Chapter 2

Data Description

This chapter discusses how to describe data by reporting measures of the center of a data set, by reporting measures of spread for a data set, and by making various visual representations of a data set. It also briefly discusses how to relate an individual data point to the data set as a whole.

2.1 Vocabulary

Make sure you understand each of the following terms well enough to use them appropriately when describing your own data:

- frequency
- histogram
- stem-and-leaf diagram
- mean (or average)
- median
- interquartile range (IQR)
- box plot
- box and whiskers plot
- variance
- standard deviation
- z-score
- outlier

2.2 Specific notes

- p. 9: A dot plot is just a histogram (with a very small bin size).
- p. 10: A frequency table is just a collection of the set of numbers needed to create a histogram.

- pp. 9-13: These pages are showing different ways to visually 'see' data.

- pp. 14-18: Here is where the text discusses measures of the center of a data set.

- p. 16: Make sure you understand how to translate the complex-looking formula for the mean into the reality of a simple operation – add the data points up and divide by the number of points.

- pp. 19-23: Measures of spread are discussed.

- p. 22: The 'technical reasons' for using $n-1$ rather than n in the calculation of the variance (and hence the standard deviation) involve degrees of freedom. If you're into math, look it up in another text. It's not that hard to understand.

- pp. 22-23: Although you will likely never have to compute the standard deviation or variance by hand, you should know how they are calculated. The standard deviation essentially measures of the average distance of the data from the mean. With that in mind, make use of a calculator or computer to determine the actual value of the variance or standard deviation for any data set.

- p. 24: The z-score is almost hidden here at the end of the chapter, but it is an important concept that will return in several guises throughout the rest of the book. So take some time right now to figure out what it is and how to calculate it.

- p. 25: The empirical rule listed here is a good one to memorize.

2.3 Additional information

One common measure of a data set which *The Cartoon Guide to Statistics* leaves out is the **mode** which is the technical name for the most common value of the data set. So, for example, in the weight data discussed at the beginning of the chapter, the data set has two modes – 150 pounds and 155 pounds. Each of those values occurs ten times in the data set, and no other value occurs more often.

Also, as a measure of spread, a very simple range can be calculated by subtracting the minimum value from the maximum value. This range turns out not to be very useful for many data sets because it includes outliers.

2.4 Exercises

Exercise 2.1: Describe the data for just the male Penn State students listed on p. 9.

(a) Report the mean, median, and mode.

(b) Make a frequency table and histogram.

(c) Make a stem-and-leaf diagram.

(d) Report the interquartile range, variance, and standard deviation.

(e) Make a box and whiskers plot.

(f) What percentage of the data points are within one standard deviation of the mean?

(g) What percentage of the data points are within two standard deviations of the mean?

(h) What is the z-score for each of the following weights: 130 lb, 150 lb, 195 lb ?

Exercise 2.2: Describe the data for just the female Penn State students listed on p. 9.

(a) Report the mean, median, and mode.

(b) Make a frequency table and histogram.

(c) Make a stem-and-leaf diagram.

(d) Report the interquartile range, variance, and standard deviation.

(e) Make a box and whiskers plot.

(f) What percentage of the data points are within one standard deviation of the mean?

(g) What percentage of the data points are within two standard deviations of the mean?

(h) What is the z-score for each of the following weights: 110 lb, 130 lb, 150 lb ?