Andrew Moore, 08/27/2021

MATH-471, Homework 1

This document was prepared using the `knitr` and `rmarkdown` packages, with R version 4.1.0 (2021-05-18).

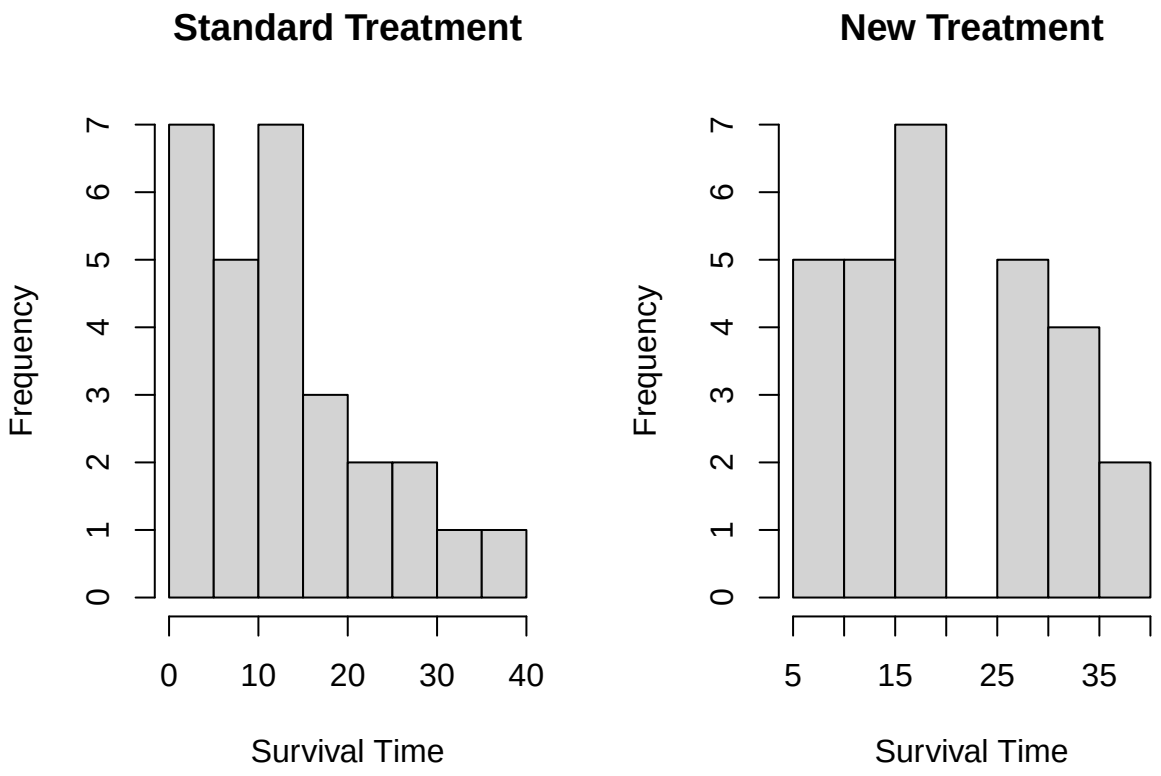
3.7

The survival times (in months) for two treatments for patients with severe chronic left ventricular heart failure are given in the following tables.

```
standard <- c(
  4, 15, 24, 10, 1, 27, 3, 14, 2, 16, 32, 7, 13, 36, 29, 6, 12, 18, 14, 15, 1,
  6, 13, 21, 20, 8, 3, 2
)

new_therapy <- c(
  5, 20, 29, 15, 7, 32, 36, 17, 15, 19, 35, 10, 16, 39, 27, 14, 10, 16, 12, 13, 16,
  9, 18, 33, 30, 29, 31, 27
)
```

a. Construct separate relative frequency histograms for the survival times of both the therapies.

```
par(mfrow = c(1, 2))
hist(standard, main = "Standard Treatment", xlab = "Survival Time")
hist(new_therapy, main = "New Treatment", xlab = "Survival Time")
```



b. Compare the two histograms. Does the new therapy appear to generate a longer survival time? Explain your answer.

