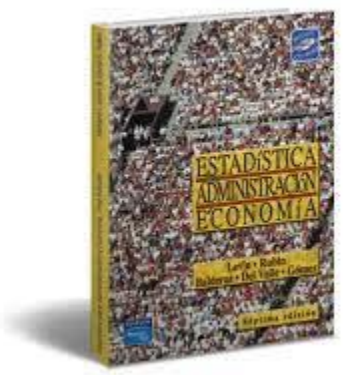


L – Cap 2 y 3

Tema 2: RESÚMENES DE DATOS Y DESCRIPCIONES ESTADÍSTICAS

Finalidad:

¿Para qué la Estadísitca?

Para presentar un conjunto de datos analizándolos y resumiendo sus características principales. Esto de denomina una **DESCRIPCIÓN ESTADÍSTICA DE DATOS**

Descripción Estadística de Datos:

Realizar un análisis de datos considerando MEDIDAS DESCRIPTIVAS

Medidas Descriptivas:

- **Medidas de Posición**
- **Medidas de Dispersión**
- **Medidas de Forma**

¿Cómo analizamos los datos?

a) Datos sin agrupar (pocos datos, $n < 30$)
b) Datos Agrupados en intervalos (muchos datos, $n > 30$)

Medidas de Posición: Dividen a los datos obtenidos en partes proporcionales, de forma que cada parte tenga el mismo número de elementos.

- Media Aritmética
- Media Ponderada
- Mediana
- Moda
- Cuartiles
- Deciles
- Percentiles

Medidas de Dispersión: Nos informan sobre cuánto se alejan del centro los valores que toman los datos a analizar.

- Desvío Estándar
- Varianza
- Rango
- Coeficiente de Variación

Medidas de Forma: Nos informan acerca de la forma de la distribución de los datos alrededor de las medidas de tendencia central.

- Asimetría
- Curtosis

a) Trabajando con datos sin agrupar

a.1) Medidas de Posición o de Tendencia Central

Media Aritmética:

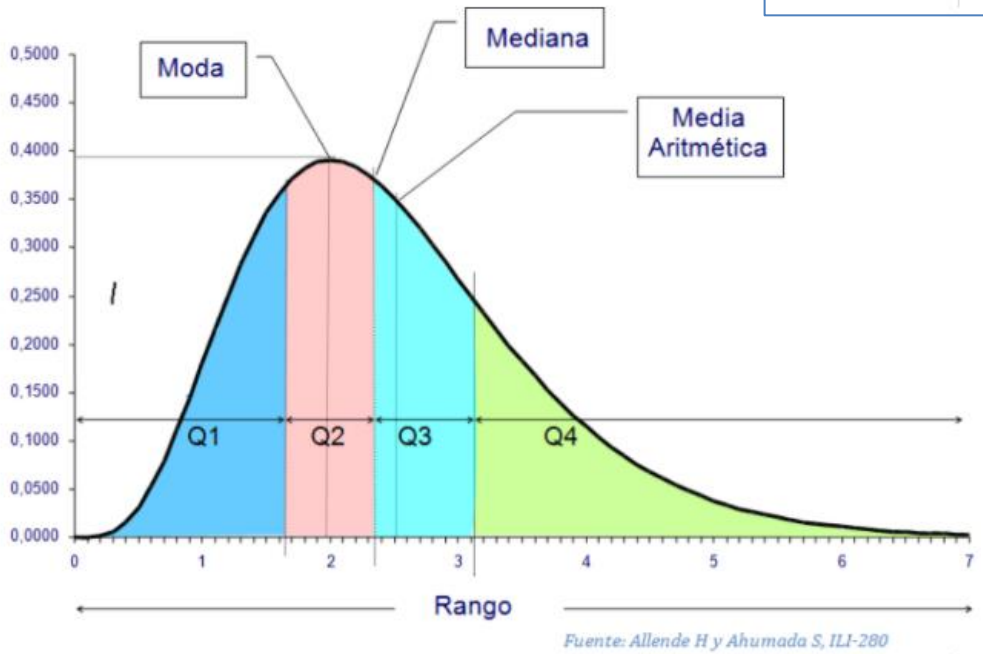
de una muestra: $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$

de una población: $\mu = \frac{\sum_{i=1}^{N} X_i}{N}$

Media Ponderada: Tienen en cuenta la importancia relativa de cada dato.

de una muestra: $\overline{X_w} = \frac{\sum_{i=1}^{n}(w \,.x)}{\sum_{i=1}^{n}(w)}$

de una población: $\mu_w = \frac{\sum_{i=1}^{N}(w \,.x)}{\sum_{i=1}^{N}(w)}$



Fuente: Allende H y Ahumada S, ILI-280

Mediana: centro de los datos.
Divide al conjunto de datos en dos partes iguales (*se deben ordenar los datos*).

$\tilde{X} = X_{\frac{n+1}{2}}$ n impar

$\tilde{X} = \frac{1}{2}\left[X_{\frac{n}{2}} + X_{\frac{n}{2}+1}\right]$ n par

DATOS ORDENADOS

Moda: Valor de mayor frecuencia. Puede ser unimodal y bimodal.
Puede no haber moda si 3 valores o más son los de mayor frecuencia.

Cuartiles, Deciles y Percentiles: Dividen los datos en cuatro, diez o cien partes iguales.
Los datos deben estar ordenados.

