

12.3 Box and Whiskers Plots

Box and Whiskers or Boxplots are most useful when comparing two or more sets of sample data. Differences in the centers and the spread of the datasets are clearly visible with a boxplot.

A boxplot gives a picture of the symmetry of a dataset and shows outliers very clearly. These features are important when deciding which summary statistics best describe the data set. For example, the boxplot can help a statistician decide whether the data is normally distributed, as we will discuss in Section 12.4.

To create the boxplot, we first split the data into fourths, or quartiles. The first quartile, q_1 , is the median of the bottom half of the data set; and the third quartile, q_3 , is the median of the upper half of the data set. If the data set includes an even number of values, then the data set can simply be split in half to find q_1 and q_3 . If the data set has an odd number of values, then the data should be split into two halves, *not including the median*.

For example, consider the data set $\{1, 2, 3, 3, 4, 5, 7, 8\}$ (problem 25, page 702).

The set has an even number of elements, so the median is $\frac{3+4}{2} = 3.5$. The top

and bottom halves also have an even number of elements, so that $q_1 = \frac{2+3}{2} = 2.5$

and $q_3 = \frac{5+7}{2} = 6$.

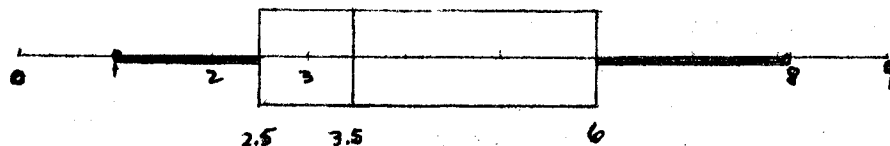
The next step in creating a boxplot of a data set is to create the set called the **five number summary**. This ordered set consists of the smallest (s), the first quartile (q_1), the median (m), the third quartile (q_3), and the largest (L) values in the data set, or the set $\{s, q_1, m, q_3, L\}$.

In our example above, the **five number summary** would be the ordered set $\{1, 2.5, 3.5, 6, 8\}$.

Notice that the **range**, discussed on page 705, is just $L - s$ from this set. A second measure of dispersion, which is easily calculated from this set, is the **Interquartile Range (IQR)**, which is $q_3 - q_1$. Note that the range gives the spread of the entire data set, and the IQR gives the spread of the middle half of the data set.

In our example, the range is $8 - 1 = 7$ and the IQR is $6 - 2.5 = 3.5$.

Finally, from the five number summary we can create a box and whiskers plot, by drawing vertical segments at the median and at q_1 and q_3 and horizontal segments connecting the tops and bottoms of the segments, to create a box around the middle half of the data. We then draw whiskers from q_1 to s and from q_3 to L .



Outliers: the length of the whiskers should be no longer than 1.5 times the IQR. So to draw the whisker above the 3rd quartile (q_3), draw it to the largest data value that is less than or equal to the value that is 1.5 IQR's above the 3rd quartile. Any data value larger than that should be marked as an outlier above, which is typically done by indicating the point with an asterisk. The process is repeated for values below the 1st quartile (q_1), with any values less than $(q_1 - 1.5 \times IQR)$ marked as outliers below.