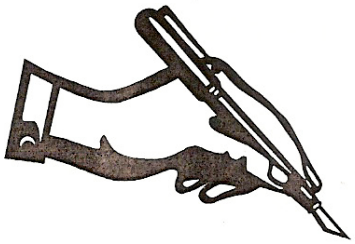
UNIVERSIDAD AUTONOMA DE MADRID

Estadística Descriptiva - EXCEL - SPSS
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

Estadística Teórica I



ESTADÍSTICA UNIDIMENSIONAL

UNIVERSIDAD AUTONOMA DE MADRID

Estadística Descriptiva - EXCEL - SPSS
Facultad Ciencias Económicas y Empresariales
Departamento de Economía Aplicada
Profesor: Santiago de la Fuente Fernández

1. En la tabla se muestran las rentas (en miles de euros) y el número de personas que las perciben:

Rentas (miles euros) $[L_i - L_{i+1})$	n_i
3 - 7	12
7 - 13	18
13 - 17	24
17 - 23	12
23 - 27	12

Se quiere obtener:

a) El polígono de frecuencias absolutas y el histograma.
b) Mediana, Percentil 75 y Moda.
c) Media Aritmética, Media Geométrica y Armónica.
d) Desviación Media (respecto a la media) y Coeficiente de Variación Media.
e) Coeficientes de Asimetría de Pearson y de Fisher.
f) Hallar la Media Aritmética y Desviación Típica utilizando un cambio de variable.
g) Coeficiente de Curtosis.
h) Concentración de la renta (curva de Lorenz, Índice de Gini).

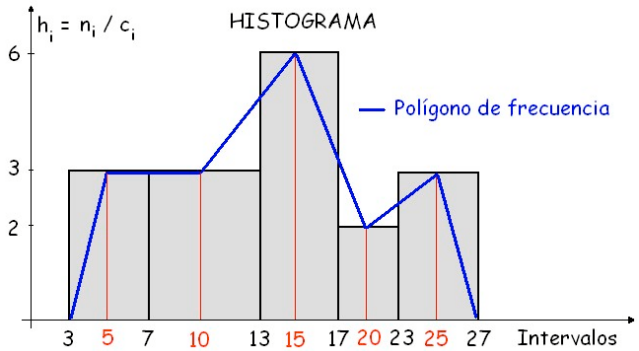
Solución:

a) El polígono de frecuencias absolutas y el histograma.- La tabla de frecuencias absolutas:

Rentas (miles euros) $[L_i - L_{i+1})$	x_i	n_i	N_i		Amplitud c_i	$h_i = \frac{n_i}{c_i}$
3 - 7	5	12	12	**39**	4	3
7 - 13	10	18	30		6	3
13 - 17	15	24	54		4	6
17 - 23	20	12	66	**58,5**	6	2
23 - 27	25	12	78		4	3

En la construcción del histograma hemos de colocar encima de cada intervalo un rectángulo cuyo área sea igual (en número) a la frecuencia absoluta de dicho intervalo, procediendo a calcular la altura h_i de cada rectángulo

$h_i = \frac{n_i}{c_i}$ donde c_i es la longitud del intervalo



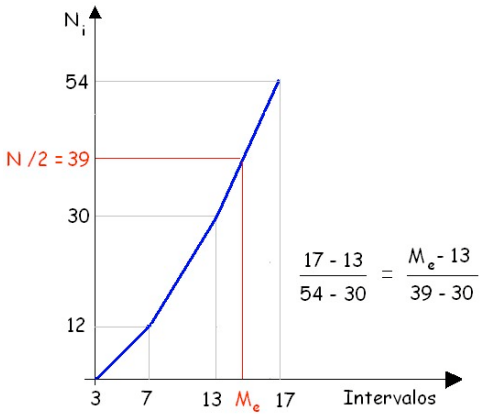
b) Mediana, Percentil 75 y Moda.

En las tablas de tipo III, los intervalos no son de amplitud constante.

Para calcular la Mediana, $\frac{N}{2}=\frac{78}{2}=39$. La observación 39 se encuentra en el intervalo [13 - 17)

Calculamos la Mediana:

$$M_e = L_i + \frac{\frac{N}{2}-N_{i-1}}{N_i-N_{i-1}}\,c_i = 13+\frac{\frac{78}{2}-30}{54-30}\,4 = 13+\frac{39-30}{54-30}\,4 = 14{,}5$$



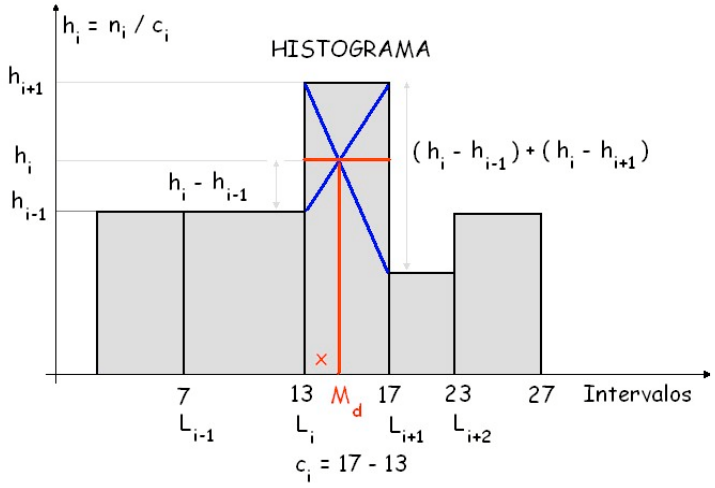
En la representación gráfica se establece una proporcionalidad entre las bases y las alturas.

- En el caso del Percentil 75, $\frac{75.78}{100}=58{,}5$

La observación 58,5 se encuentra en el intervalo [17 - 23)

$$P_{75} = L_i + \frac{\frac{75.N}{100}-N_{i-1}}{N_i-N_{i-1}}\,c_i = 17+\frac{58{,}5-54}{66-54}\,6 = 18{,}125$$

- La Moda es el intervalo de máxima frecuencia. Por tanto, el intervalo modal es [13 - 17). La posición exacta de la moda se calcula estableciendo una proporcionalidad entre las bases y las alturas.



$$\frac{c_i}{M_d - L_i} = \frac{(h_i - h_{i-1}) + (h_i - h_{i+1})}{(h_i - h_{i-1})} \Rightarrow M_d = L_i + \frac{(h_i - h_{i-1})}{(h_i - h_{i-1}) + (h_i - h_{i+1})}\,c_i$$

Cuando la amplitud de los intervalos es constante c_i = constante

$$M_d = L_i + \frac{(n_i - n_{i-1})}{(n_i - n_{i-1}) + (n_i - n_{i+1})}\,c_i$$

En nuestro caso, $M_d = L_i + \frac{(h_i - h_{i-1})}{(h_i - h_{i-1}) + (h_i - h_{i+1})} c_i$ con lo cual, $M_d = 13 + \frac{(6-3)}{(6-3)+(6-2)} 4 = 13{,}428$

La moda aproximada cuando existen distintas amplitudes: $M_d = L_i + \frac{h_{i+1}}{h_{i-1} + h_{i+1}} c_i = 13 + \frac{2}{3+2} 4 = 14{,}6$

c) Media Aritmética, Media Geométrica y Armónica.

Rentas $[L_i - L_{i+1})$	x_i	n_i	$x_i.n_i$	n_i / x_i	$\log x_i$	$n_i \log x_i$
3 - 7	5	12	60	2,4	0,6989	8,3868
7 - 13	10	18	180	1,8	1	18
13 - 17	15	24	360	1,6	1,1760	28,224
17 - 23	20	12	240	0,6	1,3010	15,612
23 - 27	25	12	300	0,48	1,3979	16,7748
		$\sum_{i=1}^{5} n_i = 78$	$\sum_{i=1}^{5} x_i.n_i = 1140$	$\sum_{i=1}^{5} n_i / x_i = 6{,}88$		$\sum_{i=1}^{5} n_i \log x_i = 86{,}9976$

- Media aritmética: $\bar{x} = \frac{\sum_{i=1}^{5} x_i.n_i}{N} = \frac{1140}{78} = 14{,}615$ (miles de euros)

- Media geométrica: $\bar{x}_G = \sqrt[N]{x_1^{n_1} x_2^{n_2} x_3^{n_3} \cdots x_k^{n_k}}$, para el cálculo se procede tomando logaritmos, con lo cual: $\log \bar{x}_G = \frac{1}{N} \log \left[x_1^{n_1} x_2^{n_2} x_3^{n_3} \cdots x_k^{n_k} \right] = \frac{1}{N} \left[\log(x_1^{n_1}) + \log(x_2^{n_2}) + \cdots + \log(x_k^{n_k}) \right] = \frac{1}{N} \sum_{i=1}^{k} n_i \log x_i$

$\bar{x}_G = 10^{\left[\frac{1}{N} \sum_{i=1}^{k} n_i \log x_i \right]}$

en consecuencia, $\log \bar{x}_G = \frac{1}{N} \sum_{i=1}^{5} n_i \log x_i = \frac{1}{78}(86{,}9976) = 1{,}115 \Rightarrow \bar{x}_G = 10^{1{,}115} = 13{,}031$ (miles de euros)

- Media armónica: $\bar{x}_A = \frac{N}{\sum_{i=1}^{k} \frac{n_i}{x_i}}$, con lo cual, $\bar{x}_A = \frac{N}{\sum_{i=1}^{5} \frac{n_i}{x_i}} = \frac{78}{6{,}88} = 11{,}337$ (miles de euros)

Obsérvese que se verifica la fórmula de Foster, para distribuciones de frecuencias con valores positivos: $\bar{x}_A \le \bar{x}_G \le \bar{x}$

d) Desviación Media (respecto a la media) y Coeficiente de Desviación Media.

Rentas $[L_i - L_{i+1})$	x_i	n_i	$x_i.n_i$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert n_i$
3 - 7	5	12	60	9,615	115,38
7 - 13	10	18	180	4,615	83,07
13 - 17	15	24	360	0,385	9,24
17 - 23	20	12	240	5,385	64,62
23 - 27	25	12	300	10,385	124,62
		$\sum_{i=1}^{5} n_i = 78$	$\sum_{i=1}^{5} x_i.n_i = 1140$		$\sum_{i=1}^{5} \lvert x_i - \bar{x} \rvert n_i = 396{,}93$

$$\bar{x} = \frac{\sum_{i=1}^{5} x_i . n_i}{N} = \frac{1140}{78} = 14{,}615$$

- La desviación media respecto a la media aritmética: $D_M(\bar{x}) = \frac{\sum_{i=1}^{k} |x_i - \bar{x}| n_i}{N}$

con lo cual, $D_M(\bar{x}) = \frac{\sum_{i=1}^{5} |x_i - \bar{x}| n_i}{N} = \frac{396{,}93}{78} = 5{,}088$

- El coeficiente de variación media respecto a la media aritmética: $CV_{D_M}(\bar{x}) = \frac{D_{M_{\bar{x}}}}{|\bar{x}|}$

$$CV_{D_M}(\bar{x}) = \frac{D_{M_{\bar{x}}}}{|\bar{x}|} = \frac{5{,}088}{14{,}615} = 0{,}3481$$

NOTA.- La desviación media respecto a la mediana: $D_M(M_e) = \frac{\sum_{i=1}^{k} |x_i - M_e| n_i}{N}$ y el coeficiente de variación media respecto a la mediana $CV_{D_M}(M_e) = \frac{D_M(M_e)}{|M_e|}$

e) Coeficientes de Asimetría de Pearson y de Fisher.

