

Tema 2: Resumen de datos y descripciones estadísticas

Finalidad: Presentar un conjunto de datos analizados desde el punto de vista estadístico.

Descripción Estadística: Realizar un análisis de datos considerando medidas descriptivas.

Medidas descriptivas:

- Medias de Posición
- Medidas de Dispersión
- Sesgo

Puede trabajarse:

- Con datos sin Agrupar (poca cantidad de datos)
- Con datos agrupados (en intervalos, muchos datos)

Medidas de Posición:

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

Medidas de Dispersión:

- Desvío Estándar
- Varianza
- Rango
- Coeficiente de Variación

Otras medidas:

- Asimetría
- Curtosis

Medidas 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:

- de una muestra: $\overline{X_w} = \frac{\sum_{i=1}^n (w \cdot x)}{\sum_{i=1}^n (w)}$
- de una población: $\mu_w = \frac{\sum_{i=1}^N (w \cdot x)}{\sum_{i=1}^N (w)}$

Mediana (centro de los datos):

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

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

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

Medidas de Posición Ponderadas:

Tienen en cuenta la importancia relativa de cada dato.

Ej: inversiones a Plazo Fijo:

Rendimiento: 7%, 8%, 9% (si coloco diferentes cantidades de dinero a plazo fijo).

$$\text{Rendimiento Medio: } \frac{7+8+9}{3} = 8\%$$

$$\text{Rendimiento Medio Ponderado: } \overline{X}_w = \frac{x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3}{w_1 + w_2 + w_3} = \frac{\sum_{i=1}^3 (x_i \cdot w_i)}{\sum_{i=1}^3 (w_i)}$$

Siendo w_i : cantidad de dinero.

Ej: Media General

Si en un proceso de muestreo queremos conocer el Promedio General de los parámetros de las k muestras distintas:

- Medias Muestrales: $\overline{X}_1; \overline{X}_2; \overline{X}_3; \dots; \overline{X}_k$

Obtenemos que: $n_1; n_2; n_3; \dots; n_k$ *datos*

$$\overline{\overline{X}} = \frac{n_1 \cdot \overline{X}_1 + n_2 \cdot \overline{X}_2 + \dots + n_k \cdot \overline{X}_k}{n_1 + n_2 + \dots + n_k} = \frac{\sum_{i=1}^k (n_i \cdot \overline{X}_i)}{\sum_{i=1}^k n_i}$$

Moda: Valor de mayor frecuencia

- Puede ser unimodal y bimodal
- Puede no haber moda. Ej: 3 valores o más con la misma frecuencia (más alta)

Cuartiles, Deciles y Percentiles (con datos ordenados):

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

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

Ej: n = 10

$$\begin{array}{cccccccccccc} 1 & 2 & (3) & 4 & 5 & | & 6 & 7 & (8) & 9 & 10 \\ & & Q_1 & & & & Q_2 & & & Q_3 & \end{array}$$

Ej: n = 12

$$\begin{array}{cccccccccccccc} 1 & 2 & 3 & | & 4 & 5 & 6 & | & 7 & 8 & 9 & | & 10 & 11 & 12 \\ & & & & Q_1 & & & & Q_2 & & & & Q_3 & & \end{array}$$

$$D_1 = X_{\frac{n+1}{10}} \quad ; \quad D_2 = X_{\frac{2 \cdot (n+1)}{10}} \quad \dots \quad n \text{ impar}$$

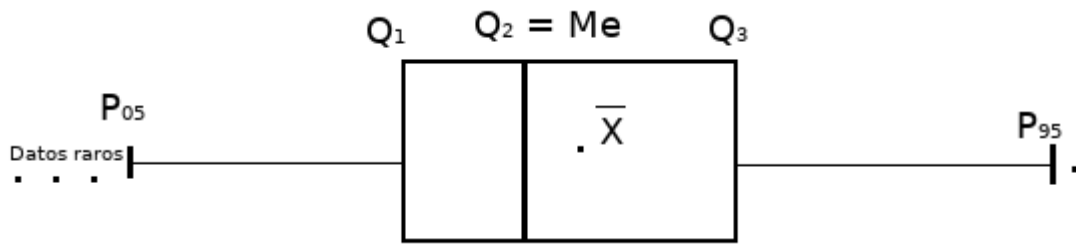
$$D_1 = \frac{1}{2} \cdot (X_{\frac{n}{10}} + X_{\frac{n}{10}+1}) \quad ; \quad D_2 = \frac{1}{2} \cdot (X_{\frac{2 \cdot n}{10}} + X_{\frac{2 \cdot n}{10}+1}) \quad \dots \quad n \text{ par}$$

