico** (outlier) “arrastra” la media hacia arriba
- Media de los seis otros valores: 10.883€
- El valor de la media puede no ser representativo del conjunto de los valores, especialmente en poblaciones o muestras pequeñas, cuando una es muy diferente de las otras

Tema 5- Descripción numérica (1)

10

1. La media: propiedades -3

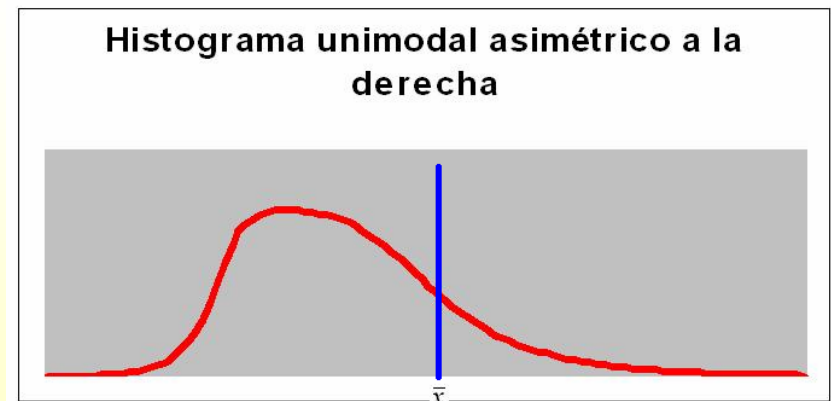
- En general, cuando el gráfico que representa la distribución de valores no es simétrico, sino sesgado, la media está desviada, en relación con la mayoría de los valores, hacia la cola más larga de la distribución
- Cuanto más sesgada es la distribución: menos representativa es la media

Tema 5- Descripción numérica (1)

11

1. La media: propiedades -3

- Cola hacia la derecha: media mayor que la mayoría de los valores

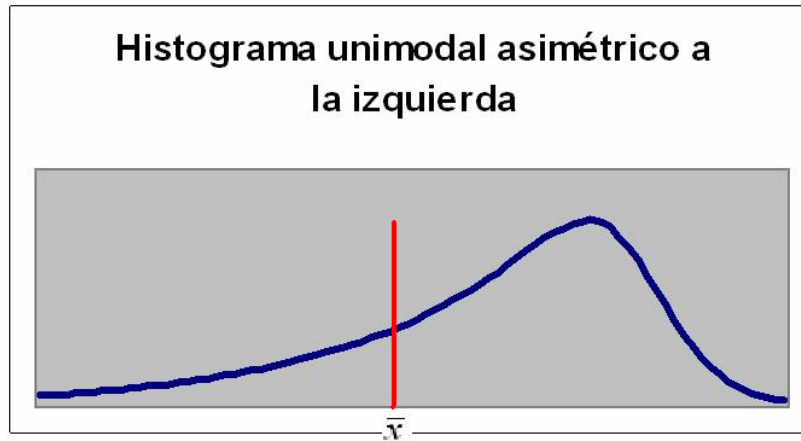


Tema 5- Descripción numérica (1)

12

1. La media: propiedades -3

- Cola hacia la izquierda: media menor que la mayoría de los valores



13

1. La media: media ponderada

- Si tenemos dos poblaciones o muestras de tamaños n_1 y n_2 , y tenemos el valor medio de una variable en ambas poblaciones \bar{x}_1 y \bar{x}_2
- Podemos calcular la media de todos los sujetos que componen las dos muestras o poblaciones, utilizando la fórmula de la **media ponderada**

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}$$

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2} = \frac{n_1 \frac{\sum x_1}{n_1} + n_2 \frac{\sum x_2}{n_2}}{n_1 + n_2} = \frac{\sum x_1 + \sum x_2}{n_1 + n_2}$$

14

1. La media: media ponderada (2)

- Lo mismo se aplica si, en lugar de 2 poblaciones o muestras, tenemos muchas más
- Por ejemplo: variable con la edad media de la población de 285 municipios de Castilla-La Mancha
- Media ponderada: sumamos cada valor multiplicado por la población del municipio y lo dividimos por la población de todos ellos

$$\bar{x} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2 + n_3 \bar{x}_3 + \dots + n_n \bar{x}_n}{n_1 + n_2 + n_3 + \dots + n_n}$$

15

2. La desviación típica

- Dos conjuntos de datos pueden tener la misma media pero ser muy distintos
 - ◆ 13, 15, 17, 21, 23, 25 (media es 19)
 - ◆ 3, 5, 7, 31, 33, 35 (media es 19)
- Diferencia: dispersión respecto a media
- Consecuencia: junto a media (posición) es necesario otro valor que exprese la dispersión.