The new therapy seems to be somewhat more effective at increasing the survival times of patients. A higher relative frequency of patients are reaching survival times of 25 months or more. Both groups of patients have the same standard deviation for survival time, but the *new therapy* group has a higher average (and median) survival time.

```
data.frame(
  Measure = c("Sample Mean", "Sample Median", "Sample Standard Deviation"),
  Standard_Therapy = round(c(mean(standard), median(standard), sd(standard)), 2),
  New_Therapy = round(c(mean(new_therapy), median(new_therapy), sd(new_therapy)), 2)
)
```

```
##                     Measure Standard_Therapy New_Therapy
## 1               Sample Mean            13.29       20.71
## 2             Sample Median            13.00       17.50
## 3 Sample Standard Deviation             9.81        9.81
```

3.26

Pushing economy and wheelchair-propulsion technique were examined for eight wheelchair racers on a motorized treadmill in a paper by Goosey and Campbell [Adapted Physical Activity Quarterly (1998) 15:36–50]. The eight racers had the following years of racing experience:

Racing experience (years): 6, 3, 10, 4, 4, 2, 4, 7

a. Verify that the mean years of experience is 5 years. Does this value appear to adequately represent the center of the data set?

$$\bar{y} = \frac{(2+3+4+4+4+6+7+10)}{8} = \frac{40}{8} = 5$$

I would say that the center is probably closer to 4.5, and that the presence of the two races with 2 and 10 years of experience are distorting what we see from the sample mean.

b. Verify that $\Sigma_i(y_i - \bar{y})^2 = \Sigma_i(y_i - 5)^2 = 46$.

We know from part a. that the sample mean, $\bar{y}$ is 5. Representing y as a column vector, we can calculate the sum:

$$y = \begin{bmatrix} 6 \\ 3 \\ 10 \\ 4 \\ 4 \\ 2 \\ 4 \\ 7 \end{bmatrix}$$

$$\Sigma_i(y_i - 5)^2 = \Sigma \begin{bmatrix} (6-5)^2 \\ (3-5)^2 \\ (10-5)^2 \\ (4-5)^2 \\ (4-5)^2 \\ (2-5)^2 \\ (4-5)^2 \\ (7-5)^2 \end{bmatrix} = \Sigma \begin{bmatrix} 1 \\ 4 \\ 25 \\ 1 \\ 1 \\ 9 \\ 1 \\ 4 \end{bmatrix} = 46$$

This value is the *sum of squared deviations* from the mean $\bar{y}$.

c. Calculate the sample variance and standard deviation for the experience data. How would you interpret the value of the standard deviation relative to the sample mean?

The sample variance, s_y^2, is the sum of squared deviations from the mean, divided by the sample size (minus 1). There are 8 racers, so our sample size is $n = 8$.

$$s_y^2 = \frac{1}{n-1}\Sigma_i(y_i - \bar{y})^2 = \frac{46}{7} \approx 6.57$$

The sample standard deviation, s_y, is the square-root of the variance.

$$s_y = \sqrt{\frac{1}{n-1}\Sigma_i(y_i - \bar{y})^2} = \sqrt{\frac{46}{7}} \approx 2.56$$

Having computed both the mean and standard deviation, I would say that the average racer in our sample has 5 years of experience, plus or minus about 5 years.