$$P_1 = X_{\frac{n+1}{100}} \quad ; \quad P_2 = X_{\frac{2 \cdot (n+1)}{100}} \quad \dots \quad n \text{ impar}$$

$$P_1 = \frac{1}{2} \cdot (X_{\frac{n}{100}} + X_{\frac{n}{100}+1}) \quad ; \quad P_2 = \frac{1}{2} \cdot (X_{\frac{2 \cdot n}{100}} + X_{\frac{2 \cdot n}{100}+1}) \quad n \text{ par}$$

Gráfico de Caja y Extensión

Rango intercuartil: $R_{IQ} = Q_3 - Q_1$



Medidas de Dispersión

- Varianza

- Muestra: $S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1}$ (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: $CV = \frac{S}{\bar{X}}$
- Población: $CV = \frac{\sigma}{\mu}$

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

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad \text{pero } \bar{X} = \frac{\sum_{i=1}^n X_i}{n}$$

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

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

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

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

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

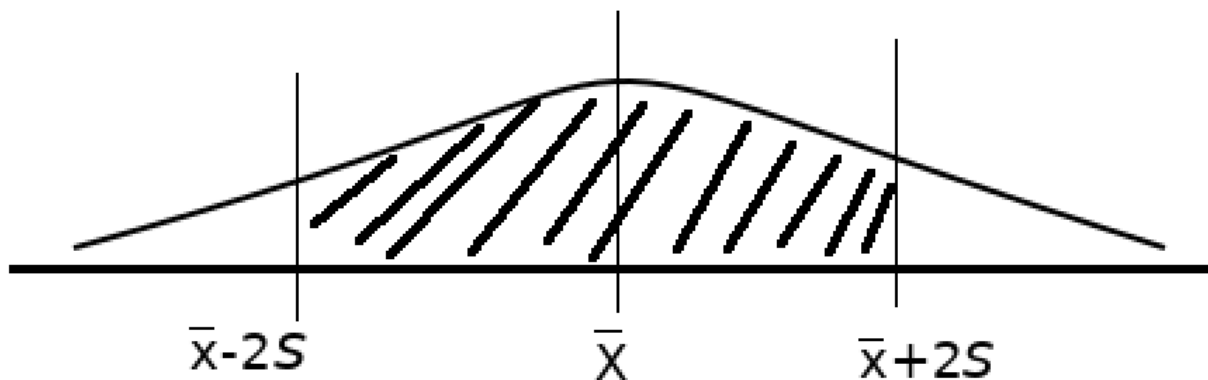
Teorema de Chebyshev

Si S es pequeño los datos están agrupados más cerca de la media.

¿Cómo definimos S pequeño o S grande?

Teorema: en relación a un conjunto de datos cualquiera (poblacional o muestral) y una constante $k > 1$ cuando menos $1 - \frac{1}{k^2}$ 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 $1 - \frac{1}{k^2} = \frac{3}{4} = 75\%$. El 75% de los datos debe estar a $\bar{X} + 2S$ y $\bar{X} - 2S$ para que la desviación se considere pequeña.

