Ejercicios T2- ANÁLISIS DE DATOS UNIDIMENSIONALES

1.- El número de clientes que acuden a un establecimiento en cierta hora han sido a lo largo del último mes ha sido:
3,4,4,5,7,1,2,3,5,4,6,10,6,3,8,2,3,4,5,6,4,8,2,9,11,15,4,2
Representar la tabla de frecuencias, el diagrama de barras y el diagrama acumulativo.

X_i (clientes)	n_i	f_i	N_i	F_i
1	1	1/28	1	1/28
2	4	4/28	5	5/28
3	4	4/28	9	9/28
4	6	6/28	15	15/28
5	3	3/28	18	18/28
6	3	3/28	21	21/28
7	1	1/28	22	22/28
8	2	2/28	24	24/28
9	1	1/28	25	25/28
10	1	1/28	26	26/28
11	1	1/28	27	27/28
15	1	1/28	28	1
	28	1		

Diagrama de barras

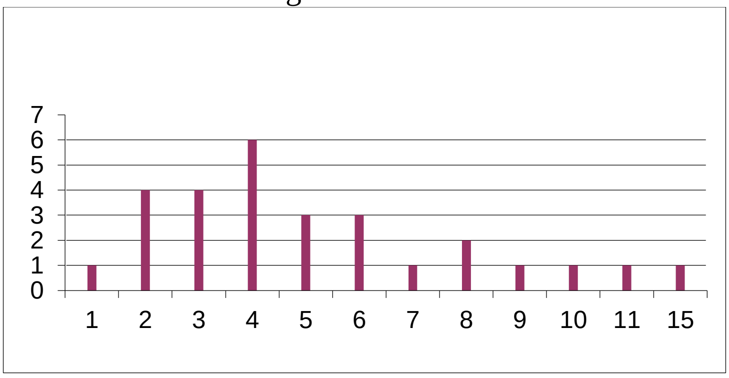
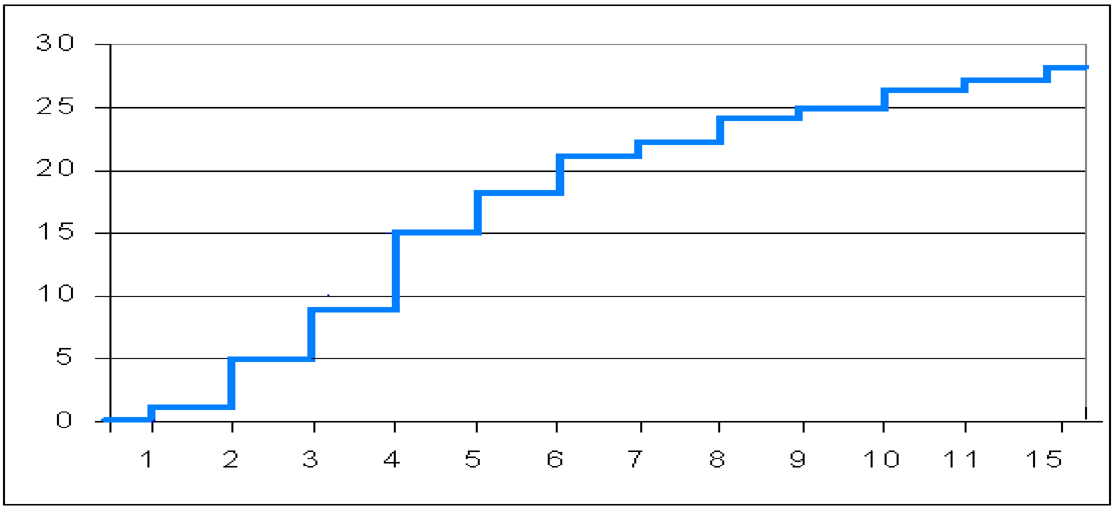