$Q_1 = X_{\frac{n+1}{4}}$; $Q_2 = X_{\frac{n+1}{2}} = \tilde{X}$; $Q_3 = X_{\frac{3.(n+1)}{4}}$ con n impar

$Q_1 = \frac{1}{2}(X_{\frac{n}{4}} + X_{\frac{n}{4}+1})$; $Q_2 = \frac{1}{2}(X_{\frac{n}{2}} + X_{\frac{n}{2}+1})$; $Q_3 = \frac{1}{2}(X_{\frac{3.n}{2}} + X_{\frac{3}{4}.(n+1)})$ con n par

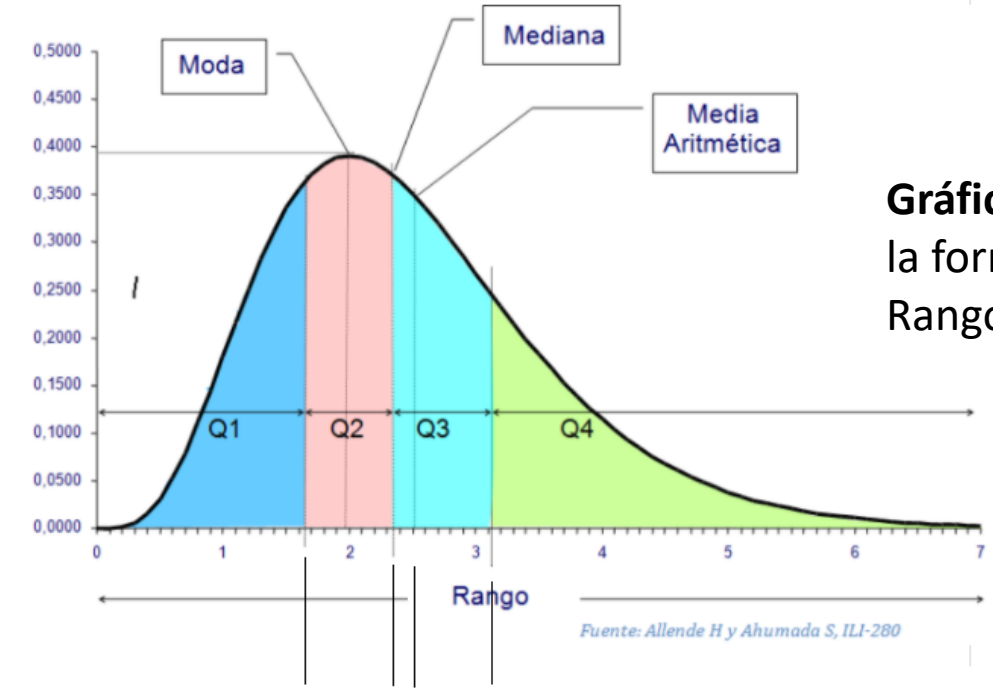
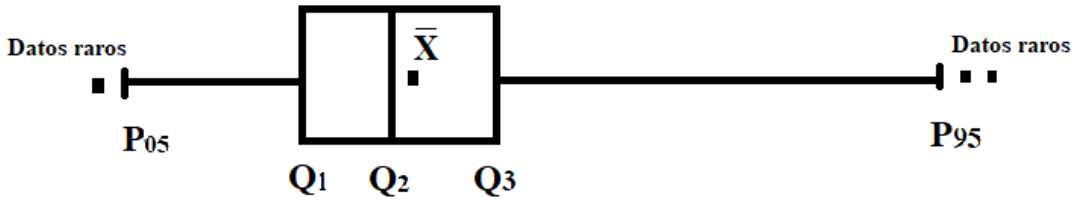


Gráfico de Caja y Extensión: Permite observar la forma de la distribución de los datos.
Rango intercuartil: $R_{IQ} = Q_3 - Q_1$



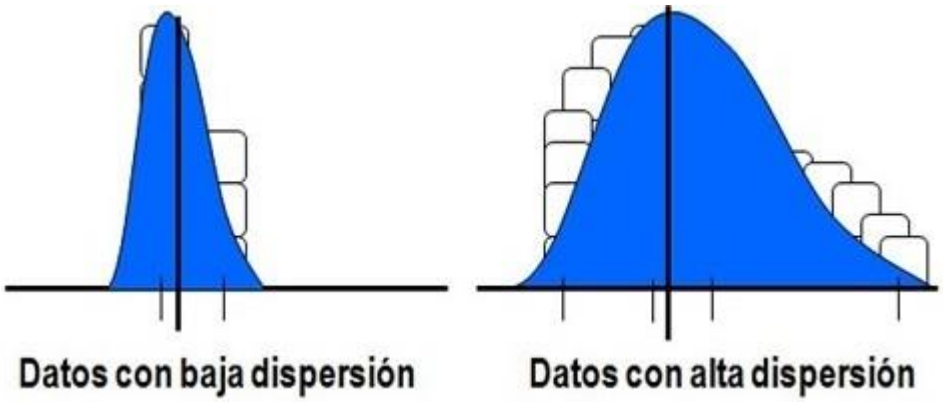
a.2) Medidas de Dispersión

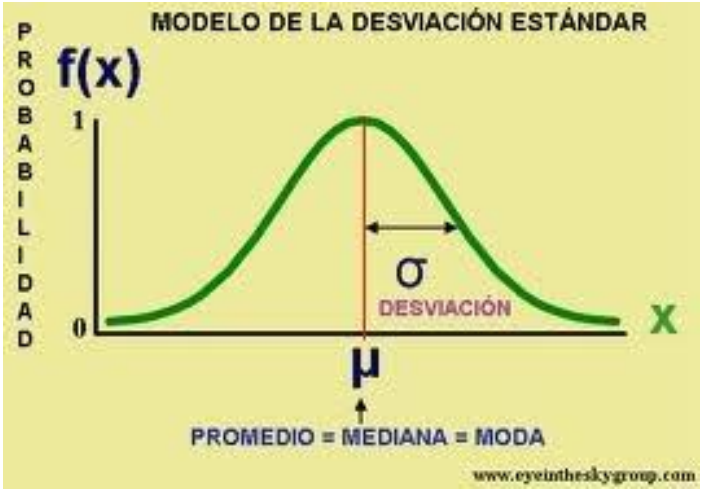
- **Varianza:**
  - Muestra: $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ (n-1 grados de libertad de los datos de la muestra)
  - Población: $\sigma^2 = \frac{\sum_{i=1}^{N}(X_i - \mu)^2}{N}$