3.30

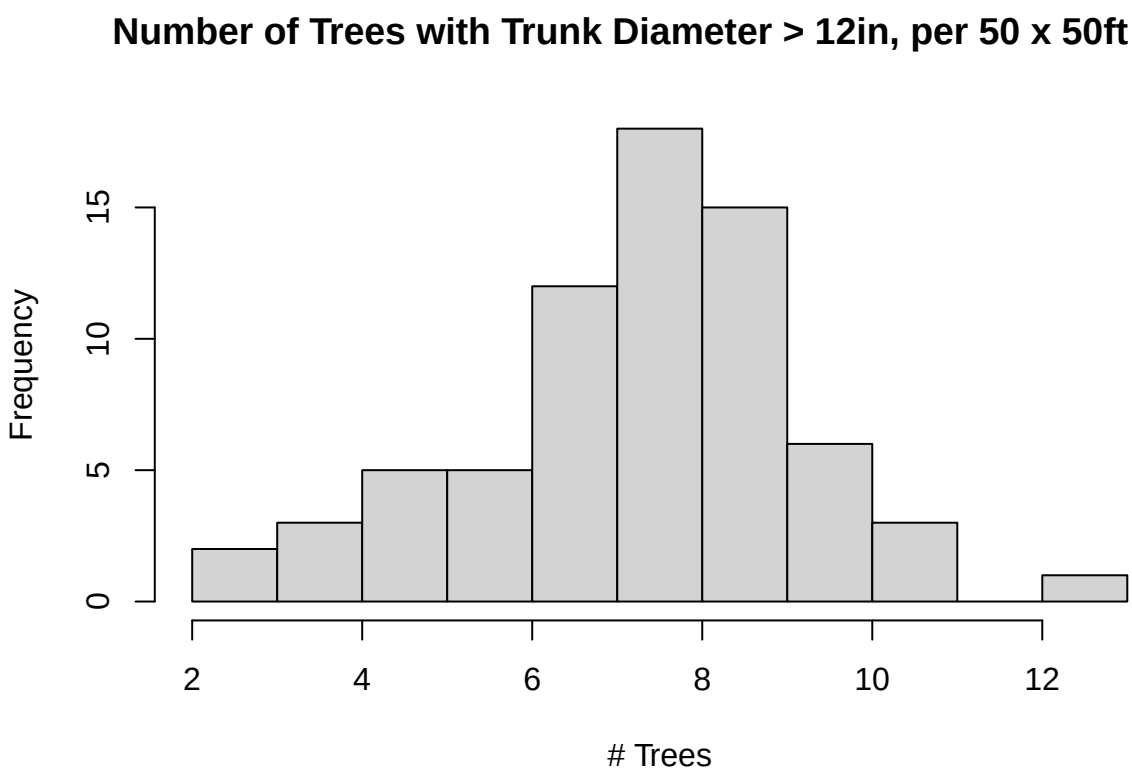
To assist in estimating the amount of lumber in a tract of timber, an owner decided to count the number of trees with diameters exceeding 12 inches in randomly selected 50 3 50-foot squares. Seventy 50 x 50 squares were randomly selected from the tract and the number of trees (with diameters in excess of 12 inches) was counted for each. The data are as follows.

```
timber <- "7 8 6 4 9 11 9 9 9 10 9 8 11 5 8 5 8 8 7 8 3 5 8 7 10 7 8 9 8 11 10 8 9 8 9 9 7 8 13 8 9 6 7

timber <- timber |>
  strsplit(" ")  |>
  unlist()       |>
  as.numeric()

print(timber)
```

```
##  [1]  7  8  6  4  9 11  9  9  9 10  9  8 11  5  8  5  8  8  7  8  3  5  8  7 10
## [26]  7  8  9  8 11 10  8  9  8  9  9  7  8 13  8  9  6  7  9  9  7  9  5  6  5
## [51]  6  9  8  8  4  4  7  7  8  9 10  2  7 10  8 10  6  7  7  8
```

a. Construct a relative frequency histogram to describe these data.

```
hist(timber, main = "Number of Trees with Trunk Diameter > 12in, per 50 x 50ft plot", xlab = "# Trees")
```



b. Calculate the sample mean $\bar{y}$ as an estimate of μ, the mean number of timber trees with diameter exceeding 12 inches for all 50 3 50 squares in the tract.

```
# sample mean
(ybar <- mean(timber))
```

```
## [1] 7.728571
```