Diagrama acumulativo

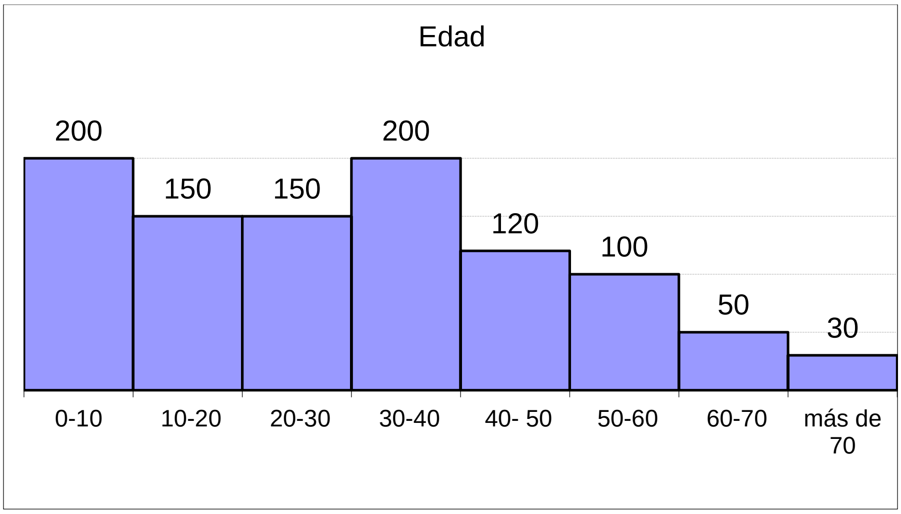


2.- Una población tiene una distribución por edades como la siguiente:

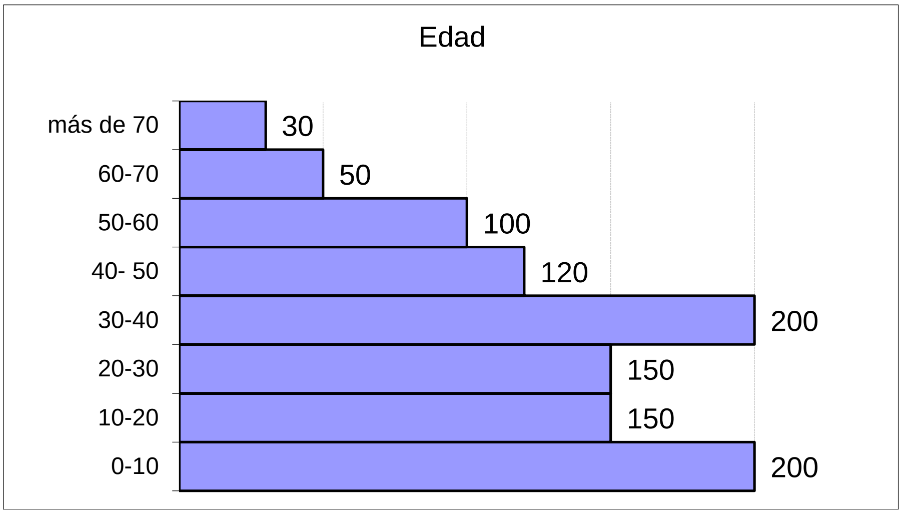
Edad	población
0-10	200
10-20	150
20-30	150
30-40	200
40- 50	120
50-60	100
60-70	50
Más de 70	30
	1000

Obtener su representación gráfica Histograma/ pirámide de población.
La media de edad de la población, la edad más frecuente, la edad mediana, varianza de la edad de esta población.

El histograma resultaría como:



Si este mismo histograma lo colocamos en posición vertical con rectángulos horizontales para cada tramo de edad obtenemos la habitual "pirámide de población":

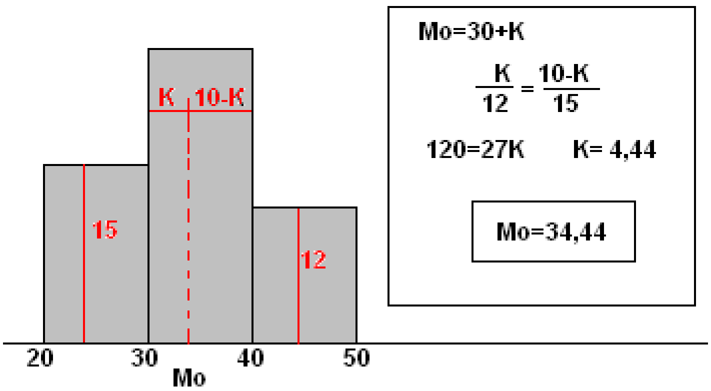


Para obtener las edad más frecuente debemos considerar que se trata de una distribución con valores agrupados una vez obtenido el "intervalo modal" habrá que determinar "internamente" el valor de la moda:

