Resultado de aprendizaje

- *Determina e interpreta las medidas de descriptivas de un conjunto de datos*

Contenidos de la presentación

- *Herramientas de estadística descriptiva*
  - *Estadígrafos de dispersión*
    - *Rango*
    - *Rango intercuartil*
    - *Varianza*
    - *Desviación estándar*
    - *Coeficiente de variación*

FORMULARIO RESUMEN DE ESTADÍSTICA DESCRIPTIVA: Análisis exploratorio de datos (AED)

I. Tablas de frecuencias: 1) Variables cuantitativas, a) discretas b) continuas

N° Clase	***Variable o Clase***	***Frecuencia Absoluta***	***Frecuencia Relativa***	***Frecuencia Absoluta Acumulada***	***Frecuencia Relativa Acumulada***
	Valor	*FA*	*FR* ($h_i = \frac{n_i}{N}$)	*FAA*	*FRA* ($H_i = \sum_{i=1}^{i} h_i$)
Enumeración	x_i	n_i	h_i	N_i	H_i
1	x_1	n_1	$h_1 = n_1/N$	$N_1 = n_1$	$H_1 = h_1$
2	x_2	n_2	$h_2 = n_2/N$	$N_2 = n_1 + n_2$	$H_2 = h_1 + h_2$
...	...	...	...	...	...
k	x_k	n_k	$h_k = n_k/N$	$N_n = \sum_{i=1}^{n} f_i = N$	$H_n = \sum_{i=1}^{n} h_i = 1$
	Total	N	1		

N° Clase	***Intervalo***		***Centro o Marco de Clase***	***Frecuencia Absoluta***	***Frecuencia Relativa***	***Frecuencia Absoluta Acumulada***	***Frecuencia Relativa Acumulada***
	Límite inferior	*Límite superior*	*MC*	*FA*	*FR*	*FAA*	*FRA*
	L_i	L_s	x_i	n_i	h_i	N_i	H_i
1	$[L_0, L_1\rangle$		$x_1 = \frac{L_0 + L_1}{2}$	n_1	$h_1 = n_1/N$	$N_1 = n_1$	$H_1 = h_1$
2	$[L_1, L_2\rangle$		$x_2 = \frac{L_1 + L_2}{2}$	n_2	$h_2 = n_2/N$	$N_2 = n_1 + n_2$	$H_2 = h_1 + h_2$
...	...		...	...	...	...	...
k	$[L_{n-1}, L_n]$		$x_n = \frac{L_{n-1} + L_n}{2}$	n_k	$h_k = n_k/N$	$N_n = \sum_{i=1}^{n} f_i = N$	$H_n = \sum_{i=1}^{n} h_i = 1$
	Total			N	1		

2) Variables cualitativas nominales y ordinales en tablas de frecuencias; por convención no presentan frecuencias acumuladas. Para las variables ordinales, las clases se estructuran de acuerdo con el orden de la variable.

II. Estadígrafos o medidas de resumen

Indicadores estadísticos o estadígrafos			
Posición		Dispersión	Forma
Tendencia central	Tendencia no central		
Media ($\bar{x}$) Mediana (me) Moda (mo)	Cuantiles o Fractiles: cuartiles (Q), quintiles (K), deciles (D), percentiles (P)	Rango Rango intercuartil (RIC)	Varianza (S^2) Desviación estándar (S) Coeficiente de variación (CV)	Asimetría Curtosis

1. Datos seriados (sin estructura de agrupamiento en tablas de frecuencias)

a) Medidas de posición de tendencia central

Media	Mediana (datos en orden ascendente)		Moda
	Datos n impar	Datos n par	
$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$	$\tilde{x}_{\frac{n+1}{2}}$	$\frac{\tilde{x}_{\frac{n}{2}} + \tilde{x}_{\frac{n}{2}+1}}{2}$	Dato que más se repite

b) Medidas de posición de tendencia no central: Cuantiles o fractiles

Cuanti*	Cuartil (Q) $i = 1,2,3$	Quintil (K) $i = 1,2,3,4$	Decil (D) $i = 1,2,3 \ldots 9$	Percentil (P) $i = 1,2,3 \ldots 99$
Posición n par	$P_{Q_i} = i \cdot \frac{n}{4}$	$P_{K_i} = i \cdot \frac{n}{5}$	$P_{D_i} = i \cdot \frac{n}{10}$	$P_{P_i} = i \cdot \frac{n}{100}$
Posición n impar	$P_{Q_i} = i \cdot \frac{n+1}{4}$	$P_{K_i} = i \cdot \frac{n+1}{5}$	$P_{D_i} = i \cdot \frac{n+1}{10}$	$P_{P_i} = i \cdot \frac{n+1}{100}$

- El valor del cuantil se obtiene desde la posición correspondiente, previo ordenamiento ascendente de los datos.
  a) Si el resultado es un número entero, se toma la posición exacta.
  b) En caso de un resultado con un número decimal, el valor del cuantil buscado será la media proporcional entre x_i y x_{i+1}. Sea un número de la forma 'e.d' donde 'e' es la parte entera y 'd' la parte decimal. Así el cuantil buscado, usando como ejemplo el cuartil $_{Q_i}$ será: $Q_i = x_i + d \cdot (x_{i+1} - x_i)$
- Fórmulas también aplican para datos discretos agrupados en tablas de frecuencias, en este caso el valor del cuantil corresponde a la clase, de la posición leída desde la frecuencia absoluta acumulada (este número acumulado debe ser igual o superior a la posición obtenida en el calculo).

c) Medidas de dispersión

Rango	Varianza muestral*	Desviación estándar	Coeficiente de variación
$R = X_{max} - X_{min}$	$S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{S^2}$	$CV(\%) = (S/\bar{x}) \cdot 100$

*Procedimiento de cálculo de la Varianza: Se calcula la diferencia de cada valor observado y la media $(x_i - \bar{x})$, cada diferencia se eleva al cuadrado, $(x_i - \bar{x})^2$ luego se suman $\sum_{i=1}^{n}(x_i - \bar{x})^2$ y se dividen por (n -1).

2. Datos agrupados en tablas de frecuencias

a) Media $\bar{x} = \frac{\sum(x_i \cdot n_i)}{n}$; Varianza $S^2 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 \cdot n_i]}{n-1}$; Desviación estándar $S = \sqrt{S^2}$

Clase - Marca de Clase	*Frecuencia Absoluta*	*media*	*varianza*
X_0	n_0	$x_0 \cdot n_0$	$(x_0 - \bar{x})^2 \cdot n_0$
X_1	n_1	$x_1 \cdot n_1$	$(x_1 - \bar{x})^2 \cdot n_1$
X_2	n_2	$x_2 \cdot n_2$	$(x_2 - \bar{x})^2 \cdot n_2$
...	...	...	...
X_i	n_i	$x_i \cdot n_i$	$(x_i - \bar{x})^2 \cdot n_i$
	n	$\sum(x_i \cdot n_i)$	$\sum_{i=1}^{n}[(x_i - \bar{x})^2 \cdot n_i]$
		$\bar{x} = \frac{\sum(x_i \cdot n_i)}{n}$	$S^2 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})^2 \cdot n_i]}{n-1}$

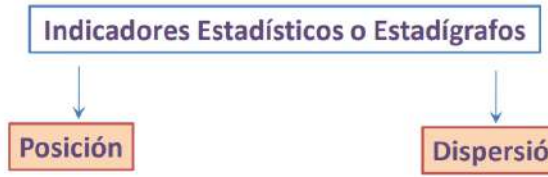
b) Cuantiles para datos agrupados en intervalos (tablas de frecuencias)