- El coeficiente de asimetría de Pearson exige el cálculo de la Moda M_d y la desviación típica σ

$$A_P = \frac{\bar{x} - M_d}{\sigma} \begin{cases} A_P > 0 & \text{Asimetría a la derecha o positiva} \\ A_P = 0 & \text{Simetría} \\ A_P < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

Este coeficiente tiene sentido cuando la moda es única

- El coeficiente de asimetría de Fisher: $A_F = \frac{m_3}{\sigma^3} \begin{cases} A_F > 0 & \text{Asimetría a la derecha o positiva} \\ A_F = 0 & \text{Simetría} \\ A_F < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$

Sabemos que, $M_d = 13{,}428$ $\quad \bar{x} = 14{,}615$

Rentas $[L_i - L_{i+1})$	x_i	n_i	$x_i . n_i$	$x_i - \bar{x}$	$(x_i - \bar{x})^2 n_i$	$(x_i - \bar{x})^3 n_i$
3 - 7	5	12	60	- 9,615	1109,38	- 10666,676
7 - 13	10	18	180	- 4,615	383,368	- 1769,243
13 - 17	15	24	360	0,385	3,557	1,369
17 - 23	20	12	240	5,385	347,979	1873,865
23 - 27	25	12	300	10,385	1294,179	13440,046
		$\sum_{i=1}^{5} n_i = 78$	$\sum_{i=1}^{5} x_i . n_i = 1140$		3138,463	2879,361

♦ Varianza, desviación típica y tercer momento respecto a la media:

$$m_2 = \sigma^2 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^2 n_i}{N} = \frac{3183{,}463}{78} = 40{,}237 \quad \Rightarrow \sigma = \sqrt{40{,}237} = 6{,}343$$

El tercer momento respecto a la media: $m_3 = \frac{\sum_{i=1}^{5}(x_i - \bar{x})^3 n_i}{N} = \frac{2879{,}361}{78} = 36{,}915$

♦ El coeficiente de asimetría de Pearson: $A_P = \frac{\bar{x} - M_d}{\sigma} = \frac{14{,}615 - 13{,}428}{6{,}343} = 0{,}187 > 0$, con lo que la distribución presenta una asimetría a la derecha o positiva.

♦ El coeficiente de asimetría de Fisher: $A_F = \frac{m_3}{\sigma^3} = \frac{36{,}915}{6{,}343^3} = 0{,}145 > 0$, con lo que la distribución presenta una asimetría a la derecha o positiva.

f) Hallar la media aritmética y desviación típica utilizando un cambio de variable.

Hacemos el cambio de variable $z_i = \frac{x_i - 15}{5}$

Rentas $[L_i - L_{i+1})$	x_i	n_i	$x_i.n_i$	z_i	$z_i.n_i$	$z_i^2.n_i$
3 - 7	5	12	60	-2	-24	48
7 - 13	10	18	180	-1	-18	18
13 - 17	15	24	360	0	0	0
17 - 23	20	12	240	1	12	12
23 - 27	25	12	300	2	24	48
		$\sum_{i=1}^{5} n_i = 78$	$\sum_{i=1}^{5} x_i.n_i = 1140$		$\sum_{i=1}^{5} z_i.n_i = -6$	$\sum_{i=1}^{5} z_i^2.n_i = 126$

♦ Media aritmética: $a_1 = \bar{z} = \frac{\sum_{i=1}^{5} z_i.n_i}{N} = \frac{-6}{78} = -0{,}0769$

Siendo $z_i = \frac{x_i - 15}{5} \Rightarrow \bar{x} = 15 + 5\bar{z} = 15 + 5(-0{,}0769) = 14{,}615$

NOTA.- $E[a + bx] = a + b\,E(\bar{x})$

♦ Varianza: $\sigma_z^2 = a_2 - a_1^2 = \frac{\sum_{i=1}^{5} z_i^2 n_i}{N} - \left[\frac{\sum_{i=1}^{5} z_i n_i}{N}\right]^2 = \frac{126}{78} - (-0{,}0769)^2 = 1{,}60947$

Desviación típica: $\sigma_z = \sqrt{1{,}60947} = 1{,}2686$

Como $z_i = \dfrac{x_i - 15}{5} \Rightarrow \sigma_x = 5\sigma_z = 5(1{,}2686) = 6{,}343$

NOTA.- $V[a+bx] = b^2 . V(x)$

g) Coeficiente de Curtosis.

La curtosis de una distribución de frecuencias es el apuntamiento que presenta el polígono de frecuencias alrededor de la media. El coeficiente de curtosis $g_2 = \dfrac{m_4}{\sigma^4} - 3$, siendo

$$m_2 = \sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N} \quad \text{y} \quad m_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N}$$

- $g_2 > 0$ Más apuntamiento que la normal: Leptocúrtica
- $g_2 = 0$ Igual apuntamiento que la normal: Mesocúrtica
- $g_2 < 0$ Menor apuntamiento que la normal: Platicúrtica

En la distribución, conocemos: $\bar{x} = 14{,}615 \qquad \sigma_x = 6{,}343$

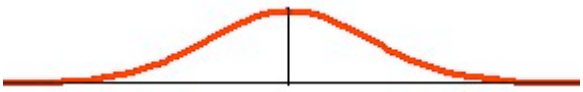
Rentas $[L_i - L_{i+1})$	x_i	n_i	$x_i.n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 n_i$	$(x_i - \bar{x})^4 n_i$
3 - 7	5	12	60	92,4482	1109,38	102560,036
7 - 13	10	18	180	21,2982	383,368	8165,039
13 - 17	15	24	360	0,1482	3,557	0,5271
17 - 23	20	12	240	28,9982	347,979	10090,747
23 - 27	25	12	300	107,848	1294,179	139574,293
		$\sum_{i=1}^{5} n_i = 78$	$\sum_{i=1}^{5} x_i.n_i = 1140$		3138,463	260390,6421

El momento de cuarto orden respecto a la media será:

$$m_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N} = \frac{260390{,}6421}{78} = 3338{,}3415$$

♦ El coeficiente de curtosis de Fisher: $g_2 = \dfrac{m_4}{\sigma^4} - 3 \Rightarrow \dfrac{3338{,}3415}{6{,}343^4} - 3 = -0{,}9377 < 0$

La distribución presenta menor apuntamiento que la normal: Platicúrtica



g) Concentración de la renta (curva de Lorenz, Índice de Gini).

Hasta este momento la palabra 'concentración' era la opuesta a 'dispersión', cuando nos ocupábamos del estudio descriptivo de los valores observados de la variable.

Con las medidas de concentración nuestro objetivo será analizar el total de los recursos repartidos entre todos los individuos que intervienen en la distribución.
Señalar que la cantidad total de los recursos no suelen estar siempre repartidos de forma equitativa, sino que, por el contrario, habrá individuos que se repartan una mayor cantidad de recursos que otros.

Para analizar la concentración necesitamos calcular las proporciones de individuos $p_i = \frac{N_i}{N}$ y de recursos acumulados $q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i} = \frac{\sum_{i=1}^{m} x_i n_i}{N\bar{x}}$. Destacar que, al estar ordenados los x_i de forma creciente, la proporción de individuos p_i siempre tiene que avanzar más rápido que la proporción de recursos repartidos q_i, es decir $p_i \geq q_i$. De este modo, la gráfica que se realiza sobre un cuadrado de lado la unidad "*curva de concentración o curva de Lorenz*" siempre estará por encima de la diagonal principal del cuadrado.

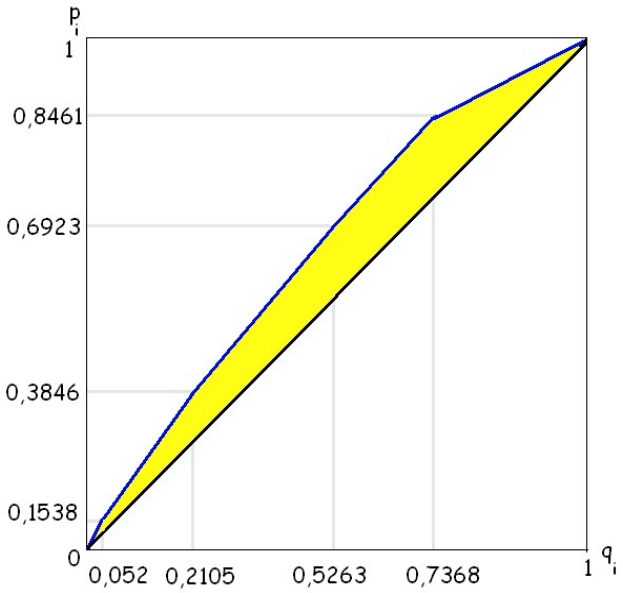
Realizamos la siguiente tabla:

Rentas $[L_i - L_{i+1})$	x_i	n_i	N_i	$p_i = N_i/N$	$x_i.n_i$	$\sum_{i=1}^{m} x_i n_i$	$q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i}$	$p_i - q_i$
3 - 7	5	12	12	0,1538	60	60	0,052	0,1018
7 - 13	10	18	30	0,3846	180	240	0,2105	0,1741
13 - 17	15	24	54	0,6923	360	600	0,5263	0,166
17 - 23	20	12	66	0,8461	240	840	0,7368	0,1093
23 - 27	25	12	78	-	300	1140	-	-
		$\sum_{i=1}^{5} n_i = 78$		$\sum_{i=1}^{4} p_i = 2,08$	1140		$\sum_{i=1}^{4} q_i = 1,53$	$\sum_{i=1}^{4} (p_i - q_i) = 0,551$

♦ La curva de concentración o curva de Lorenz

La idea de medir el área da como resultado el llamado *índice de concentración de Gini*, que se define como el doble del área comprendida entre la diagonal y la curva de Lorenz.

$$I_G = \frac{\sum_{i=1}^{k-1}(p_i - q_i)}{\sum_{i=1}^{k-1} p_i}$$



- Obsérvese que cuando $p_i = q_i \Rightarrow I_G = 0$ es el caso de equidistribución, la curva de Lorenz se encuentra sobre la diagonal del cuadrado.
- En el caso de máxima concentración (un individuo se lleva el total de los recursos), ocurre que $q_1 = q_2 = \cdots = q_{k-1} = 0$, el $I_G = 1$, la curva de Lorenz está sobre los lados del cuadrado.
- $0 \le I_G \le 1$, cuanto más equitativo sea el reparto de recursos será más cercano a cero, y más cercano a uno cuanto mayor concentración exista.
- Tanto la curva de Lorenz como el Índice de Gini se pueden presentar y calcular considerando las proporciones de individuos (p_i) y de recursos (q_i) en porcentajes.

♦ En nuestro caso, el Índice de Gini : $I_G = \frac{\sum_{i=1}^{4}(p_i - q_i)}{\sum_{i=1}^{4} p_i} = \frac{0,551}{2,08} = 0,2649$. Luego la renta, aunque no equidistribuida, no está muy concentrada.

2. Una cooperativa agrícola tiene cinco fincas en explotación. La producción de trigo y rendimientos por hectárea obtenidos son:

Fincas	Producción (Qm)	Rendimientos (Qm/Ha)
A	2400	12
B	3200	10
C	3600	15
D	4200	14
E	4400	16

¿Cuál es el rendimiento medio por hectárea para el conjunto de la cooperativa?