- **Desvío estándar**:
  - Muestra: $S = \sqrt{S^2}$
  - Población: $\sigma = \sqrt{\sigma^2}$

- **Rango:** $R = X_{max} - X_{min}$

- **Coeficiente de Variación:**
  - Muestra: $CV = \frac{S}{\overline{X}}$
  - Población: $CV = \frac{\sigma}{\mu}$





Fórmula desagregada de S^2 a nivel de los datos:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \quad \text{pero} \quad \overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad \Rightarrow \quad S^2 = \frac{\sum_{i=1}^{n}\left[X_i - \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\right]^2}{n-1}$$

cuadrado de un binomio: $$S^2 = \frac{\sum_{i=1}^{n}\left[X_i^2 - 2X_i\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) + \left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2\right]}{n-1}$$

distribución Σ: $$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n}\left[X_i\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\right] + \sum_{i=1}^{n}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}{n-1}$$

$\frac{\sum_{i=1}^{n} X_i}{n}$ es la media aritmética, por lo tanto es una Constante:

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - 2\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)\left(\sum_{i=1}^{n} X_i\right) + n\left(\frac{\sum_{i=1}^{n} X_i}{n}\right)^2}{n-1} = \frac{\sum_{i=1}^{n} X_i^2 - \frac{2\left(\sum_{i=1}^{n} X_i\right)^2}{n} + \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}$$

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1} = \frac{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)}$$

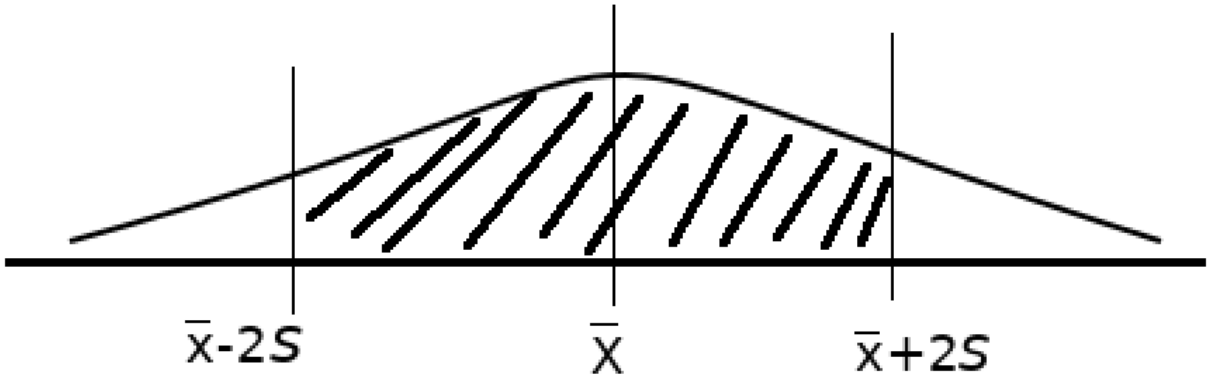
Computar: $\sum_{i=1}^{n} X_i^2$ $\sum_{i=1}^{n} X_i$

para el cálculo. Se evita pasar dos veces por todos los datos

Si *S* es pequeño los datos están agrupados más cerca de la media.
¿Cómo definimos S pequeño o *S* grande?

<u>Teorema de Chebyshev</u>: en relación a un conjunto de datos cualquiera (poblacional o muestral) y una constante k > 1 cuando menos (1 - 1/k²) de los datos debe estar dentro de k desvíos estándar a uno y otro lado de la media para que la dispersión se considere pequeña.

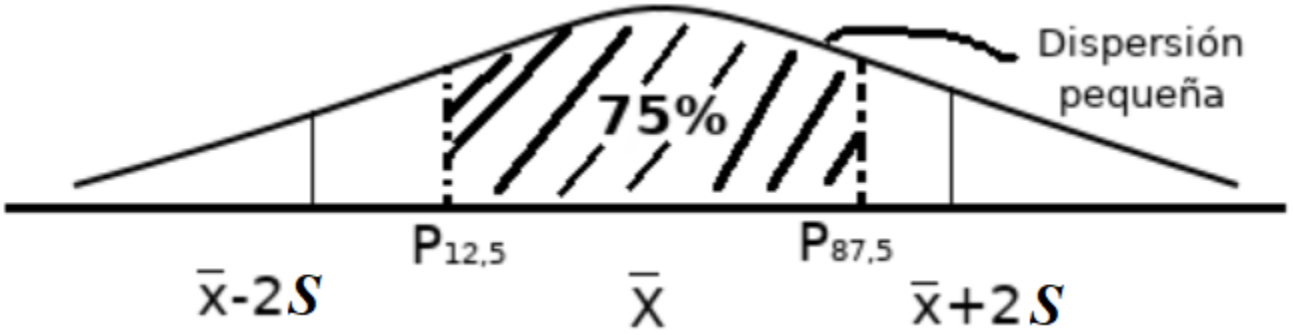
Ejemplo: si elegimos $k = 2$ entonces $1 - \frac{1}{k^2} = \frac{3}{4} = 0{,}75$. El 75% de los datos debe estar a $\overline{X} + 2S$ y $\overline{X} - 2S$ para que la desviación se considere pequeña.



