

Análisis de datos Univariante

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA





Nociones preliminares

Población estadística: Conjunto de elementos que pueden ser agrupados dentro de una determinada clase.

Muestra: Subconjunto de la población que se observa.

Individuo o Unidad estadística: Cada uno de los elementos que forman una población estadística.

Variable estadística: Conjunto de valores, medidas u observaciones comunes en toda la población, es decir, características que se pueden medir u observar.

Escalas de medida.

Variables cualitativas o Categóricas.

Escala nominal. Variable Sexo: Hombre, Mujer.

1 2

Escala ordinal. Variable Grado de aceptación de un producto:

Muy Malo, Malo, Regular, Bueno, Muy Bueno.

1 2 3 4 5

Variables cuantitativas o Numéricas.

Escala por intervalos. Variable Temperatura:

51 48 45 31 °C

Escala por ratios.

Igual a las anteriores, pero a la unidad de medición se le asigna un valor de origen verdadero de valor cero.

$$\left\{ \begin{array}{l} 1000\text{€} \xrightarrow{300\text{€ (30\%)}} 1300\text{€} \\ 3000\text{€} \xrightarrow{300\text{€ (10\%)}} 3300\text{€} \end{array} \right.$$

Variables discretas y continuas.

Nº de Hijos de una familia. Velocidad de un Automóvil.

Tabla de frecuencias

1	5	7	4	3	1	3	4	3	6	2	3	4	3	1	5	3	4	2	1	3	4	3	2	1
2	1	3	4	3	2	1	3	2	5	3	2	1	2	2	3	1	3	4	3	1	3	1	3	3
3	5	3	2	1	2	4	3	6	3	1	3	2	3	6	2	4	3	2	1	3	2	4	2	2
2	1	2	2	5	2	6	3	2	4	8	3	2	4	1	3	2	5	2	2	4	3	5	3	3
2	3	4	2	3	2	2	1	3	2	5	2	2	3	5	2	4	3	6	3	4	1	3	2	5
3	2	1	3	2	3	7	2	5	2	3	2	4	3	1	3	3	1	3	3	4	3	4	3	3
2	6	3	2	3	1	3	4	3	2	3	4	1	3	4	2	2	5	4	2	3	2	4	3	8
3	4	7	3	2	5	3	2	3	1	4	3	2	6	3	5	4	3	6	2	1	2	3	2	4

Tabla de frecuencias

Frec. Absoluta	Frec. Relativa (f.d.d. empírica)	Frec. Absoluta acumulada	Frec. Relativa aculada (f.d.D. empírica)
n_i	$f_i = \frac{n_i}{n}$	$N_i = n_i + N_{i-1}$ ($N_0=0$)	$F_i = f_i + F_{i-1}$ ($F_0=0$)

Tabla de frecuencias

Frec. Absoluta	Frec. Relativa (f.d.d. empírica)	Frec. Absoluta acumulada	Frec. Relativa aculada (f.d.D. empírica)
n_i	$f_i = \frac{n_i}{n}$	$N_i = n_i + N_{i-1}$ ($N_0=0$)	$F_i = f_i + F_{i-1}$ ($F_0=0$)

x_i	n_i	f_i	$100f_i$ (%)	N_i	F_i	$100F_i$ (%)
1	25	0.125	12.5	25	0.125	12.5
2	53	0.265	26.5	78	0.390	39.0
3	67	0.335	33.5	145	0.725	72.5
4	28	0.140	14.0	173	0.865	86.5
5	14	0.070	7.0	187	0.935	93.5
6	8	0.040	4.0	195	0.975	97.5
7	3	0.015	1.5	198	0.990	99.0
8	2	0.010	1.0	200	1	100

$n = 200$

Tabla de frecuencias

Intervalos	Marca de clase x_i^*	Frec. Absoluta	Frec. Relativa (f.d.d. empírica)	Frec. Absoluta acumulada	Frec. Relativa aculada (f.d.D. empírica)
$[a, b)$	x_i^*	n_i	$f_i = \frac{n_i}{n}$	$N_i = n_i + N_{i-1}$ ($N_0=0$)	$F_i = f_i + F_{i-1}$ ($F_0=0$)

Posibles tipos de intervalos: (a, b) , $[a, b]$, $(a, b]$, $[a, b)$.

