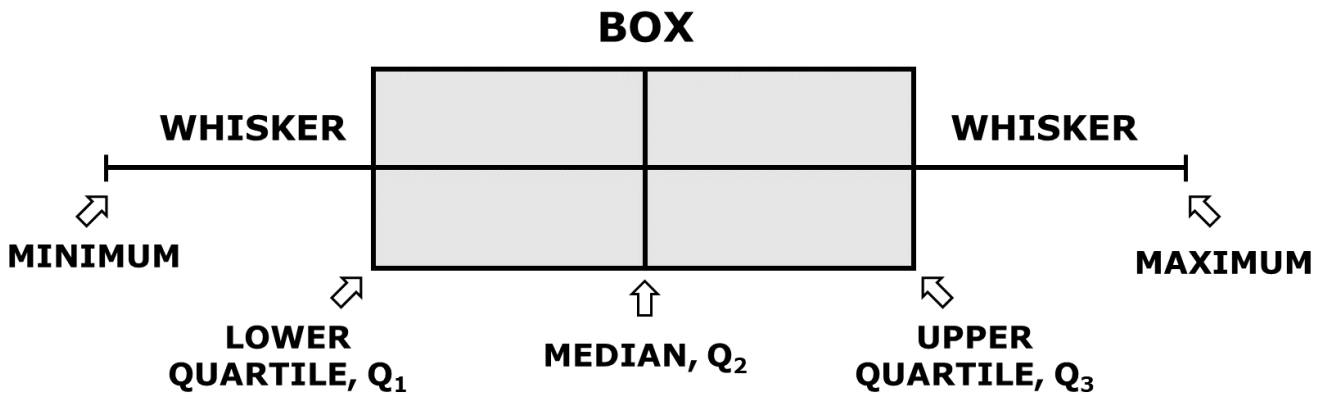
4.5.2 Visualizing the box and whisker plot

The box and whisker plot, sometimes simply called the box plot, is a type of graph that help visualize the five-number summary. It doesn't show the distribution in as much detail as histogram does, but it's especially useful for indicating whether a distribution is skewed and whether there are potential unusual observations (outliers) in the data set. A box plot is ideal for comparing distributions because the centre, spread and overall range are immediately apparent.

Figure 4.5.2.1 shows how to build the box and whisker plot from the five-number summary.

Figure 4.5.2.1 Building a box and whisker plot



In a box and whisker plot:

- The left and right sides of the box are the lower and upper quartiles. The box covers the interquartile interval, where 50% of the data is found.
- The vertical line that split the box in two is the median. Sometimes, the mean is also indicated by a dot or a cross on the box plot.
- The whiskers are the two lines outside the box, that go from the minimum to the lower quartile (the start of the box) and then from the upper quartile (the end of the box) to the maximum.
- The graph is usually presented with an axis that indicates the values (not shown on figure 4.5.2.1).
- The box and whisker plot can be presented horizontally, like in figure 4.5.2.1, or vertically.

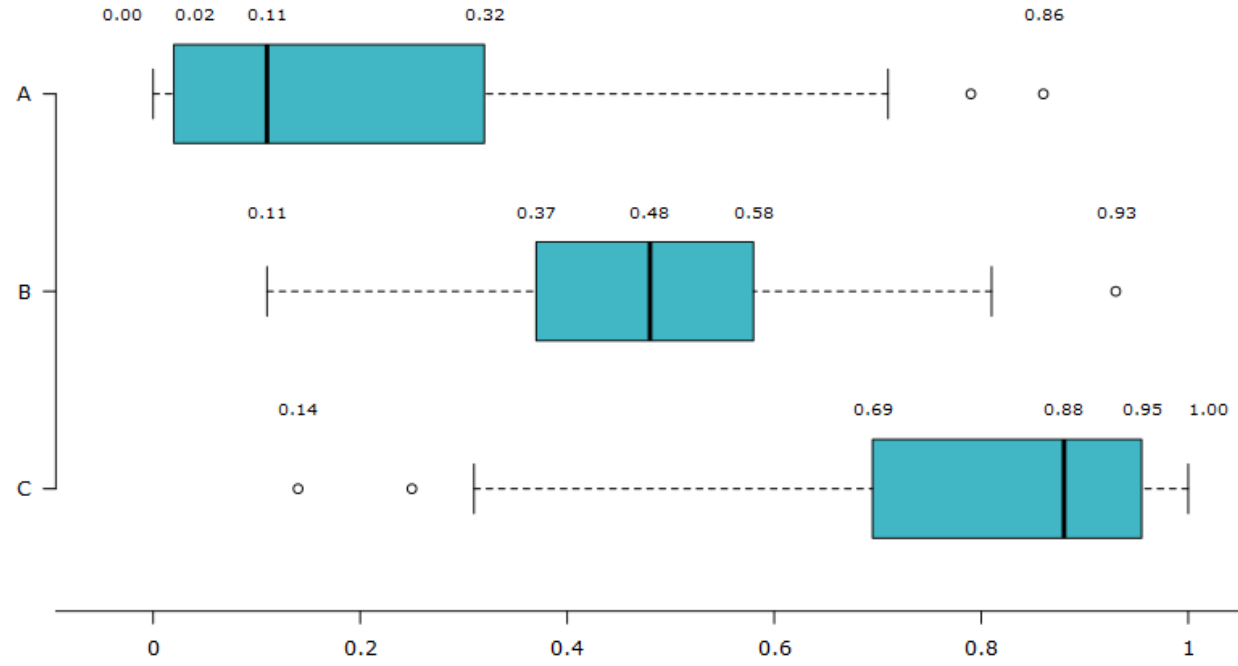
A variation of the box and whisker plot restricts the length of the whiskers to a maximum of 1.5 times the interquartile range. That is, the whisker reaches the value that is the furthest from the centre while still being inside a distance of 1.5 times the interquartile range from the lower or upper

quartile. Data points that are outside this interval are represented as points on the graph and considered potential outliers.

Example 1 – Comparison of three box and whisker plots

The three box and whisker plots of chart 4.5.2.1 have been created using R software. What can you say about the three distributions?

Chart 4.5.2.1
Box and whisker plots and five-number summaries of distributions A, B and C



- The centre of distribution A is the lowest of the three distributions (median is 0.11). The distribution is positively skewed, because the whisker and half-box are longer on the right side of the median than on the left side.

- Distribution B is approximately symmetric, because both half-boxes are almost the same length (0.11 on the left side and 0.10 on the right side). It's the most concentrated distribution because the interquartile range is 0.21, compared to 0.30 for distribution A and 0.26 for distribution C.
- The centre of distribution C is the highest of the three distributions (median is 0.88). The distribution C is negatively skewed because the whisker and half-box are longer on the left side of the median than on the right side.

All three distributions include potential outliers. Let's take distribution A, for example. The interquartile range is Q3 - Q1 = 0.32 – 0.02 = 0.30. According to the definition used by the function in R software, all values higher than Q3 + 1.5 x (Q3 - Q1) = 0.32 + 1.5 x 0.30 = 0.77 are outside the right whisker and indicated by a circle. There are two potential outliers in distribution A.