Cuartil (Q)	Quintil (K)	Decil (D)	Percentil (P)
$Q_i = L_i + \frac{i\left(\frac{n}{4}\right) - F_{(i-1)}}{f_i} \cdot a$	$K_i = L_i + \frac{i\left(\frac{n}{5}\right) - F_{(i-1)}}{f_i} \cdot a$	$D_i = L_i + \frac{i\left(\frac{n}{10}\right) - F_{(i-1)}}{f_i} \cdot a$	$P_i = L_i + \frac{i\left(\frac{n}{100}\right) - F_{(i-1)}}{f_i} \cdot a$
$i = 1,2,3$	$i = 1,2,3,4$	$i = 1,2,3,4,5,6,7,8,9$	$i = 1,2,3 \ldots 99$

Donde, además:

L_i límite inferior del intervalo que contiene al cuantil
n tamaño de la muestra
$F_{(i-1)}$ Frec. acumulada del intervalo anterior que contiene el cuantil
f_i Frecuencia absoluta del intervalo que contiene al cuantil
a = Amplitud del intervalo

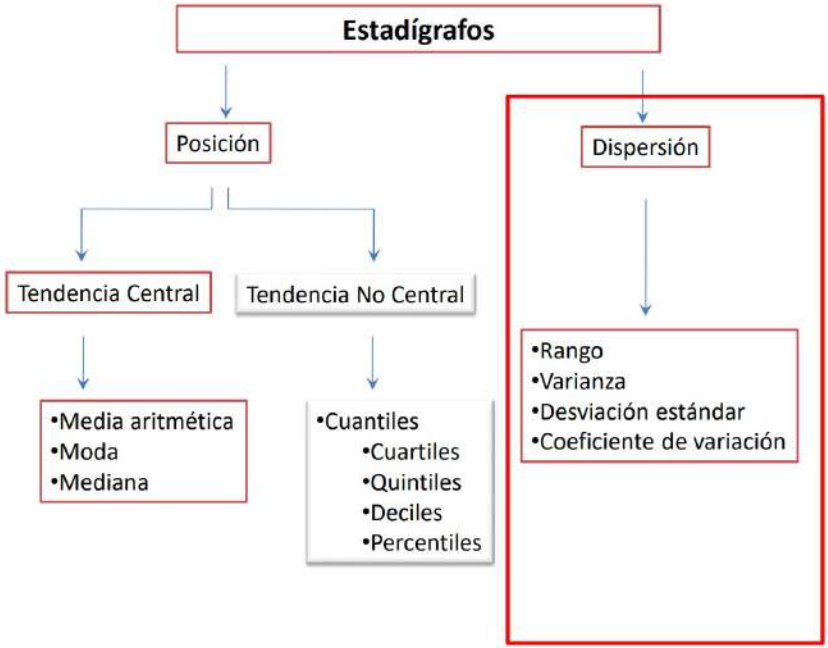
Estadística descriptiva: Estadígrafos de dispersión



- Son medidas que pretenden resumir un conjunto de valores
- Representan un centro en torno al cual se encuentra ubicado el conjunto de datos

- Tiene por finalidad cuantificar la variabilidad de los datos

  Que tan separados o disimiles son uno de otro
- Medida del "Grado de concentración o densidad" de los datos en torno a su centro de gravedad


Estadígrafos
Posición
Tendencia Central
Tendencia No Central
•Media aritmética
•Moda
•Mediana
•Cuantiles
•Cuartiles
•Quintiles
•Deciles
•Percentiles
Dispersión
•Rango
•Varianza
•Desviación estándar
•Coeficiente de variación

MEDIDAS DE DISPERSIÓN

Medida del "Grado de concentración o densidad" de los datos en torno a su centro de gravedad

- *Rango*
- *Rango intercuartil (RIC)*
- *Varianza*
- *Desvío estándar o desviación estándar*
- *Coeficiente de variación*

Rango

Es la diferencia entre el valor máximo y el valor mínimo de una serie de observaciones

$$Rango = valor\ máximo - valor\ mínimo$$

- Es fácil de comprender y obtener pero tiene limitaciones en la medición de la variabilidad:
  - Un valor alto o bajo determinaría una gran amplitud que no reflejaría la verdadera variabilidad de los datos
  - Posible solución ¿eliminar estos valores? ...este criterio es de difícil formulación e interpretación subjetiva

Rango

Presión Sanguínea

- Las presiones sistólicas (mm Hg) de seis hombres de mediana edad:

113 124 124 132 146 151 170

El rango muestral es 170 -113 = 57 mm Hg

Rango intercuartil (*RIC - RQ)*

Rango intercuartílico (IQR interquartile range)

Se refiere a la diferencia entre el tercer y el primer cuartil de una distribución

$$\mathbf{RQ = Q_3 - Q_1}$$

- *A la mitad del rango intercuartil se le conoce como desviación cuartil (DQ)*

$$DQ = RQ/2 = (Q3 - Q1)/2$$

  - *Es afectada muy poco por datos extremas*
  - *Esto lo hace una buena medida de dispersión para distribuciones sesgadas*

Se usa para construir los diagramas de caja y bigote (boxplot)

Varianza

Es la media de las diferencias con la media elevadas al cuadrado

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$$

Varianza poblacional

$$S^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Varianza muestral

Otra definición: Media de los cuadrados de las desviaciones de los datos

- Como la varianza tiene las unidades de medidas elevadas al cuadrado, estas unidades **no** son intuitivamente claras o fáciles de interpretar

Desviación estándar (D.E.)

Mide el grado de dispersión de los valores de la variable respecto a la media aritmética

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}} = \sqrt{\sigma^2}$$

Desviación estándar poblacional

$$S = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{S^2}$$

Desviación estándar muestral

Crecimiento de Crisantemos

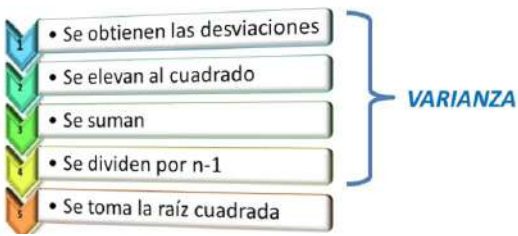
En un experimento sobre crisantemos, un botánico midió el tallo (mm):

$$76\ 72\ 65\ 70\ 82 \rightarrow \quad \overline{x} = \frac{365}{5} = 73\ mm$$

- Para obtener S^2 y S:

Observaciones: 76 72 65 70 82 $n = 5$

Observación (x_i)	
x_1	76
x_2	72
x_3	65
x_4	70
x_5	70
x_6	82



Observación (x_i)	
x_1	76
x_2	72
x_3	65
x_4	70
x_5	82
$\sum_{i=1}^{n} x_i = 365$	
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$	

1

Observación (x_i)		Desviación ($x_i - \overline{x}$)
x_1	76	
x_2	72	
x_3	65	
x_4	70	
x_5	82	
$\sum_{i=1}^{n} x_i = 365$		
$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$		