Solución:

Como se promedia una magnitud relativa, se debe utilizar la media armónica. Es decir: $\overline{x}_A = \dfrac{N}{\sum_{i=1}^{k} \dfrac{p_i}{x_i}}$

Fincas	Producción (Qm) p_i	Rendimientos (Qm/Ha) x_i	p_i / x_i
A	2400	12	200
B	3200	10	320
C	3600	15	240
D	4200	14	300
E	4400	16	275
	$N = \sum_{i=1}^{5} p_i = 17800$		$\sum_{i=1}^{5} p_i / x_i = 1335$

$$\overline{x}_A = \frac{17800}{1335} = 13{,}33 \ \text{Qm/Ha}$$

Nota.- En casos de magnitudes relativas el valor medio se tiene que calcular a través de la media armónica; transformando previamente la información se puede calcular utilizando la media aritmética. Es decir:

Fincas	Producción (Qm) p_i	Rendimientos (Qm/Ha) x_i	Superficie (Ha) $n_i = p_i / x_i$	Producción $x_i n_i$	$p_i n_i$
A	2400	12	200	2400	28800
B	3200	10	320	3200	32000
C	3600	15	240	3600	54000
D	4200	14	300	4200	58800
E	4400	16	275	4400	70400
	17800		$N = 1335$	$\sum_{i=1}^{5} x_i n_i = 17800$	$\sum_{i=1}^{5} p_i n_i = 244000$

El rendimiento medio sería: $\overline{R} = \dfrac{\sum_{i=1}^{5} x_i n_i}{N} = \dfrac{17800}{1335} = 13{,}33 \ \text{Qm/Ha}$

Adviértase que la media aritmética es la media aritmética de los rendimientos ponderados por la correspondiente superficie: $\overline{R} = \dfrac{\sum_{i=1}^{5} x_i n_i}{N} = \dfrac{17800}{200+320+240+300+275} = 13{,}33 \ \text{Qm/Ha}$

3. En la tabla adjunta se muestra la productividad (piezas/hora) de los empleados de una empresa, considerando su categoría profesional:

Categoría	Productividad (piezas/hora)
A	10
B	25
C	40
D	15

a) Hallar la productividad media en el conjunto de la empresa.
b) Hallar el tiempo medio empleado para fabricar una pieza. ¿Cuál es el número de piezas diarias en una jornada laboral de 8 horas?.

Solución:

a) La productividad media en la empresa (suponiendo que el número de empleados de cada categoría es el mismo) es la media armónica de la productividad de cada categoría, es decir:

$$\overline{x}_A = \frac{N}{\sum_{i=1}^{4} \frac{n_i}{x_i}} = \frac{4}{\frac{1}{10}+\frac{1}{25}+\frac{1}{40}+\frac{1}{15}} = \frac{4}{0{,}232} = 17{,}24 \ \text{piezas/hora}$$

b) El tiempo medio empleado a la hora $\overline{t}$ será el inverso de la media armónica:

$$\overline{t} = \frac{1}{\overline{x}_A} = \frac{1}{17{,}24 \ \text{piezas/hora}} = 0{,}058 \ \text{horas/pieza} = 0{,}058 \,.\, 60 = 3{,}48 \ \text{minutos/pieza}$$

El número de piezas diarias en una jornada de 8 horas será:

$$\overline{T} = \overline{x}_A \,.\, 8 = 17{,}24 \,.\, 8 = 137{,}92 \approx 138 \ \text{unidades}$$

4. Una distribución (x_i, n_i) presenta las siguientes características:

$$N = 20 \qquad M_d = 5 \qquad \overline{x} = 6 \qquad \sigma_x^2 = 2{,}4$$

Determinar los mismos parámetros para las distribuciones: $\begin{cases} (x_i + 3, n_i) \\ (10x_i, n_i) \end{cases}$

Solución:

a) Se define la variable $y_i = x_i + 3$, que considerando las propiedades de los estadísticos mencionados, verifica que:

- $\bar{y} = \frac{\sum (x_i + 3).n_i}{N} = \frac{\sum x_i.n_i}{N} + 3\frac{\sum .n_i}{N} = \bar{x} + 3 = 6 + 3 = 9$
- $\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2.n_i}{N} = \frac{\sum [(x_i + 3) - (\bar{x} + 3)]^2.n_i}{N} = \frac{\sum (x_i - \bar{x})^2.n_i}{N} = \sigma_x^2 = 2{,}4$
- $M_d(y) = M_d(x) + 3 = 5 + 3 = 8$

b) Se hace el cambio de variable $w_i = 10\, x_i$

- $\bar{w} = \frac{\sum 10\, x_i.n_i}{N} = 10\frac{\sum x_i.n_i}{N} = 10\,\bar{x} = 10\,.\,6 = 60$
- $\sigma_y^2 = \frac{\sum (y_i - \bar{y})^2.n_i}{N} = \frac{\sum [10\, x_i - 10\,\bar{x}]^2.n_i}{N} = 10^2\,\frac{\sum (x_i - \bar{x})^2.n_i}{N} = 100\,\sigma_x^2 = 240$
- $M_d(y) = 10\, M_d(x) = 10\,.\,5 = 50$

5. Dos distribuciones simétricas y campaniformes presentan la siguiente información:

Distribución X	Distribución Y
$M_e = 16$	$M_0 = 18$
$\sigma_x^2 = 36$	$\sigma_y^2 = 36$

¿Cuál de las dos distribuciones presenta mayor variabilidad?

Solución:

Como son distribuciones de tipo campaniforme y simétricas, la media aritmética, mediana y moda coinciden. En otras palabras, $\bar{x} = 16$ y $\bar{y} = 18$

Por otra parte, para analizar la variabilidad de las dos distribuciones no podemos recurrir a la comparación de varianzas, puede ser que ambas distribuciones se expresen en unidades diferentes, y además hay que relacionar la variabilidad con el promedio correspondiente.

En esta línea, utilizaremos el coeficiente de variación de Pearson:

$$C.V_x = \frac{\sigma_x}{\bar{x}} = \frac{6}{16} = 0{,}375 \qquad C.V_y = \frac{\sigma_y}{\bar{y}} = \frac{6}{18} = 0{,}333$$

La distribución Y presenta un coeficiente menor, por lo que tiene una dispersión relativa más pequeña.

6a. Una población se encuentra dividida en dos estratos

	Nº elementos	Media aritmética	Varianza
A	100	4	9
B	400	6	16

Hallar la media y la varianza para el conjunto de la población.

Solución:

El peso o tamaño relativo de cada estrato será: $p_i = \frac{N_i}{N}$

En este caso, $p_A = \frac{N_A}{N} = \frac{100}{500} = 0{,}2 \quad p_B = \frac{N_B}{N} = \frac{400}{500} = 0{,}8$

$$\bar{x}_A = 4 \qquad \sigma_A^2 = 9 \qquad \bar{x}_B = 6 \qquad \sigma_B^2 = 16$$

- La media total de la población será: $\bar{x} = p_A\,\bar{x}_A + p_B\,\bar{x}_B = 0{,}2.4 + 0{,}8.6 = 5{,}6$
- La varianza es: $\sigma_x^2 = p_A\,\sigma_A^2 + p_B\,\sigma_B^2 + p_A\,(\bar{x}_A - \bar{x})^2 + p_B\,(\bar{x}_B - \bar{x})^2$

$$\sigma_x^2 = 0{,}2\,.\,9 + 0{,}8\,.\,16 + 0{,}2.\,(4 - 5{,}6)^2 + 0{,}8.\,(6 - 5{,}6)^2 = 15{,}24$$

Nota.- Cuando la población se ha subdividido en k estratos o categorías, de forma que:
$N = N_1 + N_2 + \cdots + N_k$.

Estrato o Categoría	Tamaño	Media aritmética	Varianza
1	N_1	$\bar{x}_1$	σ_1^2
2	N_2	$\bar{x}_2$	σ_2^2
3	N_3	$\bar{x}_3$	σ_3^2
⋮	⋮	⋮	
k	N_k	$\bar{x}_4$	σ_k^2

El peso o tamaño relativo de cada estrato $p_i = \frac{N_i}{N}$

Los valores de la variable X se representan por x_{ij} , donde los subíndices indican el valor i-ésimo del estrato j-ésimo. La media aritmética de toda la población será entonces:

$$\bar{x} = \frac{(x_{11} + x_{21} + \cdots + x_{N_1 1}) + (x_{12} + x_{22} + \cdots + x_{N_2 2}) + \cdots + (x_{1k} + x_{2k} + \cdots + x_{N_k k})}{N} =$$

$$= \frac{\sum_{i=1}^{N_1} x_{i1} + \sum_{i=1}^{N_2} x_{i2} + \cdots + \sum_{i=1}^{N_k} x_{iN_k}}{N} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{N_j} x_{ij}}{N}$$, en función de las medias de cada estrato será:

$$\bar{x} = \frac{\sum_{i=1}^{N_1} x_{i1} + \sum_{i=1}^{N_2} x_{i2} + \cdots + \sum_{i=1}^{N_k} x_{iN_k}}{N} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j} x_{ij}}{N} = \frac{\sum_{j=1}^{k} N_j \dfrac{\sum_{i=1}^{N_j} x_{ij}}{N_j}}{N} = \sum_{j=1}^{k} p_j \bar{x}_j$$

Para calcular la varianza de toda la población, se tiene en cuenta:

$$\sigma_x^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ij}-\bar{x})^2}{N} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ij}-\bar{x}+\bar{x}_j-\bar{x}_j)^2}{N} = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}[(x_{ij}-\bar{x}_j)+(\bar{x}_j-\bar{x})^2]}{N} =$$

$$= \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ij}-\bar{x}_j)^2}{N} + \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{x}_j-\bar{x})^2}{N} + 2\,\frac{\overbrace{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ij}-\bar{x}_j).(\bar{x}_j-\bar{x})}^{=0}}{N}$$, con lo cual,

$$\sigma_x^2 = \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(x_{ij}-\bar{x}_j)^2}{N} + \frac{\sum_{j=1}^{k}\sum_{i=1}^{N_j}(\bar{x}_j-\bar{x})^2}{N} = \frac{\sum_{j=1}^{k} N_j \dfrac{\sum_{i=1}^{N_j}(x_{ij}-\bar{x}_j)^2}{N_j}}{N} + \frac{\sum_{j=1}^{k} N_j\,(\bar{x}_j-\bar{x})^2}{N}$$

entonces,

$\sigma_x^2 = \sum_{j=1}^{k} p_j \,.\, \sigma_{xj}^2 + \sum_{j=1}^{k} p_j \,.\, (\bar{x}_j - \bar{x})^2$, siendo σ_{xj}^2 la varianza del estrato o categoría j-ésimo.

Adviértase que en el caso de dos estratos o categorías:

$$\bar{x} = \sum_{j=1}^{2} p_j \bar{x}_j = p_1 \bar{x}_1 + p_2 \bar{x}_2$$

$$\sigma_x^2 = \sum_{j=1}^{2} p_j \,.\, \sigma_{xj}^2 + \sum_{j=1}^{2} p_j \,.\, (\bar{x}_j - \bar{x})^2 = p_1 \sigma_{x1}^2 + p_2 \sigma_{x2}^2 + p_1 (\bar{x}_1 - \bar{x})^2 + p_2 (\bar{x}_2 - \bar{x})^2$$

7. Dadas las observaciones de la variable (X,Y):

X \ Y	1	2	4
1	3	0	0
2	2	0	0
3	0	4	0
4	0	4	0
5	0	0	4

Determinar razonadamente:

a) La independencia o dependencia de las variables.
b) La recta de regresión de Y sobre X
c) ¿Cuál sería el valor de Y para X=6 según la regresión realizada?. ¿Es fiable el valor obtenido?.

