Medidas de variabilidad o dispersión

Material complementario

Contenidos

- Rango
- Varianza y Desviación estándar
  - Datos no agrupados
  - Datos agrupado
- Coeficiente de variación
- Error estándar de la media
- Análisis de datos
  - Sesgo y Curtosis
  - Detección de valores atípicos
  - Diagrama de caja y bigote

Medidas De Dispersión

- La segunda característica más importante que describe un conjunto de datos, es **la dispersión**
- La **dispersión** es la cantidad de variación, o diseminación en los datos. Determina si los valores están relativamente cercanos entre sí, o no
- Tiene como propósito ofrecer información adicional que permita juzgar la confiabilidad de la medida de tendencia central

Aplicaciones

Se les usa para comparar distribuciones y para calcular los errores estándar, que serán de importancia en la estadística inferencial, en las pruebas de hipótesis y en los intervalos de confianza

Rango

- Es la medida de dispersión más fácil de calcular

$$Rango = Valor \text{ maximo - Valor minimo}$$

- No están usada ya que sólo considera los valores extremos de la serie de datos
- Cuando se elimina la influencia de los valores extremos hablamos de un rango intercuartil, que corresponde a :

$$Rango\,\text{int}\,ercuartil = Q_3 - Q_1$$

Varianza

- Indica qué tan dispersos se encuentran los datos, en promedio, de la media de la población
- Para representar la varianza poblacional y la varianza muestral se utilizan los siguientes dos símbolos:
  - σ^2 - donde σ es la letra griega (sigma) al cuadrado que determinará la varianza de una población
  - s^2 - determina la varianza de la muestra analizada

La fórmula para calcular la varianza de una población está dada por la expresión:

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1} = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}\right]$$

donde:

x_i = son las observaciones que componen la población, i = 1, 2, 3,...,N

μ = la media de la población

N = El número total de elementos de la población.

σ^2 = La varianza de la población

Para calcular la varianza muestral para datos no agrupados se utiliza la misma fórmula reemplazando las variables σ^2, μ y N por s^2, $\bar{x}$ y n, respectivamente, esto es:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]$$

donde:

$\bar{x}$ - es la media muestral

x_i - son las observaciones que componen la muestra, i = 1, 2, 3,...,n

n - el número total de elementos de la muestra

s^2 - La varianza de la muestra

Para calcular la varianza muestral para datos agrupados se utiliza la fórmula:

$$s^2 = \frac{\sum_{i=1}^{k} f_i (M_i - \overline{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i M_i^2 - \frac{\left(\sum_{i=1}^{k} f_i M_i\right)^2}{n}\right]$$

donde:

$\overline{x}$ - es la media muestral

x_i - es la marca de clase i, i = 1, 2, 3,...,k

f_i - es la frecuencia absoluta del intervalo de clase i, i = 1, 2, 3,...,k

k - es el número de intervalos de clase

n - el número total de elementos de la muestra

s^2 - La varianza de la muestra

Desviación Estándar

- En la varianza, los resultados se expresan en unidades originales al cuadrado, por lo que se requiere de una medida de desviación que sea útil en unidades originales que no estén elevadas
- Esta medida es llamada desviación estándar y es la raíz cuadrada de la varianza
- Para representar la desviación estándar poblacional y la desviación estándar muestral se utilizan los siguientes dos símbolos:
  - σ- donde sigma es la letra griega que determinará la desviación estándar de una población
  - s -determina la desviación estándar de la muestra analizada

La fórmula para calcular la desviación estándar de una población está dada por la expresión:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N-1}} = \sqrt{\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - \frac{\left(\sum_{i=1}^{N} x_i\right)^2}{N}\right]}$$

donde:

x_i = son las observaciones que componen la población, i = 1, 2, 3,...,N

μ = la media de la población

N = El número total de elementos de la población

σ = La desviación estándar de la población

Para Desviación estándar muestral de datos individuales se utiliza la misma fórmula reemplazando las variables σ y N por s, $\bar{x}$ y n, respectivamente, esto es:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}$$

donde:

$\bar{x}$ - es la media muestral

x_i - son las observaciones que componen la muestra, i = 1, 2, 3,...,n

n - el número total de elementos de la muestra

s - la desviación estándar de la muestra

Para datos agrupados se utiliza la fórmula:

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (M_i - \overline{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} f_i M_i^2 - \frac{\left(\sum_{i=1}^{k} f_i M_i\right)^2}{n}\right]}$$

donde:

$\overline{x}$ - es la media muestral

M_i - es la marca de clase i, i = 1, 2, 3,...,k

f_i - es la frecuencia absoluta del intervalo de clase i, i = 1, 2, 3,...,k

k - es el número de intervalos de clase

n - el número total de elementos de la muestra

s - la desviación estándar de la muestra

Ejemplo

Ingresos mensuales en dólares					
1000	1110	1010	1070	1030	1000
1150	990	1090	1080	1150	1200
1050	1030	1120	1050	1030	1150
1230	1170	1180	1110	1160	1100
1100	1060	1130	1105	935	1210

Datos No Agrupados

n	X_i	X_i^2	X_i	X_i^2
30	935	874225	1100	1210000
n-1	990	980100	1105	1221025
29	1000	1000000	1110	1232100
	1000	1000000	1110	1232100
	1010	1020100	1120	1254400
	1030	1060900	1130	1276900
	1030	1060900	1150	1322500
	1030	1060900	1150	1322500
	1050	1102500	1150	1322500
	1050	1102500	1160	1345600
	1060	1123600	1170	1368900
	1070	1144900	1180	1392400
	1080	1166400	1200	1440000
	1090	1188100	1210	1464100
	1100	1210000	1230	1512900
		Total	**32800**	**36013050**

Varianza

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right] = \frac{1}{29}\left[36013050 - \frac{(32800)^2}{30}\right]$$

$$= \frac{1}{29}\left[36013050 - \frac{(32800)^2}{30}\right] = \frac{1}{29}\left[36013050 - 35861333.3\right] = 5231.6092$$

Desviación estándar

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}\right]}$$

$$s = \sqrt{s^2} = \sqrt{5231.6092} = 72.33$$

- Este último cálculo significa que existe una dispersión de $ 72.33 con respecto a la media

- Esta unidad de medida es congruente con la obtenida al calcular la media aritmética, por lo tanto, se pueden hacer inferencias con respecto a la población objeto de estudio a través de los intervalos de confianza

Ejemplo

- Consideremos los valores expuestos en el ejemplo anterior y definamos las clases

INT. DE CLASE	**MARCA DE CLASE M_i**	**FREC. ABS. fi**	X_i^2	fM_i	$f_iM_i^2$
(930 - 980]	955	1	912025	955	912025
(980 – 1030]	1005	7	1010025	7035	7070175
(1030 – 1080]	1055	5	1113025	5275	5565125
(1080 – 1130]	1105	8	1221025	8840	9768200
(1130 – 1180]	1155	6	1334025	6930	8004150
(1180 – 1230]	1205	3	1452025	3615	4356075
		30=n	**Total**	**32650**	**35675750**
		29= n-1			

Varianza

$$s^2 = \frac{\sum_{i=1}^{k} f_i (M_i - \bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i M_i^2 - \frac{\left(\sum_{i=1}^{k} f_i M_i\right)^2}{n}\right] = \frac{1}{29}\left[35675750 - \frac{(32650)^2}{30}\right]$$

$$= \frac{1}{29}[35675750 - 35534083.3] = 4885.057$$

Desviación estándar

$$s = \sqrt{\frac{\sum_{i=1}^{k} f_i (M_i - \bar{x})^2}{n-1}} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^{k} f_i M_i^2 - \frac{\left(\sum_{i=1}^{k} f_i M_i\right)^2}{n}\right]} =$$

$$s = \sqrt{4885.057} = 69.89$$

Datos Agrupados

- Este último cálculo significa que existe una dispersión de \$ 69.89 con respecto a la media