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

2

Observación (x_i)	Desviación ($x_i - \overline{x}$)	Desviación al cuadrado $(x_i - \overline{x})^2$
76	*76 -73= 3*	*9*
72	*72 -73= -1*	*1*
65	*65 -73= -8*	*64*
70	*70 -73= -3*	*9*
82	*82 -73= 9*	*81*

$$\sum_{i=1}^{n} x_i = 365$$

$$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$$

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

Observación (x_i)	Desviación ($x_i - \overline{x}$)	Desviación al cuadrado $(x_i - \overline{x})^2$
76	76 -73= 3	9
72	72 -73= -1	1
65	65 -73= -8	64
70	70 -73= -3	9
82	82 -73= 9	81
$\sum_{i=1}^{n} x_i = 365$	**3**	$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 164$
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$		

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

Observación (x_i)	Desviación $(x_i - \overline{x})$	Desviación al cuadrado $(x_i - \overline{x})^2$
76	76 -73= 3	9
72	72 -73= -1	1
65	65 -73= -8	64
70	70 -73= -3	9
82	82 -73= 9	81
$\sum_{i=1}^{n} x_i = 365$		$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 164$
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$		$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{164}{4} = 41$

4

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

Observación (x_i)	Desviación ($x_i - \overline{x}$)	Desviación al cuadrado $(x_i - \overline{x})^2$
76	76 -73= 3	9
72	72 -73= -1	1
65	65 -73= -8	64
70	70 -73= -3	9
82	82 -73= 9	81
$\sum_{i=1}^{n} x_i = 365$		$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 164$
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$		$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{164}{4} = 41$
		$S = \sqrt{S^2} = \sqrt{41} = 6{,}4$

5

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

Observación (x_i)	Desviación ($x_i - \overline{x}$)	Desviación al cuadrado $(x_i - \overline{x})^2$
76	76 -73= 3	9
72	72 -73= -1	1
65	65 -73= -8	64
70	70 -73= -3	9
82	82 -73= 9	81
$\sum_{i=1}^{n} x_i = 365$		$\sum_{i=1}^{n} (x_i - \bar{x})^2 = 164$
$\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{365}{5} = 73$		$S^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = \frac{164}{4} = 41$
		$S = \sqrt{S^2} = \sqrt{41} = 6{,}4$

1. Se obtienen las desviaciones
2. Se elevan al cuadrado
3. Se suman
4. Se dividen por n-1
5. Se toma la raíz cuadrada

VARIANZA

$$\textit{Datos: } 76 \quad 72 \quad 65 \quad 70 \quad 82$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{76 + 72 + 65 + 70 + 82}{5} = \frac{365}{5} = 73\ mm$$

$$S^2 = \frac{\sum_{1=1}^{n} (x_i - \bar{x})^2}{n - 1}$$

$$S^2 = \frac{(76 - 73)^2}{4} + \frac{(72 - 73)^2}{4} + \frac{(65 - 73)^2}{4} + \frac{(70 - 73)^2}{4} + \frac{(82 - 73)^2}{4} = \frac{164}{4} = 41\ mm$$

$$S = \sqrt{\frac{\sum_{1=1}^{n} (x_i - \bar{x})^2}{n - 1}} = \sqrt{S^2} \quad \rightarrow \quad S = \sqrt{41} = 6,4\ mm$$

Datos: 76 72 65 70 82

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{76 + 72 + 65 + 70 + 82}{5} = \frac{365}{5} = 73\ mm$$

$$S^2 = \frac{1}{n-1} \sum_{1=1}^{n} (x_i - \bar{x})^2$$

$$S^2 = \frac{1}{5-1} \cdot \left[(76-73)^2 + (72-73)^2 + (65-73)^2 + (70-73)^2 + (82-73)^2\right]$$

$$S^2 = \frac{1}{5-1} \cdot 164 = \frac{164}{4} = 41\ mm$$

$$S = \sqrt{\frac{\sum_{1=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{S^2} \quad \rightarrow \quad S = \sqrt{41} = 6,4\ mm$$

Datos: 76 72 65 70 82

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{76 + 72 + 65 + 70 + 82}{5} = \frac{365}{5} = 73\ mm$$

$$S^2 = \boxed{\frac{1}{n-1}} \sum_{1=1}^{n} (x_i - \bar{x})^2$$

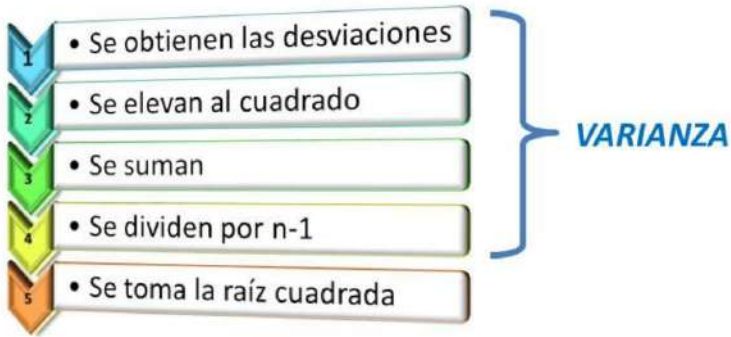
$$S^2 = \frac{1}{5-1} \cdot \left[(76-73)^2 + (72-73)^2 + (65-73)^2 + (70-73)^2 + (82-73)^2\right]$$

$$S^2 = \frac{1}{5-1} \cdot 164 = \frac{164}{4} = 41\ mm$$

$$S = \sqrt{\frac{\sum_{1=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{S^2} \quad \rightarrow \quad S = \sqrt{41} = 6,4\ mm$$

Para 4, 8, 10, 11,17

Calcular: Media, Rango, Varianza y Desviación Estándar (dejar expresado)



Considerando las cinco observaciones: 4-8-10-11-17

Xi	$Xi - \bar{X}$	$(Xi - \bar{X})^2$
4	-6	36
8	-2	4
10	0	0
11	1	1
17	7	49
$\sum_{i=1}^{5} Xi = 50$	$\sum_{i=1}^{5} (Xi - \bar{X}) = 0$	$\sum_{i=1}^{5} (Xi - \bar{X})^2 = 90$

$$\bar{X} = \frac{\sum_{i=1}^{n} Xi}{n} = \frac{\sum_{i=1}^{5} Xi}{5} = \frac{50}{5} = 10$$

$$S = \sqrt{\frac{\sum_{i=1}^{n} (Xi - \bar{X})^2}{n-1}} = \sqrt{\frac{90}{4}} = \sqrt{22.5} = 4.74$$

$$S = \sqrt{varianza} \rightarrow S = \sqrt{S^2}$$

Crecimiento de Crisantemos

La varianza muestral S^2 es la D. E. al cuadrado.

Por lo tanto para obtener la desviación estándar S, se aplica raíz cuadrada

$$\boldsymbol{S = \sqrt{varianza}}$$

- La varianza de los datos de crecimiento de crisantemos es $s^2 = 41\ mm^2$
- La desviación estándar es $S = 6{,}4\ mm$

Varianza y desviación estándar

Existe relación: ambas miden la dispersión de los valores observados con respecto a la media

- Diferencia: la varianza está dada en unidades al cuadrado, la D.E. tiene la misma unidad de medida que la media

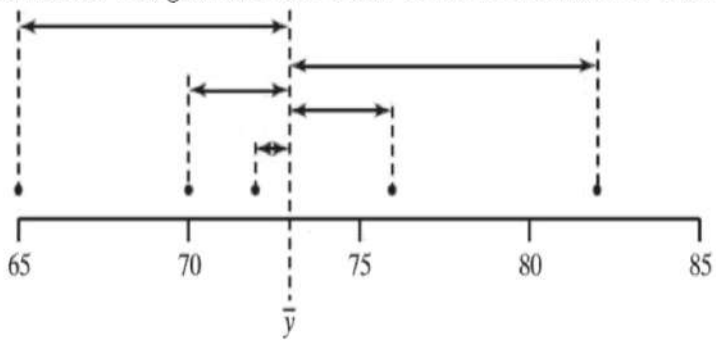
- Para la presentación de resultados se prefiere D.E. : $\mathbf{73 \pm 6,4\ mm}$

Interpretación de la Desviación Estándar (D.E.)