Intervalo	x_i^*	n_i	f_i	N_i	F_i
$[1, 2]$	1.5	78	0.39	78	0.39
$(2, 4]$	3	95	0.475	173	0.865
$(4, 6]$	5	22	0.11	195	0.975
$(6, 8]$	7	5	0.025	200	1

Medidas gráficas

Diagrama de barras (frecuencias absolutas)

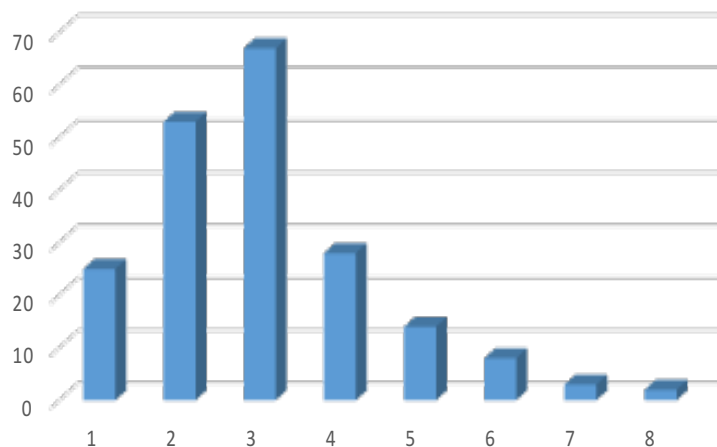


Diagrama de barras (frecuencias relativas)

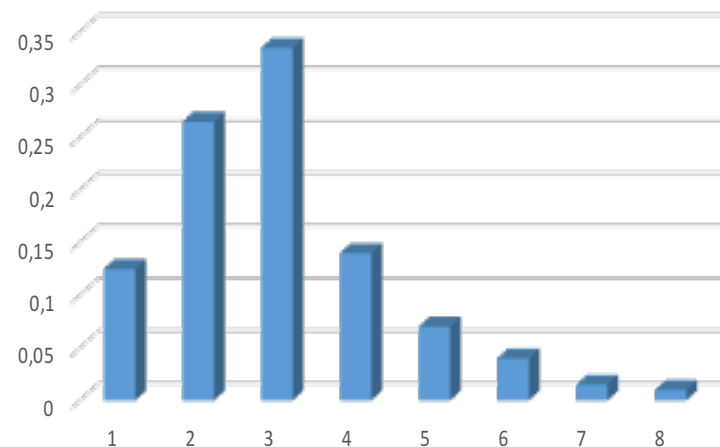
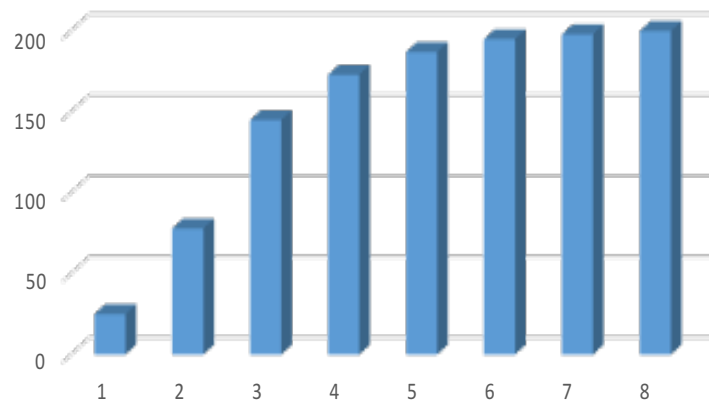
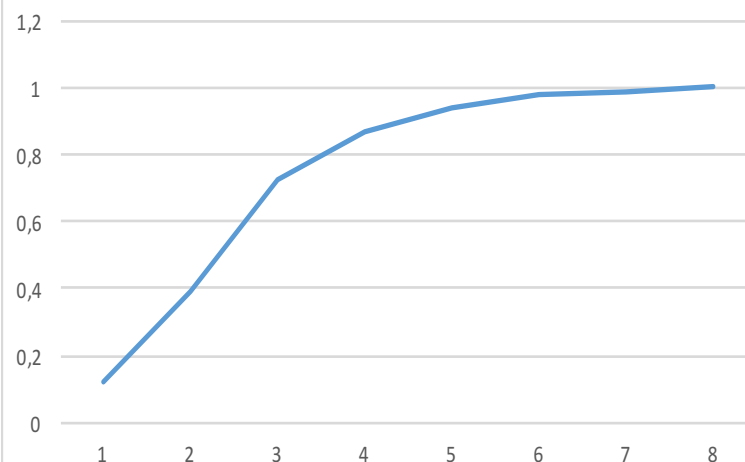


Diagrama de barras (frecuencias absolutas acumuladas)

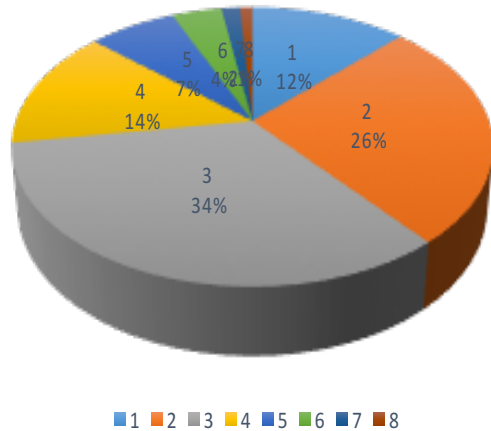


Polígono de frecuencias (acumuladas)

