UNIT 5. NUMERICAL MEASURES: DISPERSION AND SHAPE MEASURES

DISPERSION MEASURES

Dispersion measures evaluate the greater or lesser *variability* that exists in a data set. They serve to establish the dispersion of the values of a variable and to compare the existing dispersion in two different populations. In addition, they allow assessing the degree of *representativeness* of a position measure according to the magnitude of the dispersion.

The central position measures synthesize the available information, giving a value that summarizes the global behaviour of the variable. A measure of central position will be more or less representative depending on the proximity of the data to that measure of position. **Dispersion measures** allow us to know how close or far data are with respect to a measure of central position.

In short, the measures of dispersion quantify how far apart data are, either with respect to each other, or with respect to the central value that represents them.

Dispersion measures can be classified into:

- Absolute dispersion measures
    - Not referred to any measure of central tendency: range, interquartile range, decile range, percentile range.
    - Referred to a measure of central tendency: Variance and Standard Deviation.
- Relative dispersion measures
    - Not referred to any measure of central tendency: Relative and semi-interquartile ranges.
    - Referred to some measure of central tendency: Pearson's Coefficient of Variation.

Absolute dispersion measures

Ranges

The easiest way to get an initial idea of the dispersion between the data is by calculating the difference between the largest value and the smallest value: the **Range** or **Sample Range.**

$$R = x_k - x_1$$

By using only the two extreme data, anomalous or outlier observations greatly affect this measure and its value can distort the magnitude of the data dispersion. The **Interquartile Range** can be used to obtain a more reliable measure and less sensitive to outliers. This range is calculated as the difference between the third and first quartiles:

$$R_I = IQR = C_3 - C_1$$

The interquartile range measures the spread in the middle 50% of the data set. The generalization of this measure allows covering a greater percentage of data. Therefore, two measures are usually employed: (i) the interdecile range (middle 80%), and (ii) the percentile range (middle 98%), which can be calculated as follows:

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

Squared Deviations from the Mean

These measures are obtained as the average of the data distances to a central position measure. To measure the distance we use the square of the deviation. Thus, the root-mean-square deviation with respect to a measure of position P is defined by the following expression:

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

The most usual measures of central position (P=Mean, Mode and Median) allow obtaining the corresponding squared deviations. Among all of them, the squared deviation from the mean stands out, which is called **variance**.

Variance and Standard Deviation

It is denoted by S^2 and its expression, as a particular case, is given by the arithmetic mean of the squared deviations from the mean:

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

Therefore, the variance measures the variability of a set of data with respect to their arithmetic mean.

Properties:

- It is non-negative. If the variance is zero it is because all the values of the variable are equal and there is no dispersion:

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

  This gives us the guideline for its interpretation: the closer it is to 0, the smaller the dispersion of the data with respect to the arithmetic mean, and the greater the representativeness of the arithmetic mean. On the contrary, a high value of the variance indicates a considerable distance of the data from the arithmetic mean, which makes it less representative.
- It is invariant to changes of origin, but not of scale:

  $$Y = a + b \times X \Rightarrow \begin{cases} S_Y^2 = \mathrm{b}^2 \times S_X^2 \\ S_Y = |\mathrm{b}| \times S_X \end{cases}$$
- To calculate the variance, an equivalent expression (shortcut formula) is usually employed:

  $$S_X^2 = \frac{1}{\mathrm{N}}\sum_{i=1}^{k} x_i^2 \times n_i - \overline{x}^{\,2}$$
- The drawback of the variance is that it is expressed in squared units (different from those of the data), which gives problems of interpretation. That is the reason for introducing the **standard deviation**:

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

Relative dispersion measurements

Sometimes, it is required to compare the dispersion or variability between two or more distributions. The variability of the distributions can correspond to data of different kind. In addition, they can be expressed in different units or, even expressed

in the same units, their position is diverse. This type of situation requires the use of some kind of coefficients that quantify the dispersion, but in relative terms. We now introduce the relative version of the dispersion measures.

Relative and semi-interquartile ranges

They are the relative version of the ranges. The **relative range** is obtained as:

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

And the **semi-interquartile range** is defined as:

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

They are dimensionless measures. They are variant to changes in origin, but they are invariant to changes in scale.

Pearson's Coefficient of Variation

The relative version of the variance is the Pearson's Coefficient of Variation which is obtained as:

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

It expresses the standard deviation as a percentage of the mean (provided the mean is positive). It is a dimensionless measure. A value of less than 0.2 (20%) indicates that the relative dispersion is low and therefore it can be concluded that the arithmetic mean is representative. The closer the Coefficient of Variation is to zero, the lower the relative dispersion or greater homogeneity of the distribution. The arithmetic mean reaches its maximum representativeness when the Coefficient of Variation is zero. It should not be employed when the mean is zero or very close to zero.