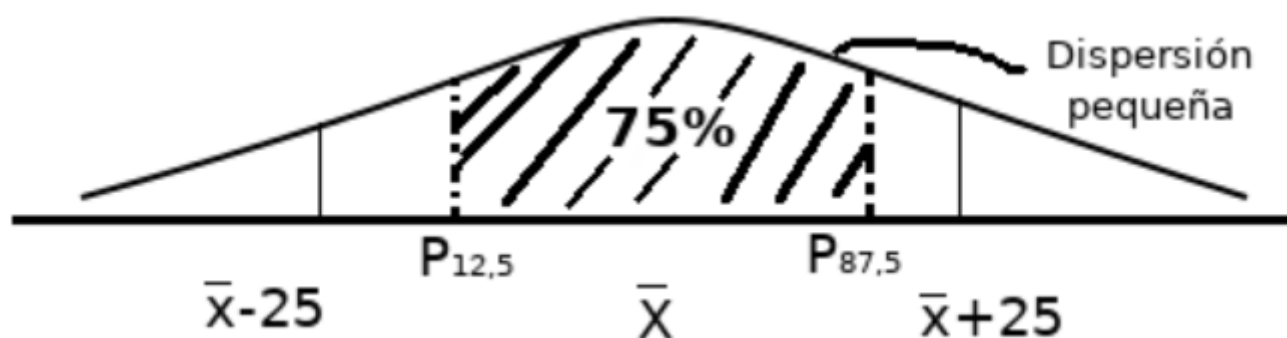


- a) Ordenar los datos
- b) Calcular $P_{12,5}$ y $P_{87,5}$

$$P_{12,5} = \frac{X_{12,5}}{100} \cdot n \quad ; \quad P_{87,5} = \frac{X_{87,5}}{100} \cdot n$$

- c) Calcular $1 - \frac{1}{k^2} \Rightarrow \text{Si } k = 2 \Rightarrow 1 - \frac{1}{4} = \frac{3}{4}$

- d) Si $P_{12,5} > \bar{X} + 2S$ Y $P_{87,5} < \bar{X} - 2S$ entonces el 75% de los datos están dentro de 2 desvíos estándar alrededor de la media y la dispersión es pequeña.



Problema:

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

- Definir Población objetivo.
- Definir variable aleatoria.
- Realizar gráfico de tallo y hojas(original).
- Realizar el gráfico de tallo y hojas ordenado.
- Calcular las medidas de posición ($\bar{X}, \tilde{X}, Q_1, Q_2, M_o$).
- Realizar el gráfico de caja y extensiones.
- Calcular las medidas de dispersión (S^2, S, R, CV).
- Calcular el coeficiente de asimetría de Pearson e indicar si existe asimetría y tipo.
- Agrupar en intervalos de clase.
- Realizar el Histograma y Polígono de frecuencias simples y relativas simples.
- Realizar la gráfica de frecuencias acumuladas simples, relativas simples y la ojiva ascendente.
- Determinar las medidas de posición para datos agrupados de forma gráfica y analítica.
- Calcular la varianza, desvío estándar y coeficiente de variación para datos agrupados.
- Calcule la asimetría de Pearson considerando las variables obtenidas para datos agrupados e indique el tipo de asimetría.
- Compare las medidas de posición y dispersión calculadas para datos sin agrupar y datos agrupados. Extraiga conclusiones de esta comparación.

Resultados:

- 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

$$\bar{X} = \sum_{i=1}^n x_i/n = \frac{70,97}{40} = 1,774 \text{ mg}$$

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

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

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

Mo : Sin moda.