Tema 5- Descripción numérica (1)

16

2. La desviación típica: posible cálculo

- Una posible idea: la media de las **desviaciones respecto a la media**

$$\frac{\sum (x_i - \bar{x})}{N}$$

- Problema: el numerador es cero (se compensan)
- Solución: elevar al cuadrado, calcular la media de los cuadrados, y hallar la raíz cuadrada

2. La desviación típica: cálculo

- Esta es la fórmula de la **desviación típica**:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N}}$$

- Lo que está dentro de la raíz cuadrada, es decir el cuadrado de la desviación típica se llama **varianza**:

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N}$$

2. La desviación típica: cálculo alternativo

- Igual que la media, cuando hay valores repetidos, la desviación típica también puede calcularse con esta otra fórmula:

$$s_c = \sqrt{\frac{(c_1 - \bar{x}_c)^2 n_1 + (c_2 - \bar{x}_c)^2 n_2 + \dots + (c_k - \bar{x}_c)^2 n_k}{N}} =$$

$$\sqrt{\frac{(c_1 - \bar{x}_c)^2 n_1}{N} + \frac{(c_2 - \bar{x}_c)^2 n_2}{N} + \dots + \frac{(c_k - \bar{x}_c)^2 n_k}{N}} =$$

$$\sqrt{(c_1 - \bar{x}_c)^2 f_1 + (c_2 - \bar{x}_c)^2 f_2 + \dots + (c_k - \bar{x}_c)^2 f_k} =$$

$$\sqrt{\sum (c_i - \bar{x}_c)^2 f_i}$$

$$s_c = \sqrt{\sum (c_i - \bar{x}_c)^2 f_i}$$

2. La desviación típica: poblaciones y muestras

- Por razones técnicas (matemáticas), cuando se calcula la desviación típica y la varianza de una muestra, en lugar de la de una población, el denominador es (N-1) en lugar de N

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$

$$s_x^2 = \frac{\sum (x_i - \bar{x})^2}{N - 1}$$

2. La desviación típica: cálculo con EXCEL

- Ejemplos con EXCEL (variable GTINE)
- Si nuestros casos son una muestra
 - =DESVEST(rango) (170.340,54)
 - =VAR(rango)
 - Comprobación: diferencias, cuadrados, sumarlos, dividir por (n-1), hallar raíz cuadrada
- Si nuestros casos son una población
 - =DESVESTP(rango) (169.201,12)
 - =VARP(rango)

Tema 5- Descripción numérica (1)

2. La desviación típica: interpretación

- Mide la dispersión: cuanto más grande, mayor dispersión.
- Pero significado no intuitivo
- Es algo así como la “media de las desviaciones respecto a la media”
- Unidades: las mismas en las que se exprese la variable (euros, metros, puntos en examen...)
- ¿Grande o pequeña? Según lo que sepamos de la variable misma
- Ejemplo: examen de 0 a 10; desviación típica de 1 o de 3

Tema 5- Descripción numérica (1)

22

2. La desviación típica: propiedades

- Siempre valor positivo $s_x \geq 0$
- Sólo valor 0 si todas las observaciones tienen el mismo valor
- Como la media, muy afectada por valores atípicos
- Sueldos de ejemplo anterior (trans.8):
 - Desviación típica incluyendo 7 valores: 66.178,5
 - Desviación típica sólo de 6 valores “normales”: 481,0

Tema 5- Descripción numérica (1)