Standardisation: Z-Score

The standardisation of a variable consists of a linear transformation carried out by subtracting the mean of the variable and dividing the difference by the standard deviation of the variable. If X is a variable with arithmetic mean $\bar{x}$ and standard deviation S_X, the values of the standardised variable Z (z-scores) are obtained by the following expression:

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

Standardised variables are variables that have been standardized to have a mean of 0 and a standard deviation of 1. Each value of the standardised variable z_i (z score) corresponds to the number of "standard deviations" a value is from the arithmetic mean.

The z-scores allow us to compare values from different populations as they are measured on a common scale. They also allow us to locate them in the frequency distribution, without needing to express the context of this distribution

SHAPE MEASURES

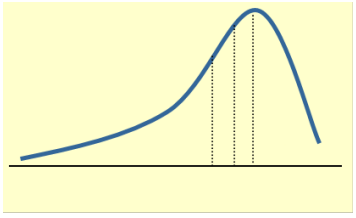
Many models assume normal distribution. The assumption of normality is necessary in many statistical hypothesis tests in inference statistics. If not fulfilled, the results obtained when analysing the data can be distorted.

Shape measures are employed to test if a frequency distribution follows a normal distribution without having to use a graphical presentation. They compare tails of the distribution between themselves or with respect to the centre of the distribution

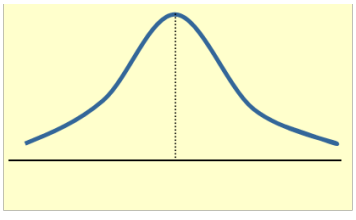
Shape measures evaluate two aspect: skewness and kurtosis.

Measures of Skewness

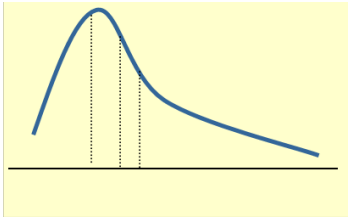
The most intuitive way to define **symmetry** is from its graphical representation, since a vertical line can be drawn and a check can be made to see if, when the figure is folded along it, both parts coincide exactly. When this does not happen, the distribution is skewed.



Skewed-left



Symmetric



Skewed-right

A distribution is said to be **Symmetric** if the portion on one side of the middle is nearly identical to the portion on the other side (like two mirror images). In this case, the mode, median and mean are located at the centre and are almost equal ($\bar{X}$ = Me = Mo).

A distribution is said to be **Skewed-right** or **Positively Skewed** if the tail extends farther to the right. In this case, the arithmetic mean is the largest of the three measures ($\bar{X}$ > Me > Mo).

A distribution is said to be **Skewed-left** or **Negatively Skewed** if the tail extends farther to the left. In this case, the arithmetic mean is the lowest of the three measures ($\bar{X}$ < Me < Mo).

Measures of skewness quantify if the observations are balanced, or approximately evenly distributed, around the centre of the frequency distribution. The most important one is the Fisher's Coefficient of Skewness

Fisher's Coefficient of Skewness

Fisher's Coefficient of Skewness is is defined as the average of the standardised deviations of the values of the distribution with respect to the arithmetic mean raised to the power of three:

$$FCS = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{x_i - \bar{x}}{S_X}\right)^3 \times n_i = \frac{1}{N}\sum_{i=1}^{k} {z_i}^3 \times n_i \quad with \quad z_i = \frac{x_i - \bar{x}}{S_X}$$

Properties:

- This coefficient is dimensionless as the numerator and denominator terms have the same units.
- The sign depends on the sign of its numerator. If its value is 0, the distribution is perfectly symmetric. If its value is positive, the distribution is skewed- right or positively skewed. Finally, if its value is negative, the distribution is skewed- left or negatively skewed.
- A Fisher's skewness coefficient is considered statistically significant if:

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

Measures of Kurtosis

These measures quantify the peakedness of the frequency distribution by comparing the centre with the tails of the distribution. Thus, the greater or lesser concentration of frequencies in the centre of the distribution will result in a more or less pointed distribution.

Measures of **kurtosis** should only be measured in bell-shaped, unimodal, and symmetric, or slightly skewed distributions.

