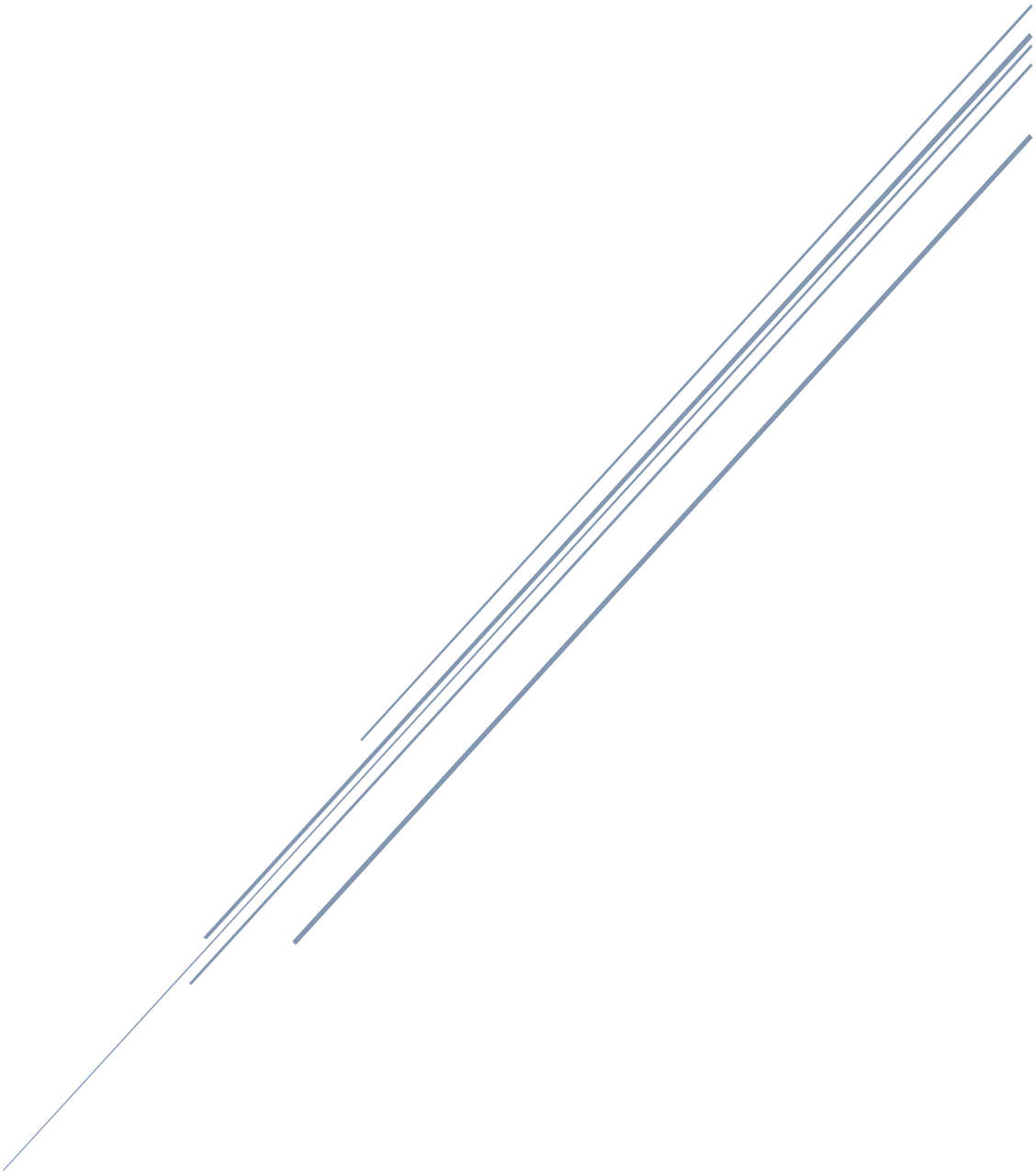
TEMA 4. DESCRIPCIÓN NUMÉRICA DE DATOS UNIVARIANTES: MEDIDAS DE POSICIÓN

Curso OCW de “Estadística Descriptiva con Excel para Grados de Ciencias Sociales”



Grupo Decisión Multicriterio Zaragoza (GDMZ)

1. MEDIDAS DE DISPERSIÓN

Las m**edidas de dispersión** evalúan la mayor o menor *variabilidad* existente en un conjunto de datos. Pero no sólo sirven para establecer la dispersión de los valores de una variable o para comparar la que existe en dos poblaciones diferentes, sino que también permiten valorar el grado de representatividad de una medida de posición a tenor de la magnitud de la dispersión.