Se puede interpretar como la distancia promedio de las observaciones respecto a la media muestral:

"En promedio cuanto se alejan los datos de la media"

La Figura muestra una gráfica de los datos donde se ha marcado cada distancia



Propiedades Varianza y desviación estándar

1. *Ambas medidas son siempre un número no negativo.*
   - *La $\boldsymbol{\sigma}$ y $\boldsymbol{\sigma^2}$ son cero sólo cuando todos los datos son iguales*
2. *Si cada dato de una muestra se aumenta o se disminuye en una constante **K** la desviación estándar y la varianza originales **no cambian**.*
3. *Si cada dato de una muestra se multiplica por una constante K, entonces las nuevas σ y σ^2 son respectivamente **K· σ y K· σ^2***

Coeficiente de variación (C.V.)

Cuando se desea hacer referencia a la relación entre el tamaño de la media y la variabilidad de las observaciones, se usa el coeficiente de variación muestral CV

Coeficiente de variación (C.V.)

Definición: Dada una muestra $x_1, x_2,\ldots,x_n$ con media $\overline{x}$ y desviación estándar S, el coeficiente de variación muestral se define como:

$$CV = S/\overline{x} \cdot 100$$

Esta medida es adimensional y permite comparar la variabilidad de características medidas en diferentes escalas

Coeficiente de variación (C.V.)

Establecer la homogeneidad o heterogeneidad de los datos mediante la D.E. requiere conocimiento y principalmente experiencia del fenómeno en estudio para una correcta interpretación

El CV es una medida útil porque mide la dispersión en forma relativa que permite una interpretación más sencilla de la variabilidad

Ejemplo: Calcular el Coeficiente de variación y concluir

En dos poblaciones A y B los pesos promedios recién nacidos y su correspondiente desviación estándar son 2515 ± 40 g para la población A y 2630 ± 380 g para la población B.

¿En cuál población los pesos al nacer son más homogéneos?

$$CV = \frac{S}{\bar{x}} * 100$$

Población A: 2515 ± 40 g ➔ 40/2515*100% = 1,6%

Población B: 2630 ± 380 g ➔ 380/2630*100% = 14,4%

La población más homogénea es la A, dado su menor CV

Marca de Clase	Frecuencia Absoluta	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
MC	FA	FR	FAA	FRA
X_i	n_i	h_i	N_i	H_i
X_1	n_1	$h_1=n_1/N$	$N_1 = n_1$	$H_1 = h_1$
X_2	n_2	$h_2=n_2/N$	$N_2 = n_1 + n_2$	$H_2 = h_1 + h_2$
...	...	...		
X_i	n_i	$H_n=n_n/N$	$F_n = \sum_{i=1}^{n} f_i = N$	$H_n = \sum_{i=1}^{n} h_i = 1$
Total	N	1		

ESTADÍGRAFOS PARA DATOS TABULADOS EN TABLAS DE FRECUENCIAS

Medidas de resumen para datos agrupados

Marca de Clase	Frecuencia Absoluta	$x_i \cdot n_i$	Frecuencia Relativa	Frecuencia Absoluta Acumulada	Frecuencia Relativa Acumulada
MC	FA		FR	FAA	FRA
X_i	n_i		h_i	N_i	H_i
X_1	n_1	$X_1 \cdot n_1$	$h_1=n_1/N$	$N_1 = n_1$	$H_1 = h_1$
X_2	n_2	$X_2 \cdot n_2$	$h_2=n_2/N$	$N_2 = n_1 + n_2$	$H_2 = h_1 + h_2$
...	...		...		
X_i	n_i	$X_i \cdot n_i$	$H_n=n_n/N$	$F_n = \sum_{i=1}^{n} f_i = N$	$H_n = \sum_{i=1}^{n} h_i = 1$
Total	N	$\sum x_i n_i$	1		

Media = Cociente entre la sumatoria del producto de cada frecuencia absoluta por su marca de clase y el número total de datos

$$\overline{x} = \frac{\sum n_i x_i}{n}$$

Moda = Valor de la clase con frecuencia absoluta más alta

Mediana = Valor de la clase donde se iguala o sobrepasa el 50% de las observaciones

Ejemplo: Número de hermanos de estudiantes de una clase

			media
Marca de Clase		***Frecuencia Absoluta***	$x_i \cdot n_i$
x_i		n_i	
x_1	0	3	$0 \cdot 3 = 0$
x_2	1	5	$1 \cdot 5 = 5$
x_3	2	5	$2 \cdot 5 = 10$
x_4	3	4	$3 \cdot 4 = 12$
x_5	4	2	$4 \cdot 2 = 8$
x_6	5	1	$5 \cdot 1 = 5$
		n	$\overline{x} = \frac{\sum x_i n_i}{n}$
		20	$\overline{x} = \frac{40}{20} = 2$

Medidas de resumen para datos agrupados

Marca de Clase	Frecuencia Absoluta	$x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot n_i$
MC	FA				
x_i	n_i				
X_1	n_1	$X_1 \cdot n_1$	$x_1 - \overline{x}$	$(x_1 - \overline{x})^2$	$(x_1 - \overline{x})^2 \cdot n_1$
X_2	n_2	$X_2 \cdot n_2$			
...	...				
X_i	n_i	$X_i \cdot n_i$			
	n	$\overline{x} = \frac{\sum x_i n_i}{n}$			$\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i$

Cálculo de la varianza

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i}{n - 1}$$

Ejemplo: Número de hermanos de estudiantes de una clase

			media
Marca de Clase		***Frecuencia Absoluta***	$x_i \cdot n_i$
x_i		n_i	
x_1	0	3	$0 \cdot 3 = 0$
x_2	1	5	$1 \cdot 5 = 5$
x_3	2	5	$2 \cdot 5 = 10$
x_4	3	4	$3 \cdot 4 = 12$
x_5	4	2	$4 \cdot 2 = 8$
x_6	5	1	$5 \cdot 1 = 5$
		n	$\overline{x} = \frac{\sum x_i n_i}{n}$
		20	$\overline{x} = \frac{40}{20} = 2$

Ejemplo: Número de hermanos de estudiantes de una clase

	Marca de Clase x_i	Frecuencia Absoluta n_i	media $x_i \cdot n_i$	$x_i - \bar{x}$
x_1	0	3	$0 \cdot 3 = 0$	$0 - 2 = -2$
x_2	1	5	$1 \cdot 5 = 5$	$1 - 2 = -1$
x_3	2	5	$2 \cdot 5 = 10$	$2 - 2 = 0$
x_4	3	4	$3 \cdot 4 = 12$	$3 - 2 = 1$
x_5	4	2	$4 \cdot 2 = 8$	$4 - 2 = 2$
x_6	5	1	$5 \cdot 1 = 5$	$5 - 2 = 3$
		n	$\bar{x} = \frac{\sum x_i n_i}{n}$	
		20	$\bar{x} = \frac{40}{20} = 2$	