a) Ordenar los datos
b) Calcular $P_{12,5}$ y $P_{87,5}$

$$P_{12,5} = X_{\frac{12,5}{100}.n} \; ; \qquad P_{87,5} = X_{\frac{87,5}{100}.n}$$

c) Calcular $1 - \frac{1}{k^2} \Rightarrow Si\ k = 2 \Rightarrow 1 - \frac{1}{4} = \frac{3}{4}$
d) Si $P_{87,5} > \overline{X} + 2S$ Y $P_{12,5} < \overline{X} - 2S$ entonces el 75% de los datos están dentro de 2 desvíos estándar alrededor de la media y la dispersión es pequeña.

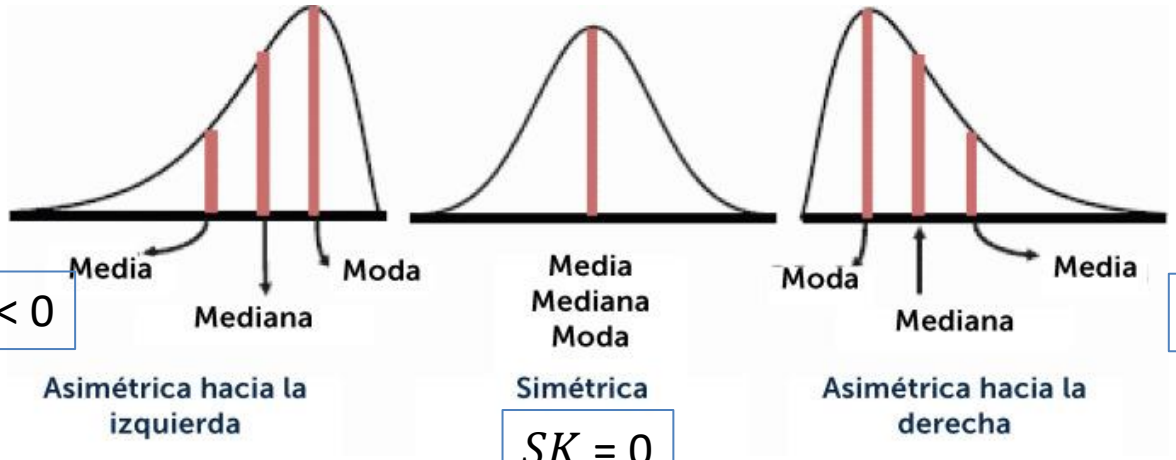


a.3) Medidas de Forma

Asimetría: Miden la mayor o menor simetría de la distribución.

Un índice de Asimetría muy utilizado es el de Pearson.

$$SK = \frac{3\left(\overline{X} - \tilde{X}\right)}{S}$$



$SK < 0$

Asimétrica hacia la izquierda

Simétrica

$SK = 0$

Asimétrica hacia la derecha

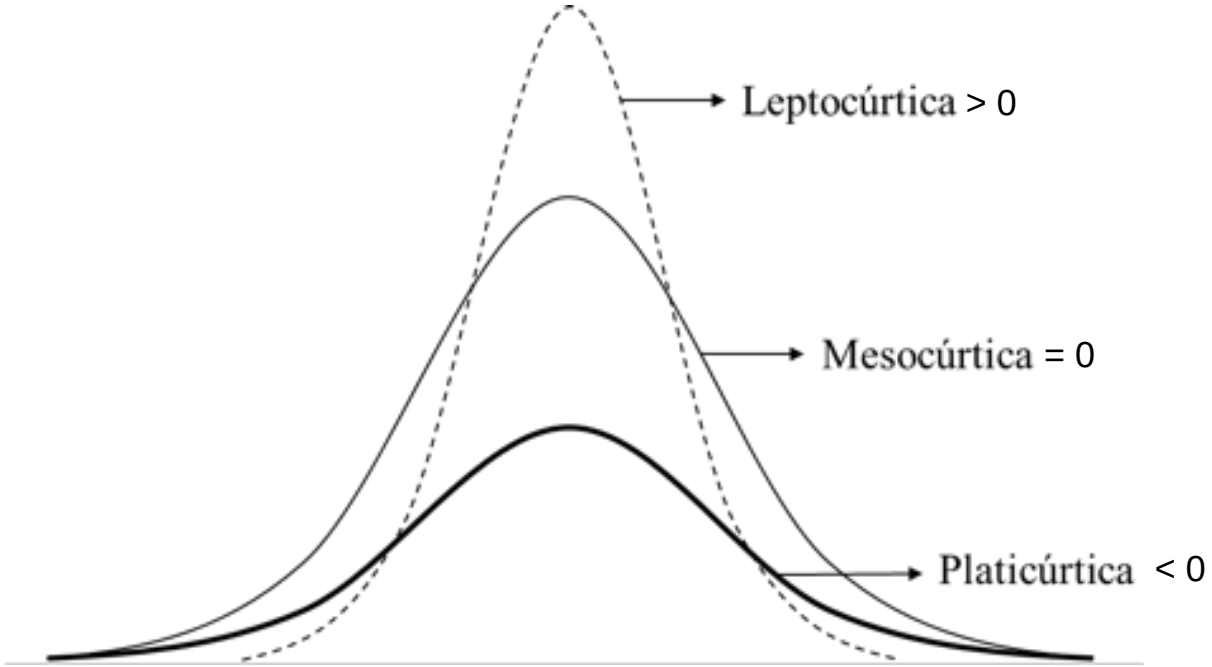
$SK > 0$

Curtosis: Miden la mayor o menor concentración de datos alrededor de la media.

