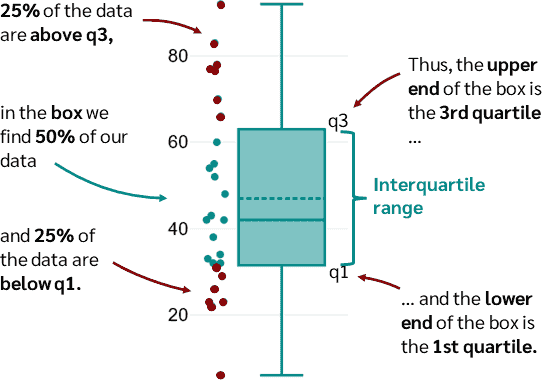
## 

## Creative Coding

## Lab #4 (Box and Whisker Plots)

**What is a box and whisker plot?**



A box and whisker plot (or boxplot) displays the ***median***, the ***quartiles***, the ***range*** of values covered by the data and any ***outliers*** which may be present. It gives a clear picture of all these features and allows a visual appreciation of ***symmetry*** or ***skewness***.

**Note on Skewness:**

We can work out the skew using the function ***skew*** in Excel.

As a general rule of thumb: If skewness is less than -1 or greater than 1, the distribution is highly skewed. If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed. If skewness is between -0.5 and 0.5, the distribution is approximately symmetric.

**Example 1:**

The β endorphin concentrations (in pmol/l) recorded for eleven runners who collapsed after the Great North Run are as follows (written in order of increasing size).

|  |
| --- |
| 66 |
| 72 |
| 79 |
| 84 |
| 102 |
| 110 |
| 123 |
| 144 |
| 162 |
| 169 |
| 414 |
|  |

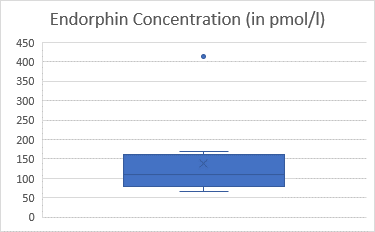
*(Data sourced from Dale, G., Fleetwood, J.A., Weddell, A., Ellis, R.D. and Sainsbury, J.R.C. (1987) Beta-endorphin: a factor in 'fun run' collapse? British Medical Journal,****294****, 1004.*

**Questions:**

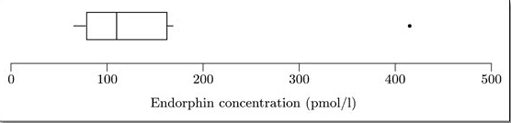
1. Calculate the summary statistics for this data set – mean, median, Q1, Q3, IQR, minimum, maximum, skew, lower adjacent value, upper adjacent value, outliers if any.
2. Construct the box plot.
3. Comment on the distribution – for example, if it is symmetrical or skewed etc.

To construct a boxplot in Excel, select the data, then choose ‘Insert’ and from the ‘Charts’ panel, choose ‘Box and Whisker’.

**Box plot for the collapsed runners:**



OR



Note that the median and quartiles are used to construct the ‘box’.

The ***‘box’*** is a rectangle with edges defined by the lower and upper quartiles; so it indicates where the ‘middle 50%’ of the data can be found. The horizontal line inside the box is located at the median. The X is located at the mean.

The ***‘whiskers’*** are constructed next. These are lines drawn parallel to the scale (so they are vertical in this case). Essentially, each whisker extends outwards from the edge of the box as far as the most extreme observation. However, as you will see in the next step, some observations may be classified as potential outliers; and in fact the whiskers extend only to cover observations which are not classified as potential outliers.

**Summary of data:**

|  |  |
| --- | --- |
| mean | 138.6364 |
| median | 110 |
| Q1 | 79 |
| Q3 | 162 |
| IQR | 83 |
| minimum | 66 |
| maximum | 414 |
| skew | 2.572002 |

**How do we find the outliers (if they exist)?**

The whiskers are drawn outwards as far as observations called ***adjacent values***. The ***lower adjacent value*** is the furthest observation which is within one and a half *iqr* (interquartile range) of the lower end of the box; and the ***upper adjacent value*** is the furthest observation which is within one and a half *iqr* of the upper end of the box. So, the interquartile range is needed to construct the whiskers.

|  |  |
| --- | --- |
| Are there any outliers? |  |
| Q1 - 1.5\*IQR | -45.5 |
| Q3 + 1.5\*IQR | 286.5 |
|  |  |
| lower adjacent value | 66 (same as minimum) |
| upper adjacent value | 169 (different to maximum) |

In this case, Q1 – 1.5\*IQR = -45.5. Therefore, we stick with the minimum which is 66 and well inside the limits.

