Carson Slater
STA 5380 Homework #3

1. Describe one categorical measurement and one quantitative measurement that is commonly taken in a field of study that interests you, such as human health or renewable energy. Give the possible outcomes of the categorical measurement. Describe the units in which the quantitative variable is measured and the range of values that it is expected to take.

   In baseball, a categorical variable of interest is the handedness of a hitter, where a hitter could be "right handed," "left handed," or a "switch hitter," where a hitter is capable of hitting from both sides of home plate.

   A commonly measured quantitative variable is spin rate, where for each major league pitch, the spin rate is recorded in revolutions per minute (rpm). Usually, a good major league fastball spins at 2350 rpms, whereas two-seam fastballs and cutters spin around 2150 rpms, and curveballs spin around 2480 rpms. Spin rates will vary between 1800 rpms and 2700 rpms. Spin rate is not the only factor in determining the movement of the ball as it approaches home plate. There are other observed variables such as velocity, spin axes, and spin efficiency.

2. Ritchey, Bazan, and Buhman did an experiment to compare flight times of several designs of paper helicopters, dropping them from the first to ground floors of the ISU Design Center. The flight times that they reported for two different designs are given in Table 1 (the units are in seconds).

Table 1: Flight times in seconds of two paper helicopter designs.

Design 1	Design 2
2.47, 2.45, 2.43, 2.67, 2.69, 2.48, 2.44, 2.71, 2.84, 2.84	3.42, 3.50, 3.29, 3.51, 3.53, 2.67, 2.69, 3.47, 3.40, 2.87

(a) Use the methods described in Chapter 4 to determine which of the two designs appears to produce the most consistent results.

A reliable way to determine which design had the most consistent results would be to compare each design's coefficient of variation for their flight times. A lower coefficient of variation is indicative of a smaller variance, scaled to their units by diving by the mean. Hence design with the lower coefficient of variation would likely generate the most consistent results.

```
d1 <- c(2.47, 2.45, 2.43, 2.67, 2.69, 2.48, 2.44, 2.71, 2.84, 2.84)
d2 <- c(3.42, 3.50, 3.29, 3.51, 3.53, 2.67, 2.69, 3.47, 3.40, 2.87)

sd(d1) / mean(d1) # coefficient of variation for design 1

## [1] 0.06386855

sd(d2) / mean(d2) # coefficient of variation for design 2

## [1] 0.1081696
```

In this case, the first design yielded more consistent results.

(b) Which design produced the longest flight times?

```
mean(d1) # mean for design 1

## [1] 2.602

mean(d2) # mean for design 2

## [1] 3.235

quantile(d1) # quantiles for design 1

##    0%   25%   50%   75%  100%
## 2.430 2.455 2.575 2.705 2.840

quantile(d2) # quantiles for design 2

##     0%    25%    50%    75%   100%
## 2.6700 2.9750 3.4100 3.4925 3.5300
```

Both the mean and median flight times for design 2 were longer than the mean and median flight times for design 1. So design 2 produced the longest flight times.

3. The time required to complete a hand performance test (HPT) is related to the risk of nursing home admission in older populations. The data in the file "HPT.csv" on the class website gives the seconds required by a random sample of 50 people to complete the HPT.

   (a) Calculate the 5 number summary of this data. Does this summary suggest a symmetric or a skewed distribution?

```
HPT <- unlist(read.csv("HPT.csv", header = FALSE))
summary(HPT)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   47.00   96.75  128.00  166.74  155.50  782.00
```

   We have a median that is 128. Based off the maximum value 782, this distribution appears to be right-skewed.

   (b) Compare the sample standard deviation with IQR.

```
# IQR
IQR(HPT)

## [1] 58.75

sd(HPT)

## [1] 139.2285
```

   The sample standard deviation (139.2285) is far greater than the inter-quartile range (58.75); there seems to be an outlier contaminating the data.