El grado de Curtosis es:

$$Cu = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^4}{S^4} - 3$$

Si este coeficiente es nulo, la distribución se dice normal (similar a la distribución normal de Gauss) y recibe el nombre de **mesocúrtica**. Si el coeficiente es positivo, la distribución se llama **leptocúrtica**, más puntiaguda que la anterior. Hay una mayor concentración de los datos en torno a la media. Si el coeficiente es negativo, la distribución se llama **platicúrtica** y hay una menor concentración de datos en torno a la media. sería más achatada que la primera.



Ejemplo:

Se ha obtenido una muestra del contenido de nicotina (en miligramos) de 40 cigarrillos seleccionados al azar de una empresa tabacalera.

1.24	2.08	1.79	1.58	1.67	1.69	0.72	2.31
1.51	2.55	1.88	1.75	1.37	1.63	1.09	2.46
1.47	1.40	2.03	2.09	0.85	1.92	1.70	1.93
1.68	1.64	1.85	2.37	1.79	1.82	2.11	1.86
1.64	1.69	1.97	2.17	1.74	1.75	2.28	1.90

a) Definir Población objetivo.
b) Definir variable aleatoria.
c) Realizar gráfico de tallo y hojas(original).
d) Realizar el gráfico de tallo y hojas ordenado.
e) Calcular las medidas de posición ($\overline{X}, \tilde{X}, Q_1, Q_2, M_o$).
f) Realizar el gráfico de caja y extensiones.
g) Calcular las medidas de dispersión (S^2, S, R, CV).
h) Calcular el coeficiente de asimetría de Pearson e indicar si existe asimetría y tipo.
i) Calcular la Curtosis e indicar tipo.

Solución:

a) La población objetivo es la población de cigarrillos que fabrica la empresa tabacalera.

b) La variable aleatoria es el contenido de nicotina de los cigarrillos en miligramos (mg).

c) Gráfico de Tallo y Hojas original:

0	
0+	85 72
1	24 47 40 37 09
1+	51 68 64 64 69 79 88 85 97 58 75 67 79 74 69 63 92 82 75 70 90 80 93
2	08 03 09 37 17 11 28 31 46
2+	55

d) Gráfico de Tallo y Hojas ordenado:

0	
0+	72 85
1	09 24 37 40 47
1+	51 58 63 64 64 67 68 69 69 70 74 75 75 79 79 82 85 86 88 90 92 93 97
2	03 08 09 11 17 28 31 37 46
2+	55

e) Medidas y posición

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{70{,}97}{40} = 1{,}774\ mg$$

$$\tilde{X} = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{20} + X_{21}}{2} = \frac{1{,}75\ +\ 1{,}79}{2} = 1{,}77\ \ mg$$

$$Q_1 = \frac{X_{\frac{n}{4}} + X_{\frac{n}{4}+1}}{2} = \frac{X_{10} + X_{11}}{2} = \frac{1{,}63 + 1{,}64}{2} = 1{,}635\ \ mg$$

$$Q_3 = \frac{X_{\frac{3n}{4}} + X_{\frac{3n}{4}+1}}{2} = \frac{X_{30} + X_{31}}{2} = \frac{1{,}97 + 2{,}03}{2} = 2{,}000\ \ mg$$

Mo : Sin moda.

f) Gráfico de Caja y Extensiones:

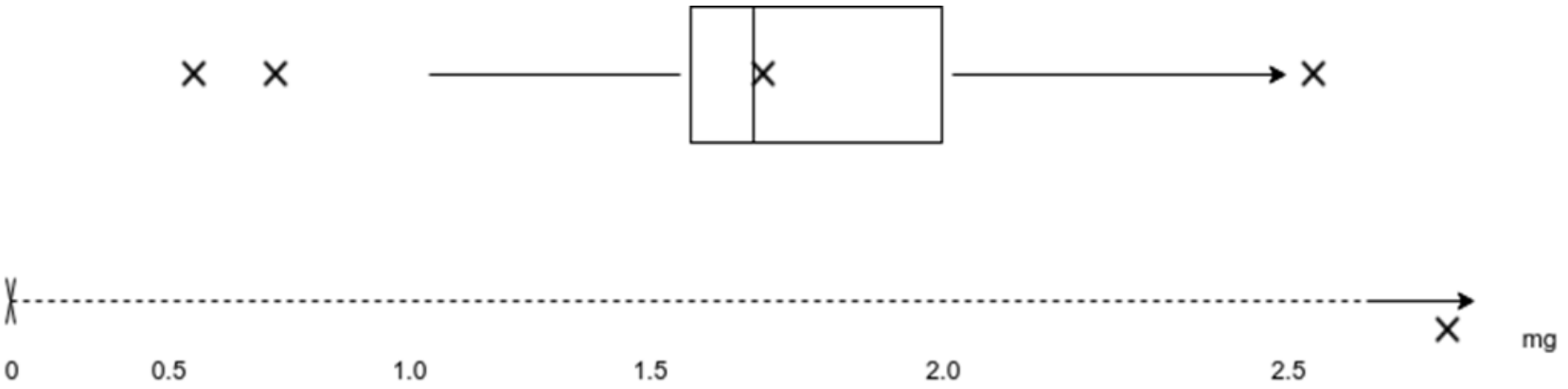
$Q_1 = 1{,}635\ mg$ $\quad Q_3 = 2{,}000\ mg \longrightarrow IQ = Q_3 - Q_1 = 0{,}365\ mg$