Fisher's kurtosis coefficient

The most important kurtosis coefficient, defined by Fisher, is calculated as follows:

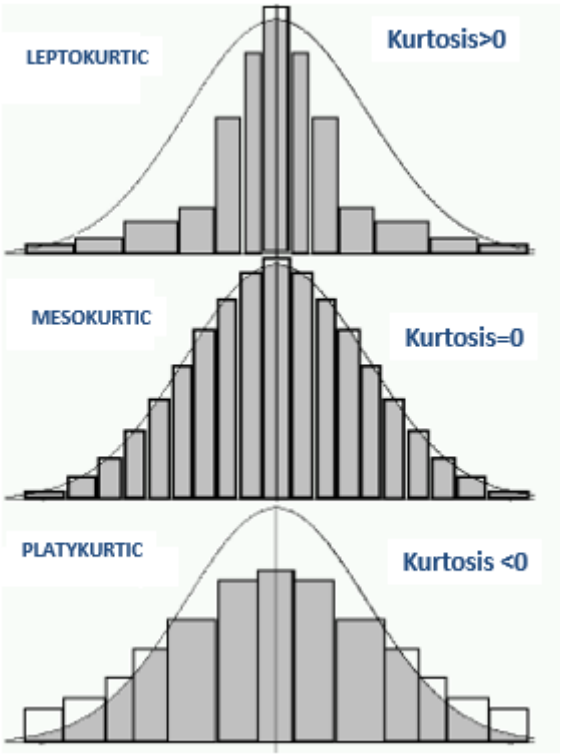
$$CK = \frac{1}{N}\sum_{i=1}^{k}\left(\frac{x_i - \bar{x}}{S_X}\right)^4 \times n_i - 3 = \frac{1}{N}\sum_{i=1}^{k} {z_i}^4 \times n_i - 3 \quad with \quad z_i = \frac{x_i - \bar{x}}{S_X}$$

This coefficient is defined in relative terms and is calculated taking as a reference the one corresponding to the normal distribution, which is the mathematical reference model, for which the coefficient is 0.

- $CK = 0$, the peakedness is similar to that of the normal distribution (mesokurtic distribution).
- $CK > 0$, more peakedness than the normal distribution (leptokurtic distribution).
- $CK < 0$, less peakedness than the normal distribution (platikurtic distribution).

A Fisher kurtosis coefficient is considered to be statistically significant if:

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



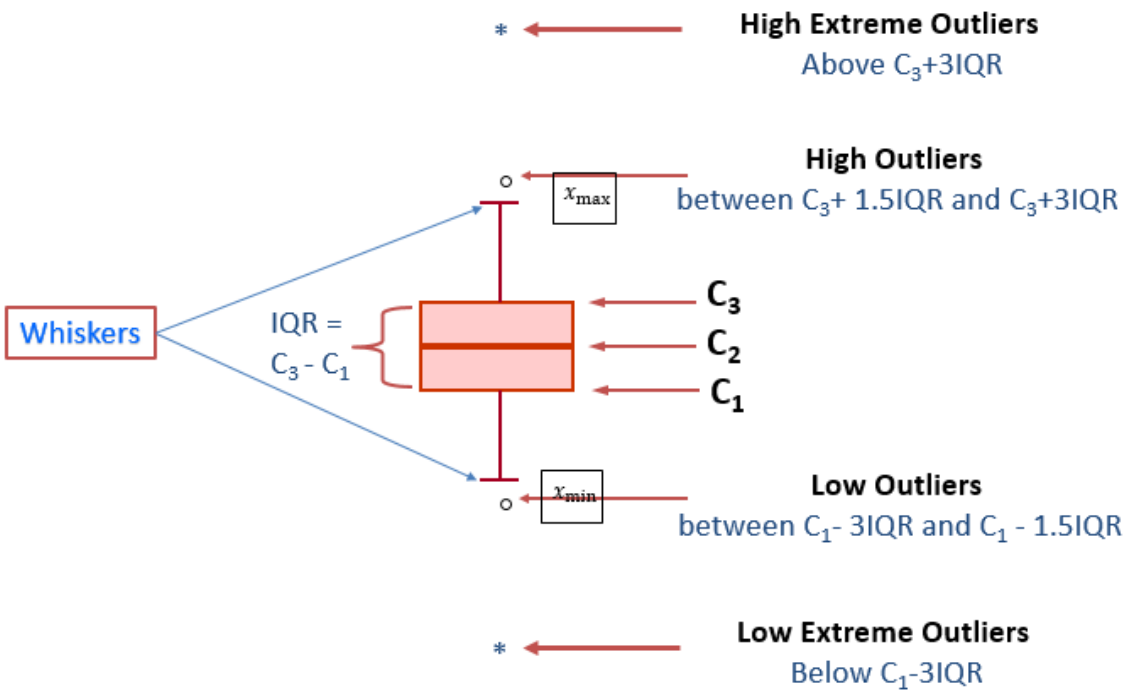
All the coefficients of both skewness and kurtosis, being relative measures, are invariant to changes in origin and scale. Skewness and kurtosis do not depend on the units, nor on the origin.

The coefficient of skewness and the coefficient of kurtosis are invariant for linear changes (because they are relative measures). They are also dimensionless measures.