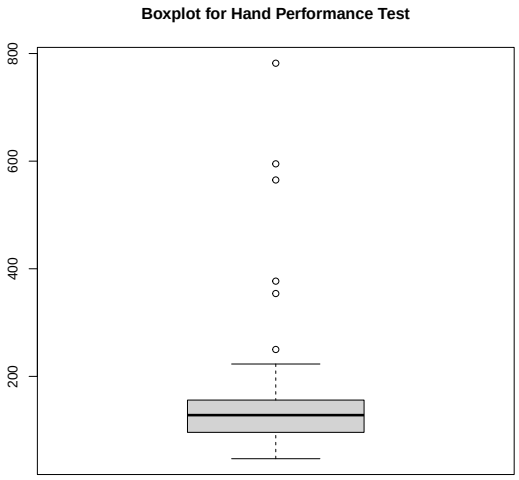
   (c) How does a 10% trimmed mean compare with the sample mean? Does this comparison suggest any outliers?

```
mean(HPT)

## [1] 166.74

mean(HPT, trim = .1)

## [1] 132.425
```

   The mean is 166.74, and the 10% trimmed mean in 132.425. This result in part (c) and the prior result in part (b) suggests there is an outlier contaminating the data, skewing the mean to the right.

   (d) Make a boxplot. Are there any outliers in the data using the $1.5 \times$ IQR rule?

```
boxplot(HPT,
        main = "Boxplot for Hand Performance Test")
```

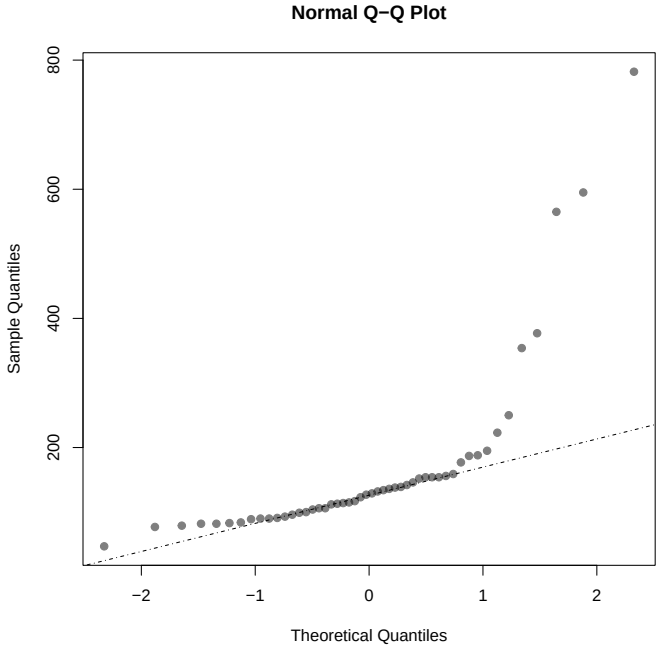



```
subset(HPT, !(median(HPT) - 1.5*IQR(HPT) < HPT &
              median(HPT) + 1.5*IQR(HPT) > HPT))

##  V14 V116 V119 V120 V135 V141 V145
##  223  782  354  250  377  565  595
```

There are indeed observations that break the $1.5 \times$ IQR rule.

(e) Make a normal quantile plot. Do the data appear to be normally distributed? If not, how do they deviate from normality?

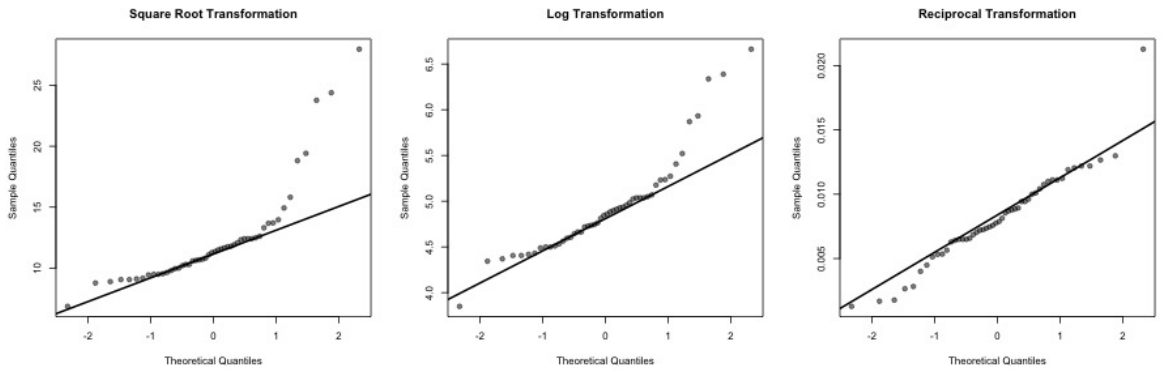
```
qqnorm(HPT, pch = 19, col = alpha("black", 0.5))
qqline(HPT, lty = 4)
```



This distribution deviates from a normal distribution with extra heaviness in the tails, particularly the right tail.