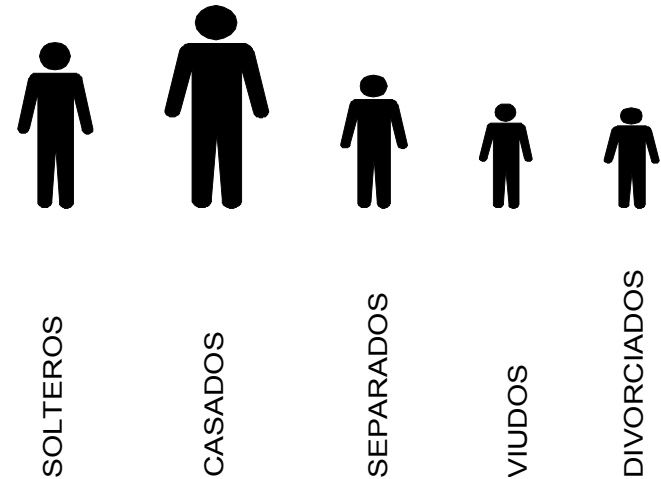
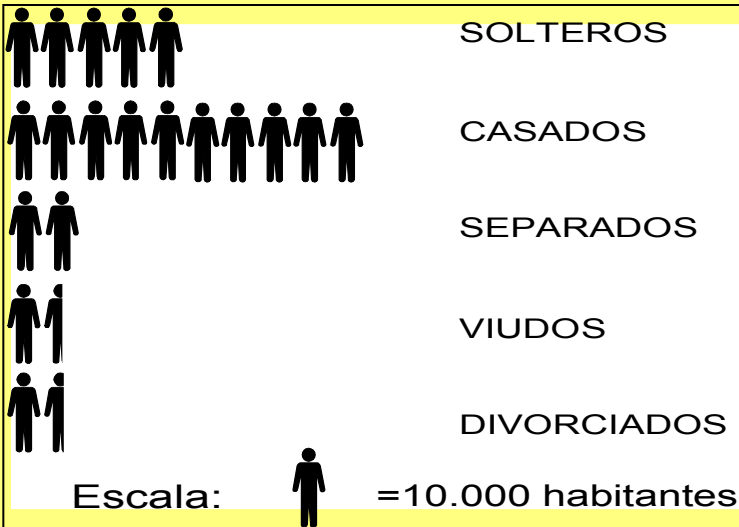
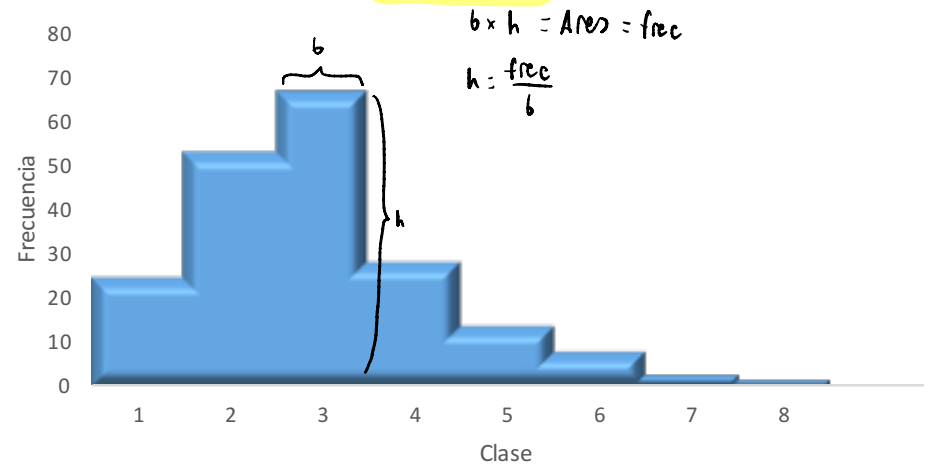


Medidas gráficas

Diagrama de pastel

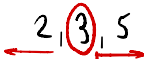
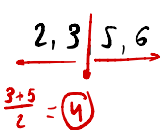


Histograma



Pictogramas

Medidas de centralización (Datos sin agrupar)

Moda (m_o)	Valor de máxima frecuencia
Mediana (m_e)	<p> $n = \text{impar} \rightarrow m_e = x_{\left(\frac{n+1}{2}\right)}$  </p> <p> $n = \text{par} \rightarrow m_e = \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2}$  </p>
Media (\bar{x})	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Localización: Cuartiles, deciles, percentiles	<p>C_1 = Valor que deja el 25% de los datos a su izquierda.</p> <p>C_2 = Igual que C_1 con el 50% de los datos.</p> <p>C_3 = Igual que C_1 con el 75% de los datos.</p>

Supongamos que se ha realizado un estudio de población en el que sólo se mide la edad de cada individuo. Los resultados, tras analizar una muestra de 25 individuos son:

18 20 30 42 19 55 62 18 41 42 22 35 62
71 47 64 25 75 19 26 42 45 23 24 25

se ordenan los datos originales:

18 18 19 19 20 22 23 24 25 25 26 30 35
41 42 42 42 45 47 55 62 62 64 71 75

Edad	n_i	f_i	N_i	F_i
18	2	0,08	2	0,08
19	2	0,08	4	0,16
20	1	0,04	5	0,20
22	1	0,04	6	0,24
23	1	0,04	7	0,28
24	1	0,04	8	0,32
25	2	0,08	10	0,40
26	1	0,04	11	0,44
30	1	0,04	12	0,48
35	1	0,04	13	0,52
41	1	0,04	14	0,56
42	3	0,12	17	0,68
45	1	0,04	18	0,72
47	1	0,04	19	0,76
55	1	0,04	20	0,80
62	2	0,08	22	0,88
64	1	0,04	23	0,92
71	1	0,04	24	0,96
75	1	0,04	25	1,00
	25	1,00		

✓ Moda: $m_o = 42$

✓ Mediana: $m_e = 35$

✓ Media: $\bar{x} = \frac{18 + 18 + \dots + 71 + 75}{25} = 38,08$

$$\bar{x} = \frac{1}{25}(18 \times 2 + 19 \times 2 + \dots + 42 \times 3 + \dots + 75 \times 1) = 38,08$$

✓ Cuartiles: $C_1 = \frac{22 + 23}{2} = 22,5$

$$C_3 = \frac{47 + 55}{2} = 51$$

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA

EP
SC

Medidas de dispersión (Datos sin agrupar)



Rango (R)	$R = X_{(n)} - X_{(1)}$ <i>= grande - pequeño</i>
Varianza (S^2)	$S^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n} \sum_{i=1}^k (X_i - \bar{X})^2 n_i = \left(\frac{1}{n} \sum_{i=1}^n X_i^2 \right) - \bar{X}^2 = \left(\frac{1}{n} \sum_{i=1}^k X_i^2 n_i \right) - \bar{X}^2$
Desviación típica (S)	$S = +\sqrt{S^2}$
Cuasi-varianza (\bar{S}^2)	$\bar{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2 = \frac{1}{n-1} \sum_{i=1}^k (X_i - \bar{X})^2 n_i$
Cuasi-desviación típica (\bar{S})	$\bar{S} = +\sqrt{\bar{S}^2}$
Relación entre (S^2) y (\bar{S}^2)	$n S^2 = (n-1) \bar{S}^2$
Error estándar (e.s.)	$e.s. = \frac{S}{\sqrt{n-1}} = \frac{\bar{S}}{\sqrt{n}}$
Coefficiente de variación (CV)	$CV = \frac{S}{\bar{X}}$

Supongamos que se ha realizado un estudio de población en el que sólo se mide la edad de cada individuo. Los resultados, tras analizar una muestra de 25 individuos son:

18 20 30 42 19 55 62 18 41 42 22 35 62
71 47 64 25 75 19 26 42 45 23 24 25

se ordenan los datos originales:

18 18 19 19 20 22 23 24 25 25 26 30 35
41 42 42 42 45 47 55 62 62 64 71 75

Edad	n_i	f_i	N_i	F_i
18	2	0,08	2	0,08
19	2	0,08	4	0,16
20	1	0,04	5	0,20
22	1	0,04	6	0,24
23	1	0,04	7	0,28
24	1	0,04	8	0,32
25	2	0,08	10	0,40
26	1	0,04	11	0,44
30	1	0,04	12	0,48
35	1	0,04	13	0,52
41	1	0,04	14	0,56
42	3	0,12	17	0,68
45	1	0,04	18	0,72
47	1	0,04	19	0,76
55	1	0,04	20	0,80
62	2	0,08	22	0,88
64	1	0,04	23	0,92
71	1	0,04	24	0,96
75	1	0,04	25	1,00
	25	1,00		

✓ Moda: $m_o = 42$

✓ Mediana: $m_e = 35$

✓ Media: $\bar{x} = \frac{18 + 18 + \dots + 71 + 75}{25} = 38,08$

$$\bar{x} = \frac{1}{25}(18 \times 2 + 19 \times 2 + \dots + 42 \times 3 + \dots + 75 \times 1) = 38,08$$

✓ Cuartiles: $C_1 = \frac{22 + 23}{2} = 22,5$

$$C_3 = \frac{47 + 55}{2} = 51$$

✓ Rango:

$$R = 75 - 18 = 57$$

✓ Coeficiente de Variación:

$$CV = \frac{17,708}{38,08} = 0,4650 \Rightarrow 46,5\%$$

✓ Varianza, Desviación $S^2 = 313,593$ $S = 17,708$

Típica, Cuasi-Varianza y $\bar{S}^2 = 326,66$ $\bar{S} = 18,073$
Cuasi-Desviación Típica:

Medidas de dispersión (Datos sin agrupar)