Observamos así que hay dos intervalos modales : el primero 0-10 años y el cuarto: 30-40 años:

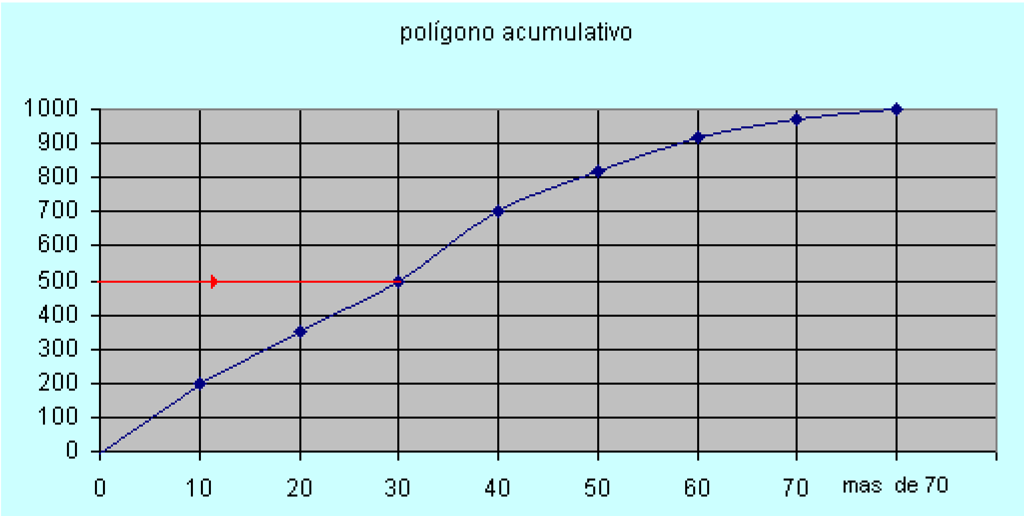
Para determinar las dos modas consideramos que ésta estará proporcionalmente más cerca del intervalo adyacente de mayor densidad de frecuencia (mayor altura del histograma) y proporcionalmente más alejada del intervalo adyacente de menor densidad de frecuencia:

Esto quiere decir que para la primera moda ésta será "obviamente" 10 años (el extremo superior del intervalo modal).Y para la segunda moda tendremos:



De forma que las dos **modas** son **10** y **34,4 años** como edades más frecuentes.
Respecto a la mediana es fácil comprobar, tanto en la tabla de frecuencias acumuladas como en el polígono acumulativo que la frecuencia acumulada alcanza el valor de la mitad de las observaciones (500) justo en el extremo superior del intervalo mediano (20-30) por lo que la mediana es 30 años.

Edad	población	Ni
0-10	200	200
10-20	150	350
20-30	150	500
30-40	200	700
40- 50	120	820
50-60	100	920
60-70	50	970
más de 70	30	1000
	1000	



Para obtener, por último, la media, la varianza y otros indicadores numéricos de la distribución utilizaremos la tabla de frecuencias con las marcas de clase:

Edad	marcas de clase x_i	población n_i	$x_i n_i$	x_i^2	$x_i^2 n_i$
0-10	5	200	1000	25	5000
10-20	15	150	2250	225	33750
20-30	25	150	3750	625	93750
30-40	35	200	7000	1225	245000
40- 50	45	120	5400	2025	243000
50-60	55	100	5500	3025	302500
60-70	65	50	3250	4225	211250
más de 70	75	30	2250	5625	168750
		1000	30400		1303000
				a_2=	1303
		media=	**30,4**	**Varianza=**	**378,84**
				Dtípica=	**19,46**

3.- Las rentas familiares anuales de dos países se distribuyen según las características siguientes:

PAIS	RENTA FAM. MEDIA	VARIANZA	DESV. TIPICA	NUM. FAMILIAS
A	30000 €	81000000	9000 €	10 MILLONES
B	35000 €	90000000	9486,83 €	15 MILLONES

a) ¿En cuál de los dos países hay mayor variabilidad en la renta familiar anual?
b) ¿Cuál es la renta media de la agregación de los dos países?
c) Una familia que ingrese al año 45000 € ¿sería más "rica" en términos relativos en el país A o en el B?
d) Una familia que este en el percentil 75 de renta en el país A es más rica o menos (en relación al país en el que vive) que otra familia del país B que esté en el percentil 60.