23

2. La desviación típica: regla de Chebychev

- **Para cualquier conjunto de datos**, la proporción de observaciones que distan menos de m desviaciones típicas de la media, es **como mínimo**:
$$1 - \frac{1}{m^2}$$
- m puede ser un número no entero (1,5; 2,3)
- Por ejemplo, la proporción de datos cuyo valor dista de la media menos de 2 veces la desviación típica será:
$$1 - \frac{1}{2^2} = 1 - \frac{1}{4} = 1 - 0,25 = 0,75$$
- Es decir, el 75% de los datos de cualquier serie tienen un valor que dista de la media menos de dos veces la desviación típica

Tema 5- Descripción numérica (1)

24

2. La desviación típica: regla de Chebychev

- La proporción de datos cuyo valor dista de la media menos de 3 veces la desviación típica será:

$$1 - \frac{1}{3^2} = 1 - \frac{1}{9} = 1 - 0,111 = 0,88$$

- La proporción de datos cuyo valor dista de la media menos de 4 veces la desviación típica será:

$$1 - \frac{1}{4^2} = 1 - \frac{1}{16} = 1 - 0,062 = 0,93$$

2. La desviación típica: “regla empírica”

- Cuando el histograma de los datos tiene aproximadamente la forma de una campana:**

- Aproximadamente el 68% de los datos caen entre

$$\bar{x} \pm s_x$$

- Aproximadamente el 95% de los datos caen entre

$$\bar{x} \pm 2s_x$$

- Todos o casi todos los datos caen entre

$$\bar{x} \pm 3s_x$$

2. La desviación típica: “regla empírica”

- Se llama “regla empírica” porque es derivada de la observación de lo que “suele suceder” en la práctica
- Por eso formulada en términos de “aproximadamente”
- Sólo sirve para datos con distribuciones más o menos de campana
- Otros datos, con distribuciones sesgadas: no funciona

3. El coeficiente de variación

- Problema de la desviación típica: varía con el valor absoluto de la variable
- Difícil comparar desviaciones típicas de variables con valores muy distintos
 - Ejemplo: G1 y G2 en ficheros HOGARES (desvest(rango))
 - Es “mucho” o “poco”??

3. El coeficiente de variación (2): concepto y cálculo

- Solución: una medida de dispersión independiente de los valores absolutos
- Coeficiente de variación:

$$CV_x = \frac{s_x}{|\bar{x}|}$$

- En EXCEL: calcular por separado desviación típica y media y dividir

4. El coeficiente de asimetría

- Otro rasgo interesante sobre una variable: simétrica o asimétrica
- Medición simetría: examinando diferencias entre valores y media
- Para hacer el número más manejable: las diferencias se dividen entre el valor de la desviación típica
- Forma de agregar información sobre el signo (positivo o negativo) y el tamaño: usando el cubo de las diferencias

4. El coeficiente de asimetría (2): cálculo

- Coeficiente de asimetría: “media” de los cubos de las diferencias entre los valores y la media, divididos por la desviación típica

$$CA_x = \frac{\sum \left(\frac{x_i - \bar{x}}{s_x} \right)^3}{N} = \frac{1}{N} \sum \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

- (Libro: otra fórmula, que da el mismo resultado)

4. El coeficiente de asimetría (2): cálculo

- Como con la media y la desviación típica, cuando tenemos muchos casos con el mismo valor, podemos simplificar el cálculo usando esta fórmula

$$CA_c = \sum \left(\frac{c_i - \bar{x}_c}{s_c} \right)^3 f_i$$

4. El coeficiente de asimetría (2): cálculo

■ Coeficiente de asimetría

$$CA_c = \frac{\left(\frac{c_1 - \bar{x}_c}{s_c}\right)^3 n_1 + \left(\frac{c_2 - \bar{x}_c}{s_c}\right)^3 n_2 + \dots + \left(\frac{c_k - \bar{x}_c}{s_c}\right)^3 n_k}{N}$$

$$= \frac{\left(\frac{c_1 - \bar{x}_c}{s_c}\right)^3 n_1}{N} + \frac{\left(\frac{c_2 - \bar{x}_c}{s_c}\right)^3 n_2}{N} + \dots + \frac{\left(\frac{c_k - \bar{x}_c}{s_c}\right)^3 n_k}{N}$$

$$= \left(\frac{c_1 - \bar{x}_c}{s_c}\right)^3 f_1 + \left(\frac{c_2 - \bar{x}_c}{s_c}\right)^3 f_2 + \dots + \left(\frac{c_k - \bar{x}_c}{s_c}\right)^3 f_k$$

