Sheet 1: Data Collection and Descriptive Statistics

NOTE: All results should be rounded to two decimal places unless otherwise stated. If a number or result has fewer decimal places, it is okay to keep fewer.

Exercise 1

[D., p. 50, Exercise 79]

Exercise 2

In the 2017 New York City Housing and Vacancy Survey, randomly selected households were asked about their borough, their monthly rent, their monthly income, the number of bedrooms in their unit, the presence of person elevators, among many other variables.

(a) Indicate which of the listed variables are numerical and which are categorical. If they are numerical, are they continuous or discrete? If they are categorical, list the corresponding categories (= possible values).

(b) Describe one procedure that would have made sure that the households were, at least, approximately, selected according to a simple random sampling.

(c) For the following three alternative procedures, explain why they would have very likely resulted in biased samples:

(1) Ask randomly selected Columbia students and faculty.

(2) Ask randomly selected NYC residents on Facebook and Instagram.

(3) Ask randomly selected people at randomly selected NYC supermarkets.

Exercise 3

[D, Section 1.3, Exercise 34ab]

Exercise 4

[D, Section 1.2, Exercise 19]

Exercise 5

[D, Section 1.4, Exercise 44]

Exercise 6

[D, Section 1.4, Exercise 51]

Exercise 7

Draw a boxplot for the discoveries data that we discussed in class. To do so, first determine the lower quartile, the median, the upper quartile, the LOB, the UOB, the set of outliers, the smallest data point that is not an outlier and the largest data point that is not an outlier.

Exercise 8

For each landmass exceeding 10,000 square miles, the following table lists its name along with the logarithm of its area in thousands of square miles. You can use that $\widetilde{x} = 3.71$, $Q_1 = 2.99$ and $Q_3 = 5.21$.

(a) Construct a histogram with class intervals $]2,3],]3,4],\ldots,]9,10]$ and relative frequency on the vertical axis. Is the histogram unimodal, bimodal, or multimodal?

(b) Determine the LOB, the UOB, the set of outliers, the smallest data point that is not an outlier and the largest data point that is not an outlier.

(c) Based on this information, draw a boxplot of the data.

```
##           Africa       Antarctica             Asia        Australia
##         9.350624         8.612503         9.740262         7.995644
##     Axel Heiberg           Baffin            Banks           Borneo
##         2.772589         5.214936         3.135494         5.634790
##          Britain          Celebes            Celon             Cuba
##         4.430817         4.290459         3.218876         3.761200
##            Devon        Ellesmere           Europe        Greenland
##         3.044522         4.406719         8.228177         6.733402
##           Hainan       Hispaniola          Hokkaido           Honshu
##         2.564949         3.401197         3.401197         4.488636
##          Iceland          Ireland             Java           Kyushu
##         3.688879         3.496508         3.891820         2.639057
##            Luzon       Madagascar          Melville         Mindanao
##         3.737670         5.424950         2.772589         3.583519
##         Moluccas      New Britain       New Guinea  New Zealand (N)
##         3.367296         2.708050         5.723585         3.784190
##  New Zealand (S)     Newfoundland    North America    Novaya Zemlya
##         4.060443         3.761200         9.147401         3.465736
##  Prince of Wales         Sakhalin    South America      Southampton
##         2.564949         3.367296         8.823942         2.772589
##      Spitsbergen          Sumatra           Taiwan         Tasmania
##         2.708050         5.209486         2.639057         3.258097
## Tierra del Fuego            Timor        Vancouver         Victoria
##         2.944439         2.564949         2.484907         4.406719
```