Momentos respecto del origen de orden r	$a_r = \frac{1}{n} \sum_{i=1}^n x_i^r$
Momentos respecto de la media de orden r	$m_r = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^r$

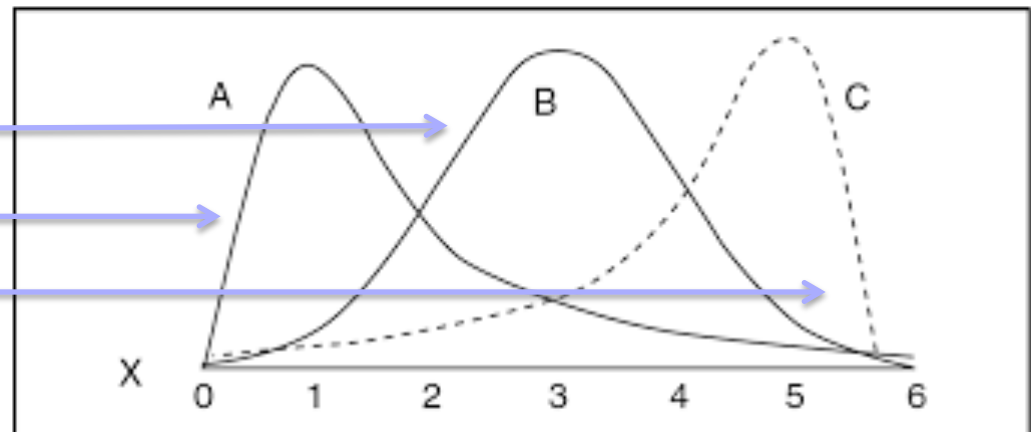
Medidas de forma (Datos sin agrupar)

Coeficiente de asimetría (φ_1)	$\varphi_1 = \frac{m_3}{S^3}$ $\varphi_1 = 0 \rightarrow$ Distribución simétrica $\varphi_1 < 0 \rightarrow$ D. Asimétrica a la izquierda $\varphi_1 > 0 \rightarrow$ D.A. a la derecha
Coeficiente de curtosis (φ_2)	$\varphi_2 = \frac{m_4}{S^4} - 3$ $\varphi_2 = 0 \rightarrow$ Mesocurtica $\varphi_2 > 0 \rightarrow$ Leptocurtica $\varphi_2 < 0 \rightarrow$ Platicurtica
Datos no normales (outliers)	$Z_i = \frac{X_i - \bar{X}}{S}$ Si $Z_i \in [-2, 2]$ datos normales

Distribución simétrica

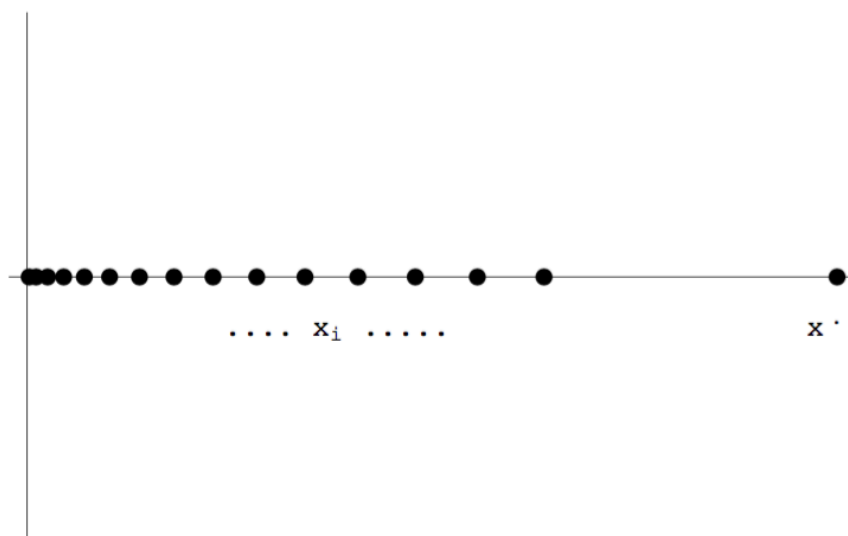
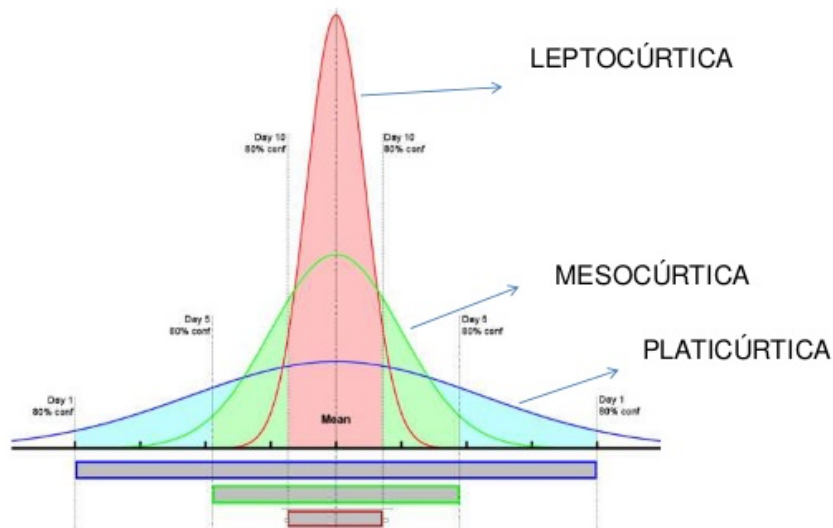
Distribución asimétrica a la derecha

