

The very first thing that has to be written is (depending on the size) an introduction, or an abstract, or a comprehensive title, or a statement of the objective, or ... of the report :

Objective : Using a **SPREADSHEET** software running on a computer,
work out **STATISTICAL MEASURES** and **COMPARE** sets of data.

We will be using *OpenOffice calc v3.2*

The number of viewers of each episode of the first four seasons of LOST, an American drama TV series, are shown in four different tables.

Open a blank spreadsheet, enter the data in four separate columns.

Use the « Chart wizard » to represent each set of data by a separate scatter plot.

Use the « Function wizard » to compute the mean, median, 1st and 3rd quartiles for each distribution.

Use all these tools and values to compare the four seasons. Which season was the most successful, which one pleased – and kept – its viewers the most ? The answer should be detailed, motivated and nuanced.

	Season 1	Season 2	Season 3	Season 4
Mean	18.38	18.92	13.75	14.62
Median	18.3	19.1	12.65	14.1
1 st quartile	17.18	16.4	12.13	13.4
2 nd quartile	18.3	19.1	12.65	14.1
3 rd quartile	19.05	21.45	15.7	15.4
interquartile range	1.88	5.05	3.58	2

Static observations

Chronology is withdrawn from the study, because of setting the data in increasing order of size.

The highest mean indicates the season globally the most successful: season 2 has been the most watched, and considering that « to be watched » is equivalent to « to be successful », seems to be the most successful.

The absolute maximum for all 4 seasons is 23.5 M and corresponds to the 1st episode of season 2.

The means for seasons 3 and 4 are significantly smaller than those for the first two seasons, it shows that maybe viewers of season 2 did not like it as much as season 1, and many stopped watching.

Comparing mean and median.

Mean and median are not the same measurement.

It is true that both hint about the sizes of the values that may be found within the distribution.

However, it is comparing mean and median that brings the intended piece of information.

If both are the same, it indicates that the values are globally balanced about the mean.

If the median is bigger than the mean, it indicates that values are unbalanced : half the values or more lie above the mean, and most important, the more values lie above the mean, the further the smallest values are from the mean.

As for seasons of LOST, the median (18.3) for season 1 is closer to its mean (18.38) than median for season 2 (19.1) is to its mean (18.92), and this indicates a wider spread for the values of season 2.

The interquartile range measures how close one to each other are the central half of all the figures. The smaller the spread, the more grouped together are the central half of the figures.

Computing the range between the minimum and maximum values brings some drawback : outliers (unnatural values because of mistake, accident, ...) may change this measure of the spread into misleading value.

By computing the interquartile range, we get rid of these outliers.

The interquartile range is the smallest for season 1, and this may indicate that season 1 kept its viewers the most.

At that point though, we still don't know whether there were more viewers at the beginning or at the end of the season.

Chronological observations

The chart for season 2 indicates that the number of viewers drops after eight episodes.

The number of viewers climbed steadily until episode eight of season 2, then decreased.

This may indicate that season 1 was the most appreciated, and this led many to watch the next season.

Trend line for season 1 slopes from lower left to upper right, though slightly, and this indicates that the number of viewers steadily increased until the end of the season.

Trend lines for the last three seasons are decreasing, they kept losing viewers.

For all seasons, its graphs shows that there were more viewers for the last episode than for the one before.

About the mean

The spreadsheet offers a « mean » function named « average ».

The genuine name for this function is « arithmetic mean », and it is defined from a sum as :

$$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n}$$

Other means exist, for instance, the « geometric mean » is defined from a product as :

$$\bar{x} = \sqrt[n]{\prod_{i=1}^{i=n} x_i}$$

About the graphs

When creating a graph, it fits in a rectangle with default width and length.

The default unit is the maximum value within the set of figures, divided by the length of the corresponding side of the rectangle.

When there are more than one set of data, and one graph is needed for each set, the default behaviour of the creating process often produces misleading graphs.

One must be careful reshaping the frames in order to set a common unit for all graphs.

An easy way to avoid misleading graphs may be drawing all graphs within a single frame, so that common units are used on both axes for all sets of data.

About the mode

Season 1 data show 6 modal values, but the « mode » function is unable to display more than one, and the first modal encountered (the smallest) is displayed.

Season 4 data gathers 13 distinct values, so there is no mode, and the function « mode » displays an error instead of displaying no mode.

The mode is of no significant use for this study.

[About the quartiles](#)

Suppose the distribution is ordered in increasing order.

The 1st quartile is found, according to the French cursus, through the following steps :

Divide the total number of values by 4, round down to the nearest whole number.

The first quartile is the value at that position in the distribution.

For season 1, there are 24 values, divided by 4 is 6. The sixth value happens to be 17.1.

But we observe that the software computes 17.18 for that 1st quartile.

In fact, the quartiles are weighted means of the two values closest to a quarters way through.

For something different, we observe that the median is the same as the 2nd quartile.