Q3 + 1.5\*IQR = 286.5. In this case we use 169 instead of the maximum = 414. Why? Because 414 is higher than 286.5 and so outside the limits. Therefore we say that 414 is an outlier.

**Spread and Shape:**

The length of the box represents the interquartile range and the lengths of the whiskers relative to the length of the box give an idea of how stretched out the rest of the values are. Thus, these aspects of the diagram give an idea of the dispersion of the data set.

The unusually large value in this data set is clearly shown and the median gives an assessment of the centre. These particular data are not symmetric since symmetric data will produce a box plot which is symmetric about the median. They are right skewed as the right-hand section of the box is longer than the left (or in this case the top is longer than the bottom). The skewness is 2.572 from Excel.

However, it should be taken into account that this particular data set has only eleven values, and this is too small a number to infer anything definite about any underlying structure.

**Example 2:**

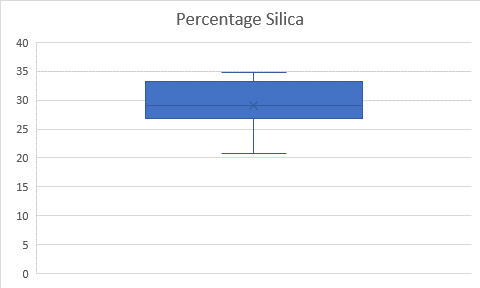
The table below contains data on the percentage of silica found in 22 chondrite meteors. The data are given in order of increasing size.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 20.77 | 22.56 | 22.71 | 22.69 | 26.39 | 27.08 | 27.32 | 27.33 |
| 27.57 | 27.81 | 28.69 | 29.36 | 30.25 | 31.89 | 32.88 | 33.23 |
| 33.28 | 33.4 | 33.52 | 33.83 | 33.95 | 34.82 |  |  |

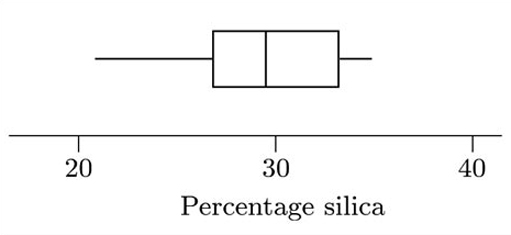
*(Source: Good, I.J. and Gaskins, R.A. (1980) Density estimation and bump-hunting by the penalized likelihood method exemplified by scattering and meteorite data. J. American Statistical Association,****75****, 42-56.)*

1. Calculate the summary statistics for this data set – mean, median, Q1, Q3, IQR, minimum, maximum, skew, lower adjacent value, upper adjacent value, outliers if any.
2. Construct the box plot.
3. Comment on the distribution – for example, if it is symmetrical or skewed etc.

**Boxplot for silica content of chondrite meteors**



OR

**

**Summary of data:**

|  |  |
| --- | --- |
| mean | 29.15136 |
| median | 29.025 |
| Q1 | 26.9075 |
| Q3 | 33.31 |
| IQR | 6.4025 |
| minimum | 20.77 |
| maximum | 34.82 |
| skew | -0.45304 |

|  |  |
| --- | --- |
| Are there any outliers? |  |
| Q1 - 1.5\*IQR | 17.30375 |
| Q3 + 1.5\*IQR | 42.91375 |
|  |  |
| lower adjacent value | 20.77 |
| upper adjacent value | 34.82 |

This boxplot is clearly not symmetrical. However, the pattern of its skewness is not straightforward.

The sample skewness is negative (= −0.446), indicating that the data are left-skew. To some extent the boxplot reflects this: the left whisker is considerably longer than the right, indicating that the smaller values are more spread out than are the larger values.

However, the box gives a different impression. The box corresponds to the middle half of the data values, and the line denoting the median divides this into two parts, each corresponding to one-quarter of the data. In this case, the left part of the box is shorter than the right part – i.e. right-skewed rather than left-skewed.

In this example, the sample skewness correlates with the pattern suggested by the whiskers of the boxplot (left-skew), rather than with that suggested by the box. Essentially, this occurs because all the values in the data set are used to calculate the sample skewness; and the calculation involves a sum of powers of values, so that the sample skewness is particularly affected by the more extreme values in the data set. In a boxplot, the whiskers correspond to the more extreme values. In this boxplot, the whiskers suggest that the data are left-skew, matching the sample skewness.