Las medidas de posición central sintetizan la información disponible, dando un valor que resume el comportamiento global de la variable. Una medida de posición central será más o menos representativa en función de la proximidad de los datos a dicha medida de posición. Las **medidas de dispersión** nos permiten conocer lo cerca o lejos que se encuentran los datos respecto a una medida de posición central.

En definitiva, las medidas de dispersión van a cuantificar lo separados que están los datos; bien entre sí, bien con respecto del valor central que los representa.

Las medidas de dispersión se pueden clasificar en:

- Medidas de dispersión absoluta
    - No hacen referencia a ninguna medida de tendencia central
        - **Recorridos** muestral, intercuartílico, decil y percentil
    - Hacen referencia a una medida de tendencia central
        - **Desviaciones cuadráticas**: Varianza y Desviación Típica
- Medidas de dispersión relativa
    - No hacen referencia a ninguna medida de tendencia central:
        - **Recorridos** relativo y semi-intercuartílico
    - Hacen referencia a alguna medida de tendencia central
        - **Coeficiente de Variación de Pearson**

1.1. Medidas de dispersión absolutas

1.1.1. Recorridos

La forma más sencilla de tener una idea inicial de la dispersión entre los datos es calculando la diferencia entre el valor máximo y el mínimo: el **Rango o Recorrido muestral.**

$$Re = x_k - x_1$$

Al utilizar sólo los dos datos extremos, esta medida se ve muy afectada por observaciones anómalas o atípicas y su valor puede distorsionar la magnitud de la dispersión entre el grueso de los datos. Para obtener una medida más fiable y menos

sensible a datos atípicos, se pueden recurrir al **Recorrido Intercuartílico**, que se calcula como la diferencia entre el tercer y primer cuartil:

$$R_I = C_3 - C_1$$

Dentro de este recorrido estarán comprendidos el 50% de los datos centrales. Esta medida puede generalizarse, para abarcar un mayor porcentaje de datos, dando lugar a los recorridos decil y también percentil:

$$R_D = D_9 - D_1 \qquad R_P = P_{99} - P_1$$

1.1.2. Desviaciones Cuadráticas Medias

Éstas medidas se obtienen como el promedio de las distancias de los datos a una medida de posición central. Para medir la distancia empleamos el cuadrado de la desviación. Así, la desviación cuadrática media respecto de una medida de posición *P* se calcula como:

$$D_P^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - P)^2 \times n_i$$

Considerando las diferentes medidas de posición central usuales (Media, Moda y Mediana), se obtienen las correspondientes expresiones. Entre todas ellas destaca la desviación cuadrática respecto de la media, que se denomina **varianza.**

1.1.3. Varianza y Desviación típica

Se denota por $\mathbf{S^2}$ y su expresión, como caso particular, viene dada por:

$$S^2 = \frac{1}{N}\sum_{i=1}^{k}(x_i - \overline{x})^2 \times n_i$$

Por lo tanto, la varianza mide la variabilidad de un conjunto de datos respecto de la media aritmética de los mismos.

Propiedades:

- Es no negativa y si es cero es porque todos los valores de la variable coinciden y no hay dispersión:

  $$S^2 \geq 0 \qquad S^2 = 0 \Rightarrow x_i = \bar{x}\ \forall x_i$$

  Esto nos da la pauta para su interpretación: cuanto más próxima sea a 0, menor será la dispersión de los datos respecto de la media aritmética, teniendo ésta mayor representatividad. Por el contrario, un valor elevado de la varianza indica

un alejamiento considerable de los datos respecto de la media aritmética, lo que la hace menos representativa.