Solución:

a) Se completa la tabla:

X \ Y	1	2	4	n_{x_i}
1	3	0	0	3
2	2	0	0	2
3	0	4	0	4
4	0	4	0	4
5	0	0	4	4
n_{y_j}	5	8	4	17

$$a_{11} = \frac{\sum_{i=1}^{5}\sum_{j=1}^{3} n_{ij}}{N} = \frac{1}{17}(1.1.3 + 2.1.2 + 3.2.4 + 4.2.4 + 5.4.4 = 8,412$$

Las distribuciones marginales de X e Y vienen reflejadas en la siguiente tabla:

x_i	n_{x_i}	$x_i.n_{x_i}$	$x_i^2.n_{x_i}$
1	3	3	3
2	2	4	8
3	4	12	36
4	4	16	64
5	4	20	100
	$\sum_{i=1}^{5} n_{x_i} = 17$	$\sum_{i=1}^{5} x_i.n_{x_i} = 55$	$\sum_{i=1}^{5} x_i^2.n_{x_i} = 211$

$$a_{10} = \bar{x} = \frac{\sum_{i=1}^{5} x_i.n_{x_i}}{N} = \frac{55}{17} = 3,235$$

$$a_{20} = \frac{\sum_{i=1}^{5} x_i^2.n_{x_i}}{N} = \frac{211}{17} = 12,412$$

$$\sigma_x^2 = a_{20} - a_{10}^2 = 12,412 - 3,235^2 = 1,95$$

Y_j	n_{y_j}	$y_j.n_{y_j}$	$y_j^2.n_{y_j}$
1	5	5	5
2	8	16	32
4	4	16	64
	$\sum_{j=1}^{3} n_{y_j} = 17$	$\sum_{j=1}^{3} y_j.n_{y_j} = 37$	$\sum_{j=1}^{3} y_j^2.n_{y_j} = 101$

$$a_{01} = \bar{y} = \frac{\sum_{j=1}^{3} y_j.n_{y_j}}{N} = \frac{37}{17} = 2,176$$

$$a_{02} = \frac{\sum_{j=1}^{3} y_j^2.n_{y_j}}{N} = \frac{101}{17} = 5,941$$

$$\sigma_y^2 = a_{02} - a_{01}^2 = 5,941 - 2,176^2 = 1,21$$

Las variables (X,Y) son independientes cuando la covarianza $m_{11} = a_{11} - a_{10}.a_{01} = 0$, en este caso,

$m_{11} = 8,42 - 3,235.2,176 = 1,38 \neq 0 \;\; \mapsto$ las variables X e Y no son independientes.

b) La recta de regresión de X sobre Y viene dada por la expresión: $y - \overline{y} = \frac{m_{11}}{\sigma_x^2}(x - \overline{x})$, con lo que,

$y - 2,176 = \frac{1,38}{1,95}(x - 3,235) \;\; \Rightarrow \;\; y = -0,089 + 0,7.x$ (recta de regresión de Y sobre X)

$\beta_{Y/X} = \frac{m_{11}}{\sigma_x^2} = \frac{1,38}{1,95} = 0,7$ coeficiente de regresión o pendiente de la recta de Y sobre X, al ser $\beta_{Y/X} > 0$ la recta es creciente.

c) Cuando $x = 6$, el valor de y según la recta de regresión será: $y = -0,089 + 0,7.6 = 4,111$

Para analizar la fiabilidad del valor obtenido recurrimos al coeficiente de correlación ρ (otras personas, lo hacen con el coeficiente de determinación ρ^2).

El coeficiente de determinación se define como el producto de los coeficientes de regresión:

$$\rho^2 = \beta_{Y/X}.\beta_{X/Y} = \frac{m_{11}}{\sigma_x^2}.\frac{m_{11}}{\sigma_y^2} = \frac{m_{11}^2}{\sigma_x^2.\sigma_y^2}$$

en nuestro caso, $\rho^2 = \frac{1,38^2}{1,95.1,21} = 0,80$. Dado que $0 \leq \rho^2 \leq 1$, y que el coeficiente de determinación está próximo a uno, podemos concluir que la fiabilidad de los resultados es muy grande.

Si hubiéramos optado por el coeficiente de correlación: $\rho = \frac{m_{11}}{\sigma_x.\sigma_y}$, se tendría:

$\rho = \frac{1,38}{\sqrt{1,95.1,21}} = 0,898$. Dado que $-1 \leq \rho \leq 1$, diríamos que la fiabilidad de los resultados es muy grande, estando en correlación positiva.

NOTA.- Para hallar el valor medio de la distribución condicionada X/Y=2

La distribución de frecuencias sería:

X \ Y=2	2	$n_{x_i/Y=2}$	$x_i.n_{x_i/Y=2}$
1	0	0	0
2	0	0	0
3	4	4	12
4	4	4	16
5	0	0	0
		8	28

$$\overline{x}_{Y=2} = \frac{\sum_{i=1}^{5} x_i.n_{x_i/Y=2}}{N} = \frac{28}{8} = 3,5$$

8. Una vacuna antitetánica se administró a una muestra de cuarenta y dos personas. Posteriormente, a las cinco horas de su inyección, se tomó la temperatura a las mismas, obteniendo:

Temperatura en grados	37	37,2	37,5	38	38,1	38,5	39
Número de personas	1	5	15	6	10	5	0

Se pide:

a) Media geométrica.
b) Hallar la mediana.
c) Coeficiente de variación media (tomando como parámetro la media).
d) Coeficiente de asimetría de Pearson.

Solución:

a) Cálculo de la media geométrica

x_i		n_i	N_i	$\log x_i$	$n_i . \log x_i$
37		1	1	1,568	1,568
37,2		5	6	1,571	7,853
37,5	M_e	15	**21**	1,574	23,610
38		6	27	1,580	9,479
38,1		10	37	1,581	15,809
38,5		5	42	1,585	7,927
39		0	42	1,591	0
		42			66,247

Media geométrica: $\bar{x}_G = \sqrt[N]{x_1^{n_1}\, x_2^{n_2}\, x_3^{n_3} \cdots x_k^{n_k}}$, para el cálculo se procede tomando logaritmos, con lo cual: $\log \bar{x}_G = \frac{1}{N}\log\left[x_1^{n_1}\, x_2^{n_2}\, x_3^{n_3} \cdots x_k^{n_k}\right] = \frac{1}{N}\left[\log(x_1^{n_1}) + \log(x_2^{n_2}) + \cdots + \log(x_k^{n_k})\right] = \frac{1}{N}\sum_{i=1}^{k} n_i \log x_i$

$\bar{x}_G = 10^{\left[\frac{1}{N}\sum_{i=1}^{k} n_i \log x_i\right]}$

en consecuencia, $\log \bar{x}_G = \frac{1}{N}\sum_{i=1}^{7} n_i \log x_i = \frac{1}{42}(66{,}247) = 1{,}577 \Rightarrow \bar{x}_G = 10^{1{,}577} = 37{,}757$ grados

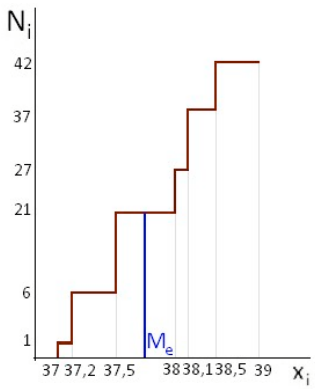
b) Cálculo de la mediana

$N/2 = 42/2 = 21$

En el diagrama de frecuencias acumuladas, observamos que 21 se encuentra en la columna de la frecuencia absoluta acumulada N_i.

Por tanto, la mediana M_e será el punto medio entre 37,5 y 38.

Es decir, $M_e = 37{,}75$



c) Coeficiente de variación media (tomando como parámetro la media)

Lo primero que hemos de calcular es la media aritmética: $\bar{x} = \frac{\sum_{i=1}^{7} x_i . n_i}{N} = \frac{1587}{42} = 37{,}78$

x_i	n_i	$x_i . n_i$	$\lvert x_i - \bar{x} \rvert . n_i$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 . n_i$
37	1	37	0,78	0,608	0,60
37,2	5	186	2,9	0,336	1,68
37,5	15	562,5	4,2	0,078	1,18
38	6	228	1,32	0,048	0,29
38,1	10	381	3,2	0,102	1,02
38,5	5	192,5	3,6	0,518	2,59
39	0	0	0	1,488	0
	42	1587	16		7,36

La desviación media D_M respecto a la media aritmética viene dada por la expresión:

$$D_M(\bar{x}) = \frac{\sum_{i=1}^{7} \lvert x_i - \bar{x} \rvert n_i}{N} = \frac{16}{42} = 0{,}38$$

Luego el coeficiente de variación media respecto a la media, $CV_{D_M}(\bar{x}) = \frac{D_{M_{\bar{x}}}}{\lvert \bar{x} \rvert} = \frac{0{,}38}{37{,}78} = 0{,}01$

d) Coeficiente de asimetría de Pearson

$$A_P = \frac{\bar{x} - M_d}{\sigma} \begin{cases} A_P > 0 & \text{Asimetría a la derecha o positiva} \\ A_P = 0 & \text{Simetría} \\ A_P < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

Este coeficiente tiene sentido cuando la moda es única

La moda $M_d = 37{,}5$ y la desviación típica $\sigma = \sqrt{\frac{\sum_{i=1}^{7} (x_i - \bar{x})^2 . n_i}{N}} = \sqrt{\frac{7{,}36}{42}} = 0{,}42$

por tanto, $A_P = \frac{\bar{x} - M_d}{\sigma} = \frac{37{,}78 - 37{,}5}{0{,}42} = 0{,}67$ por lo que la distribución es asimétrica a la derecha.

9. Para dos empresas, A y B, del sector de hostelería, las distribuciones de los salarios mensuales entre sus empleados, en cientos de euros, son las siguientes:

Empresa A	
Salarios	Número de empleados
6,5 - 7,5	10
7,5 - 8,5	15
8,5 - 9,5	40
9,5 - 10,5	25
10,5 - 11,5	10

Empresa B	
Salarios	Número de empleados
8,5 - 9,5	10
9,5 - 10,5	15
10,5 - 11,5	40
11,5 - 12,5	25
12,5 - 13,5	10

Se pide:

a) Para qué empresa es mayor el salario medio mensual.
b) Para cuál de ellas resulta más representativo el salario medio.
c) ¿Cuál sería el salario que define el 25% más alto de la banda salarial, en ambas empresas?.

Solución:

a) Para qué empresa es mayor el salario medio mensual.

Empresa A					
Salarios	x_i	n_i	N_i	$x_i.n_i$	$x_i^2.n_i$
6,5 - 7,5	7	10	10	70	490
7,5 - 8,5	8	15	25	120	960
8,5 - 9,5	9	40	65	360	3240
9,5 - 10,5	10	25	90	250	2500
10,5 - 11,5	11	10	100	110	1210
		100		910	8400

$$a_1 = \overline{x} = \frac{\sum_{i=1}^{5} x_i.n_i}{N} = \frac{910}{100} = 9{,}1$$

$$a_2 = \frac{\sum_{i=1}^{5} x_i^2.n_i}{N} = \frac{8400}{100} = 84$$

$$\sigma_x^2 = a_2 - a_1^2 = 84 - 9{,}1^2 = 1{,}19 \quad \mapsto \quad \sigma_x = 1{,}091$$

Empresa B					
Salarios	y_j	n_j	N_j	$y_j.n_j$	$y_j^2.n_j$
8,5 - 9,5	9	10	10	90	810
9,5 - 10,5	10	15	25	150	1500
10,5 - 11,5	11	40	65	440	4840
11,5 - 12,5	12	25	90	300	3600
12,5 - 13,5	13	10	100	130	1690
		100		1110	12440

$$a_1 = \overline{y} = \frac{\sum_{j=1}^{5} y_j.n_j}{N} = \frac{1110}{100} = 11{,}1$$

$$a_2 = \frac{\sum_{i=1}^{5} y_j^2.n_j}{N} = \frac{12440}{100} = 124{,}40$$

$$\sigma_y^2 = a_2 - a_1^2 = 124{,}4 - 11{,}1^2 = 1{,}19 \quad \mapsto \quad \sigma_y = 1{,}091$$

La empresa B presenta un salario medio mensual mayor.

b) Para cuál de ellas resulta más representativo el salario medio.