$\tilde{X} = 1{,}77\ mg$ $\quad$ Dato Raro $X_r^+ = Q_3 + 1{,}5.IQ$

$\overline{X} = 1{,}774\ mg$ $\quad X_r^+ = 2{,}000 + 1{,}5.0{,}365 = 2{,}5475\ mg$

$X_{max} => X_r \rightarrow Dato\ raro\ X/X > 2{,}5475\ mg$

$X_{min} = Q_1 - 1{,}5.IQ = 1{,}088\ ;\ X/X < 1{,}088\ mg$



g) Medidas de dispersión:

$$S^2 = \frac{n\sum_{i=1}^{n} X_i^2 \; - \; (\sum_{i=1}^{n} X_i \,)^2}{n(n-1)} \; = \; \frac{40\,.\,131{,}8343 \; - \; (70{,}97)^2}{40\;(39)} = 0{,}152 \;\; mg^2$$

$$S \; = \; \sqrt{S^2} \; = \; 0{,}39\; mg \qquad\qquad R \; = \; X_{max} \; - \; X_{min} \; = \; 1{,}83\; mg$$

$$CV \; = \frac{S}{\overline{X}} \; = \; \frac{0{,}39}{1{,}774} \; = \; 0{,}2198$$

h) Coeficiente de Asimetría de Pearson y Curtosis:

$$SK \; = \; 3\frac{(\,\overline{X} - \tilde{X}\,)}{S} \; = \; 3\;\frac{(1{,}774 \; - \; 1{,}77)}{0{,}39} \; = \; 0{,}031$$

Leve asimetría positiva ➞ SK > 0

$$Cu = \frac{1}{n}\frac{\sum_{i=1}^{n}\left(x_i - \overline{X}\right)^4}{S^4} - 3 = \frac{1}{40}\frac{3{,}2474}{0{,}0231} - 3 = 0{,}5145 > 0$$ Leptocúrtica

b) Trabajando con datos agrupados

b.0) Agrupar en Intervalos de Clase: *IC*

Número de *IC*: *NIC*

$$\left.\begin{array}{l} NIC = 5.log_{10}(n) \\ NIC = \sqrt{n} \end{array}\right\} \quad 5 \leq NIC \leq 15$$

Ancho del *IC*: *A*

$$A = \frac{R}{NIC}$$

Para el ejemplo de los cigarrillos:

$$\left.\begin{array}{l} NIC = 5\,log_{10}n = 8{,}01 \\ NIC = \sqrt{N} = \sqrt{40} = 6{,}32 \end{array}\right\} \quad 5 \leq NIC \leq 15$$

$NIC = 5\,log_{10}n = 8{,}01$	*NIC* = 7 intervalos de clase
$NIC = \sqrt{N} = \sqrt{40} = 6{,}32$	

$$A = \frac{R}{NIC} = \frac{X_{max} - X_{min}}{NIC} = \frac{2{,}55 - 0{,}75}{7} = 0{,}257 \quad \rightarrow 0{,}30$$

Diagrama de tallo y hojas ordenado

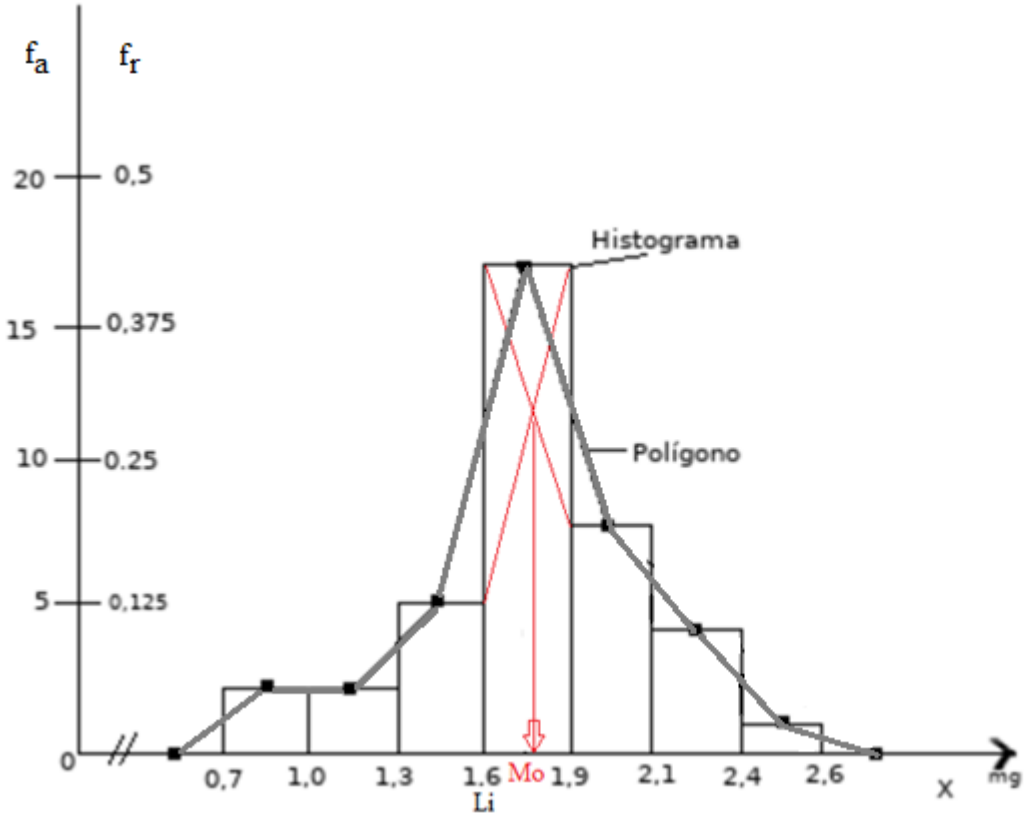
Sabiendo que:
NIC = 7 y *A = 0,3*
con los datos
originales armamos
La tabla de Frecuencias