Ejemplo: Número de hermanos de estudiantes de una clase

Marca de Clase x_i	Frecuencia Absoluta n_i	media $x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$
0	3	$0 \cdot 3 = 0$	$0 - 2 = -2$	4
1	5	$1 \cdot 5 = 5$	$1 - 2 = -1$	1
2	5	$2 \cdot 5 = 10$	$2 - 2 = 0$	0
3	4	$3 \cdot 4 = 12$	$3 - 2 = 1$	1
4	2	$4 \cdot 2 = 8$	$4 - 2 = 2$	4
5	1	$5 \cdot 1 = 5$	$5 - 2 = 3$	9

Ejemplo: Número de hermanos de estudiantes de una clase

		media			**varianza**
Marca de Clase x_i	**Frecuencia Absoluta** n_i	$x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot n_i$
0	3	$0 \cdot 3 = 0$	$0 - 2 = -2$	4	$4 \cdot 3 = 12$
1	5	$1 \cdot 5 = 5$	$1 - 2 = -1$	1	$1 \cdot 5 = 5$
2	5	$2 \cdot 5 = 10$	$2 - 2 = 0$	0	$0 \cdot 5 = 0$
3	4	$3 \cdot 4 = 12$	$3 - 2 = 1$	1	$1 \cdot 4 = 4$
4	2	$4 \cdot 2 = 8$	$4 - 2 = 2$	4	$4 \cdot 2 = 8$
5	1	$5 \cdot 1 = 5$	$5 - 2 = 3$	9	$9 \cdot 1 = 9$
					$\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i$
					38

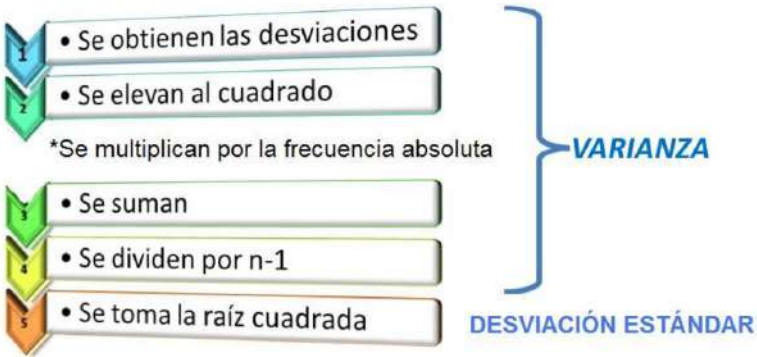
Ejemplo: Número de hermanos de estudiantes de una clase

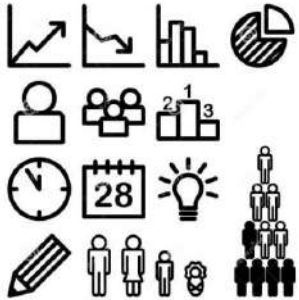
		media			varianza
Marca de Clase x_i	Frecuencia Absoluta n_i	$x_i \cdot n_i$	$x_i - \overline{x}$	$(x_i - \overline{x})^2$	$(x_i - \overline{x})^2 \cdot n_i$
0	3	$0 \cdot 3 = 0$	$0 - 2 = -2$	4	$4 \cdot 3 = 12$
1	5	$1 \cdot 5 = 5$	$1 - 2 = -1$	1	$1 \cdot 5 = 5$
2	5	$2 \cdot 5 = 10$	$2 - 2 = 0$	0	$0 \cdot 5 = 0$
3	4	$3 \cdot 4 = 12$	$3 - 2 = 1$	1	$1 \cdot 4 = 4$
4	2	$4 \cdot 2 = 8$	$4 - 2 = 2$	4	$4 \cdot 2 = 8$
5	1	$5 \cdot 1 = 5$	$5 - 2 = 3$	9	$9 \cdot 1 = 9$
					$\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i$
					38

$$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2 \cdot n_i}{n - 1} = \frac{38}{19} = 2$$

Obtención Varianza y Desviación estándar para datos tabulados o agrupados

Procedimiento para su obtención
Primeramente se debe haber obtenido la media

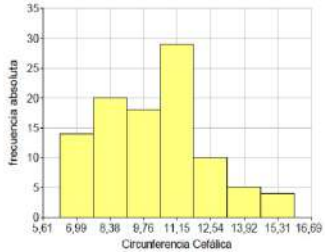




ESTADÍSTICA DESCRIPTIVA

Representaciones gráficas

Gráficos de frecuencias



Los gráficos obtenidos a partir de una tabla de frecuencias permiten visualizar la información contenida en las tablas de manera rápida y sencilla, mostrando con mayor claridad el comportamiento de los datos

Medidas de forma de un histograma

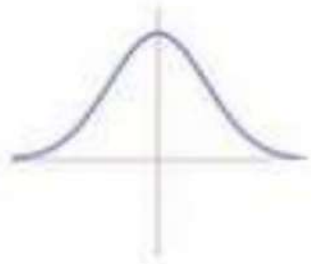
Es la apariencia externa de la distribución de frecuencias y se representa por el aspecto gráfico fundamentalmente

Dentro de la forma se incluye:

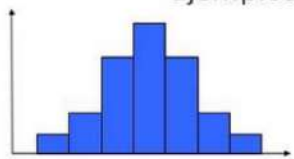
- *la* ***simetría o asimetría*** *de la curva*
- *y el* ***grado de aplanamiento*** *de la curva*

Son medidas relativas: cocientes o razones (no tienen unidad de medida).

*Una **distribución es simétrica** cuando la curva que la representa es exactamente igual a ambos lados del punto de referencia.*



Ejemplos de curvas simétricas



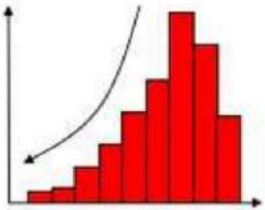
Distribución simétrica y unimodal

Distribución uniforme

Asimetría

Distribución asimétrica negativa

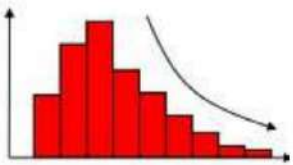
$\bar{x} < Md < Mo$



Asimetría negativa: si los datos se concentran hacia valores altos de la variable (derecha)

Distribución asimétrica positiva

$\bar{x} > Md > Mo$



Asimetría positiva: si los datos se concentran hacia valores bajos de la variable (izquierda)

Coeficiente de Asimetría de Bowley

Se basa en la posición de los cuartiles (Q) y la mediana (Me):

$$A_B = \frac{Q_3 + Q_1 - 2\,Me}{Q_3 - Q_1}$$

- *En una distribución* ***simétrica*** Q_3 *estará a la misma distancia de la mediana que* Q_1*, así* $A_B = 0$
- $A_B > 0$ *: distribución* ***positiva***
- $A_B < 0$ *: distribución* ***negativa***

Coeficiente de Asimetría de Pearson

$$A_P = \frac{\overline{x} - Mo(X)}{S_x}$$

siendo $\overline{x}$ la media, $Mo(X)$ la moda y S_x la desviación típica

- *Si **A_P=0**: la distribución es **simétrica***
- *Si **A_P>0**: distribución **asimetría positiva**; la media es mayor que la moda*
- *Si **A_P<0**: distribución **asimetría negativa**; la media es menor que la moda*

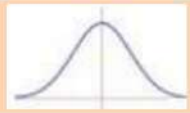
Coeficiente de Asimetría de Fisher

$$A_F = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^3 \cdot n_i}{N \cdot S_x^3}$$

Siendo x_i uno de los datos o,
en datos agrupados en intervalos, la marca de clase,
$\overline{x}$ la media, n_i la frecuencia absoluta de x_i o de cada intervalo i
i S_x la desviación típica