En una distribución, la medida óptima que critica la representatividad de la media aritmética es la desviación típica, de forma que cuanto mayor sea ésta, menos representativa es la media aritmética.

Cuando se desea comparar la representatividad de dos medias aritméticas de dos distribuciones diferentes, no tiene sentido compararlas con una medida de dispersión absoluta (desviaciones

típicas). Se debe emplear una medida de dispersión relativa, medida sin dimensiones y que no depende de las unidades empleadas en las distribuciones que deseamos comparar.

Una medida de dispersión relativa es el coeficiente de variación de Pearson: $C.V = \frac{\sigma}{\bar{x}}$

En consecuencia,

Empresa A: $C.V_x = \frac{\sigma_x}{\bar{x}} = \frac{1{,}091}{9{,}1} = 0{,}1199$

Empresa B: $C.V_y = \frac{\sigma_y}{\bar{y}} = \frac{1{,}091}{11{,}1} = 0{,}0983$

$C.V_y < C.V_x$

La empresa B presenta una dispersión más pequeña, siendo más homogénea respecto a los salarios y su media aritmética será más representativa.

c) ¿Cuál sería el salario que define el 25% más alto de la banda salarial, en ambas empresas?.

Para conocer qué salario define el 25% más alto de la banda salarial, se calcula el percentil 75 (o el tercer cuartil). Para ello, se debe averiguar que intervalo lo contiene, que será el primero cuya frecuencia absoluta acumulada sea superior a $(75.N/100 = 75.100/100 = 75)$

Empresa A: $P_{75} = L_i + \frac{\frac{75.N}{100} - N_{i-1}}{N_i - N_{i-1}} c_i = 9{,}5 + \frac{75-65}{90-65} 1 = 9{,}9$

Empresa B: $P_{75} = L_i + \frac{\frac{75.N}{100} - N_{i-1}}{N_i - N_{i-1}} c_i = 11{,}5 + \frac{75-65}{90-65} 1 = 11{,}9$

El 25% de los salarios más altos en las dos empresas son, respectivamente, 990 euros y 1190 euros.

Observando los salarios, lógicamente se podría haber previsto que $P_{75}(B) = P_{75}(A) + 2$

10. En dos regiones diferentes se determinaron las siguientes distribuciones de la renta (expresados en 10.000 euros):

Región A	
Niveles de renta	Número de individuos
0,5 - 1,5	345
1,5 - 2,5	225
2,5 - 3,5	182
4,5 - 6,5	56
6,5 - 10	32

Región B	
Niveles de renta	Número de individuos
0,5 - 1,5	583
1,5 - 2,5	435
2,5 - 3,5	194
4,5 - 6,5	221
6,5 - 10	67

a) ¿Depende el índice de concentración de Gini de los individuos incluidos en cada nivel?
b) Determinar la concentración de la renta para el conjunto de las dos regiones. Dibujar la curva de Lorenz correspondiente.
c) ¿Qué parte de la renta percibe el 5% del personal mejor pagado en la región A?.
d) ¿Qué porcentaje de individuos percibe el 50% de la renta en la región B?.

Solución:

a) ¿Depende el índice de concentración de Gini de los individuos incluidos en cada nivel?

Región A

Renta	x_i	n_i	$x_i.n_i$	$\sum_{i=1}^{m} x_i n_i$	N_i	$p_i = N_i/N$	$q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i}$	$p_i - q_i$
0,5 - 1,5	1	345	345	345	345	0,4107	0,1722	0,2386
1,5 - 2,5	2	225	450	795	570	0,6786	0,3967	0,2819
2,5 - 3,5	3,5	182	637	1432	752	0,8952	0,7146	0,1807
4,5 - 6,5	5,5	56	308	1740	808	0,9619	0,8683	0,0936
6,5 - 10	8,25	32	264	2004	840	1	1	0
		840	2004			3,9464		0,7947

Región B

Renta	x_i	n_i	$x_i.n_i$	$\sum_{i=1}^{m} x_i n_i$	N_i	$p_i = N_i/N$	$q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i}$	$p_i - q_i$
0,5 - 1,5	1	583	583	583	583	0,3887	0,1495	0,2392
1,5 - 2,5	2	435	870	1453	1018	0,6787	0,3725	0,3061
2,5 - 3,5	3,5	194	679	2132	1212	0,8080	0,5466	0,2614
4,5 - 6,5	5,5	221	1215,5	3347,5	1433	0,9553	0,8583	0,0971
6,5 - 10	8,25	67	552,75	3900,25	1500	1	1	0
		1500	3900,25			3,8307		0,9037

El índice de concentración de Gini: $I_G = \frac{\sum_{i=1}^{n-1} (p_i - q_i)}{\sum_{i=1}^{n-1} p_i}$ $\quad 0 \le I_G \le 1$

Región A: $I_G(A) = \frac{\sum_{i=1}^{4} (p_i - q_i)}{\sum_{i=1}^{4} p_i} = \frac{0,7947}{2,9474} = 0,27$ $\quad$ Región B: $I_G(B) = \frac{\sum_{i=1}^{4} (p_i - q_i)}{\sum_{i=1}^{4} p_i} = \frac{0,9037}{2,8307} = 0,32$

Como $I_G(A) \neq I_G(B)$, con los mismo niveles de renta, el índice de Gini depende del número de empleados en cada nivel.

b) Determinar la concentración de la renta para el conjunto de las dos regiones. Dibujar la curva de Lorenz correspondiente.

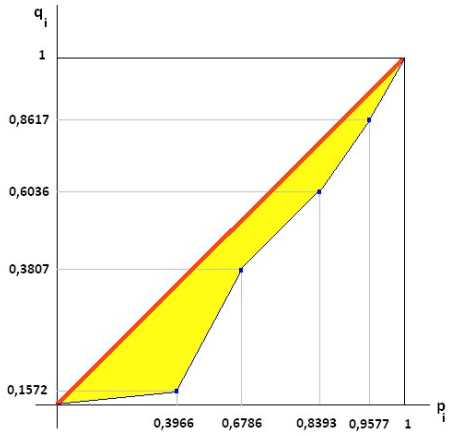
Distribución de renta para las dos regiones

Renta	x_i	n_i	$x_i.n_i$	$\sum_{i=1}^{m} x_i n_i$	N_i	$p_i = N_i/N$	$q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i}$	$p_i - q_i$
0,5 - 1,5	1	928	928	928	928	0,3966	0,1572	0,2394
1,5 - 2,5	2	660	1320	2248	1588	0,6786	0,3807	0,2979
2,5 - 3,5	3,5	376	1316	3564	1964	0,8393	0,6036	0,2357
4,5 - 6,5	5,5	277	1523,5	5087,5	2241	0,9577	0,8617	0,0960
6,5 - 10	8,25	99	816,75	5904,25	2340	1	1	0
		2340	5904,25			3,8722		0,8690

El índice de concentración de Gini es: $I_G = \frac{\sum_{i=1}^{4}(p_i - q_i)}{\sum_{i=1}^{4} p_i} = \frac{0,8690}{2,8722} = 0,30$

Se puede verificar que $I_G(A) < I_G < I_G(B)$, y que $I_G \neq I_G(A) + I_G(B)$

La curva de concentración de Lorenz plasma su coherencia con el índice de Gini calculado, puesto que cuanto más próxima esté la curva a la diagonal principal menor será la concentración, y en consecuencia, mejor será la distribución de la renta.



c) ¿Qué parte de la renta percibe el 5% del personal mejor pagado en la región A?

Región A

Renta	$\%p_i = (N_i/N).100$	$\%q_i = \frac{\sum_{i=1}^{m} x_i n_i}{\sum_{i=1}^{k} x_i n_i}.100$
0,5 - 1,5	41,07	17,22
1,5 - 2,5	67, 86	39,67
2,5 - 3,5	89, 52	71,46
	95	84,08
4,5 - 6,5	96,19	86,83
6,5 - 10	100	100

El 5% de individuos mejor pagados será la que va del tramo del 95% al 100% (columna p_i).

Para determinar el porcentaje de renta que le correspondería al 95% de los individuos mejor pagados, bajo la hipótesis de linealidad, se puede establecer la relación siguiente en porcentajes:

$$\frac{96,19-89,52}{86,83-71,46}=\frac{95-89,52}{x} \quad \mapsto \quad x=\frac{84,23}{6,67}=12,62$$

en consecuencia, el 95% de los individuos percibiría una renta de $71,46\%+12,62\%=84,08\%$

Al 5% de los individuos mejor pagados le corresponde un porcentaje de la nómina (columna q_i):

$$100\%-84,08\%=15,92\% \text{ de la renta}$$

d) ¿Qué porcentaje de individuos percibe el 50% de la renta en la región B?

Región B

Renta	$\%p_i=(N_i/N).100$	$\%q_i=\frac{\sum_{i=1}^{m}x_i n_i}{\sum_{i=1}^{k}x_i n_i}.100$
0,5 - 1,5	39,66	15,72
1,5 - 2,5	67,86	38,07
	x	50
2,5 - 3,5	83,93	60,36
4,5 - 6,5	95,77	86,17
6,5 - 10	100	100

En la tabla se observa que al 67,86 % de los individuos le corresponde el 38,07 % de la renta, y al 83,93 % de individuos el 60,36 % de la renta.

En consecuencia, el 50 % de la renta estará distribuida entre un conjunto de individuos situado entre el 67,86 % y el 83,93 %.

Bajo la hipótesis de linealidad, se establece la relación en porcentajes:

$$\frac{83,93-67,86}{60,36-38,07}=\frac{x-67,86}{50-38,07} \quad \mapsto \quad x=67,86+\frac{191,72}{22,29}=76,46$$

Por tanto, el 50 % de la renta se reparte entre el 76,46 % de los individuos.

11. En la tabla adjuntan aparecen las notas obtenidas por diez alumnos:

Matemáticas	2	10	8	6	4	3	8	6	2	1
Estadística	3	8	7	5	4	4	6	6	1	1

a) Si otro alumno obtiene un 5 en cada asignatura, en relación con la clase. ¿en cuál de ellas ha sacado mejor nota?.
b) Hallar la recta de regresión que explica la nota de estadística en función de la nota de matemáticas?.
c) Cuando un alumno consiga un 6 en matemáticas, ¿qué nota es previsible que obtenga en estadística?. ¿Con qué grado de fiabilidad?.

Solución:

a) Para conocer la posición relativa de la calificación de un nuevo estudiante respecto a las calificaciones de los diez estudiantes, es necesario conocer el número de unidades de desviación típica que se ha separado de la media en cada una de las dos asignaturas. Para ello, se utiliza la variable tipificada.

En este sentido, llamando X ="calificación de matemáticas" e Y = "calificación en estadística", si el nuevo alumno obtiene (x=5, y=5), para comparar en qué asignatura ha obtenido mejor calificación tendríamos que tipificar: $\frac{x-\overline{x}}{\sigma_x}$ e $\frac{y-\overline{y}}{\sigma_y}$, el valor más alto obtenido será el de mejor calificación.