- Esta unidad de medida es congruente con la obtenida al calcular la media aritmética, por lo tanto, se pueden hacer inferencias con respecto a la población objeto de estudio a través de los intervalos de confianza

Coeficiente De Variación

- Es la dispersión relativa existente entre la desviación estándar y la media aritmética de los datos
- Este coeficiente está dado como el cociente resultante de dividir la desviación estándar entre la media:

$$C.V. = \frac{S}{\overline{X}}$$

- El coeficiente de variación se puede expresar como porcentaje

- Esta medida de variabilidad expresa la desviación estándar por unidad experimental como una medida general del experimento.

- De esta forma se puede comparar entre dos o más coeficientes de variación, y observar cuál muestra tiene mayor variabilidad.

Interpretación del C.V

Valor del coeficiente de variación (%)	Interpretación del coeficiente	
	Variabilidad	Estabilidad
Igual a 0	Nula	Muy alta
Mayor de 0 hasta 20	Baja	Alta
Mayor de 20 hasta 60	Moderada	Moderada
Mayor de 60 hasta 90	Alta	Baja
Mayor de 90	Muy alta	Nula

En el ejemplo de los ingresos de las familias, el coeficiente de variación es calculado a continuación.

$$C.V. = \frac{S}{\overline{X}} = \frac{72.33}{1093.33} = 0.06615$$

$$PCV = 0.06615(100) = 6.62\%$$

Este resultado implica una **variación baja**, lo cual se traduce a que la variable presenta una **buena estabilidad** en su comportamiento, por lo tanto, las estimaciones que se deriven de ella podrán considerarse **confiables**

Error estándar de la media

- Estadígrafo vinculado con el error relacionado con la obtención de una media muestral
- Muy importante en el desarrollo de intervalos de confianza empleando una distribución t - student

$$EE(\bar{X}) = \frac{s}{\sqrt{n}}$$

Análisis de datos

En Estadística, la información debe ser de mayor precisión y fiabilidad posible. Debe existir una depuración de los datos experimentales.

ERRORES EN LAS OBSERVACIONES MUESTRALES

- Errores o variables que pueden existir en X(M):
  - Variabilidad de la fuente o inherente: comportamiento natural de los datos
  - Errores del Medio: Cuando no se dispone de la técnica adecuada o cuando no existe un procedimiento para realizar la transformación de una forma exacta. Ej: Redondeo forzoso con variables continuas
- Error del experimentador:
  - Error de la Información: cuando un modelo o estructura matemática no es adecuada o precisa a la población, o al considerar información o hipótesis iniciales incorrectas
  - Error de Planificación: cuando no se delimita correctamente la población
  - Error de realización: por una valoración errónea de los elementos de M(es decir, el paso de la información de un instrumento a otro, Ej: de la libreta al ordenador.)

Sesgo

El sesgo es el grado de asimetría o falta de la misma de una distribución de frecuencia, por lo que numéricamente se puede calcular como:

$$sesgo = \frac{3(\bar{x} - mediana)}{s}$$

$$\text{Si } sesgo < 0 \quad \text{la curva esta sesgada a la izquierda}$$

$$\text{Si } sesgo > 0 \quad \text{la curva esta sesgada a la derecha}$$

$$\text{Si sesgo } = 0 \quad \text{la curva es simétrica}$$

Coeficiente de Simetría de FISHER

$$\gamma_1 = \frac{\bar{x}^3}{S^3} = \frac{1}{S^3} * \sum_{i=1}^{k} \left(x_i - \bar{x} \right)^3 * f_i$$