- Si A_F**=0**: la distribución es **simétrica**
- Si A_F**>0**: distribución **positiva**; se 'alarga' a valores mayores que la media
- Si A_F**<0**: distribución **negativa;** se 'alarga' a valores menores que la media

Curtosis



Indica el grado de apuntamiento o achatamiento del gráfico. *Se toma como referencia la curva normal o de campana*

Los indicadores de curtosis, **miden el nivel de concentración de datos en la región central.**

Curtosis

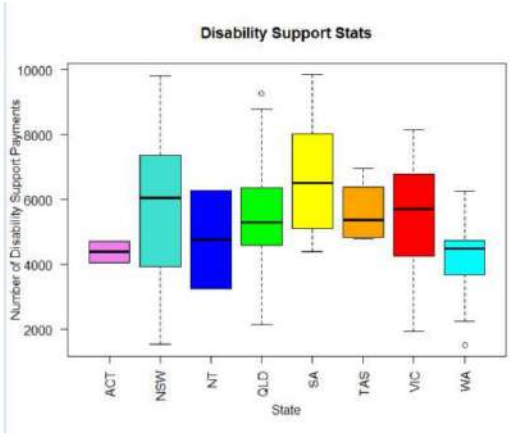
$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4}{N \cdot S_x^4} - 3$$

siendo $\overline{x}$ la media y S_x la desviación típica

- Para datos agrupados

$$Curtosis = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^4 \cdot n_i}{N S_x^4} - 3$$

n_i la frecuencia absoluta de x_i o de cada intervalo i



Representaciones gráficas: Grafico de cajas o Boxplot

Diagrama de cajas o Box-Plot

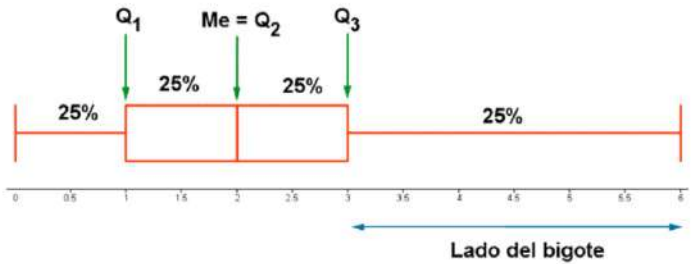
El AED implica el uso de técnicas estadísticas para identificar patrones que pueden estar ocultos en un grupo de números

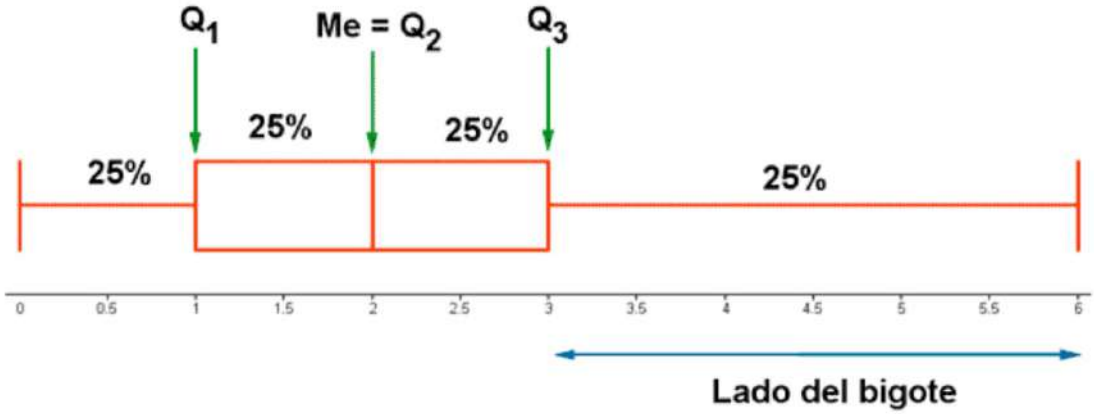
- *La técnica Boxplot se utiliza para resumir visualmente y comparar grupos de datos*
- *Usa la mediana, los cuartiles y mínimos y máximos para transmitir el nivel, la dispersión y la simetría de una distribución de datos*

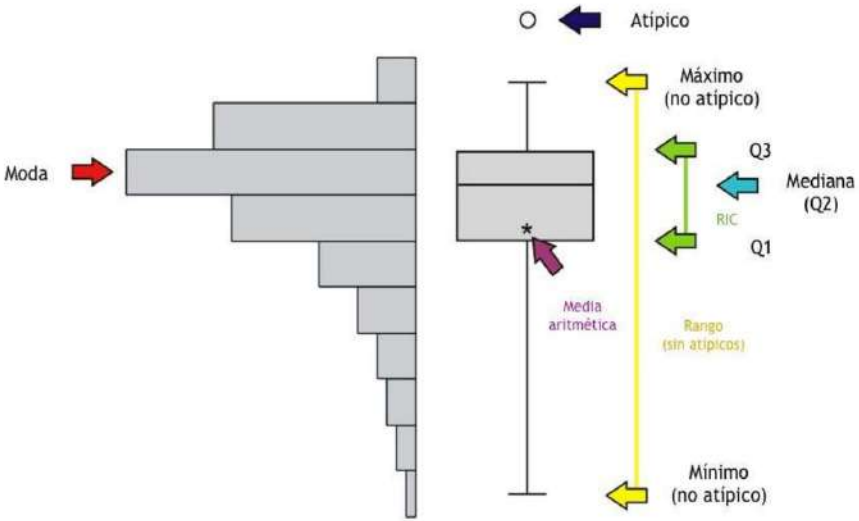
Diagrama de cajas

- *Se pueden identificar datos atípicos*
- *Una gran ventaja es que se puede construir fácilmente a mano*

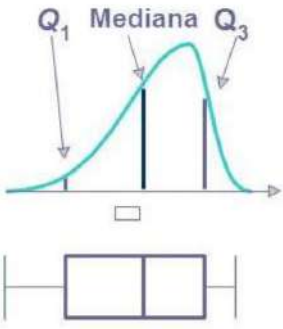
Es una herramienta que puede mejorar el razonamiento sobre información cuantitativa



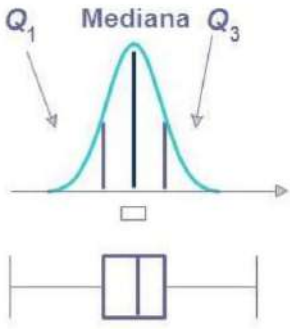

Q_1
Me = Q_2
Q_3
25%
25%
25%
25%
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
5.5
6
Lado del bigote

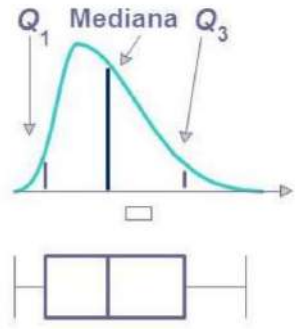

Atípico
Máximo
(no atípico)
Moda
Q3
Mediana
(Q2)
RIC
Q1
Media
aritmética
Rango
(sin atípicos)
Mínimo
(no atípico)