- Es invariante ante cambios de origen, pero no de escala:

$$Y = a + b \times X \;\Rightarrow \begin{cases} S_Y^2 = \mathrm{b}^2 \times S_X^2 \\ S_Y = |\mathrm{b}| \times S_X \end{cases}$$

- Para el cálculo de la varianza se emplea la expresión equivalente:

$$S_X^2 = \frac{1}{\mathrm{N}} \sum_{i=1}^{k} x_i^2 \times n_i - \overline{x}^{\,2}$$

- El inconveniente de la varianza es que viene expresada en unidades cuadráticas, motivo por el cual se introduce la **desviación típica**:

$$S = \sqrt{S^2} = \sqrt{\frac{1}{N} \sum_{i=1}^{k} (x_i - \overline{x})^2 \times n_i}$$

1.2. Medidas de dispersión relativas

En ocasiones, se requiere comparar la dispersión o variabilidad existente entre dos o más distribuciones. Éstas pueden corresponder a datos de diferente índole, además de poder estar expresadas en distintas unidades, o aún expresadas en las mismas unidades, su posición es diversa. Este tipo de situaciones requieren utilizar algún tipo de coeficientes que cuantifiquen la dispersión, pero en términos relativos. Introducimos a continuación la versión relativa de las medidas de dispersión.

1.2.1. Recorridos relativo y semi-intercuartílico

Son la versión relativa de los recorridos. El **Recorrido relativo** se obtiene como:

$$R_r = \frac{\mathrm{Re}}{x_{máx}} = \frac{x_{máx} - x_{\min}}{x_{máx}}$$

Y el **Recorrido semi-intercuartílico** viene dado por:

$$R_{SI} = \frac{(C_3 - C_1)}{(C_3 + C_1)}$$

Son medidas adimensionales. No son invariantes ante cambios de origen, pero sí de escala.

1.2.2. Coeficiente de Variación

La versión relativa de la varianza es el Coeficiente de Variación de Pearson que se obtiene como:

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

Es una medida adimensional que si es menor que 0.2 (20%) indica que la dispersión relativa es baja y por ende se puede concluir que la media aritmética es representativa. En caso contrario, no lo será. Cuanto más próximo es a 0, menor dispersión relativa o mayor homogeneidad presenta la correspondiente distribución y cuando se anula es cuando la media aritmética alcanza su máxima representatividad. Sin embargo, cuando la media aritmética es cero no debe utilizarse.

1.3 Tipificación de una variable

La tipificación de una variable consiste en transformarla linealmente restándole su media y dividiéndola por su desviación típica. Si X es una variable con media $\bar{x}$ y desviación típica S_X, los valores de la variable tipificada Z se obtienen mediante:

$$z_i = \frac{x_i - \bar{x}}{S_X}$$

La media de una variable tipificada vale cero y su desviación típica uno. Cada valor de la variable tipificada z_i corresponde al número de "desviaciones" en que el valor está separado respecto de la media aritmética.

Los valores tipificados se pueden comparar directamente al estar situados en una escala común, y aquél que resulte más alto (en valor absoluto) señalará al dato que es más alejado o atípico respecto de su distribución.

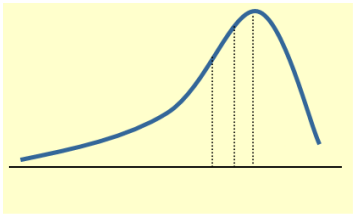
2. MEDIDAS DE FORMA

Además de las medidas de posición y las de dispersión parece lógico conocer otros aspectos acerca de cómo está distribuida la frecuencia. Ello queda perfectamente reflejado en la forma o apariencia gráfica que adopta la distribución de frecuencias.

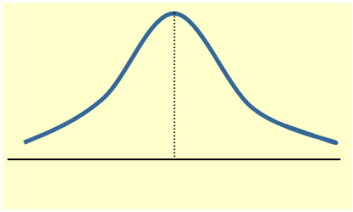
Las **Medidas de Forma**, como su nombre indica, son unas magnitudes que evaluar numéricamente el perfil de la distribución sin necesidad de realizar su representación gráfica. Las más importantes son las de asimetría y las de apuntamiento o curtosis.

2.1. Medidas de Asimetría

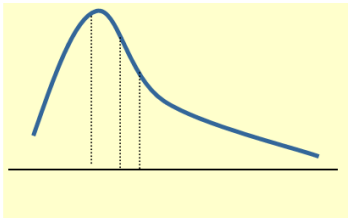
La forma más intuitiva de definir la **simetría** es a partir de su representación gráfica ya que puede trazarse una línea vertical y comprobar si al doblar por ella la figura, ambas partes coinciden exactamente. Cuando esto no ocurre, la distribución es asimétrica.



