**A1We Using Excel to create a box plot**

We have already discussed techniques for visually representing data (see histograms and frequency polygons). In this section, we present another important method, called box plots (also known as box and whisker plot). A **box plot** is a graphical method of displaying the symmetry or skewness in a data set. It shows a measure of central location (the median), two measures of dispersion (the range and inter-quartile range), the skewness (from the orientation of the median relative to the quartiles) and potential outliers.

**Example**

Consider the student results obtained in a statistics examination as presented in Table 1.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 73 | 73 | 73 | 73 | 73 | 73 | 73 | 73 | 73 | 73 |
| 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 | 78 |
| 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 |
| 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 | 75 |

Table 1

Figure 1 illustrates the box-and-whisker plot for the quantitative marks example, where the summary statistics are as follows: first quartile Q1 = 66, minimum = 61, median = 70, maximum = 82 and third quartile Q3 = 74.



Figure 1

The box-and-whisker plot shows that the lowest 25% of the statistics marks are less spread out than the highest 25% of the distribution. The plot also shows that the other half are approximately equally spread out. This corresponds to the five-number-summary analysis in the previous section.

**To identify symmetry**

The box plot is interpreted as follows:

If the median within the box is not equidistant from the whiskers (or hinge), then the data is skewed. The box plot indicates right skewness because the distance between the median and the highest value is greater than the distance between the median and the lowest value. Furthermore, the top whisker (distance between Q3 and maximum) is longer than the lower whisker (distance between Q1 and minimum).

**To identify outliers**

The box plot is interpreted as follows:

The minimum and maximum points (or whiskers) are identified and enable identification of any extreme values (or outliers). A simple rule to identify an outlier (or suspected outlier) is that the whisker (maximum value - minimum value) should be no longer than three times the length of the box (Q3 – Q1). In this case the value of maximum – minimum is 21 and 3\*(Q3 – Q1) is 24. The conclusion is that extreme values are not present in the data set.

In conclusion, the middles 50% of the distribution is somewhat right skewed.

To create a box-and-whisker plot, two different Excel methods can be used, as well as a SPSS method.

**Excel Solution**

**Method 1** Excel standard chart method – this will re-create Figure 2

Input data into Cells D4:E9 and highlight D4:E9 (include the labels)



Figure 2

Select Insert > Insert Line or Area Chart



Figure 3

Choose option 4 Line with Markers



Figure 4

Click on chart and Select Design > Select Data > click on Switch Row/Column



Figure 5

Click OK



Figure 6

Now we can add the data point legend to the chart.

Click on the chart and Select Add Chart Element

Choose Legend and select Right



Figure 7

Remove line through legend

Right-click on a data point on the chart and Select Format Data Series

Choose No Line



Figure 8

Repeat for all data points on the chart.



Figure 9

Add lines to the data points

Click on Chart and Select Design > Add Chart Elements

Choose Lines > High-Low Lines



Figure 10

Add box to the data points

Click on Chart and Select Design > Add Chart Elements

Choose Up/Down Bars > Up/Down Bars



Figure 11

The final step is the add the chart, vertical axis title, and horizontal axis title.

Click on chart and Select Design > Add Chart Element

Select Axis Titles > Primary Vertical and call it ‘Stats mark’.

Click on chart and Select Design > Add Chart Element

Select Axis Titles > Primary Horizontal and call it ‘Statistic’.

Click on chart and Select Design > Add Chart Element

Select Chart Title > Above Chart and call it ‘Box plot for statistics data’.

Finally, I’ve removed the chart border, reduced the vertical axis from 0 – 90 to 50 – 90, and changed the box colour to black to enable clear identification on the box plot graph illustrated in figure 12.



Figure 12

From the box plot we observe that the statistics data set is approximately symmetrical.