a) La variabilidad de ambas distribuciones sólo podrá compararse si utilizamos un indicador de variabilidad relativa como el coeficiente de variación de Pearson, por ejemplo:

$V = \frac{S}{\bar{x}}$ que en los dos casos que nos ocupan quedaría como:

$$V_A = \frac{S_A}{\bar{x}_A} = \frac{9000}{30000} = 0{,}3 \qquad y \qquad V_B = \frac{S_B}{\bar{x}_B} = \frac{9486{,}83}{35000} = 0{,}2710$$

por lo que en el país A la variabilidad de las rentas familiares **es mayor en términos relativos** , aun cuando en términos absolutos fuera menor (desviación típica más baja)

b) La renta media de la agregación de ambos países vendrá dada por la expresión:

$$\bar{x} = \frac{N_A.\bar{x}_A + N_B.\bar{x}_B}{N_A + N_B} = \frac{30000 \cdot 10000000 + 35000 \cdot 15000000}{25000000} = 33000\, euros$$

c) Para saber si una familia con unos ingresos de 45000 € sería más rica en uno u otro país deberemos comparar el valor "tipificado" según una y otra distribución:

El valor tipificado correspondiente a 45000 en la primera distribución sería:

$$Z_A = \frac{45000 - \bar{x}_A}{S_A} = \frac{45000 - 30000}{9000} = 1{,}667$$

El valor tipificado correspondiente a 45000 en la segunda distribución sería:

$$Z_B = \frac{45000 - \bar{x}_B}{S_B} = \frac{45000 - 35000}{9486{,}83} = 1{,}054$$

Por lo tanto esa familia sería **relativamente más rica** en el segundo país.

d) Una familia que esté en el percentil 75 deja por debajo al 75 % de las familias de la distribución a la que pertenece (sea ésta cual fuere) mientras que una familia que esté en percentil 60 deja al 60 % por debajo de ella.

Así pues en relación a su propio país la familia del país A sería más rica.
(Posiblemente en términos absolutos ocurra lo contrario)

4.- Dada una variable estadística cuya media es 100 y cuya desviación típica es 25 ¿qué transformación habrá que hacer para que la variable transformada tenga media 10 y desviación típica 5 ?

Por un lado tenemos que la variable original , x, es tal que su media es 100 y su desviación típica es 25

Y realizamos un transformación a y = a + bx pretendiendo que su media sea 10 y su desviación típica 5

La media de y vendrá dada por : $\bar{y} = a + b\bar{x} = a + 100b$

y la desviación típica vendrá dada por : $S_y = bS_x = 25b$

por lo que tendremos que : $\left.\begin{array}{l} \bar{y} = 10 = a + 100b \\ S_y = 5 = 25b \end{array}\right\} \quad \left. b = \frac{1}{5} \right\} 10 = a + \frac{100}{5} \rightarrow a = -10$

esto es, la transformación es: $y = -10 + \frac{x}{5} = \frac{x - 50}{5}$

5.-Sabiendo que los momentos ordinarios de orden 1, 2,3, y 4 de cierta distribución son a_1=10, a_2=181, a_3=930 y a_4= 1200. Obtener media, varianza, m_3, m_4 coeficiente de asimetría y coeficiente de curtosis.

La media es el momento ordinario de primer orden: $\bar{x} = a_1 = 10$

La varianza es el momento central de segundo orden y basándonos en la expresión de recurrencia entre momentos centrales y ordinarios ; esto es $m_r = \left(a - \bar{x}\right)^{(r)} = \sum_{i=0}^{r} (-1)^i \cdot \binom{r}{i} a_{r-i} \cdot \bar{x}^i$

Tendremos que:

$$S^2 = m_2 = \left(a - \bar{x}\right)^{(2)} = \sum_{i=0}^{2} (-1)^i \cdot \binom{2}{i} a_{2-i} \cdot \bar{x}^i = \binom{2}{0} a_2 - \binom{2}{1} a_1 \bar{x} + \binom{2}{2} \bar{x}^2 = a_2 - 2a_1\bar{x} + \bar{x}^2 =$$
$$= a_2 - 2\bar{x}^2 + \bar{x}^2 = a_2 - \bar{x}^2 = 181 - 10^2 = 81$$