Asimetría a Izquierda



Simetría



Asimetría a Derecha

Una distribución es **simétrica** respecto de un promedio, si ocurre que hay un mismo número de datos equidistantes y con idéntica frecuencia a ambos lados del eje de simetría. Una distribución es **asimétrica a la derecha** cuando las frecuencias descienden más lentamente por la derecha que por la izquierda. Una distribución es **asimétrica a la izquierda** cuando las frecuencias descienden más lentamente por la izquierda que por la derecha.

2.1.1. Coeficiente de Asimetría de Fisher

El Coeficiente de Asimetría de Fisher viene dado por:

$$C.A.F = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{x_i - \bar{x}}{S_X}\right)^3 \times n_i = \frac{1}{N}\sum_{i=1}^{k} {z_i}^3 \times n_i \quad con \quad z_i = \frac{x_i - \bar{x}}{S_X}$$

Propiedades:

- Este coeficiente es adimensional al aparecer en las mismas unidades los términos del numerado y denominador.
- El signo depende del de su numerador. Si su valor es 0 la distribución es perfectamente simétrica. Si su valor es positivo la distribución presenta asimetría a derecha. Finalmente, si su valor es negativo la distribución presenta asimetría a izquierda.
- Se considera que un coeficiente de asimetría de Fisher es significativo estadísticamente, si en valor absoluto, es superior a $2\sqrt{(6/N)}$, es decir:

$$|CAF| > 2\sqrt{\frac{6}{N}}$$

2.2. Medidas de Curtosis o Apuntamiento

Estas medidas analizan el perfil más o menos puntiagudo de la distribución, comparando la zona central con las colas de la distribución. Así, la mayor o menor

concentración de frecuencias alrededor de la media y en la zona central de la distribución dará lugar a una distribución más o menos apuntada.

Las medidas de **apuntamiento** o **curtosis** sólo deberían medirse en distribuciones campaniformes, unimodales y simétricas o con ligera asimetría**.**

2.2.1. Coeficiente de Curtosis de Fisher

El coeficiente de apuntamiento más importante debido a Fisher se calcula como:

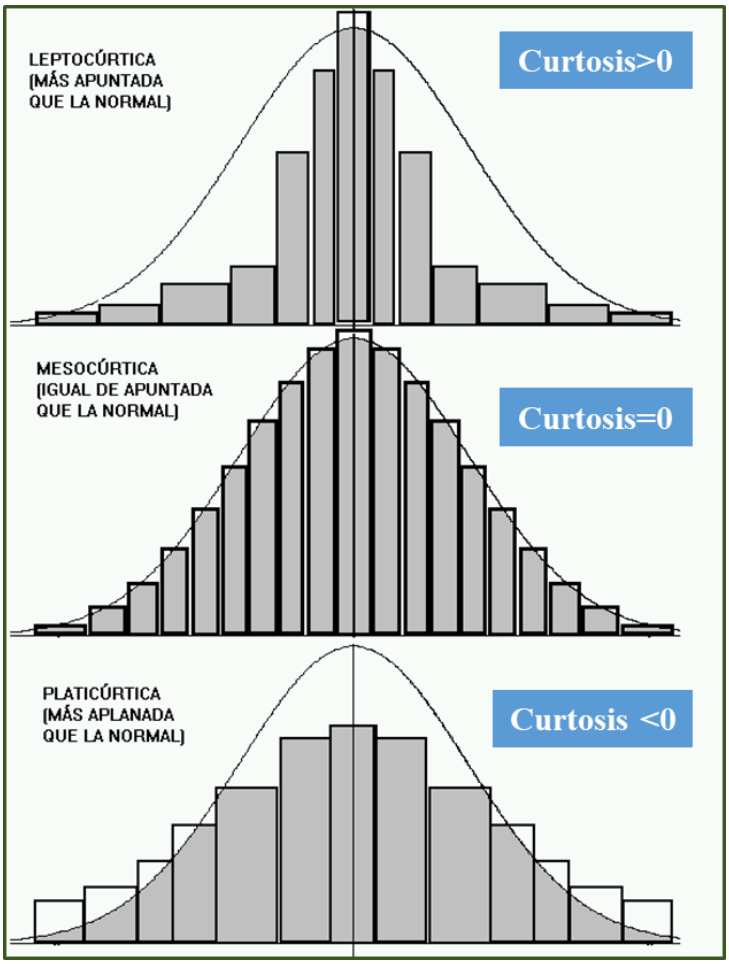
$$C.K. = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{x_i - \bar{x}}{S_X}\right)^4 \times n_i - 3 = \frac{1}{N}\sum_{i=1}^{k} {z_i}^4 \times n_i - 3 \quad con \quad z_i = \frac{x_i - \bar{x}}{S_X}$$

Este coeficiente se define en términos relativos y se calcula tomando como referencia el correspondiente a la curva normal que es el modelo matemático de referencia, de gran aplicabilidad y con buenas propiedades, y para el cual el coeficiente vale 0.