c. Calculate s for the data. Construct the intervals $(\bar{y} + s)$, $(\bar{y} + 2s)$, and $(\bar{y} + 3s)$. Count the percentages

of squares falling in each of the three intervals, and compare these percentages with the corresponding percentages given by the Empirical Rule.

```
# sample standard deviation
(s <- sd(timber))
```

```
## [1] 1.984881
```

```
# calculate the 3 intervals
(ybar_plus_s  <- ybar + 1 * c(-s, s))
```

```
## [1] 5.743691 9.713452
```

```
(ybar_plus_2s <- ybar + 2 * c(-s, s))
```

```
## [1]  3.75881 11.69833
```

```
(ybar_plus_3s <- ybar + 3 * c(-s, s))
```

```
## [1]  1.773929 13.683214
```

```
library(dplyr, include.only = "between")  # bring in the between() function

# empirical rule predicts 68.27%
sum(between(timber, ybar_plus_s[1], ybar_plus_s[2])) / length(timber)
```

```
## [1] 0.7142857
```

```
# empirical rule predicts 95.45%
sum(between(timber, ybar_plus_2s[1], ybar_plus_2s[2])) / length(timber)
```

```
## [1] 0.9571429
```

```
# empirical rule predicts 99.73%
sum(between(timber, ybar_plus_3s[1], ybar_plus_3s[2])) / length(timber)
```

```
## [1] 1
```

3.35

Consumer Reports in its May 1998 issue provides cost per daily feeding for 28 brands of dry dog food and 23 brands of canned dog food. Using the Minitab computer program, the following side-by-side boxplot for these data was created.

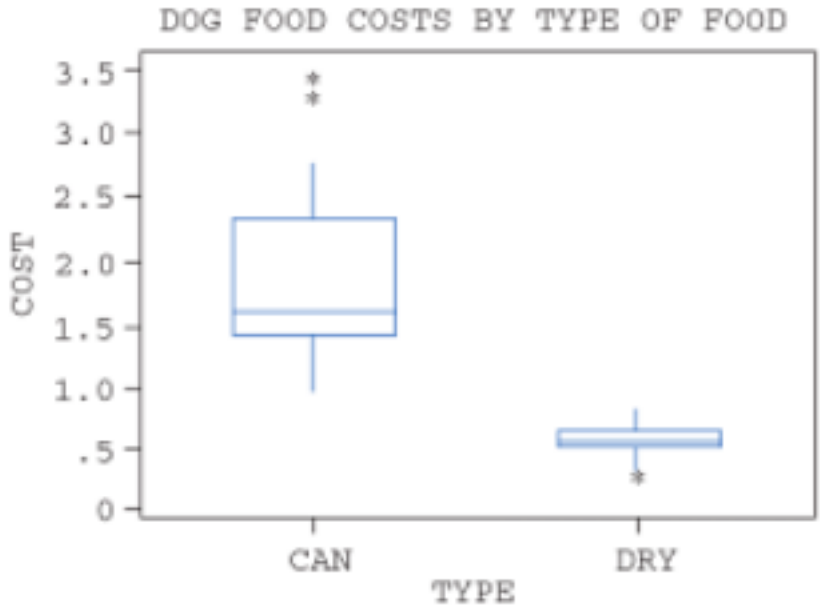


Figure 1: Retrieved from textbook

a. From these graphs, determine the median, lower quartile, and upper quartile for the daily costs of both dry and canned dog food.

Visually inspecting the boxplots:

Measure	Can	Dry
Median	1.60	0.5
Lower Quartile	1.49	0.49
Upper Quartile	2.4	0.7

b. Comment on the similarities and differences in the distributions of daily costs for the two types of dog food.

The two types of dog food appear very different from each other in terms of price. Canned food is both more expensive generally, but also has a much higher level of variability in the level it's sold at. Dry food is consistently cheaper. Both types have a few outliers, but they reflect the overall pattern of each product (i.e., the two expensive outliers are found in canned food, whereas the the inexpensive outlier is found among the dry food).