Medidas de forma



Asimetría negativa



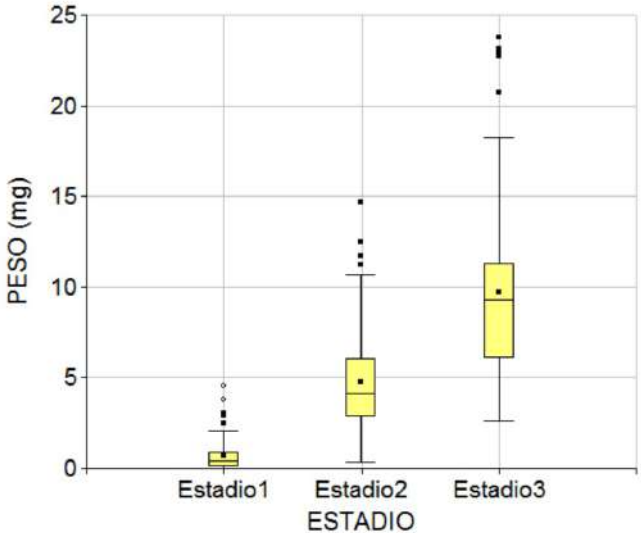


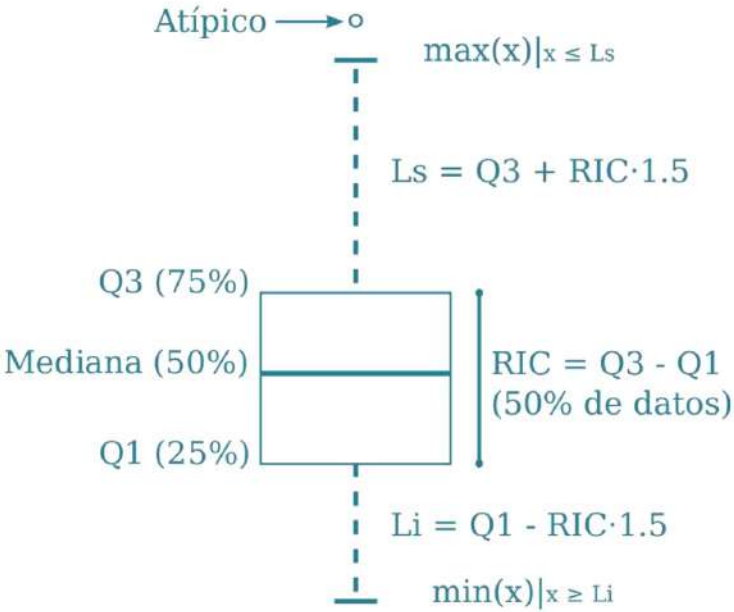
Asimetría positiva

· *Peso (mg) de 100 larvas de cada estadio de una polilla*

Estadio 1			Estadio 2			Estadio 3		
0.47	2.87	0.06	2.40	4.85	3.09	22.74	7.96	10.03
0.05	0.24	0.63	3.48	4.46	9.22	3.63	11.19	4.54
0.25	0.00	0.86	3.69	10.67	5.28	8.17	15.34	10.88
1.43	0.00	0.00	5.35	1.75	2.25	9.82	5.14	4.68
0.49	0.28	0.04	3.01	0.92	2.19	7.59	11.01	5.32
4.52	0.39	0.00	1.98	1.46	3.97	8.33	7.48	14.40
2.92	1.06	0.47	1.88	4.51	4.15	12.49	10.19	10.83
0.14	0.11	0.12	12.47	2.35	2.81	7.74	10.95	5.54
1.76	1.00	0.07	11.24	5.47	3.75	23.73	12.87	9.75
0.18	0.01	2.94	5.43	4.07	0.73	6.79	13.67	6.51
0.69	0.37	0.92	7.29	14.67	2.59	8.28	7.56	9.93
0.00	0.56	0.03	3.88	1.40	3.83	6.46	9.12	9.10
0.20	1.20	0.01	4.19	5.07	2.92	11.99	10.93	11.80
0.75	0.40	0.05	3.34	3.43	6.40	14.52	22.87	15.05
3.02	3.77	0.76	11.69	9.01	5.50	18.25	4.57	12.49
0.29	0.28	0.39	2.98	6.09	7.22	13.62	11.30	5.48
1.68	0.46	1.06	1.36	5.31	5.60	8.74	8.56	6.68
0.37	0.31	0.84	2.97	9.54	4.29	8.53	3.93	10.45
0.06	0.84	0.12	1.93	7.55	4.68	9.61	23.12	11.35
0.72	0.91	0.51	3.84	8.33	2.32	2.83	5.44	9.58
0.09	0.23	1.87	2.33	2.89	3.93	13.69	14.41	5.56
0.10	0.06	0.75	3.02	4.64	5.11	10.83	2.63	8.52
0.69	0.27	0.03	5.02	9.59	3.03	8.10	6.52	7.73
0.00	1.87	1.80	6.25	7.13	3.46	9.49	17.35	7.02
0.77	1.26	0.56	9.29	3.29	2.05	3.16	10.24	5.56
0.10	0.82	0.85	2.83	7.16	1.67	10.64	12.34	16.14
0.14	0.00	0.05	6.31	0.35	4.45	5.13	6.81	10.95
0.90	0.00	0.05	1.61	2.81	3.47	10.18	4.17	5.22
0.00	1.57	0.53	5.89	9.33	5.76	4.18	8.38	11.05
1.25	0.04	0.02	6.49	3.01	1.75	6.04	4.87	20.70
2.50	0.36	0.01	8.35	6.65	1.97	17.87	5.46	10.24
2.05	0.01	0.04	4.22	6.44	9.41	5.97	10.45	7.97
1.82	0.20		2.95	5.94		5.18	17.90	
1.76	0.00		2.61	5.43		10.19	3.44	

Box-Plot



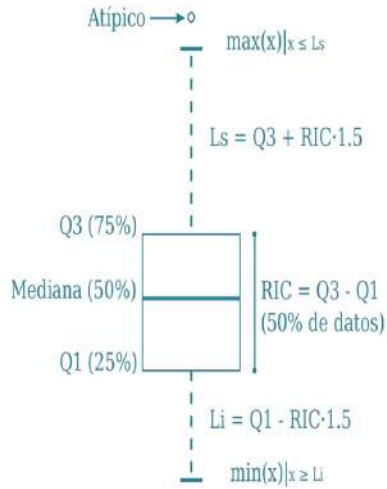

Atípico
max(x)|x ≤ Ls
Ls = Q3 + RIC·1.5
Q3 (75%)
Mediana (50%)
Q1 (25%)
RIC = Q3 - Q1
(50% de datos)
Li = Q1 - RIC·1.5
min(x)|x ≥ Li

Procedimiento para elaborar un diagrama de caja

1. *Ordenar los datos y obtener el valor mínimo, máximo, y los cuartiles Q_1, Q_2 y Q_3*

2. *Dibujar un rectángulo con extremos Q_1 y Q_3 e indicar la posición de la mediana (Q_2), mediante una línea horizontal*

3. *Obtener el rango intercuartil (RIC) que es la diferencia entre Q_3 y Q_1.*



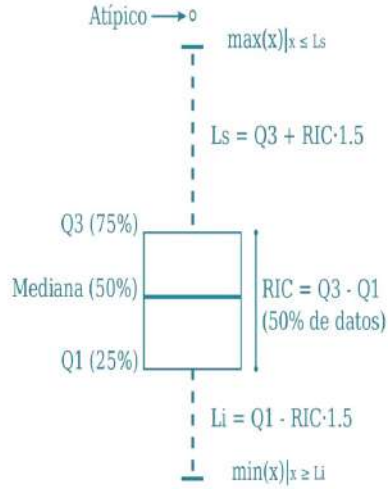
Procedimiento para elaborar un diagrama de caja