33

4. El coeficiente de asimetría (2): con EXCEL

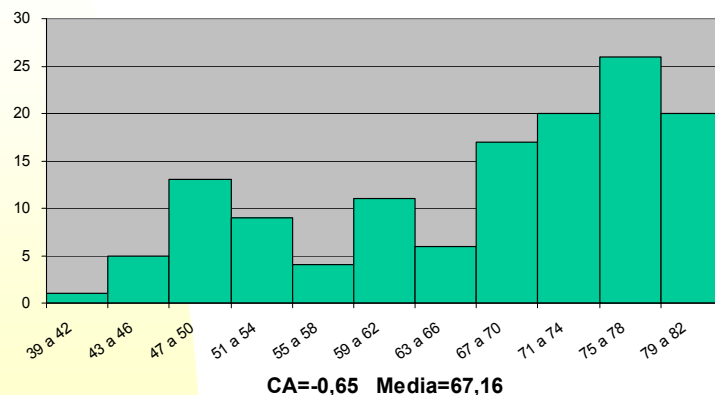
○ =COEFICIENTE.ASIMETRIA (rango)
(OJO: fórmula ligeramente diferente)

$$CA_x = \frac{N}{(N-1)(N-2)} \sum \left(\frac{x_i - \bar{x}}{s_x} \right)^3$$

Tema 5- Descripción numérica (1)

4. El coeficiente de asimetría (3): ejemplos

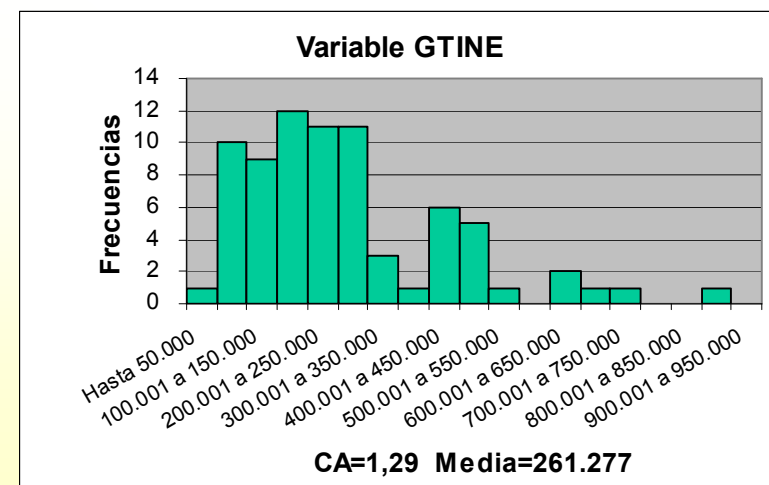
Esperanza de vida (ESPM en fichero PAISES)



Tema 5- Descripción numérica (1)

35

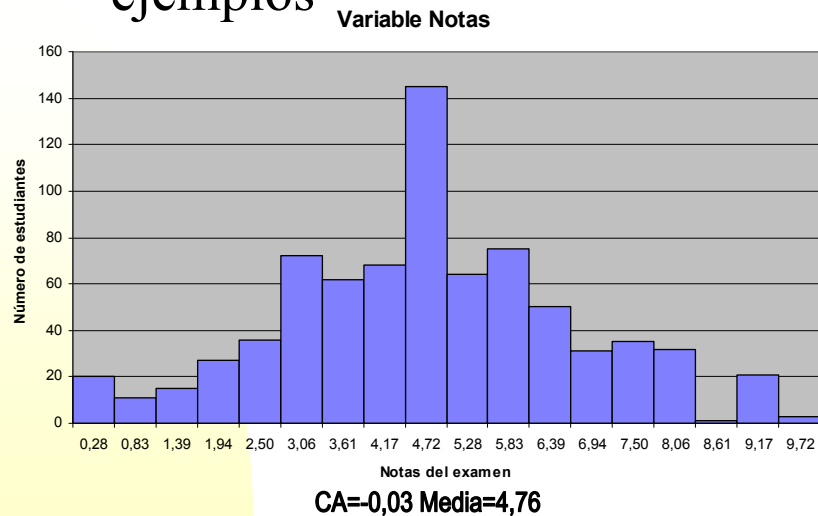
4. El coeficiente de asimetría (4): ejemplos