3.44

> The College of Dentistry at the University of Florida has made a commitment to develop its entire curriculum around the use of self-paced instructional materials such as videotapes, slide tapes, and syllabi. It is hoped that each student will proceed at a pace commensurate with his or her ability and that the instructional staff will have more free time for personal consultation in student–faculty interaction. One such instructional module was developed and tested on the first 50 students proceeding through the curriculum. The following measurements represent the number of hours it took these students to complete the required modular material.

```
times <- "16 8 33 21 34 17 12 14 27 6 33 25 16 7 15 18 25 29 19 27 5 12 29 22 14 25 21 17 9 4 12 15 13

times <- times   |>
  strsplit(" ")  |>
  unlist()       |>
  as.numeric()   |>
  sort()

print(times)
```

```
##  [1]  4  4  5  5  5  5  5  6  6  7  8  9  9  9 10 11 11 12 12 12 13 14 14 15 15
## [26] 15 16 16 16 17 17 17 18 19 21 21 21 22 23 25 25 25 26 27 27 29 29 33 33 34
```

a. Calculate the mode, the median, and the mean for these recorded completion times.

```
# finding the mode
stem(times)
```

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 44
##   0 | 555556678999
##   1 | 011222344
##   1 | 55566677789
##   2 | 11123
##   2 | 55567799
##   3 | 334
```

Based on the stem & leaf plot, it appears that **5** is the modal value.

```
median(times)
```

```
## [1] 15
```

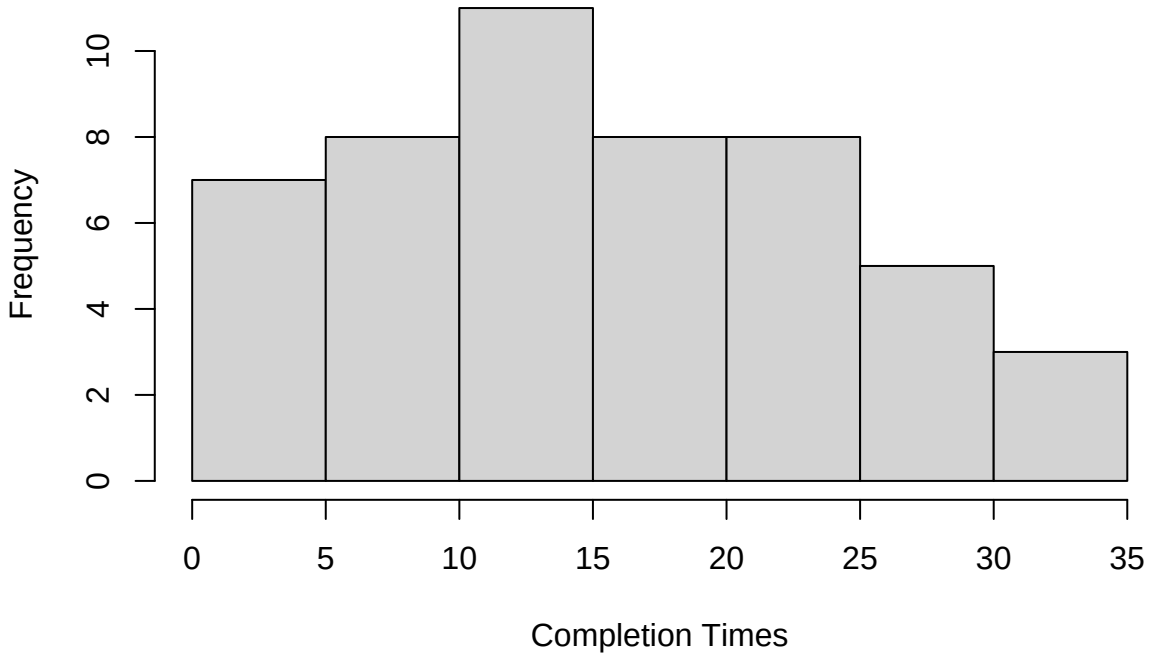
```
mean(times)
```

```
## [1] 15.96
```

b. Guess the value of s.

Using a histogram to get a visual sense of the data...