Es necesario calcular la media aritmética y la varianza de cada una de las dos distribuciones: matemáticas (X) y estadística (Y).

$$a_1 = \overline{x} = \frac{\sum_{i=1}^{10} x_i}{N} = \frac{2+10+8+6+4+3+8+6+2+1}{10} = 5$$

$$a_2 = \frac{\sum_{i=1}^{10} x_i^2}{N} = \frac{4+100+64+36+16+9+64+36+4+1}{10} = 33{,}4$$

$$\sigma_x^2 = a_2 - a_1^2 = 33{,}4 - 5^2 = 8{,}4 \qquad \mapsto \qquad \sigma_x = \sqrt{8{,}4} = 2{,}898$$

$$a_1 = \overline{y} = \frac{\sum_{i=1}^{10} y_i}{N} = \frac{3+8+7+5+4+4+6+6+1+1}{10} = 4{,}5$$

$$a_2 = \frac{\sum_{i=1}^{10} y_i^2}{N} = \frac{9+64+49+25+16+16+36+36+1+1}{10} = 25{,}3$$

$$\sigma_y^2 = a_2 - a_1^2 = 25{,}3 - 4{,}5^2 = 5{,}05 \qquad \mapsto \qquad \sigma_y = \sqrt{5{,}05} = 2{,}247$$

El nuevo alumno respecto al grupo:

Matemáticas: $\frac{x-\overline{x}}{\sigma_x} = \frac{5-5}{2{,}898} = 0$ Estadística: $\frac{y-\overline{y}}{\sigma_y} = \frac{5-4{,}5}{2{,}247} = 0{,}22$

Ha obtenido mejor calificación en Estadística que supera la media, en unidades de desviación típica.

b) La recta de regresión de Y sobre X, aplicando el método de los mínimos cuadrados, viene dada por la expresión: $y - \overline{y} = \frac{m_{11}}{\sigma_x^2}(x - \overline{x})$, donde la covarianza $m_{11} = a_{11} - \overline{x}\,\overline{y}$

El momento $a_{11} = \frac{\sum_{i=1}^{10} x_i . y_i}{N} = \frac{2.3+10.8+8.7+6.5+4.4+3.4+8.6+6.6+2.1+1.1}{10} = 28{,}7$

con lo que, la covarianza $m_{11} = a_{11} - \bar{x}.\bar{y} = 28{,}7 - 5.4{,}5 = 6{,}2$

La recta de regresión de Y sobre X: $y - 4{,}5 = \frac{6{,}2}{8{,}4}(x-5) \quad \Rightarrow \quad y = 0{,}81 + 0{,}738\,x$

c) El objetivo más importante de la regresión es la perdición del comportamiento de una variable para un valor determinado de la otra. La fiabilidad de la predicción, en principio, será tanto mejor cuanto mayor sea la correlación entre las variables. En consecuencia, una medida de la bondad de la predicción podría venir dada por el coeficiente de determinación ρ^2 $(0 \le \rho^2 \le 1)$, o por el coeficiente de correlación ρ $(-1 \le \rho \le 1)$.

Para obtener la calificación en estadística (Y) de un alumno que ha obtenido un 6 en matemáticas (X) recurrimos a la recta de regresión $y = 0{,}81 + 0{,}738\,x$, sustituyendo $x = 6$:

$$y = 0{,}81 + 0{,}738.6 = 5{,}238 \text{ calificación en estadística}$$

La fiabilidad viene dada por el coeficiente de determinación:

$$\rho^2 = \frac{m_{11}^2}{\sigma_x^2.\sigma_y^2} = \frac{6{,}2^2}{8{,}4.5{,}05} = 0{,}91$$

La recta de regresión explica el 91% las notas de estadística en función de las matemáticas.

12a. Un curso se encuentra dividido en tres grupos con los siguientes datos:

Grupo	Número alumnos	Nota media	Varianza
1	30	7	1,2
2	40	6	1,6
3	50	5,5	0,8

Se pide:

a) Nota media para todo el curso.
b) Coeficientes de variación de cada grupo.
c) ¿Cuál es el grupo más homogéneo?
d) Varianza de todas las notas del curso.

Solución:

a)

				Entre los grupos		Dentro de los grupos	
Grupos	N_i	$\bar{x}_i$	$N_i.\bar{x}_i$	$\bar{x}_i-\bar{x}$	$N_i.(\bar{x}_i-\bar{x})^2$	σ_i^2	$N_i.\sigma_i^2$
1	$N_1=30$	$\bar{x}_1=7$	210	0,958	27,533	$\sigma_1^2=1,2$	36
2	$N_2=40$	$\bar{x}_2=6$	240	- 0,042	0,071	$\sigma_2^2=1,6$	64
3	$N_3=50$	$\bar{x}_3=5,5$	275	- 0,542	14,688	$\sigma_3^2=0,8$	40
	N = 120		$\sum_{i=1}^{3} N_i.\bar{x}_i=725$		$\sum_{i=1}^{3} N_i.(\bar{x}_i-\bar{x})^2=42,29$		$\sum_{i=1}^{3} N_i.\sigma_i^2=140$

La media aritmética $\bar{x}$ para todo el curso será: $\bar{x}=\dfrac{\sum_{i=1}^{3} N_i.\bar{x}_i}{N}=\dfrac{725}{120}=6,042$

b) Los coeficientes de variación de Pearson para cada grupo son:

Grupo 1: $C.V_1=\dfrac{\sigma_1}{\bar{x}_1}=\dfrac{\sqrt{1,2}}{7}=0,156$

Grupo 2: $C.V_2=\dfrac{\sigma_2}{\bar{x}_2}=\dfrac{\sqrt{1,6}}{6}=0,21$

Grupo 3: $C.V_3=\dfrac{\sigma_3}{\bar{x}_3}=\dfrac{\sqrt{0,8}}{5,5}=0,163$

c) El grupo más homogéneo será aquel que tenga un coeficiente de variación de Pearson menor. En este sentido, el Grupo 1 es el más homogéneo.

d) La varianza para todo el curso será: $\sigma^2=\underbrace{\frac{1}{N}.\sum_{i=1}^{3} N_i.\sigma_i^2}_{\text{dentro de los grupos}}+\underbrace{\frac{1}{N}.\sum_{i=1}^{3} N_i.(\bar{x}_i-\bar{x})^2}_{\text{entre grupos}}$

varianza del curso, $\sigma^2=\dfrac{1}{N}.\sum_{i=1}^{3} N_i.\sigma_i^2+\dfrac{1}{N}.\sum_{i=1}^{3} N_i.(\bar{x}_i-\bar{x})^2=\dfrac{140}{120}+\dfrac{42,29}{120}=1,522$

13a. Los salarios mensuales de los trabajadores de Muebles Quintana, según sus categorías son:

Categoría	Número trabajadores	Salario medio (euros)	Moda (euros)	Desviación típica (euros)
A	8	2100	1750	490
B	12	630	560	140

Se pide:

a) Salario medio y desviación típica de todos los trabajadores de la empresa.
b) Si un trabajador de categoría A gana 1610 euros y otro trabajador de categoría B gana 560 euros. ¿cuál de los dos trabajadores tuvo mejor posición en su grupo?
c) ¿Cuál es el salario más frecuente de todos los trabajadores de la empresa?
d) Si en otra empresa similar el salario medio de sus trabajadores es de 1050 euros, con una desviación típica de 525 euros. ¿qué empresa tiene una distribución de salarios más homogénea?

Solución:

a) Salario medio y desviación típica de todos los trabajadores de la empresa.

				Entre los grupos		Dentro de los grupos	
	N_i	$\bar{x}_i$	$N_i.\bar{x}_i$	$\bar{x}_i-\bar{x}$	$N_i.(\bar{x}_i-\bar{x})^2$	σ_i^2	$N_i.\sigma_i^2$
A	$N_1=8$	$\bar{x}_1=2100$	16800	882	6223392	$\sigma_1^2=490^2$	1920800
B	$N_2=12$	$\bar{x}_2=630$	7560	-588	4148928	$\sigma_2^2=140^2$	235200
	N = 20		$\sum_{i=1}^{2}N_i.\bar{x}_i=$ $=24360$		$\sum_{i=1}^{2}N_i.(\bar{x}_i-\bar{x})^2=$ $=10372320$		$\sum_{i=1}^{2}N_i.\sigma_i^2=$ $=2156000$

La empresa se encuentra estructurada en dos categorías o estratos. El salario medio de la empresa será:

$$\bar{x}=\frac{\sum_{i=1}^{2}N_i.\bar{x}_i}{N}=\frac{24360}{20}=1218 \text{ euros/mes}$$

La varianza de la empresa (varianza total) se descompone en la varianza dentro de cada categoría y varianza entre categorías, es decir: $\sigma^2=\underbrace{\frac{1}{N}.\sum_{i=1}^{2}N_i.\sigma_i^2}_{\text{dentro de los estratos}}+\underbrace{\frac{1}{N}.\sum_{i=1}^{2}N_i.(\bar{x}_i-\bar{x})^2}_{\text{entre estratos}}$ 626416

Varianza de la empresa: $\sigma^2=\frac{1}{N}.\sum_{i=1}^{2}N_i.\sigma_i^2+\frac{1}{N}.\sum_{i=1}^{2}N_i.(\bar{x}_i-\bar{x})^2=\frac{2156000}{20}+\frac{10372320}{20}=626416 \text{ euros}^2$

Desviación típica de la empresa: $\sigma=\sqrt{626416}=791{,}46$ euros

b) Si un trabajador de categoría A gana 1610 euros y otro trabajador de categoría B gana 560 euros. ¿cuál de los dos trabajadores tuvo mejor posición en su grupo?

Al tratarse de dos observaciones procedentes de dos distribuciones con características diferentes, es necesario pasar a una única distribución en la que sea posible comparar las dos observaciones.
En este sentido, se pasa a la tipificación de las mismas, es decir:

$$z_A=\frac{1610-2100}{490}=-1 \qquad z_B=\frac{560-630}{140}=-0{,}5$$

siendo $z_B>z_A$, se deduce que el empleado de la empleado de categoría B obtuvo una mejor posición en su grupo que el empleado de categoría A.

Señalar que, por ser los dos valores tipificados negativos, refleja que su salario está por debajo del salario medio en ambos casos.

c) ¿Cuál es el salario más frecuente de todos los trabajadores de la empresa?

El salario más frecuente de todos los trabajadores será el que tenga mayor frecuencia absoluta.

La moda para los trabajadores de la categoría A es 1750, mientras que para los trabajadores de la categoría B es 560. No obstante, desconocemos cuáles son las frecuencias absolutas de los diferentes

salarios de los trabajadores, no pudiendo por tanto determinar con exactitud la moda para toda la empresa.

d) Si en otra empresa similar el salario medio de sus trabajadores es de 1050 euros, con una desviación típica de 525 euros. ¿qué empresa tiene una distribución de salarios más homogénea?

Como se pretende comparar la homogeneidad entre dos distribuciones, se debe calcular el coeficiente de variación de Pearson para ambas distribuciones:

Muebles Quintana: $C.V = \frac{\sigma}{\bar{x}} = \frac{791,46}{1218} = 0,65$

Otra Empresa: $C.V = \frac{\sigma}{\bar{x}} = \frac{525}{1050} = 0,50$

La otra empresa, al tener un coeficiente de variación de Pearson menor, presenta una distribución de salarios más homogénea.

CAMBIO DE ORIGEN Y DE ESCALA DE UNA VARIABLE ESTADÍSTICA

Sea una variable estadística X con media $\bar{x} = \frac{\sum_{i=1}^{k} x_i . n_i}{N}$ y varianza $\sigma_x^2 = \frac{\sum_{i=1}^{k} (x_i - \bar{x})^2 . n_i}{N}$

Si efectuamos un cambio de origen y de escala sobre la variable X, esto es, construimos otra variable $Y = aX + b$, siendo $a > 0$ y b constantes (multiplicar X por una constante "a" es efectuar un cambio de escala y sumar una constante "b" es realizar un cambio de origen).