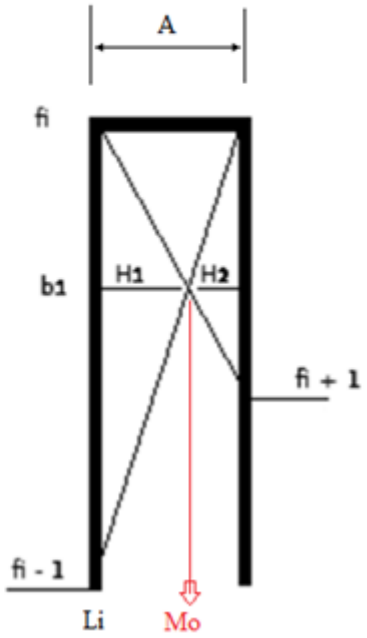
0	
0+	72 85
1	09 24 37 40 47
1+	51 58 63 64 64 67 68 69 69 70 74 75 75 79 79 82 85 86 88 90 92 93 97
2	03 08 09 11 17 28 31 37 46
2+	55

Tabla de Frecuencias

IC	X_{pm}	f_a	f_r	$f_r\%$	F_a	F_r	$F_r\%$
[0,7 - 1,0)	0,85	2	0,05	5	2	0,050	5
[1,0 - 1,3)	1,15	2	0,05	5	4	0,100	10
[1,3 - 1,6)	1,45	5	0,125	12,5	9	0,225	22,5
[1,6 - 1,9)	1,75	17	0,425	42,5	26	0,650	$65 - Q_1, \tilde{X}$
[1,9 - 2,2)	2,05	9	0,225	22,5	35	0,875	$82,5 - Q_3$
[2,2 - 2,5)	2,35	4	0,100	10,0	39	0,975	95
[2,5 - 2,8]	2,65	1	0,025	2,5	40	1,000	100
		40	1,000	100			

Histograma y Polígono de frecuencias simples y relativas simples - Polígono de frecuencia