4. *Calcular los límites admisibles superior e inferior, Li= límite inferior y Ls= límite superior.*

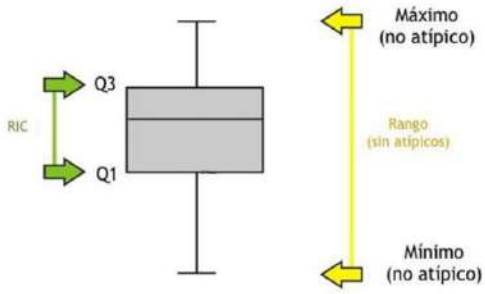
Límites admisibles o vallas

$$LI = Q_1 - 1{,}5 \cdot RIC$$

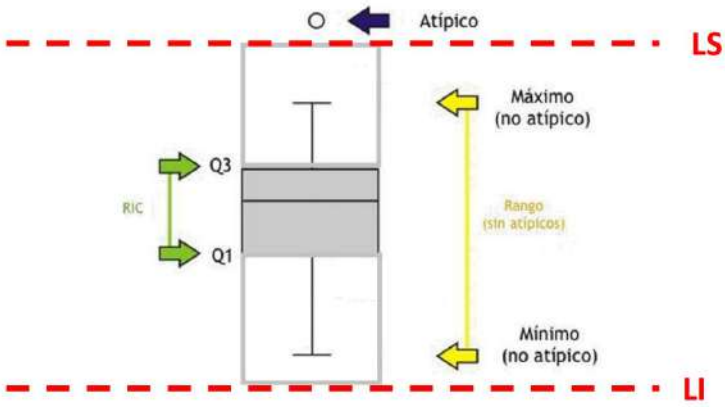
$$LS = Q_3 + 1{,}5 \cdot RIC$$



Límites admisibles: $LI = Q_1 - 1{,}5 \cdot RIC$ $\quad LS = Q_3 + 1{,}5 \cdot RIC$



Límites admisibles: $LI = Q_1 - 1{,}5 \cdot RIC$ $\quad LS = Q_3 + 1{,}5 \cdot RIC$

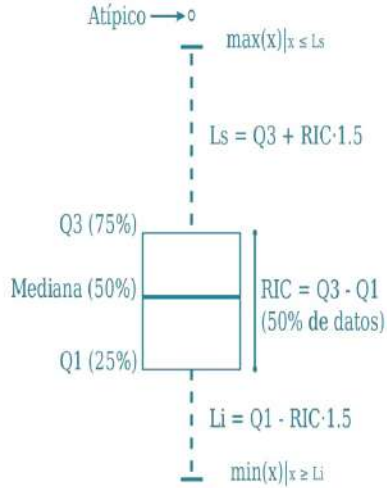


Procedimiento para elaborar un diagrama de caja

4. *Calcular los límites admisibles superior e inferior, Li y Ls.*

5. *Dibujar una línea que va desde cada extremo del rectángulo central hasta el valor más alejado no atípico, es decir, que está dentro del intervalo (Li, Ls).*

6. *Identificar todos los datos que están fuera del intervalo marcándolos como atípicos.*

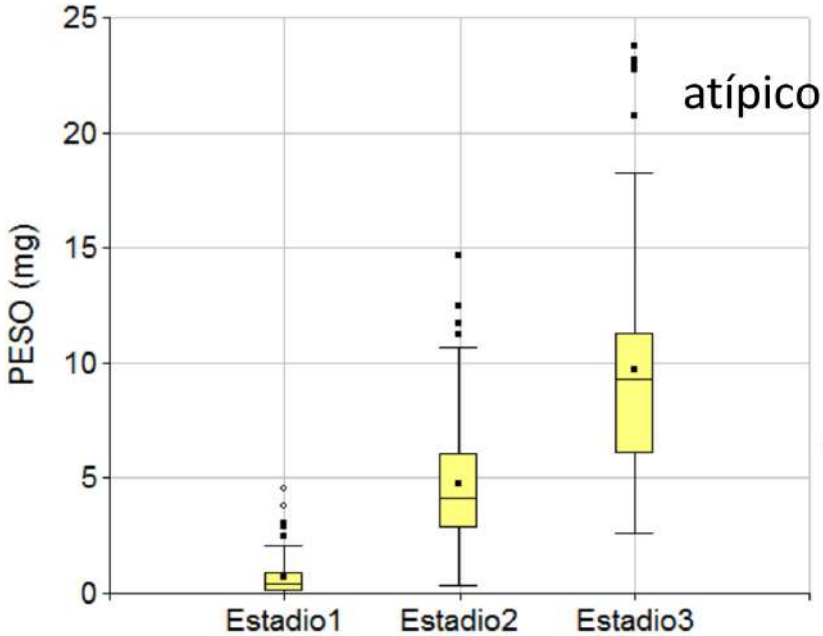


Valores atípicos leves y extremos

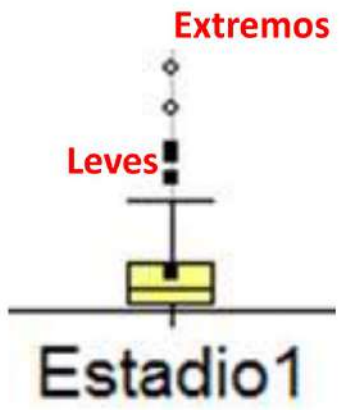
- *Se considera un valor atípico **LEVE** el que se encuentra 1,5 veces RIC de distancia a los cuartiles 1 o 3*
- *Y atípico **EXTREMO** aquel que se encuentra a 3 veces esa distancia*

Leve:	$q < Q_1 - 1{,}5 \cdot RIC$	**o**	$q > Q_3 + 1{,}5 \cdot RIC$
Extremo:	$q < Q_1 - 3 \cdot RIC$	**o**	$q > Q_3 + 3 \cdot RIC$

RIC= rango intercuartil



atípico leve vs extremo

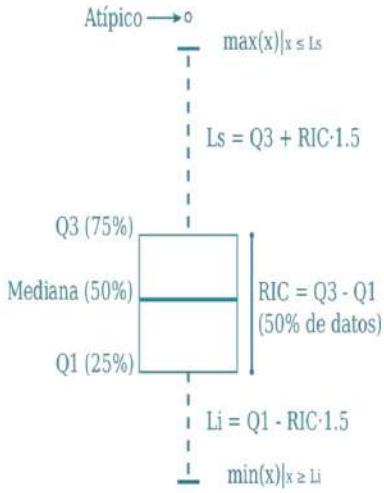


Ejercicio de obtención de un diagrama de caja

Construir un diagrama de cajas a partir de los datos entregados:

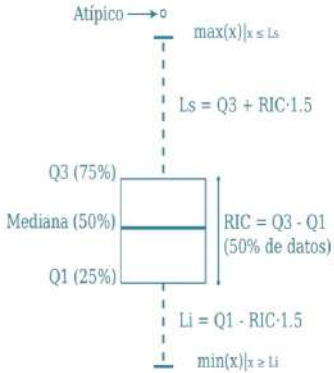
peso de ratones de laboratorio (g)

35	40	31	50	53	40	49	28	30	42	57	35	45	46	46



Ejercicio de obtención de un diagrama de caja

1°	2°	3°	4°	5°	6°	7°	8°	9°	10°	11°	12°	13°	14°	15°
28	30	31	35	35	40	40	42	45	46	46	49	50	53	57



Ejercicio de obtención de un diagrama de caja