En definitiva, para cada dato x_i hay un $y_i = ax_i + b$, con la misma frecuencia absoluta n_i. De tal modo, tenemos las tablas de frecuencias:

x_i	n_i
x_1	n_1
x_2	n_2
⋮	⋮
x_k	n_k
	N

y_i	n_i
y_1	n_1
y_2	n_2
⋮	⋮
y_k	n_k
	N

Entonces, la media aritmética, la varianza, y el coeficiente de variación de Pearson de la nueva variable serán:

📖 $\bar{y} = \frac{\sum_{i=1}^{k} y_i . n_i}{N} = \frac{1}{N} \sum_{i=1}^{k} (ax_i + b) . n_i = a \frac{1}{N} \sum_{i=1}^{k} x_i . n_i + b \frac{1}{N} \sum_{i=1}^{k} n_i = a\bar{x} + b$

- La media se ve afectada por el mismo cambio de origen y de escala efectuada sobre la variable.

📖 $\sigma_y^2 = \frac{\sum_{i=1}^{k}(y_i - \bar{y})^2.n_i}{N} = \frac{\sum_{i=1}^{k}(ax_i + b - a\bar{x} - b)^2.n_i}{N} = a^2 \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2.n_i}{N} = a^2 \sigma_x^2$

- La varianza no se ve afectada por el cambio de origen pero si por el cambio de escala efectuado sobre la variable.

📖 El coeficiente de variación de Pearson $C.V_y = \frac{\sigma_y}{\bar{y}} = \frac{a\sigma_x}{a\bar{x} + b}$.

- Si se efectúa un cambio de escala $(b = 0)$, se tiene: $C.V_y = \frac{\sigma_y}{\bar{y}} = \frac{a\sigma_x}{a\bar{x}} = C.V_x$

  El cambio de escala no afecta al coeficiente de variación.

- Si solo se efectúa un cambio de origen $(a = 1)$, queda: $C.V_y = \frac{\sigma_y}{\bar{y}} = \frac{\sigma_x}{\bar{x} + b} \neq C.V_x$

  El cambio de origen si afecta al coeficiente de variación.

1. En la tabla adjunta se reflejan dos operaciones que una empresa ha realizado con una compañía con sede en Gran Bretaña:

Importe (miles de euros)	Cambio de la libra
270	259
187	262,1

¿Qué promedio debe utilizar para conocer el cambio medio de dichas operaciones?. Razone la respuesta.

Solución:

Cambio de la libra: x_i	Importe (euros): n_i	n_i/x_i
259	270.000	270.000/259 = 1042,471
262,1	187.000	187.000/262,1 = 713,468
	N = 457.000	$\sum_{i=1}^{2} n_i/x_i = 1755,939$

270 miles euros son $\frac{270.000}{259} = 1042,471$ libras

187 miles euros son $\frac{187.000}{259} = 713,468$ libras

$$\overline{x}_A = \frac{N}{\sum_{i=1}^{2} \frac{n_i}{x_i}} = \frac{457.000}{1755,939} = 260,26 \text{ euros/libra}$$

2. Una empresa quiere saber qué porcentaje de trabajadores recibe el 50% de la masa salarial y para ello utiliza la mediana de la distribución de rentas. ¿Es correcto?. Razone la respuesta.

Solución: No sería correcto porque la mediana de la distribución de rentas es aquella cantidad tal que el 50% del número de individuos percibe una renta menor o igual que ella.

3. El coeficiente de variación de Pearson:

a) Permite comparar distribuciones, únicamente si tienen el mismo número de elementos.
b) No varia al efectuar un cambio de origen
c) Carece de unidades de medida
d) Ninguna de las respuestas es correcta

Solución: Carece de unidades de medida.

4. En una distribución simétrica se verifica que:
a) La media coincide con la moda en todos los casos
b) El rango depende del número de observaciones
c) La mediana coincide con la moda en todos los casos
d) Ninguna de las respuestas es correcta

Solución: Ninguna de las respuestas es correcta.

5. Cuál de las siguientes afirmaciones es verdadera:
a) La media es un estadístico que no utiliza toda la información muestral
b) La mediana no se ve afectada por los valores extremos
c) La media no se ve afectada por los valores extremos
d) Ninguna de las respuestas es correcta.

Solución: La mediana no se ve afectada por los valores extremos.

6. Si el coeficiente de curtosis de Fisher es mayor que cero:
a) La distribución es platicúrtica
b) La distribución es mesocúrtica
c) La distribución es leptocúrtica
d) Ninguna de las respuestas es correcta

Solución: La distribución es leptocúrtica.

7. Cuál de las siguientes variables es de tipo discreto:
a) Tiempo de espera del ave
b) Distancia entre las capitales de provincia
c) El número de viviendas existentes en Madrid
d) Ninguna de las respuestas es correcta

Solución: El número de viviendas existentes en Madrid.

8. En una distribución $\bar{x}=4$ y la $\sigma_x^2=16$. Definimos una nueva distribución $y=2x+1$. denotando por C.V. el coeficiente de variación de Pearson. ¿Cuál de las siguientes respuestas es correcta?:

a) $C.V_x=C.V_y$ b) $C.V_x>C.V_y$ c) $2C.V_x=C.V_y$ d) Ninguna de las respuestas es correcta.

Solución: $C.V_y=\frac{\sigma_y}{\bar{y}}=\frac{2\sigma_x}{2\bar{x}+1}=\frac{\sigma_x}{\bar{x}+\frac{1}{2}}<\frac{\sigma_x}{\bar{x}} \Rightarrow C.V_x>C.V_y$

9. Si el coeficiente de asimetría de Fisher es menor que cero, la distribución es:
   a) Asimétrica negativa o a la izquierda
   b) Asimétrica positiva o a la derecha
   c) Simétrica
   d) Ninguna de las respuestas es correcta

Solución: Asimétrica negativa o a la izquierda

El coeficiente de simetría de Fisher se define como

$$A_F = \frac{m_3}{\sigma^3} \begin{cases} A_F > 0 & \text{Asimetría a la derecha o positiva} \\ A_F = 0 & \text{Simetría} \\ A_F < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

10. Si el coeficiente de asimetría de Pearson es mayor que cero, la distribución es:
   a) Asimétrica negativa o a la izquierda
   b) Asimétrica positiva o a la derecha
   c) Simétrica
   d) Ninguna de las respuestas es correcta

Solución: Asimétrica positiva o a la derecha

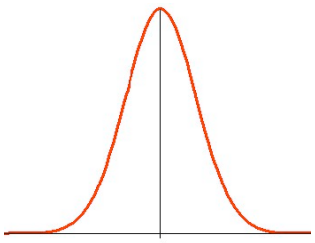
El coeficiente de simetría de Pearson se define como

$$A_P = \frac{\bar{x} - M_d}{\sigma} \begin{cases} A_P > 0 & \text{Asimetría a la derecha o positiva} \\ A_P = 0 & \text{Simetría} \\ A_P < 0 & \text{Asimetría a la izquierda o negativa} \end{cases}$$

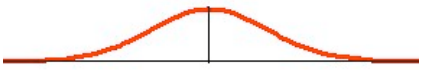
Tiene sentido obtener este coeficiente cuando la moda es única.

11. Concepto de curtosis.

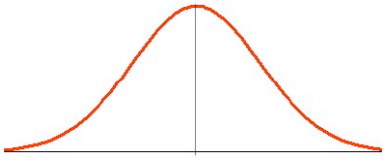
Solución: La curtosis de una distribución de frecuencias es el apuntamiento que presenta el polígono de frecuencias alrededor de la media. Si está muy apuntado diremos que la distribución es leptocúrtica, si poco apuntado platicúrtica, y si el apuntamiento es intermedio mesocúrtica.



Leptocúrtica



Platicúrtica



Mesocúrtica

El coeficiente $g_2 = \frac{m_4}{\sigma^4}$ donde $m_2 = \sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{N}$ y $m_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^4 n_i}{N}$

indica el apuntamiento de forma de la distribución comparándola con la distribución normal (Campana de Gauss), donde se tiene que:

- $g_2 > 3$ Más apuntamiento que la normal: Leptocúrtica
- $g_2 = 3$ Igual apuntamiento que la normal: Mesocúrtica
- $g_2 < 3$ Menor apuntamiento que la normal: Platicúrtica

12. Señale las ventajas e inconvenientes de la media aritmética como medida de posición de una distribución.

Solución: $a_1 = \bar{x} = \dfrac{\sum_{i=1}^{k} x_i n_i}{N}$

- Se puede calcular en todas las variables, es decir siempre que las observaciones sean cuantitativas.
- Para su cálculo se utilizan todos los valores de la distribución.
- Es única para cada distribución de frecuencias.
- Tiene un claro significado, ya que al ser el centro de gravedad de la distribución representa todos los valores observados.
- El principal inconveniente es que es un valor muy sensible a los valores extremos, con lo que en las distribuciones con gran dispersión de datos puede llegar a perder totalmente su significado.

13. Señale las ventajas e inconvenientes de la media geométrica como medida de posición de una distribución.

Solución: $\bar{x}_G = \sqrt[N]{x_1^{n_1} x_2^{n_2} x_3^{n_3} \cdots x_k^{n_k}}$ donde $\bar{x}_G = \text{antilog} \dfrac{\sum_{i=1}^{k} n_i \log x_i}{N}$

- Es más representativa que la media aritmética cuando la variable evoluciona de forma acumulativa con efectos multiplicativos.
- Cuando existe, es decir cuando la variable no tiene valores negativos, y cuando está definida, es decir cuando la distribución no tiene valores nulos, su valor está definida de forma objetiva y es único.
- Para su cálculo se tiene en cuenta todos los valores de la distribución.
- Los valores extremos tienen menor influencia que en la media aritmética.

Los principales inconvenientes:

- Mayor complicación en los cálculos.
- Su indefinición (da números con naturaleza imaginaria) cuando tiene valores negativos y cuando una observación toma el valor nulo.

14. Señale las ventajas e inconvenientes de la media armónica como medida de posición de una distribución. ¿En que casos es conveniente utilizarla?

Solución: Es una medida estadística que se utiliza cuando que se desean promediar rendimientos, velocidades, productividades, etc. Sólo se puede calcular cuando no hay observaciones iguales a cero.

Las principales ventajas son:

- Es más representativa que otras medidas en los casos de obtener promedios de velocidades, rendimientos, productividades, etc.
- Está definida de forma objetiva y es única.
- Su cálculo es sencillo y se tienen en cuenta todos los valores de la distribución.
- Los valores extremos tienen menor influencia que en la media aritmética.

El principal inconveniente se produce cuando se utilizan variables con valores muy pequeños; en estos casos, sus inversos pueden aumentar casi hasta el infinito, eliminando el efecto del resto de los valores. Por esta misma razón, no es posible calcularla cuando algún valor es cero, ya que se produce una indeterminación matemática.

15. Indique razonadamente cómo se comporta la media aritmética ante un cambio de escala y un cambio de origen en una variable.

Solución:

Supongamos que sobre una variable X_i efectuamos un cambio de origen y de escala: $Y_i = aX_i + b$ (multiplicar por 'a' es un cambio de escala y sumar 'b' es un cambio de origen). La media aritmética de Y_i sería:

$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{k} Y_i n_i = \frac{1}{N}\sum_{i=1}^{k} (aX_i + b)\, n_i = \frac{a}{N}\sum_{i=1}^{k} X_i\, n_i + \frac{b}{N}\sum_{i=1}^{k} n_i = a\overline{X} + b$$

Es decir, la media aritmética queda afectada por el mismo cambio de origen y de escala.

16. Defina los conceptos estadísticos de población, marco estadístico, muestra e individuo o unidad estadística.

Solución:

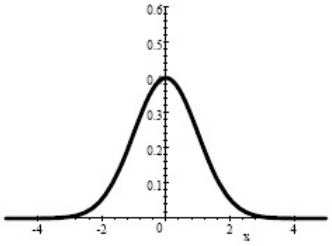
- Población.- Es el conjunto de elementos que cumplen una determinada característica.
- Marco estadístico.- Es el conjunto de información (ficheros, listados, etc.) que permite identificar a todos los individuos de la población. Es la base informativa que empleamos para seleccionar la muestra. En el marco estadístico no siempre está contenido todo el universo (por las omisiones, duplicaciones, unidades mal clasificadas, etc.)
- Muestra.- Cualquier subconjunto de individuos pertenecientes a una población determinada.
- Individuo o Unidad de investigación.- Cada uno de los elementos de la población.