Por tanto la desviación típica es 9

El momento central de tercer orden vendrá dado por:

$$m_3 = \left(a - \bar{x}\right)^{(3)} = \sum_{i=0}^{3} (-1)^i \cdot \binom{3}{i} a_{3-i} \cdot \bar{x}^i = \binom{3}{0} a_3 - \binom{3}{1} a_2 \bar{x} + \binom{3}{2} a_1 \bar{x}^2 - \binom{3}{3} \bar{x}^3 =$$
$$= a_3 - 3a_2\bar{x} + 3a_1\bar{x}^2 - \bar{x}^3 = a_3 - 3a_2\bar{x} + 3\bar{x}^3 - \bar{x}^3 = a_3 - 3a_2\bar{x} + 2\bar{x}^3 =$$
$$= 930 - 3 \cdot 181 \cdot 10 + 2 \cdot 10^3 = -2500$$

En cuanto al momento central de cuarto orden tendremos:

$$m_4 = \left(a - \bar{x}\right)^{(4)} = \sum_{i=0}^{4} (-1)^i \cdot \binom{4}{i} a_{4-i} \cdot \bar{x}^i = \binom{4}{0} a_4 - \binom{4}{1} a_3 \bar{x} + \binom{4}{2} a_2 \bar{x}^2 - \binom{4}{3} a_1 \bar{x}^3 + \binom{4}{4} \bar{x}^4 =$$
$$= a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 4a_1\bar{x}^3 + \bar{x}^4 = a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 4\bar{x}^4 + \bar{x}^4 = a_4 - 4a_3\bar{x} + 6a_2\bar{x}^2 - 3\bar{x}^4 =$$
$$= 1200 - 4 \cdot 930 \cdot 10 + 6 \cdot 181 \cdot 10^2 - 2 \cdot 10^4 = 42600$$

Por último los coeficientes de asimetría y curtosis se obtienen respectivamente a partir de los anteriores momentos como:

$g_1 = \frac{m_3}{S^3} = \frac{-2500}{9^3} = -3{,}42$ (distribución asimétrica por la izquierda) y

$g_2 = \frac{m_4}{S^4} - 3 = \frac{42600}{9^4} - 3 = 3{,}49$ (distribución más cúrtica que la normal : leptocúrtica)

6.-A partir de una variable estadística X hemos construido otra U tal que U = 5 + 10 X . El coeficiente de Asimetría de la variable U resulto ser 0 y su coeficiente de curtosis 1.2 ¿Cuánto valdrán el coeficiente de asimetría y el de curtosis para la variable original, X ?

Como los coeficientes de asimetría y de curtosis no dependen de las transformaciones lineales (cambios de origen y escala) , los coeficientes de asimetría y curtosis de la variable X serán, respectivamente 0 y 1,2.

(Este hecho puede observarse de que $g_1(x)=\frac{m_3(x)}{S_x^3}=\frac{m_3(u)/10^3}{S_u^3/10^3}=\frac{m_3(u)}{S_u}=g_1(u)$ y

$g_2(x)=\frac{m_4(x)}{S_x^4}-3=\frac{m_4(u)/10^4}{S_u^4/10^4}-3=\frac{m_4(u)}{S_u^4}-3=g_2(u)$)

7.- Un grupo de 60 adolescentes tienen unos gastos semanales expresados en euros tal como se indican en la tabla:

X	n
20	5
35	10
50	30
65	10
80	5
	60

Obtener el gasto medio la varianza , ver si la distribución es simétrica y analizar su curtosis

Para resolver el problema podemos reducir la complejidad aritmética de los cálculos realizando un cambio (lineal) de variable (un cambio de origen y de unidad). Estos cambios afectarán a la media y a la varianza pero no a los coeficientes de asimetría y de curtosis.