- Si $CK = 0$ el apuntamiento es similar al de la normal (distribución mesocúrtica)
- Si $CK > 0$ el apuntamiento es superior al de la normal (distribución Leptocúrtica)
- Si $CK < 0$ el apuntamiento es inferior al de la normal (distribución Platicúrtica)

Se considera que un coeficiente de curtosis de Fisher es significativo estadísticamente, si:

$$|CK| > 4\sqrt{\frac{6}{N}}$$

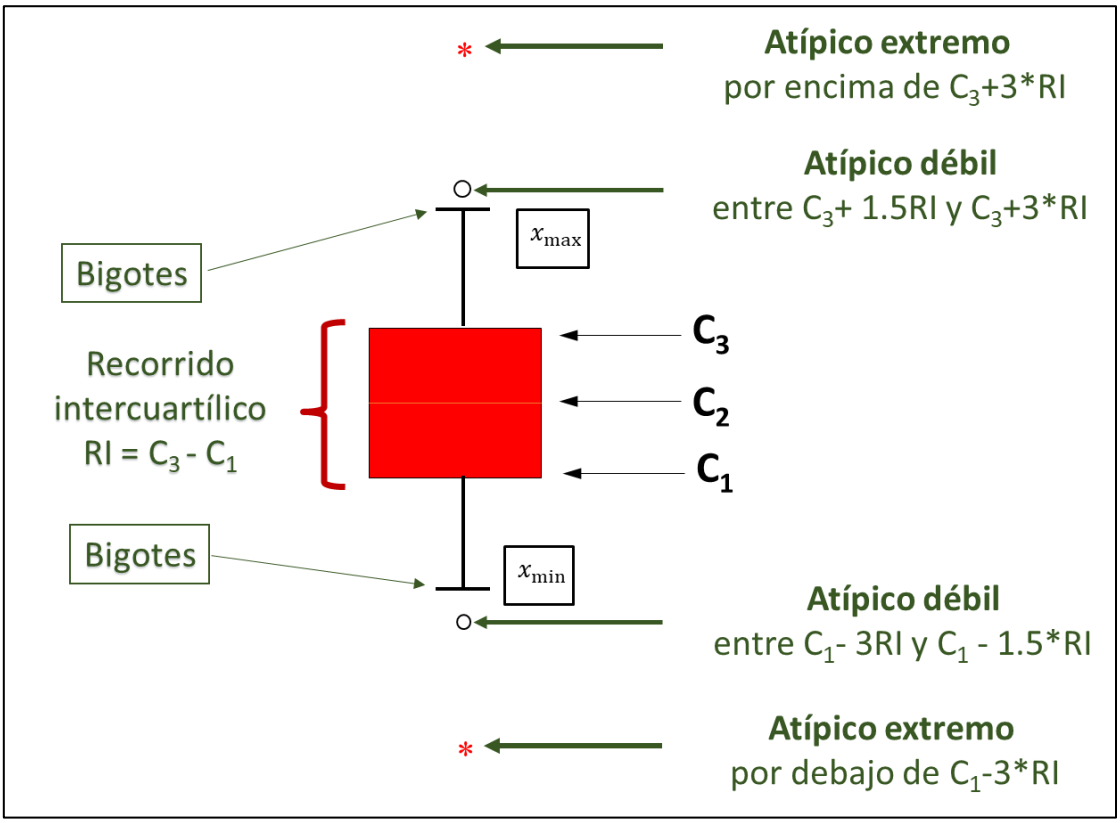


Todos los coeficientes tanto de asimetría como de apuntamiento, al ser medidas relativas, son invariantes frente a cambios de origen y escala. La asimetría y la curtosis no dependen de las unidades, ni del origen.