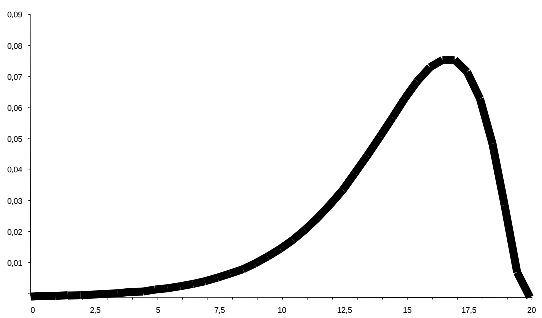
Si: $\gamma_1 = 0$ **Situación de Simetría**

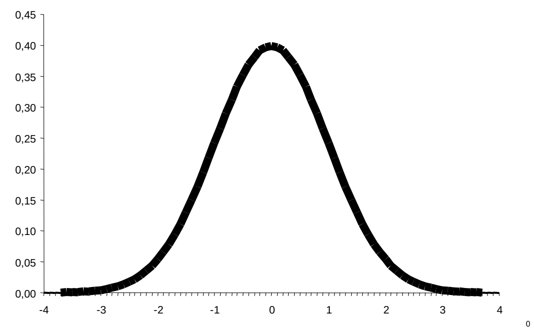
$\gamma_1 > 0$ **Situación de Simetría a la Derecha**

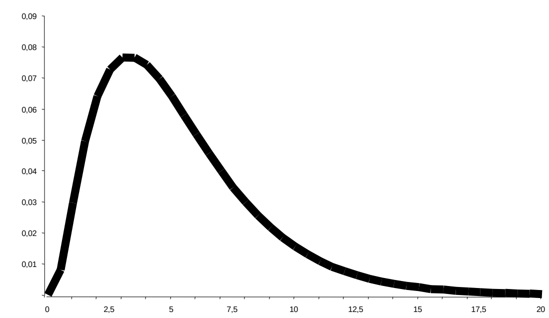
$\gamma_1 < 0$ **Situación de Asimetría a la Izquierda**

Medida Adimensional(Sin Medida)

Coeficiente de Simetría de Fisher $\gamma_1 = \dfrac{m_3}{S^3}$ → Sesgo.







$\gamma_1 < 0$

Distribución, tiende a concentrarse en Valores Altos de la Variable

Mo > Me > MA

$\gamma_1 = 0$

Distribución, es simétrica respecto a la Media

Mo = MA = Me

$\gamma_1 > 0$

Distribución, tiende a concentrarse en Valores Bajos de la Variable

Mo < Me < MA

Coeficiente de Simetría de PEARSON

$$A_s = \frac{\bar{x} - Mo}{S}$$

Si:	$A_s = 0$	**Situación de Simetría**
	$A_s > 0$	**Situación de Simetría a la Derecha**
	$A_S < 0$	**Situación de Asimetría a la Izquierda**

Medida Adimensional(Sin Medida)

Curtosis

La curtosis es el grado de esbeltez de una distribución de frecuencia, por lo que numéricamente se puede calcular como el **Coeficiente de Curtosis de Fisher** de la Variable estadística x el cual se define como:

Interpretación:

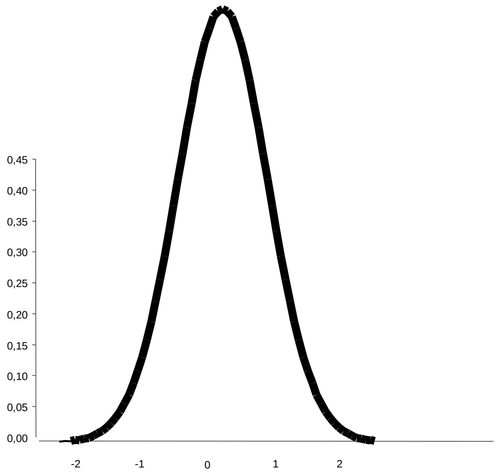
$$\gamma_2 = \frac{1}{S^4} * \sum_{i=1}^{k} \left(x_i - \underline{x}\right)^4 * f_i - 3$$