Si hacemos el cambio: $u=\frac{x-50}{15}$ (y por lo tanto $x=15u+50$) una vez obtenidas la media y la varianza de u tendremos que : $\bar{x}=15\bar{u}+50$ y $S_x^2=15^2S_u^2$

Procedamos:

u	n	un	u^2	u^2n	u^3	u^3n	u^4	u^4n
-2	5	-10	4	20	-8	-40	16	80
-1	10	-10	1	10	-1	-10	1	10
0	30	0	0	0	0	0	0	0
1	10	10	1	10	1	10	1	10
2	5	10	4	20	8	40	16	80
	60	0		60		0		180

De forma que:

$$\bar{u}=\frac{\sum_i u_i n_i}{N}=\frac{0}{60}=0 \qquad a_2(u)=\frac{\sum_i u_i^2 n_i}{N}=\frac{60}{60}=1 \qquad S_u^2=a_2(u)-\bar{u}^2=1$$

$$a_3(u)=\frac{\sum_i u_i^3 n_i}{N}=\frac{0}{60}=0 \qquad m_3(u)=(a-\bar{u})^{(3)}=a_3(u)-3a_2(u).\bar{u}+2.\bar{u}^3=0-3.1.0+2.0=0$$

$$g_1=\frac{m_3(u)}{S_u^3}=\frac{0}{1^3}=0$$

$$a_4(u)=\frac{\sum_i u_i^4 n_i}{N}=\frac{180}{60}=3$$

$$m_4(u)=(a-\bar{u})^{(4)}=a_4(u)-4a_3(u).\bar{u}+6.a_2(u).\bar{u}^2-3.\bar{u}^4=3-4.0.0+6.1.0-3.0=3$$

$$g_2=\frac{m_4(u)}{S_u^4}-3=\frac{3}{1^4}-3=0$$

y por lo tanto la media y la varianza de la variable original (los gastos) quedarán como:

$$\bar{x}=15\bar{u}+50=15.0+50=50\text{ €} \qquad S_x^2=15^2S_u^2=225.1=225\text{ €}^2$$

el coeficiente de asimetría y el de curtosis serían invariantes y pr lo tanto ambos iguales a cero

Con los valores originales los cálculos darían como resultado :

x	n	xn	x^2	x^2n	x^3	x^3n	x^4	x^4n		Media	50
20	5	100	400	2000	8000	40000	160000	800000		a^2	2725
35	10	350	1225	12250	42875	428750	1500625	15006250		S^2	225
50	30	1500	2500	75000	125000	3750000	6250000	1,88E+08		S	15
65	10	650	4225	42250	274625	2746250	17850625	1,79E+08		a_3	158750
80	5	400	6400	32000	512000	2560000	40960000	2,05E+08		m_3	0
	60	3000		163500		9525000		5,87E+08		g_1	0
										a_4	9776875
										m_4	151875
										g_2	0

8.- Sabemos que en una distribución determinada sus distintos momentos ordinarios son :
a_1=2; a_2=6; a_3=12; a_4=3; a_5=10 . Obtener el momento central de 5º orden: m_5

$$m_5=(a-\bar{x})^{(5)}=a_5-5.a_4.\bar{x}+10.a_3.\bar{x}^2-10.a_2.\bar{x}^3+5.a_1\bar{x}^4-\bar{x}^5=$$
$$=a_5-5.a_4.\bar{x}+10.a_3.\bar{x}^2-10.a_2.\bar{x}^3+4.\bar{x}^5=$$
$$=10-5.3.2+10.12.2^2-10.6.2^3+4.2^5=10-30+480-480+128=108$$

9.- Las cifras de ventas diarias de cierto comercio han oscilado este mes entre 100 € y 10000€ . Sabiendo esto ¿Cuál de las siguientes afirmaciones puede ser cierta?

a) $\bar{x} = 5500 \quad S = 0$
b) $\bar{x} = 55 \quad S = 100$
c) $\bar{x} = 5500 \quad S = -50$
d) $\bar{x} = 5500 \quad S = 50$

La media debe ser un valor del rango de variabilidad de la variable: un valor comprendido entre 100 y 10000 Euros, pero como sabemos que las ventas han **oscilado** entre esos valores la varianza no puede ser cero (y la desviación típica tampoco). (Si la varianza es cero ello significa que la variable sólo tomo un único valor). La única opción posible es la **d**

10.- Sabiendo que una distribución es simétrica y que su desviación típica es 1 y su momento central de cuarto orden es 22 ¿Cuál de las siguientes afirmaciones es cierta y por qué?

a) La distribución es mesocúrtica
b) La distribución es platicúrtica
c) La distribución es leptocúrtica
d) No podemos determinar su curtosis

Su coeficiente de curtosis será: $g_2 = \frac{m_4(u)}{S_u^4} - 3 = \frac{22}{1^4} - 3 = 19 > 0$

Por lo tanto es leptocúrtica .