Distribución asimétrica a la izquierda



Medidas de forma (Datos sin agrupar)

Coeficiente de asimetría (φ_1)	$\varphi_1 = \frac{m_3}{S^3}$ $\varphi_1 = 0 \rightarrow$ Distribución simétrica $\varphi_1 < 0 \rightarrow$ D. Asimétrica a la izquierda $\varphi_1 > 0 \rightarrow$ D.A. a la derecha
Coeficiente de curtosis (φ_2)	$\varphi_2 = \frac{m_4}{S^4} - 3$ $\varphi_2 = 0 \rightarrow$ Mesocurtica $\varphi_2 > 0 \rightarrow$ Leptocurtica $\varphi_2 < 0 \rightarrow$ Platicurtica
Datos no normales (outliers)	$Z_i = \frac{X_i - \bar{X}}{S}$ Si $Z_i \in [-2, 2]$ datos normales



Supongamos que se ha realizado un estudio de población en el que sólo se mide la edad de cada individuo. Los resultados, tras analizar una muestra de 25 individuos son:

18 20 30 42 19 55 62 18 41 42 22 35 62
71 47 64 25 75 19 26 42 45 23 24 25

se ordenan los datos originales:

18 18 19 19 20 22 23 24 25 25 26 30 35
41 42 42 42 45 47 55 62 62 64 71 75

Edad	n_i	f_i	N_i	F_i
18	2	0,08	2	0,08
19	2	0,08	4	0,16
20	1	0,04	5	0,20
22	1	0,04	6	0,24
23	1	0,04	7	0,28
24	1	0,04	8	0,32
25	2	0,08	10	0,40
26	1	0,04	11	0,44
30	1	0,04	12	0,48
35	1	0,04	13	0,52
41	1	0,04	14	0,56
42	3	0,12	17	0,68
45	1	0,04	18	0,72
47	1	0,04	19	0,76
55	1	0,04	20	0,80
62	2	0,08	22	0,88
64	1	0,04	23	0,92
71	1	0,04	24	0,96
75	1	0,04	25	1,00
	25	1,00		

✓ Moda: $m_o = 42$

✓ Mediana: $m_e = 35$

✓ Media: $\bar{x} = \frac{18 + 18 + \dots + 71 + 75}{25} = 38,08$

$$\bar{x} = \frac{1}{25}(18 \times 2 + 19 \times 2 + \dots + 42 \times 3 + \dots + 75 \times 1) = 38,08$$

✓ Cuartiles: $C_1 = \frac{22 + 23}{2} = 22,5$

$$C_3 = \frac{47 + 55}{2} = 51$$

✓ Rango:

$$R = 75 - 18 = 57$$

✓ Coeficiente de Variación:

$$CV = \frac{17,708}{38,08} = 0,4650 \Rightarrow 46,5\%$$

✓ Varianza, Desviación $S^2 = 313,593$ $S = 17,708$

Típica, Cuasi-Varianza y $\bar{S}^2 = 326,66$ $\bar{S} = 18,073$
Cuasi-Desviación Típica:

✓ Medidas de Forma: $\varphi_1 = 0,199$; $\varphi_2 = -0,986$

Medidas de centralización (Datos agrupados)

Moda (m_o) (Intervalo modal)	Intervalo de mayor frecuencia
Mediana (m_e)	$m_e = L_i + \frac{\frac{n}{2} - N_{i-1}}{n_i} a_i$ <p> L_i = Limite inferior del intervalo mediano. N_{i-2} = Frecuencia absoluta acumulada del intervalo anterior al mediano. n = Número total de datos. n_i = Frecuencia absoluta del intervalo mediano. a_i = amplitud del intervalo mediano. </p>
Media (\bar{x})	$\bar{X} = \frac{1}{n} \sum_{i=1}^k n_i x_i$
Localización: Cuartiles, deciles, percentiles.	$C_k = L_i + \frac{\frac{n}{4}k - N_{i-1}}{n_i} a_i$

Medidas de dispersión (Datos agrupados)

Varianza (S^2)	$S^2 = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2$
Desviación típica (S)	$S = \sqrt{S^2}$
Cuasi-varianza (\bar{S}^2)	$\bar{S}^2 = \frac{1}{n-1} \sum_{i=1}^k n_i (x_i - \bar{x})^2$
Cuasi-desviación típica (\bar{S})	$\bar{S} = \sqrt{\bar{S}^2}$
Momentos respecto al origen de orden r	$a_r = \frac{1}{n} \sum_{i=1}^k n_i x_i^r$
Momentos respecto a la media de orden r	$m_r = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^r$

Supongamos que se ha realizado un estudio de población en el que sólo se mide la edad de cada individuo. Los resultados, tras analizar una muestra de 25 individuos son:

18 20 30 42 19 55 62 18 41 42 22 35 62
71 47 64 25 75 19 26 42 45 23 24 25

Datos agrupados.