Tema 5- Descripción numérica (1)

36

4. El coeficiente de asimetría (4): ejemplos



Tema 5- Descripción numérica (1)

37

5. Descripción numérica con distribuciones de frecuencias

- Lo hecho hasta ahora: suponiendo que tenemos todos los datos originales
- Pero a veces, sólo tenemos la distribución de frecuencias de una variable continua
- ¿Cómo hacer una descripción numérica?
- Solución general: suponer que todos los casos de cada clase tienen como valor la marca de clase, c_i

Tema 5- Descripción numérica (1)

38

5. Descripción numérica con distribuciones de frecuencias: fórmulas (1)

- Podemos calcular la media por el procedimiento simplificado que hemos visto en la transparencia 7
- Sólo que aquí la marca de clase no es el valor exacto de los casos, sino el valor central de la clase

$$\bar{x}_c = \sum c_i f_i$$

Tema 5- Descripción numérica (1)

39

5. Distribuciones de frecuencias: ejemplo sobre la media

Distribución de frecuencias variable GTINE			
Marca de clase	Absolutas	Relativas	
c	n	f	cf
		0	
25000	1	0,01333333	333,333333
75000	10	0,13333333	10000
125000	9	0,12	15000
175000	12	0,16	28000
225000	11	0,14666667	33000
275000	11	0,14666667	40333,3333
325000	3	0,04	13000
375000	1	0,01333333	5000
425000	6	0,08	34000
475000	5	0,06666667	31666,6667
525000	1	0,01333333	7000
575000	0	0	0
625000	2	0,02666667	16666,6667
675000	1	0,01333333	9000
725000	1	0,01333333	9666,6667
775000	0	0	0
825000	0	0	0
875000	1	0,01333333	11666,6667
925000	0	0	0
	75	1	264333,333

40

5. Descripción numérica con distribuciones de frecuencias: observaciones importantes

- El resultado no es idéntico al obtenido con los datos “reales” (trans. 6, con datos reales de GTINE: 261.277)
- El resultado variará según el número de clases de la distribución de frecuencias

5. Descripción numérica con distribuciones de frecuencias: desviación típica

- Desviación típica
- Retomamos la fórmula vista en la transparencia 19: usamos marcas de clase y frecuencias relativas
- La marca de clase sustituye a los valores reales

$$s_c = \sqrt{\sum (c_i - \bar{x}_c)^2 f_i}$$

5. Descripción numérica con distribuciones de frecuencias: la desviación típica

Resultado: 170.742

Comparar con transparencia 21

c_i	n	f	cf	$c_i - \bar{x}_c$	$(c_i - \bar{x}_c)^2 f_i$
	0				
25000	1	0,0133	333,333	-239333,333	763739259,3
75000	10	0,1333	10000,000	-189333,333	4779614815
125000	9	0,1200	15000,000	-139333,333	2329653333
175000	12	0,1600	28000,000	-89333,333	1276871111
225000	11	0,1467	33000,000	-39333,333	226909629,6
275000	11	0,1467	40333,333	10666,667	16687407,41
325000	3	0,0400	13000,000	60666,667	147217777,8
375000	1	0,0133	5000,000	110666,667	163294814,8
425000	6	0,0800	34000,000	160666,667	2065102222
475000	5	0,0667	31666,667	210666,667	2958696296
525000	1	0,0133	7000,000	260666,667	905961481,5
575000	0	0,0000	0,000	310666,667	0
625000	2	0,0267	16666,667	360666,667	3468811852
675000	1	0,0133	9000,000	410666,667	2248628148
725000	1	0,0133	9666,667	460666,667	2829517037
775000	0	0,0000	0,000	510666,667	0
825000	0	0,0000	0,000	560666,667	0
875000	1	0,0133	11666,667	610666,667	4972183704
	75	1	264333,333		
				Varianza	2915288889
				Desv típica	170742,1708