```
hist(times, xlab = "Completion Times", main = "")
```



I would guess that the value of s is around 7?

c. Compute s by using the shortcut formula and compare your answer to that of part (b).

The approximate formula for s is:

$$s = \frac{range}{4} = \frac{30}{4} = 7.5$$

Fairly close, but I was a little below the approximation.

d. Would you expect the Empirical Rule to describe adequately the variability of these data? Explain.

Without having done the calculations, **yes,** I would assume the Empirical Rule would adequately describe the variability of these data. The data are not bimodal, and while there is a minor positive skew, they're otherwise symmetrical. The Empirical Rule is robust to minor deviations from symmetry, as long as the distribution is roughly mound shaped.

3.58

The most widely reported index of the performance of the New York Stock Exchange (NYSE) is the Dow Jones Industrial Average (DJIA). This index is computed from the stock prices of 30 companies. When the DJIA was invented in 1896, the index was the average price of 12 stocks. The index was modified over the years as new companies were added and dropped from the index and was also altered to reflect when a company splits its stock. The closing New York Stock Exchange (NYSE) prices for the 30 components (as of June 19, 2014) of the DJIA are given in the following table.

```
# retrieved from the textbook
nyse <- read.csv("../data/3.58.txt")

print(nyse)
```

```
##                          company  price
## 1                         3M Co 144.41
## 2            American Express Co  94.72
## 3                       AT&T Inc  35.31
## 4                      Boeing Co 132.41
## 5                 Caterpillar Inc 107.28
## 6                    Chevron Corp 130.73
## 7               Cisco Systems Inc  24.63
## 8  E.I. Dupont de Nemours and Co  67.55
## 9                 ExxonMobil Corp 101.85
## 10           General Electric Co  26.90
## 11       Goldman Sachs Group Inc 169.52
## 12                Home Depot Inc  80.10
## 13                    Intel Corp  29.99
## 14                           IBM 182.95
## 15             Johnson & Johnson 103.41
## 16        JP Morgan Chase and Co  57.36
## 17               McDonald's Corp 101.60
## 18                Merck & Co Inc  58.27
## 19                Microsoft Corp  41.45
## 20                      Nike Inc  75.43
## 21                        Pfizer  29.55
## 22           Procter & Gamble Co  80.28
## 23              The Coca-Cola Co  41.76
## 24       Travelers Companies Inc  95.51
## 25      United Technologies Corp 117.09
## 26       United Health Group Inc  79.88
## 27    Verizon Communications Inc  49.47
## 28                       Visa Inc 208.30
## 29           Wal-Mart Stores Inc  76.25
## 30                Walt Disney Co  83.69
```

a. Compute the average price of the 30 stock prices in the DJIA.

```
mean(nyse$price)
```

```
## [1] 87.58833
```

b. The DJIA is no longer an average; the name includes the word "average" only for historical reasons. The index is computed by summing the stock prices and dividing by a constant, which is changed when stocks are added or removed from the index and when stocks split.

$$DJIA = \frac{\Sigma_{j=1}^{30} y_i}{C}$$

where y_i is the closing price for stock i and $C = 0.155625$. Using the stock prices given, compute the DJIA for June 19, 2014.

```
sum(nyse$price) / 0.155625
```

```
## [1] 16884.5
```

c. The DJIA is a summary of data. Does the DJIA provide information about a population using sampled data? If so, to what population? Is the sample a random sample?

The DJIA is a curated performance summary of stocks traded in US stock exchanges. It is not a random sample currently, and was not originally generated or conceived as a random sample of stocks being traded.