#### 

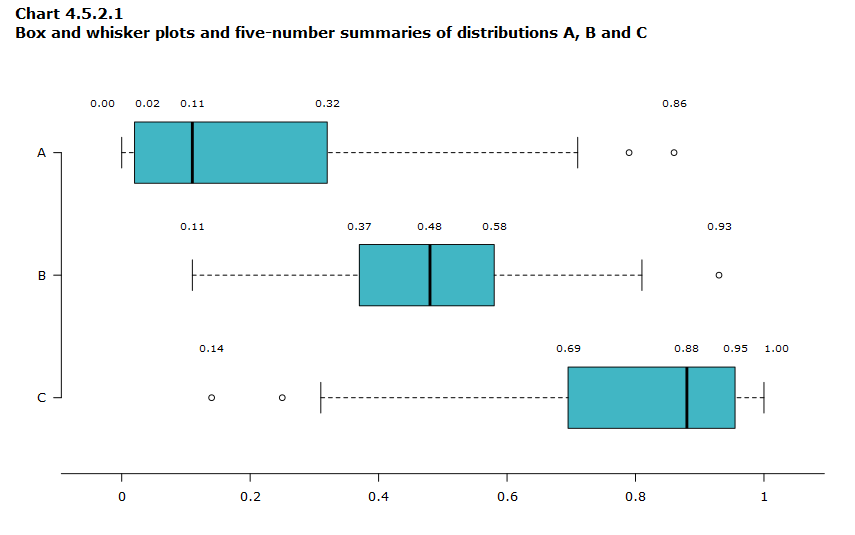
#### **Comparing Boxplots**

**Guidelines:**

1. Compare the respective medians, to compare location.
2. Compare the interquartile ranges (that is, the box lengths), to compare dispersion.
3. Look at the overall spread as shown by the adjacent values. (This is another aspect of dispersion.)
4. Look for signs of skewness. If the data do not appear to be symmetric, does each batch show the same kind of asymmetry?
5. Look for potential outliers.

After discussing these features, general conclusions should be summarized briefly.

**Example – Comparison of three box and whisker plots:**



|  |  |  |  |
| --- | --- | --- | --- |
| **Measurement** | **Distribution A** | **Distribution B** | **Distribution C** |
| **Minimum** | **0.00** | **0.11** | **0.14** |
| **Lower quartile (Q1)** | **0.02** | **0.37** | **0.69** |
| **Median (Q2)** | **0.11** | **0.48** | **0.88** |
| **Upper quartile (Q3)** | **0.32** | **0.58** | **0.95** |
| **Maximum** | **0.86** | **0.93** | **1.00** |

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.

**Question 1:**

The table below contains data on the sizes (numbers of children) of the completed families of two samples of mothers in Ontario. One sample of mothers had had fewer years of education than the other sample (six years or less for mothers in the first sample, and seven years or more for those in the other sample).

**Mothers educated for six years or less:**

14 13 4 14 10 2 13 5 0 0 13 3 9 2 10 11 13 5 14

**Mothers educated for seven years or more:**

0 4 0 2 3 3 0 4 7 1 9 4 3 2 3 2 16 6 0 13 6 6 5 9 10 5 4 3 3 5 2 3 5 15 5

*(Source: Keyfitz, N. (1953) A factorial arrangement of comparisons of family size. American J. Sociology,****53****, 470–480.)*

Based on this dataset:

1. Calculate the summary statistics in Excel.
2. Create the box plots in Excel.
3. Use the box plots to compare the distributions of the two data sets.

**Question 2:**

In a study of memory recall times, a series of stimulus words was shown to a subject on a computer screen. For each word, the subject was instructed to recall either a pleasant or an unpleasant memory associated with that word. Successful recall of a memory was indicated by the subject pressing a bar on the computer keyboard. The table below shows the recall times (in seconds) for twenty pleasant and twenty unpleasant memories.

Of key interest in this study was whether pleasant memories could be recalled more easily and quickly than unpleasant ones.

|  |  |
| --- | --- |
| **Pleasant memory** | **Unpleasant memory** |
| 1.07 | 1.45 |
| 1.17 | 1.67 |
| 1.22 | 1.9 |
| 1.42 | 2.02 |
| 1.63 | 2.32 |
| 1.98 | 2.35 |
| 2.12 | 2.43 |
| 2.32 | 2.47 |
| 2.56 | 2.57 |
| 2.7 | 3.33 |
| 2.93 | 3.87 |
| 2.97 | 4.33 |
| 3.03 | 5.35 |
| 3.15 | 5.72 |
| 3.22 | 6.48 |
| 3.42 | 6.9 |
| 4.63 | 8.68 |
| 4.7 | 9.47 |
| 5.55 | 10 |
| 6.17 | 10.93 |

*(Source: Dunn, G. and Master, D. (1982) Latency models: the statistical analysis of response times. Psychological Medicine,****12****, 659–665.)*

Based on this dataset:

1. Calculate the summary statistics in Excel.
2. Create the box plots in Excel.
3. Use the box plots to compare the distributions of recall times for the two types of memory.