5. Descripción numérica con distribuciones de frecuencias: coeficiente de asimetría

- Coeficiente de asimetría
- Utilizamos la fórmula de la transparencia 32
- Usando las marcas de clase en lugar de los valores

$$CA_c = \sum \left(\frac{c_i - \bar{x}_c}{s_c} \right)^3 f_i$$

5. Descripción numérica con distribuciones de frecuencias: el coeficiente de asimetría

c	n	f	cf	ci-xi	(ci-xi)^2*fi	((ci-xi)/sc)^3	(((ci-xi)/sc)^3)*f
	0						
25000	1	0,0133	333,333	-239333,333	763739259,3	-2,7541474	-0,03672196
75000	10	0,1333	10000,000	-189333,333	4779614815	-1,3635116	-0,18180155
125000	9	0,1200	15000,000	-139333,333	2329653333	-0,5434288	-0,06521146
175000	12	0,1600	28000,000	-89333,333	1276871111	-0,1432248	-0,02291597
225000	11	0,1467	33000,000	-39333,333	226909629,6	-0,0122253	-0,00179305
275000	11	0,1467	40333,333	10666,667	16687407,41	0,00024382	3,57598E-05
325000	3	0,0400	13000,000	60666,667	147217777,8	0,04485677	0,001794271
375000	1	0,0133	5000,000	110666,667	163294814,8	0,27228774	0,003630503
425000	6	0,0800	34000,000	160666,667	2065102222	0,83321092	0,066656874
475000	5	0,0667	31666,667	210666,667	2958696296	1,8783005	0,125220033
525000	1	0,0133	7000,000	260666,667	905961481,5	3,55823067	0,047443076
575000	0	0,0000	0,000	310666,667	0	6,02367562	0
625000	2	0,0267	16666,667	360666,667	3468811852	9,42530954	0,251341588
675000	1	0,0133	9000,000	410666,667	2248628148	13,9138066	0,185517422
725000	1	0,0133	9666,667	460666,667	2829517037	19,6398411	0,261864547
775000	0	0,0000	0,000	510666,667	0	26,754087	0
825000	0	0,0000	0,000	560666,667	0	35,4072188	0
875000	1	0,0133	11666,667	610666,667	4972183704	45,7499104	0,609998805
	75	1	264333,333				
				Varianza	2915288889	Coefasimetr	1,245058889
				Desv típica	170742,1708		

45

5. Descripción numérica con distribuciones de frecuencias: coeficiente de asimetría (2)

■ Resultado: 1,245

■ Con todos los datos (transparencia 36): nos salía 1,29

Tema 5- Descripción numérica (1)

46

Resumen

■ Conceptos y fórmulas de:

- ◆ La media
- ◆ La desviación típica
- ◆ La varianza
- ◆ El coeficiente de variación
- ◆ El coeficiente de asimetría
- ◆ Todos ellos para distribuciones de frecuencias

Tema 5- Descripción numérica (1)

47

Ejercicios recomendados

■ Del libro:

- ◆ 4.1
- ◆ 4.5 a) y c)
- ◆ 4.6
- ◆ 4.7
- ◆ 4.10
- ◆ 4.13 (Hay siete respuestas para a) y 1 para b)), suponiendo que son números distintos, claro

Tema 5- Descripción numérica (1)

48

Ejercicios recomendados

■ Ejercicios de exámenes

- ◆ Feb02: 3abcd, 5
- ◆ Jun02: 3abcd, 6
- ◆ Feb03: 3abcd, 4
- ◆ Sep03: 2abcd, 3
- ◆ Feb04: 5, 7
- ◆ Jul04: 5abcd, 6
- ◆ Feb05: 4, 6
- ◆ Jun05: 4, 6
- ◆ Feb06: 3abc, 4b
- ◆ Jun06: 3abc, 5
- ◆ Ene07: 3abc, 4
- ◆ Jul07: 3abc
- ◆ Ene 08: 3ab, 4
- ◆ Jul 08: 3ab, 4