Edad (años)	n_i	x_i^*
De 18 a 20	5	19
De 21 a 25	5	23
De 26 a 35	3	30,5
De 36 a 45	5	40,5
De 46 a 60	2	53
De 61 a 75	5	68
$n = 25$		

Tabla de frecuencias.

Edad (años)	x_i^*	n_i	f_i	N_i	F_i
De 18 a 20	19	5	0,2	5	0,2
De 21 a 25	23	5	0,2	10	0,4
De 26 a 35	30,5	3	0,12	13	0,52
De 36 a 45	40,5	5	0,2	18	0,72
De 46 a 60	53	2	0,08	20	0,8
De 61 a 75	68	5	0,2	25	1
		$n = 25$	1		

Edad	n_i	f_i	N_i	F_i
18	2	0,08	2	0,08
19	2	0,08	4	0,16
20	1	0,04	5	0,20
22	1	0,04	6	0,24
23	1	0,04	7	0,28
24	1	0,04	8	0,32
25	2	0,08	10	0,40
26	1	0,04	11	0,44
30	1	0,04	12	0,48
35	1	0,04	13	0,52
41	1	0,04	14	0,56
42	3	0,12	17	0,68
45	1	0,04	18	0,72
47	1	0,04	19	0,76
55	1	0,04	20	0,80
62	2	0,08	22	0,88
64	1	0,04	23	0,92
71	1	0,04	24	0,96
75	1	0,04	25	1,00
	25	1,00		

considerando los datos originales,

se ordenan los datos originales:

18 18 19 19 20 22 23 24 25 25 26 30 35
41 42 42 42 45 47 55 62 62 64 71 75

$$m_e = 35$$

$$\bar{x} = \frac{18 + 18 + \dots + 71 + 75}{25} = 38,08$$

$$\bar{x}_{20\%} = \frac{22 + 25 + \dots + 55}{15} = 34,93$$

$$C_1 = \frac{22 + 23}{2} = 22,5$$

$$C_3 = \frac{47 + 55}{2} = 51$$

$$S^2 = 313,593 \quad S = 17,708$$

$$\bar{S}^2 = 326,66 \quad \bar{S} = 18,073$$

$$\varphi_1 = 0,199 \quad ; \quad \varphi_2 = -0,986$$

considerando los datos agrupados

intervalo mediano es el tercero

$$m_e = 26 + 9 \frac{12,5 - 10}{3} = 33,4\hat{9}$$

$$\bar{x} = \frac{19(5) + 23(5) + \dots + 68(5)}{25} = 38$$

$$C_1 = 21 + (4) \frac{6,25 - 5}{5} = 22$$

$$C_3 = 46 + (14) \frac{18,75 - 18}{2} = 51,25$$

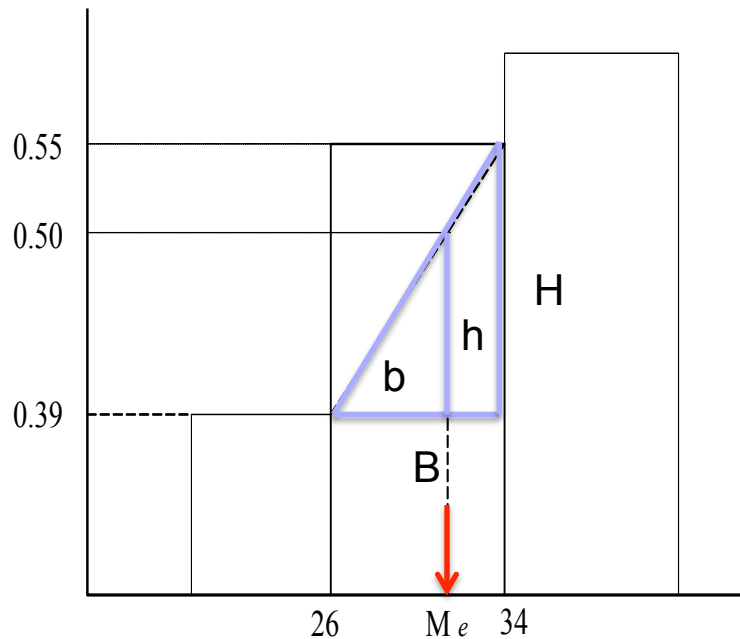
$$D_m = 13,2 \quad S^2 = 323,2$$

$$\bar{S}^2 = 336,66 \quad \bar{S} = 18,348$$

$$S = 17,977 \quad e.s. = 4,80$$

$$C_v = 0,473 \quad R_i = 51 - 22,5 = 28,5$$

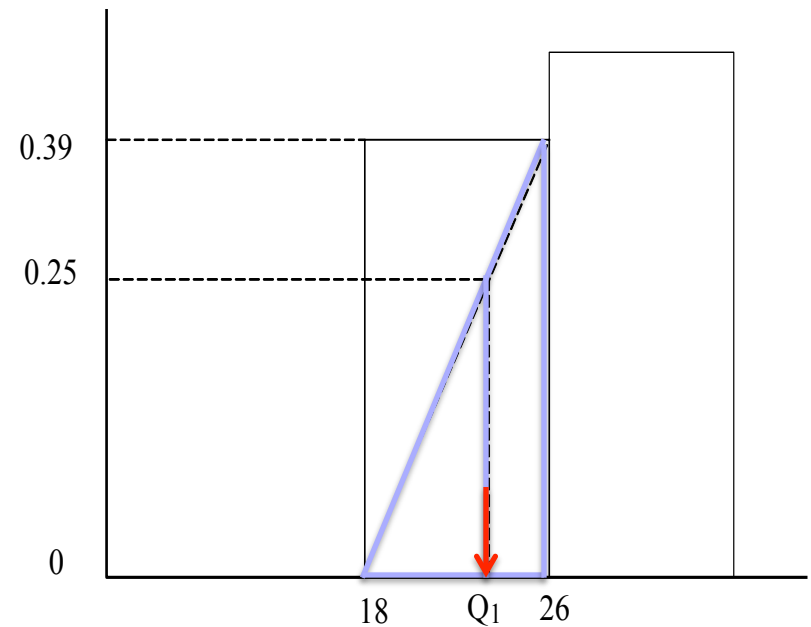
Cálculo de cuantiles (Datos agrupados en intervalos)



$$\frac{b}{B} = \frac{h}{H}$$

$$\frac{M_e - 26}{34 - 26} = \frac{0.50 - 0.39}{0.55 - 0.39}$$

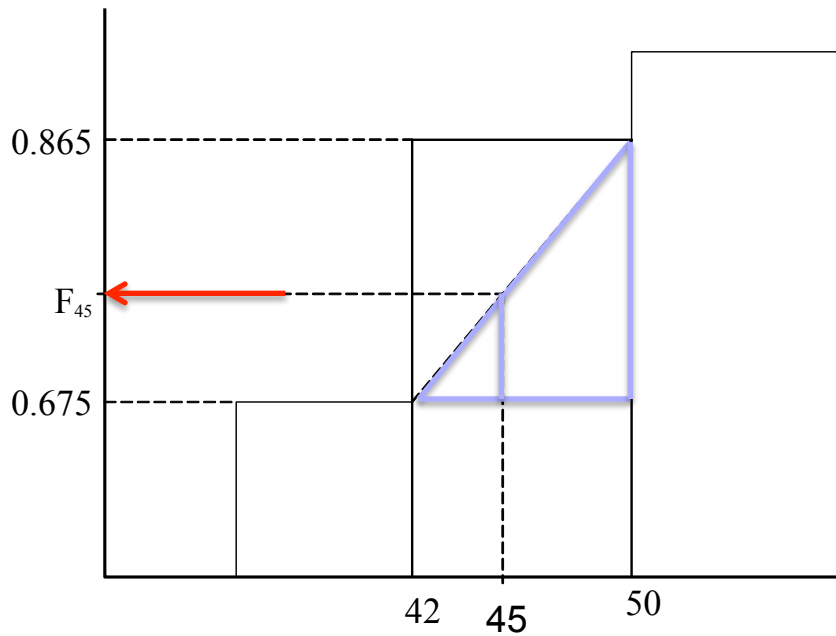
$$M_e = 26 + \frac{0.50 - 0.39}{0.55 - 0.39} (34 - 26) = 31.5$$



$$\frac{Q_1 - 18}{26 - 18} = \frac{0.25 - 0}{0.39 - 0}$$

$$Q_1 = 18 + \frac{0.25}{0.39} (26 - 18) = 23.13$$

Cálculo de cuantiles (problema inverso) (Datos agrupados en intervalos)



Porcentaje de individuos que tienen una edad inferior a 45 años

$$\frac{45 - 42}{50 - 42} = \frac{F_{x=45} - 0.675}{0.865 - 0.675}$$

$$F_{x=45} = 0.675 + \frac{45 - 42}{50 - 42} (0.865 - 0.675) = 0.746$$

Análisis de datos Univariante

ESCUELA POLITÉCNICA
SUPERIOR DE CÓRDOBA

Universidad de Córdoba

DEPARTAMENTO DE ESTADÍSTICA