f) Gráfico de Caja y Extensiones

$$Q_1 = 1,635 \text{ mg}$$

$$IQ = Q_3 - Q_1 = 0,365 \text{ mg}$$

$$Q_3 = 2,000 \text{ mg}$$

$$\tilde{X} = 1,77 \text{ mg}$$

$$\text{Dato Raro } X_r^+ = Q_3 + 1,5 \cdot IQ$$

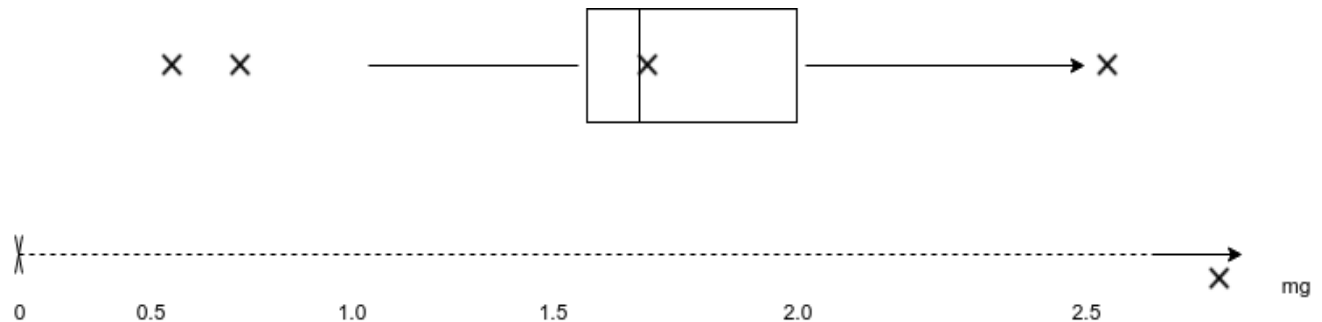
$$\bar{X} = 1,774 \text{ mg}$$

$$X_r^+ = 2,000 + 1,5 \cdot 0,365$$

$$X_r^+ = 2,5475 \text{ mg}$$

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

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



g) Medidas de dispersión

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

$$S^2 = 0,152 \text{ mg}^2$$

$$R = X_{max} - X_{min} = 1,83 \text{ mg}$$

$$S = \sqrt{S^2} = 0,39 \text{ mg}$$

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

h) Coeficiente de Asimetría de Pearson

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

Leve asimetría positiva $\rightarrow SK > 0$

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

i) Agrupar en intervalos de clase

$$NIC = 5 \log_{10} n = 8,01$$

$$NIC = \sqrt{N} = \sqrt{40} = 6,32$$

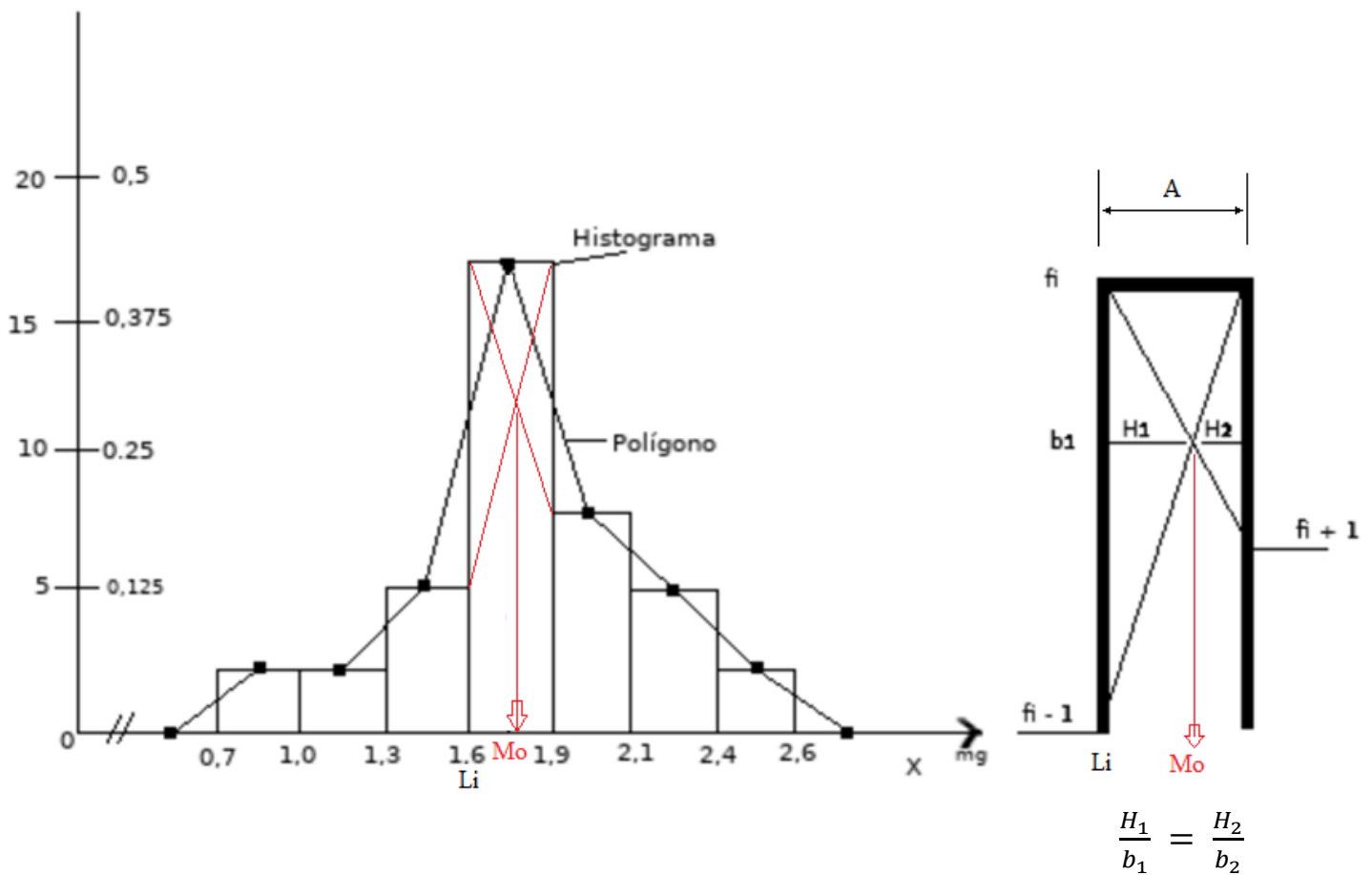
$NIC = 5 \log_{10} n = 8,01$	$NIC = 7 \text{ intervalos de clase}$
$NIC = \sqrt{N} = \sqrt{40} = 6,32$	

