

UNIT 5

Data Diagnostics

Objectives:

- To identify potential outliers with quantitative data
- To construct a modified boxplot

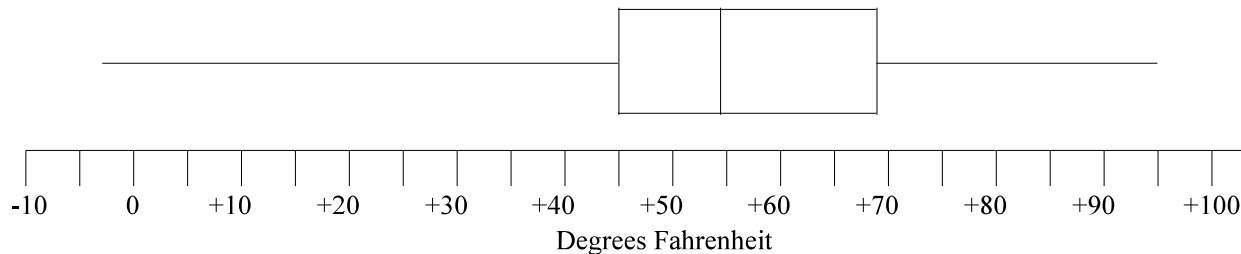
An initial and often useful step in analyzing quantitative data is to decide if there are any outliers. An *outlier* is a observation whose value is deemed to be either unusually high or unusually low relative to the other observations in the data set. Such observations may arise from some anomaly in the data collection process, or from some very unusual characteristic of the distribution of the variable, or from an error in recording data, etc. Since one or more outliers can potentially have a strong effect on the results of data analysis (i.e., we have already seen how the mean is sensitive to a few unusually large, or unusually small, observations), we need to consider carefully how to identify potential outliers and whether or not a potential outlier should be removed from the data; it is generally not wise simply to remove an observations from a data set simply because at first glance it looks like it does not belong.

There is no universal method for identifying outliers that will be best in all circumstances. Experience has suggested the following rough rule of thumb: any value which is a distance of more than $1.5(IQR)$ below Q_1 or above Q_3 is a potential outlier. Rather than throwing such an observation away, the next step, if possible, should be an investigation into the reason why the observation appears to be unusual. If there was an error, it might be possible to correct the error. If we find some other reason for the unusual value, we may obtain some insight into the distribution of the variable under consideration, or we may find some justification for not including the observation in data analysis.

To illustrate, let us imagine that the high temperature in a city is recorded in degrees Fahrenheit for each day in March, with the results displayed as the following ordered array for the sake of convenience:

-3 27 30 36 40 41 45 45 48 50 51 51 51 52 54 55 56 58 59 62 65 68 69 72 74 76 78 79 79 95

From this ordered array, it is easy to obtain the five-number summary, which (as you should verify) is -3, 45, 54.5, 69, 95. From the five-number summary, we construct the following boxplot:

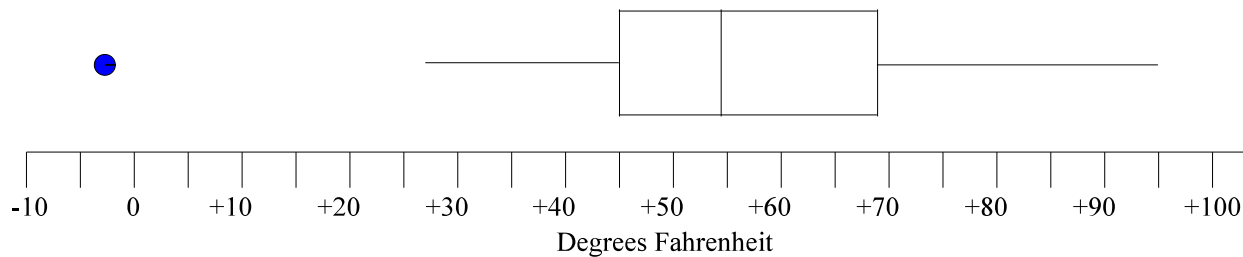


It is easy to verify that $IQR = 69 - 45 = 24$ and $1.5(IQR) = 36$. The two extreme observations are the temperature of $Min = -3$ degrees Fahrenheit and the temperature of $Max = 95$ degrees Fahrenheit. Since $Min = -3$ is more than a distance of $1.5(IQR) = 36$ below $Q_1 = 45$, our rule of thumb suggests that the temperature of -3 is a potential outlier. Since 27 is not more than a distance of $1.5(IQR) = 36$ below $Q_1 = 45$, our rule of thumb suggests that the temperature of 27 is not a potential outlier. Also, since $Max = 95$ is not more than a distance of $1.5(IQR) = 36$ above $Q_3 = 69$, our rule of thumb suggests that the temperature of 95 is not a potential outlier. Consequently, the temperature of -3 degrees Fahrenheit is the only potential outlier.