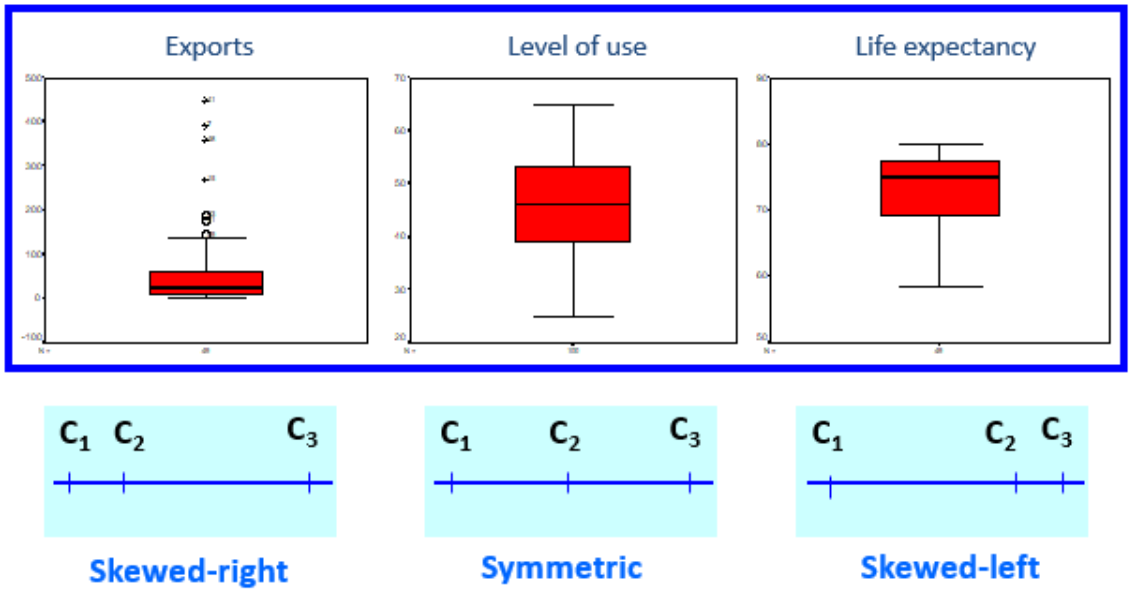
BOX-PLOTS

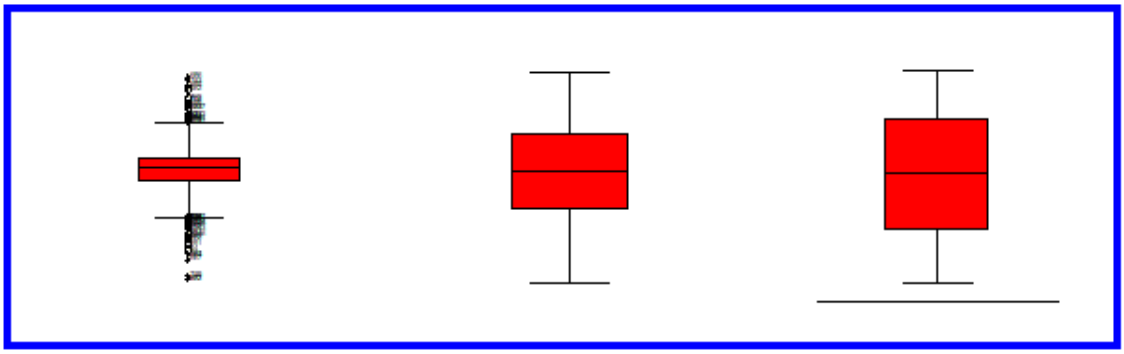
A box-plot is a chart with a central box indicating the range in which the middle 50% of the data is located. Its extremes are, therefore, the first and third quartiles of the distribution. Inside the box the position of the median is represented by a line. The lines coming out of the edges of the box are the so-called whiskers and reach the minimum and maximum values after the outliers have been removed.

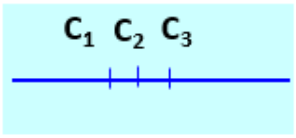
Outliers are indicated punctually using special symbols beyond the whiskers. A data is considered a weak outlier if its value is more than 1.5 times and less than 3 times the interquartile range from the edge of the box. A data is considered a strong or extreme outlier if its value is more than 3 times the interquartile range away from the box.



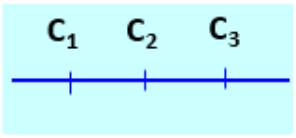
In view of the appearance of the box plot, some aspects of the numerical description of the distribution can also be concluded, such as the degree of dispersion (based on the magnitude of the range and the interquartile range), the skewness (based on the position of the median with respect to the quartiles) or the peakedness of the distribution.



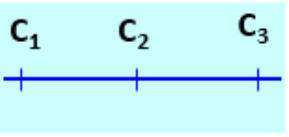




Leptokurtic



Mesokurtic



Platykurtic