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

$$A \rightarrow 0,30$$

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
[1,9 - 2,2)	2,05	9	0,225	22,5	35	0,875	82,5 - Q_2
[2,2 - 2,4)	2,35	4	0,100	10,0	39	0,975	95
[2,4 - 2,8)	2,65	1	0,025	2,5	40	1,000	100
		40	1,000	100			

j) Histograma y Polígono de frecuencias simples y relativas simples - Polígono de frecuencias



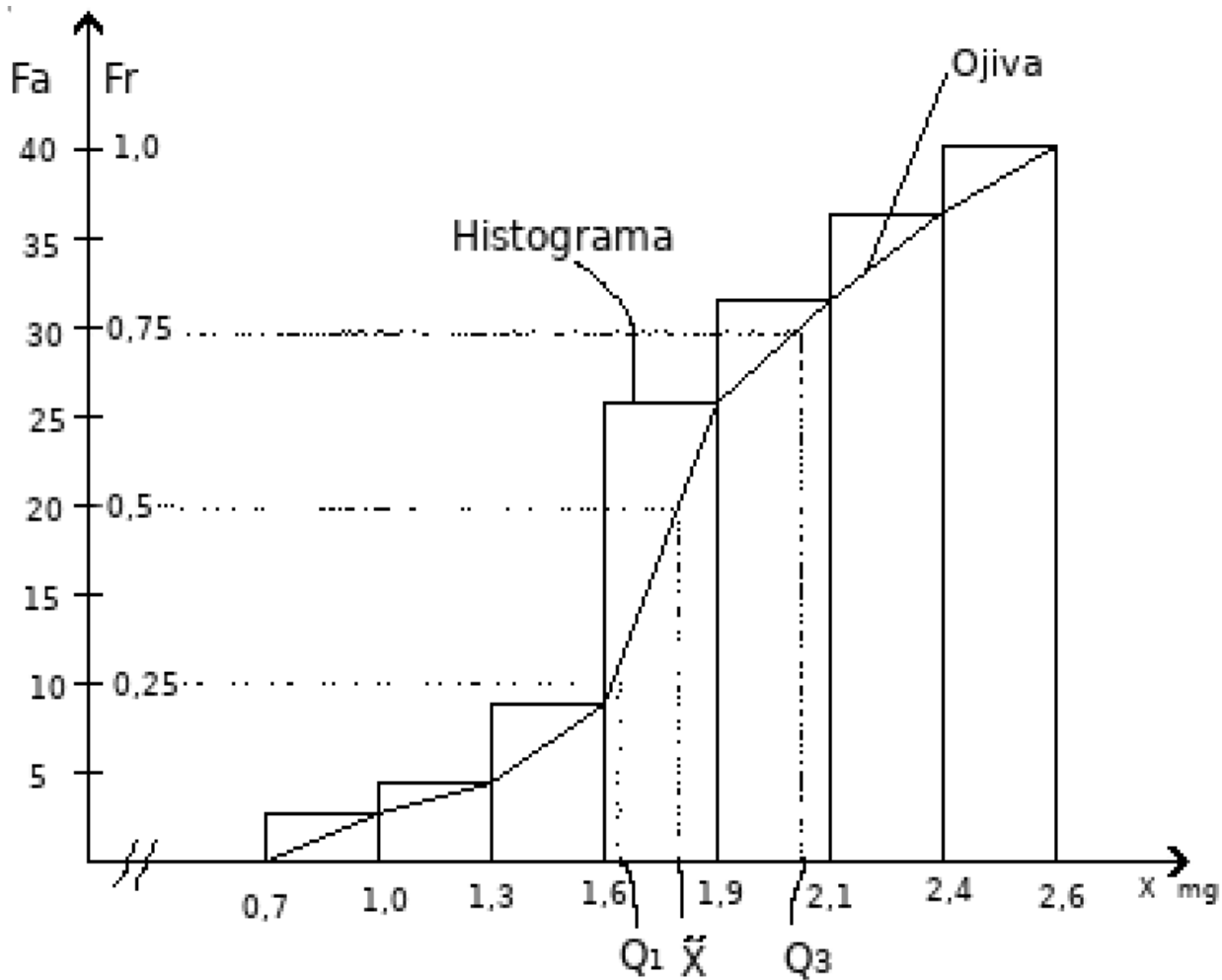
$$b_1 = f_i - (f_i - 1) \quad - \quad H_2 = \frac{H_1 \frac{b_1}{b_2}}{A - H_1}$$