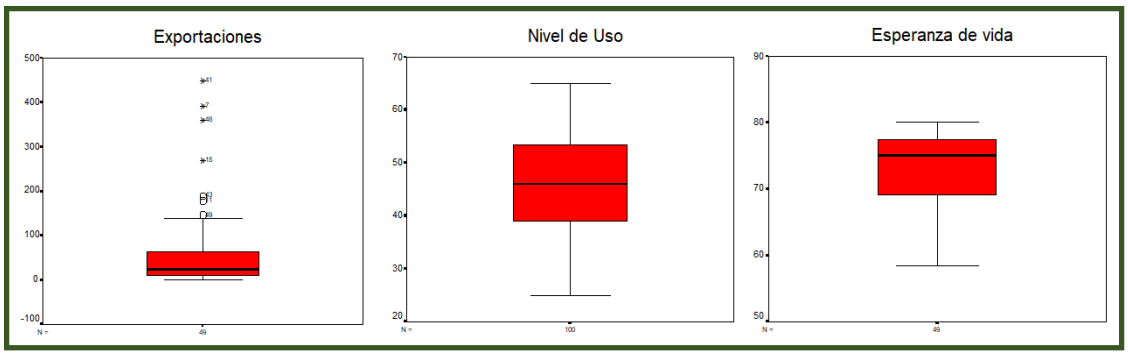
3. DIAGRAMAS DE CAJA (BOX-PLOT)

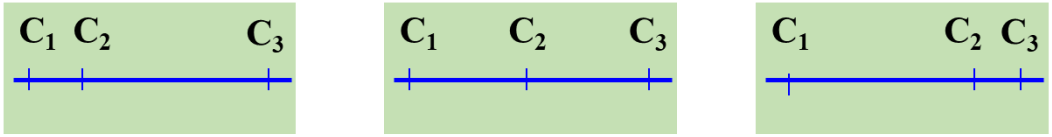
Es un gráfico con una caja central indicando el rango en el que se concentra el 50% central de los datos. Sus extremos son, por lo tanto, el primer y tercer cuartil de la distribución. En el interior de la caja se representa la posición de la Mediana mediante una línea. Las líneas que salen de los bordes de la caja son los llamados bigotes y llegan hasta los valores mínimo y máximo una vez han sido eliminados los datos atípicos.

Los valores atípicos (outliers) se indican puntualmente utilizando símbolos especiales más allá de los bigotes. Se considera que un dato es atípicos débil si su valor se encuentra a una distancia mayor de 1,5 veces y menor de 3 veces el recorrido intercuartílico respecto al borde de la caja. Un dato se considera atípico fuerte o extremo si su valor dista de la caja más de 3 veces el recorrido intercuartílico.

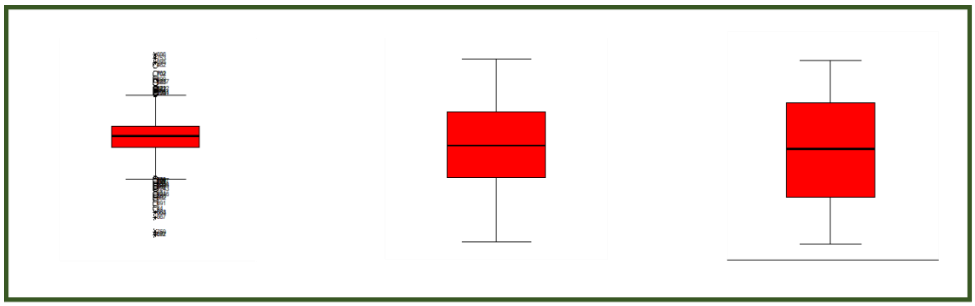


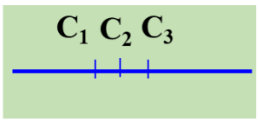
A la vista de la apariencia del Box-Plot pueden concluirse además algunos aspectos relativos a la descripción numérica de la distribución tales como el grado de dispersión (en base a la magnitud del recorrido y del recorrido intercuartílico), la asimetría (en base la posición la mediana respecto de los cuartiles) o el apuntamiento de la distribución.



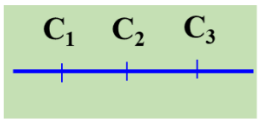


Asimetría a Derecha **Simetría** **Asimetría a Izquierda**

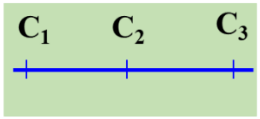




Leptocúrtica



Mesocúrtica



Platicúrtica