17. Defina el concepto y significado de las medidas de curtosis de una distribución estadística.

Solución:

Las medidas de curtosis o apuntamiento tratan de estudiar la distribución de frecuencias en la zona media. El mayor o menor número de valores de la variable alrededor de la media dará lugar a una distribución más o menos apuntada.

Para estudiar el apuntamiento comparamos el perfil de la distribución (polígono de frecuencias o histograma) con la denominada *Campana de Gauss* de ecuación: $y = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$



Para ello, se utiliza el coeficiente de curtosis de Fisher: $g_2 = \frac{m_4}{\sigma^4} - 3$

Según el valor de esta expresión, tendremos una distribución mesocúrtica cuando $g_2 = 0$, leptocúrtica cuando $g_2 > 0$, o platicúrtica cuando $g_2 < 0$

18. ¿En qué casos es preferible, por ser más representativa, utilizar la media geométrica en lugar de la media aritmética?

Solución:

Cuando los valores a promediar tengan entre sí una relación multiplicativa en lugar de aditiva. Por ejemplo, en las tasas de crecimiento.

- Ejemplo: Las tasas de crecimiento de una determinada magnitud a lo largo de cuatro períodos de tiempo han sido, respectivamente, 1,2 ; 1,5 ; 1,1 ; 1,3 <<esto quiere decir que la magnitud ha aumentado sucesiva y respectivamente el 20%, el 50%, el 10%, y el 30%>>.

  La tasa media habrá sido la media geométrica: $\bar{x}_G = \sqrt[4]{(1,2)(1,5)(1,1)(1,3)} \cong 1,27$

19. Si a una variable X_i la sometemos al mismo tiempo a un cambio de origen 0 y a un cambio de escala C, ¿cuál o cuáles de las afirmaciones son falsas o correctas?

a) Los cambios de origen afectan a la media aritmética
b) Los cambios de escala afectan a la media aritmética
c) La varianza y la desviación típica sólo se ven afectados por los cambios de escala

Solución: Las tres afirmaciones son correctas.

$$\text{Sea } Y_i = \frac{X_i - 0}{C} \text{ entonces, } \bar{Y} = \frac{1}{N}\sum\left(\frac{X_i - 0}{C}\right)n_i = \frac{1}{C}\left[\frac{1}{N}\sum X_i n_i - \frac{0}{N}\sum n_i\right] = \frac{\bar{X} - 0}{C}$$

Es decir, la media aritmética se ve afectada por los cambios de origen y de escala.

$$Var(\bar{Y}) = \frac{1}{N}\sum(Y_i - \bar{Y})^2 n_i = \frac{1}{N}\sum\left(\frac{X_i - 0}{C} - \frac{\bar{X} - 0}{C}\right)^2 n_i = \frac{1}{N}\sum\left(\frac{X_i - \bar{X}}{C}\right)^2 n_i = \frac{1}{C^2}Var(\bar{X})$$

La varianza se ve afectada por los cambios de escala. En consecuencia, también la desviación típica.

20. Tenemos una distribución con los siguientes datos expresados en euros: 1, 8, 9 y 85. Indique a simple vista si la media aritmética es representativa para esta distribución. ¿Qué debería hacerse para valorar adecuadamente esta representatividad?. ¿Qué medidas deberían calcularse?.

Solución:

- A simple vista no parece que la media aritmética sea representativa, puesto que el valor de 85 euros se aleja mucho de los otros tres.
- Para valorar adecuadamente la representatividad hay que calcular el coeficiente de variación de Pearson: $C.V = \frac{\sigma}{\bar{x}}$:

$a_1 = \bar{x} = \frac{1+8+9+85}{4} = 25{,}75 \qquad a_2 = \frac{1+64+85+7225}{4} = 1483{,}75$

$\sigma = \sqrt{1483{,}75 - (25{,}75)^2} = 34{,}36$ con lo cual, $C.V = \frac{34{,}36}{25{,}75} = 1{,}33$

al ser el C.V. mayor que la unidad, debemos descartar la media aritmética como parámetro adecuado.

21. Defina las medidas de simetría y apuntamiento de una distribución de frecuencias.

Solución:

Coeficiente de asimetría de Fisher: $g_1 = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{\left[\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i\right]^{\frac{3}{2}}}$

Coeficiente de asimetría de Pearson: $A_P = \frac{\bar{x} - M_d}{\sigma}$

Coeficiente de asimetría de Bowley: $A_B = \frac{Q_3 + Q_1 - 2M_e}{Q_3 - Q_1}$

Coeficiente de asimetría de Excel: $A_{excel} = \frac{N}{(N-1)(N-2)}\sum_{i=1}^{N}\left[\frac{x_i - \bar{x}}{\sigma}\right]^3$

En todos los casos, si el coeficiente es positivo hay asimetría a la derecha, si es negativo hay asimetría a la izquierda, y si es cero la distribución es simétrica.

- Respecto a las medidas de apuntamiento:

Coeficiente de Fisher: $g_2 = \frac{m_4}{\sigma^4} - 3 \begin{cases} > 0 & \text{Más apuntamiento que la normal: Leptocúrtica} \\ = 0 & \text{Igual apuntamiento que la normal: Mesocúrtica} \\ < 0 & \text{Menor apuntamiento que la normal: Platicúrtica} \end{cases}$

Coeficiente de Excel: $C_{excel} = \left[\frac{N(N-1)}{(N-1)(N-2)(N-3)} \sum \frac{(x_i - \bar{x})^4}{\sigma} \right] - \left[\frac{3(N-1)^2}{(N-2)(N-3)} \right]$

$$C_{excel} \begin{cases} > 0 & \text{Más apuntamiento que la normal: Leptocúrtica} \\ = 0 & \text{Igual apuntamiento que la normal: Mesocúrtica} \\ < 0 & \text{Menor apuntamiento que la normal: Platicúrtica} \end{cases}$$

22. ¿Qué coeficiente compara la forma de una distribución cualquiera con una distribución normal?

a) El coeficiente de asimetría de Fisher
b) El coeficiente de variación de Pearson
c) El coeficiente de curtosis de Fisher
d) Ninguna de las anteriores

Solución: El coeficiente de curtosis de Fisher.

23. ¿De qué depende el coeficiente de variación de Pearson?

a) Promedio considerado
b) El signo del numerador de dicho coeficiente
c) Siempre tiene signo positivo
d) Ninguna de los anteriores

Solución: Promedio considerado.

24. Multiplicando por cuatro los valores de una serie $X_i = x_1, x_2, \cdots, x_n$ se obtiene la serie $Y_i = y_1, y_2, \cdots, y_n$. ¿Cuál de las siguientes afirmaciones es correcta?

a) Ambas series tienen la misma varianza
b) Ambas tienen el mismo coeficiente de variación
c) Ambas tienen la misma media
d) Ninguna de las anteriores

Solución: Ambas tienen el mismo coeficiente de variación.

25. Dentro de las tareas a desarrollar en la etapa de definición de objetivos en una investigación estadística podemos encontrar:

a) Recogida de datos
b) Tratamiento de los datos
c) Diseño del cuestionario
d) Ninguna de las anteriores

Solución: Ninguna de las anteriores.

26. ¿En qué ocasiones no debe utilizarse la media armónica?

a) Valores muy pequeños de la variable
b) Cuando existen valores de la variable igual a cero
c) Las respuestas (a) y (b) son correctas
d) Ninguna de las anteriores

Solución: Las respuestas (a) y (b) son correctas.

27. En una distribución unidimensional, el momento de orden uno respecto a la media $m_1 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})n_i$ es igual a:

a) 0
b) $\bar{x}$
c) Depende de los valores de x
d) Ninguna respuesta es correcta

Solución: 0

28. En una distribución de frecuencias, el segundo cuartil coincide con la mediana:

a) Si la distribución es creciente
b) Si la media aritmética es igual a la mediana
c) En todos los casos
d) Ninguna respuesta es correcta

Solución: En todos los casos.

29. En tres empresas del mismo grupo se dan las siguientes cifras de producción total y productividad media por empleado:

Empresa	A	B	C
Producción total (unidades)	200	350	400
Producción por empleado	0,5	0,7	0,8

¿Cuál de las respuestas corresponde a la productividad media?

a) $\approx 1{,}47$
b) $\approx 0{,}66$
c) $\approx 0{,}68$
d) Ninguna respuesta es correcta

Solución: $\approx 0{,}68$. $\bar{x}_A = \frac{200+350+400}{\frac{200}{0{,}5}+\frac{350}{0{,}7}+\frac{400}{0{,}8}} = 0{,}678 \cong 0{,}68$

30. En la distribución unidimensional adjunta, ¿qué medida de posición central debe utilizarse?

x_i	-3	-2	-1	1	2	3
n_i	1	5	1	1	5	1

a) Media
b) Asimetría
c) Moda
d) Mediana

Solución: Moda.

31. La media y la varianza de una serie de observaciones, respectivamente, son 0 y 4. Si doblamos el valor de cada observación, la media y la varianza serán:

a) 0 y 8
b) 0 y 4
c) 0 y 16
d) Ninguna respuesta es correcta

Solución: 0 y 16.

32. En una distribución se conoce $m_4 = 4{,}23$ (momento de orden 4 respecto a la media) y $\sigma^2 = 1{,}2$. Según estos datos, la distribución es:

a) Platicúrtica
b) Mesocúrtica
c) Leptocúrtica
d) Simétrica

Solución: Platicúrtica. $g_2 = \frac{m_4}{\sigma^4} - 3 = \frac{4{,}23}{1{,}2^2} - 3 = -0{,}0625 < 0$

33. Si el coeficiente de variación de Pearson de una variable X es igual a 2 y su media es 4. ¿Cuál es la desviación típica de la variable $Y = (X/8) - 0{,}5$?

a) 4 b) 1 c) 2 d) 0

Solución: 1

$$C.V_x = 2 = \frac{\sigma_x}{4} \Rightarrow \sigma_x = 8 \Rightarrow \sigma_x^2 = 64$$

$$Var(Y) = Var\left(\frac{X}{8} - 0{,}5\right) = Var\left(\frac{X}{8}\right) = \frac{\sigma_x^2}{64} = \frac{64}{64} = 1 \Rightarrow \sigma_y = 1$$

34. ¿Cuál es el coeficiente de asimetría de Fisher de la distribución adjunta?

x_i	1	2	3	4
n_i	3	6	4	2

a) $\approx 0{,}4$ b) $\approx 0{,}25$ c) $\approx 0{,}12$ d) Otra respuesta

Solución: 0,25 $$g_1 = \frac{m_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^3 n_i}{\left[\frac{1}{N}\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i\right]^{\frac{3}{2}}} = \frac{0{,}21}{(0{,}89)^{\frac{3}{2}}} = 0{,}25$$

35. Cuando en una población la concentración de renta es máxima:

a) El índice de Gini es igual a 1
b) La curva de Lorenz es la diagonal que va desde el punto (0,0) al (100,100)
c) Las respuestas (a) y (b) son correctas
d) Ninguna de las respuestas es correcta

Solución: El índice de Gini es igual a 1

36. La curva de Lorenz se encuentra tanto más alejada de la diagonal cuanto:

a) Menores sean las diferencias $(p_i - q_i)$
b) Mayores sean las diferencias $(p_i - q_i)$
c) Más próximos estén los valores de p_i y q_i $(p_i = q_i)$
d) Ninguna de las anteriores

Solución: Mayores sean las diferencias $(p_i - q_i)$