$$b_2 = f_i - (f_i + 1)$$

$$A = H_1 + H_2 \quad 0 = H_1 \frac{b_1}{b_2} - (A - H_1) \quad 0 = H_1 \left(\frac{b_1}{b_2} + 1 \right) - A \quad H_1 = \frac{A}{\frac{b_1}{b_2} + 1}$$

$$H_1 = Mo - Li = \frac{A}{\frac{f_i - (f_i - 1)}{f_i - (f_i + 1)} + 1} \quad Mo = Li + \frac{A}{\frac{f_i - (f_i - 1)}{f_i - (f_i + 1)} + 1} \quad Mo = 1,6 + \frac{0,3}{\frac{17-5}{17-9} + 1} = 1,72 \text{ mg}$$

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



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

l) Medidas de Posición para datos agrupados

$$\bar{X} = \sum_{i=1}^k (X_{PMi} \cdot fr_i) = 1,785 \text{ mg}$$

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

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

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

m) Medidas de dispersión para datos agrupados

$$S^2 = \frac{n \sum_{i=1}^k (X_{PM}^2 \cdot f a_i) - [\sum_{i=1}^k (X_{PM} \cdot f a_i)]^2}{n(n-1)} = \frac{40(133,6) - (71,5)^2}{40(39)} = 0,149 \text{ mg}$$

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

$$CV = \frac{S}{\bar{X}} = \frac{0,385}{1,7875} = 0,2153$$

n) Asimetría y Curtosis

$$SK = \frac{3(\bar{X} - \tilde{X})}{S} = \frac{3(1,7875 - 1,794)}{0,417} = -0,046$$

Levemente asimétrica negativa. $SK < 0$

$$Cu = \frac{1}{n} \frac{\sum_{i=1}^k [(X_{PM} - \bar{X})^4 \cdot f a_i]}{S^4} - 3 = \frac{1}{40} \frac{2,92855}{0,0222} - 3 = 0,2978 > 0 \rightarrow \text{Leptocúrtica}$$

o) Comparación de medidas de posición y dispersión

	Datos sin agrupar	Datos agrupados
\bar{X}	1,774	1,785
M_o	Sin Moda	1,72
Q_1	1,635	1,618
Q_3	2,000	2,030
\tilde{X}	1,77	1,794
SK	0,031	-0,046
Cu	0,5145	0,2978
S^2	0,152	0,149
S	0,390	0,385
CV	0,2198	0,2153
R	1,83	2,1