$$\frac{H_1}{b_1} = \frac{H_2}{b_2}$$

$$b_1 = fi - (f_{i-1})$$

$$b_2 = fi - (f_{i+1})$$

$$A = H_1 + H_2$$

$$H_2 = H_1 \frac{b_2}{b_1}$$

$$H_2 = A - H_1$$

$$0 = H_1 \frac{b_2}{b_1} - (A - H_1)$$

$$0 = H_1 \left(\frac{b_2}{b_1} + 1\right) - A$$

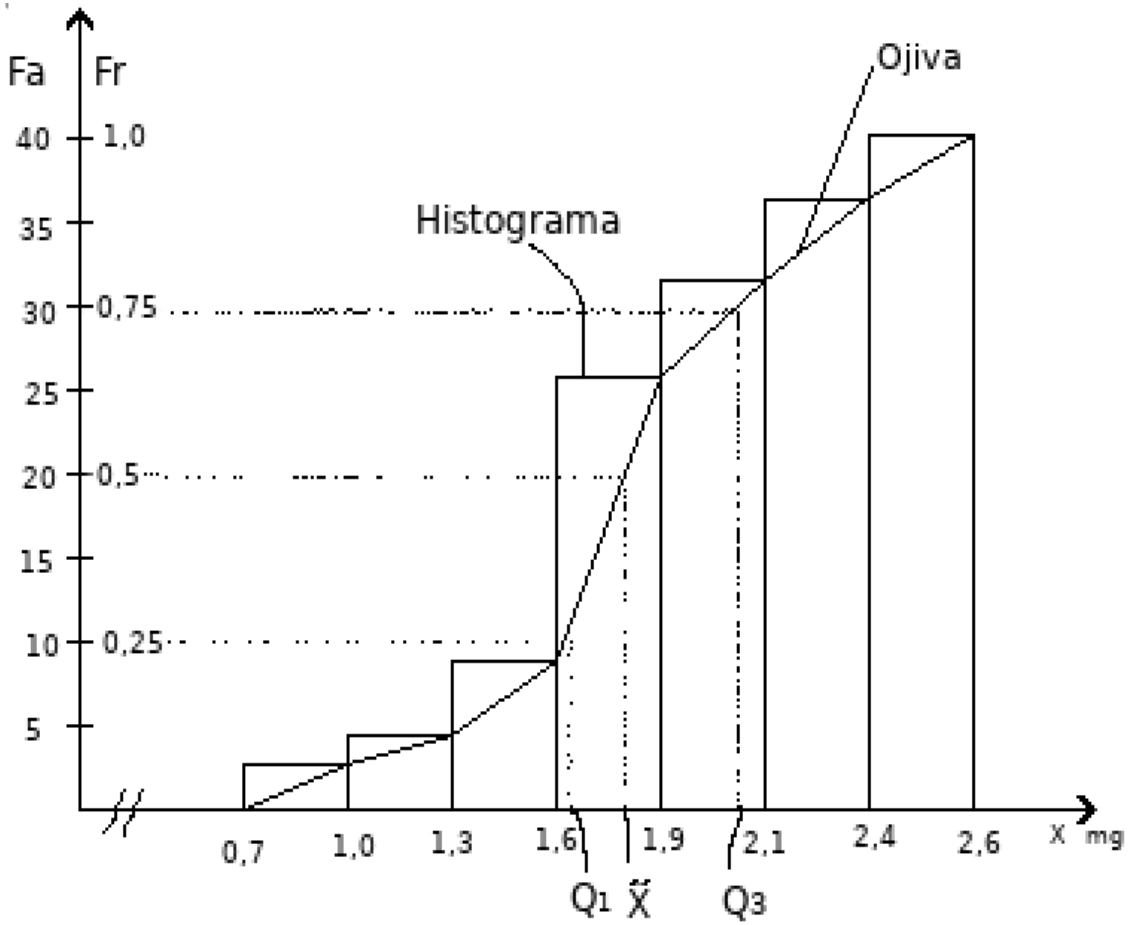
$$H_1 = \frac{A}{\frac{b_2}{b_1} + 1}$$

$$H_1 = Mo - Li = \frac{A}{\frac{fi - f_{i+1}}{fi - f_{i-1}} + 1}$$

$$Mo = Li + \frac{A}{\frac{fi - f_{i+1}}{fi - f_{i-1}} + 1}$$

$$Mo = 1{,}6 + \frac{0{,}3}{\frac{17 - 9}{17 - 5} + 1} = 1{,}78\ mg$$

Gráfico de Frecuencias acumuladas simples y acumuladas relativas simples - Ojiva.



$$\frac{\tilde{X}-L_i}{A}=\frac{0{,}5-Fr_{(i-1)}}{Fr_i-Fr_{(i-1)}} \quad ; \quad \frac{Q_1-L_i}{A}=\frac{0{,}25-Fr_{(i-1)}}{Fr_i-Fr_{(i-1)}} \quad ; \quad \frac{Q_3-L_i}{A}=\frac{0{,}75-Fr_{(i-1)}}{Fr_i-Fr_{(i-1)}}$$

b.1) Medidas de Posición o de Tendencia Central

$$\overline{X} = \sum_{i=1}^{k} (X_{PMi} \,.\, fr_i) \;=\; 1{,}785 \;\; mg$$

$$\tilde{X} = \; L_i \; + \frac{0{,}5 \; - \; Fr_{(i\,-\,1)}}{Fr_i \; - \; Fr_{(i\,-\,1)}} \,.\, A \; = \; 1{,}6 \; + \frac{0{,}5 \; - \; 0{,}225}{0{,}65 \; - \; 0{,}225} \,.\, 0{,}3 \; = \; 1{,}794 \;\; mg$$

$$Q_1 \; = \; L_i \; + \frac{0{,}25 \; - \; Fr_{(i\,-\,1)}}{Fr_i \; - \; Fr_{(i\,-\,1)}} \,.\, A \; = \; 1{,}6 \; + \frac{0{,}25 \; - \; 0{,}225}{0{,}65 \; - \; 0{,}225} \,.\, 0{,}3 \; = \; 1{,}618 \;\; mg$$

$$Q_3 \; = \; L_i \; + \frac{0{,}75 \; - \; Fr_{(i\,-\,1)}}{Fr_i \; - \; Fr_{(i\,-\,1)}} \,.\, A \; = \; 1{,}9 \; + \frac{0{,}75 \; - \; 0{,}65}{0{,}875 \; - \; 0{,}65} \,.\, 0{,}3 \; = \; 2{,}03 \;\; mg$$

$$Mo = L_i + \frac{A}{\frac{fi - f_{i+1}}{fi - f_{i-1}} + 1} = 1{,}6 + \frac{0{,}3}{\frac{17 - 9}{17 - 5} + 1} = 1{,}78\, mg$$

b.2) Medidas de Dispersión

$$S^2 = \frac{n \sum_{i=1}^{k}(X_{PM}^2 . fa_i) - \left[\sum_{i=1}^{k}(X_{PM} . fa_i)\right]^2}{n\,(n-1)} = \frac{40\,(133{,}6) - (71{,}5)^2}{40\,(39)} = 0{,}149\ mg^2$$

$$S = \sqrt{S^2} = 0{,}385\ mg$$

$$CV = \frac{S}{\overline{X}} = \frac{0{,}385}{1{,}785} = 0{,}2157$$

a.3) Medidas de Forma

Asimetría: $$SK = \frac{3\,(\overline{X} - \tilde{X})}{S} = \frac{3\,(1{,}785 - 1{,}794)}{0{,}385} = -0{,}070$$

SK < 0 ➞ Levemente asimétrica negativa

Curtosis: $$Cu = \frac{1}{n}\frac{\sum_{i=1}^{k}\left[\left(X_{PM} - \overline{X}\right)^4 . fa_i\right]}{S^4} - 3 = \frac{1}{40}\frac{2{,}92855}{0{,}0222} - 3 = 0{,}29792$$

$Cu > 0 \quad Leptocurtica$

Comparación de estadísticos entre los análisis de datos sin agrupar y datos agrupados

	Datos sin agrupar	Datos agrupados
$\overline{X}$	1,774	1,785
M_o	Sin Moda	1,78
Q_1	1,635	1,618
Q_3	2,000	2,030
$\tilde{X}$	1,77	1,794
SK	0,031	-0.070
Cu	0,5145	0,2978
S^2	0,152	0,149
S	0,390	0,385
CV	0,2198	0,2157
R	1,83	2,1