If we have verified that the temperature of -3 is not an error in recording data, we may still decide to exclude this temperature from future data analysis. If there were some special, rare circumstances that gave rise to the unusually low temperature, we may need to assess how this unusually low value will affect our conclusions. Since no strict rule for removing an observation from a data set exists, ultimately, one's judgment plays an important role in this decision, which should be made only after careful consideration.

We can indicate when data contain potential outliers by modifying the boxplot of the data. Instead of extending the lines from the box all the way to the minimum and all the way to the maximum, we stop the lines at the largest and smallest values which are not potential outliers; all potential outliers are then designated with dots. A boxplot constructed in this way is called a *modified boxplot*. If we were to construct a modified boxplot of the high temperatures in the city recorded for each day in March, the lower line would extend only down to 27 degrees Fahrenheit instead of all the way down to -3 degrees Fahrenheit; a dot would be used to

designate -3 as a potential outlier. The modified boxplot would look as follows:



Self Test Problem 5-1. Figure 2-7 displays a boxplot of orange weights for each of five types of oranges. Consider the Type B oranges and the Type C oranges.

- (a) From the boxplot for type B oranges, obtain the five-number summary, and explain why there must be at least two potential outliers.
- (b) From the boxplot for type C oranges, obtain the five-number summary, and explain why there are no potential outliers.

**Figure 5-1
Rope Breaking
Strengths (lbs.)**

6	02
7	5
8	4
9	368
10	22568
11	0347
12	48

Self Test Problem 5-2. Figure 5-1 is a stem-and-leaf display of the breaking strength in pounds for several pieces of a type of rope.

- (a) Obtain the five-number summary and the interquartile range.
- (b) Identify all potential outliers.
- (c) Construct a modified boxplot.

Answers to Self Test Problems

- 5-1** (a) The five-number summary is 3.2, 5.4, 6.0, 6.6, 8.8; $IQR = 6.6 - 5.4 = 1.2$. Since $Min = 3.2$ is more than a distance of $1.5(IQR) = 1.8$ below $Q_1 = 5.4$, and $Max = 8.8$ is more than a distance of $1.5(IQR) = 1.8$ above $Q_3 = 6.6$, the rule of thumb suggests that each of 3.2 and 8.8 is a potential outlier. (b) The five-number summary is 4.6, 5.3, 6.0, 7.4, 8.8; $IQR = 7.4 - 5.3 = 2.1$. Since $Min = 4.6$ is not more than a distance of $1.5(IQR) = 3.15$ below $Q_1 = 5.3$, and $Max = 8.8$ is not more than a distance of $1.5(IQR) = 3.15$ above $Q_3 = 7.4$, the rule of thumb suggests that neither the minimum nor the maximum is a potential outlier; consequently, no other observation can possibly be an outlier.
- 5-2** (a) The five-number summary is 60, 93, 103.5, 113, 128; $IQR = 113 - 93 = 20$. (b) Since 60 and 62 are each more than a distance of $1.5(IQR) = 30$ below $Q_1 = 93$, each of 60 and 62 is a potential outlier. Since 128 is not more than a distance of $1.5(IQR) = 30$ above $Q_3 = 113$, there are no other potential outliers. (c) The modified boxplot should be constructed with a box ranging from 93 to 113 with the line dividing the box at 103.5; the lines from extending from the box should stop at 75 and 128, and dots should be placed above 60 and 62.

Summary

An initial and often useful step in analyzing data is to decide if there are any *outliers*. An outlier is an observation whose value is deemed to be either unusually high or unusually low relative to the other observations in the data set. Experience has suggested that any value which is a distance of more than $1.5(IQR)$ below Q_1 or above Q_3 is a potential outlier; however, there is no universal method for identifying outliers that will be best in all circumstances. Once we have verified that an outlier did not result from an error in recording data, we may then try to see what information the potential outlier provides about the distribution of the variable under consideration, and we must decide whether or not there is any justification for not including the potential outlier in data analysis.

A *modified boxplot* can be used to display outliers. Instead of extending the lines from the box all the way to the minimum and all the way to the maximum, we stop the lines at the largest and smallest values which are not potential outliers; all potential outliers are then designated with dots.