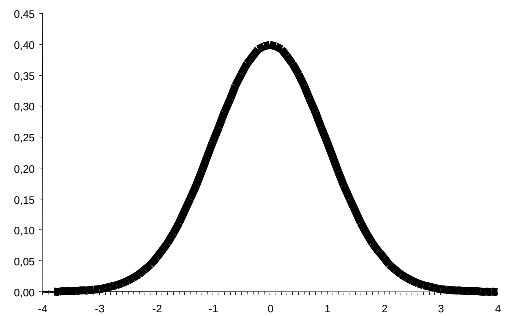
γ_2 **= 0** — **Mesocúrtica**

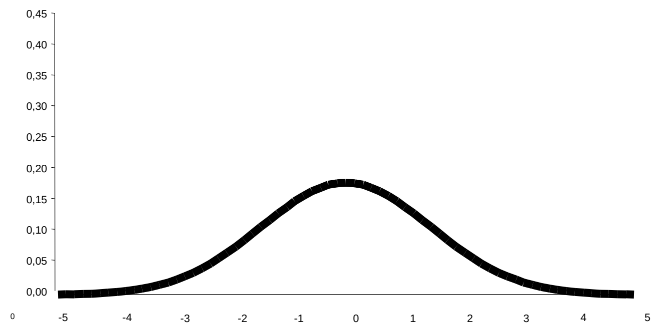
γ_2 **> 0** — **Leptocúrtica**

γ_2 **< 0** — **Platicúrtica**

Coeficiente $\gamma_2 = \dfrac{m_4}{S^4} - 3$ → Curtosis







$\gamma_2 < 0$

Distribución tiende a concentrarse alrededor de la Media.

Variancia Pequeña

Aguzada

$\gamma_2 = 0$

Distribución "Normal"

$\gamma_2 > 0$

Distribución tiende a dispersarse

Variancia grande

Achatada

Detección de valores atípicos

En las observaciones pueden aparecer valores extraños o anómalos:

- Observación Atípica: es aquel valor de X(M) que presenta una gran variabilidad de tipo inherente
- Observación Errónea: es el valor que presenta un gran error del medio y/o un gran error del experimentador.

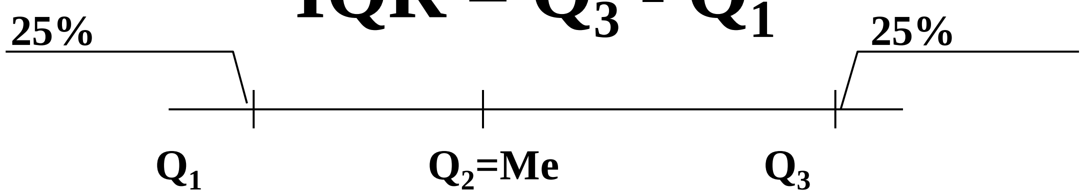
Outlier: es aquella observación que siendo atípica y/o errónea, tiene un comportamiento muy diferente respecto de los datos, frente al análisis que se desea realizar sobre las observaciones experimentales

Inlier: es aquella observación atípica y/o errónea que no tiene el comportamiento de Outlier. Es decir, se comporta casi igual o igual que los datos de nuestro análisis

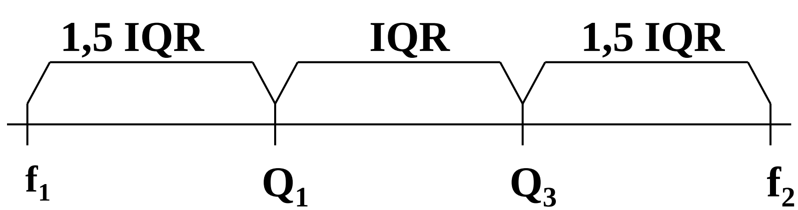
Recorrido Intercuartílico

- Las **vallas interiores** de la variable estadística x:

$$IQR = Q_3 - Q_1$$



$$f_1 = Q_1 – 1{,}5\ IQR \qquad [f_1,f_2] \qquad f_2 = Q_3 + 1{,}5\ IQR$$



- Las **vallas exteriores** de la variable estadística X:

$$F_1 = Q_1 - 3\ IQR \qquad [F_1,F_2] \qquad F_2 = Q_3 + 3\ IQR$$

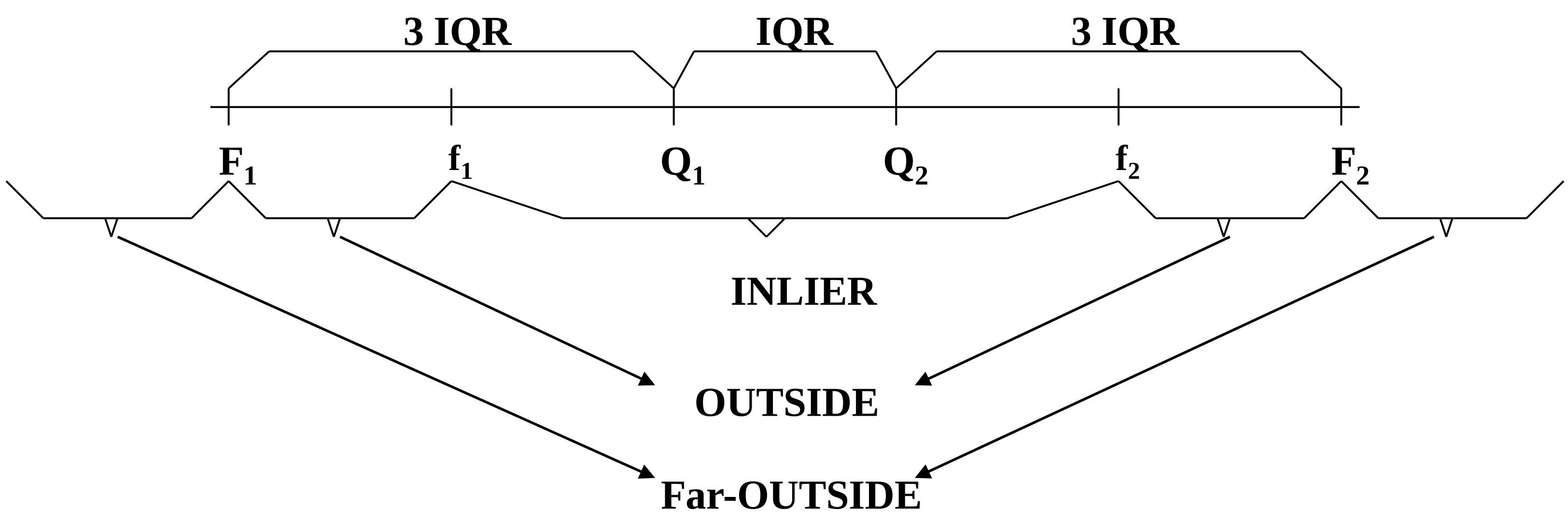


Diagrama de caja y bigote

- Es una forma de resumir en una gráfica los datos
- La base de este diagrama es el cálculo de la mediana y los cuartiles Q_1 y Q_3. También se usa el IQR
- Al utilizar este diagrama se puede identificar valores atípicos sin necesidad de cálculos complejos

Pasos a seguir

- Se traza un rectángulo con los extremos en el primer y tercer cuartil. Esta zona contiene el 50% de los datos
- En la caja se traza una línea recta vertical en el lugar de la mediana
- Se ubican los límites mediante el rango intercuartílico. Los límites están a 1.5·IQR abajo de Q_1 y a 1.5·IQR arriba de Q_3. Si los datos están fuera de estos límites se consideran atípicos
- Las líneas punteadas se llaman bigotes de la caja y se trazan desde los extremos de ésta hasta los límites
- Se marcan con un asterisco las localizaciones de valores atípicos

Representación visual para describir, simultáneamente, varias características importantes tales como

- Centro
- Dispersión
- Desviación de la asimetría
- Identificación de las observaciones (valores atípicos)