(f) Transform the data in three ways: $\sqrt{\text{HPT}}$, $\log(\text{HPT})$, and $1/\text{HPT}$. Which transformation gives the most nearly normal distribution?

```
root_HPT <- sqrt(HPT)
log_HPT <- log(HPT)
inv_HPT <- 1/HPT
par(mfrow = c(1, 3))
qqnorm(root_HPT, pch = 19, col = alpha("black", 0.5),
       main = "Square Root Transformation")
qqline(root_HPT, lty = 4)
qqnorm(log_HPT, pch = 19, col = alpha("black", 0.5),
       main = "Log Transformation")
qqline(log_HPT, lty = 4)
qqnorm(inv_HPT, pch = 19, col = alpha("black", 0.5),
       main = "Reciprocal Transformation")
qqline(inv_HPT, lty = 4)
```



It appears that the reciprocal transformation is the most similar to a normal distribution.

4. The article "The Effect of Experimental Error on the Determination of the Optimum Metal-Cutting Conditions" by Ermer and Wu (*The Journal of Engineering for Industry*, 1967) contains a dataset gathered in a study of tool life in a turning operation. The cutting speed, measured in sfpm, and the tool life, measured in minutes, are recorded for each tool.

   (a) Which variable would be considered the explanatory and which the response? (See page 346 of the textbook for a short description of the difference between these types of variables.)

   In this study, the response variable would be the tool life, as this data was gathered in a study of tool life in a turning operation. This implies the explanatory variable would be the cutting speed.

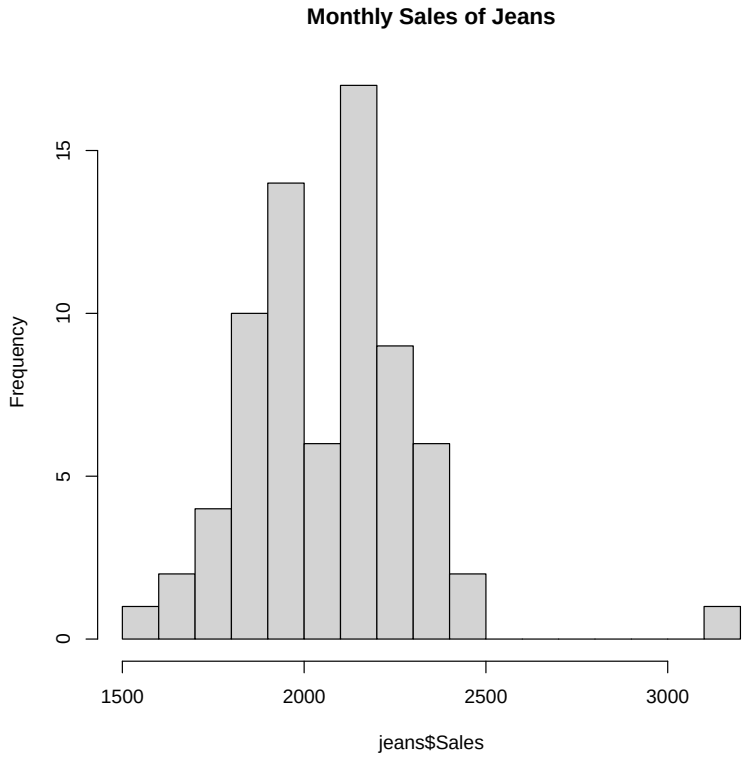
   (b) The correlation between these two variables is -0.85. Interpret what this means in the context of the problem.

   A correlation of $\rho = -0.85$ would indicate there is a strong negative linear relationship between a tool's life and the cutting speed of the tool. This would lend to the idea that a tool's high cutting speed is **associated** (*not causal*) with a shorter lifespan.

5. The monthly sales of pairs of jeans (in 1000's) over six years in the United Kingdom are given in the file "jeans.csv" on the class website.

   (a) Make a histogram of the sales data. Comment on the shape of the distribution and any outliers.

```
jeans <- read.csv("jeans.csv")
hist(jeans$Sales, breaks = "fd",
                  main = "Monthly Sales of Jeans")
```



   This distribution is interesting because it is mostly bell-shaped, but when the binwidths get smaller it is bimodal. If you assume its unimodal, then it is barely right-skewed. Also there is one major outlier.

   (b) What format are the year and month variables when this dataset is loaded into `R`? Convert them to a date object using the `lubridate` package.

```
typeof(jeans$Month); typeof(jeans$Year)

## [1] "integer"
## [1] "integer"
```

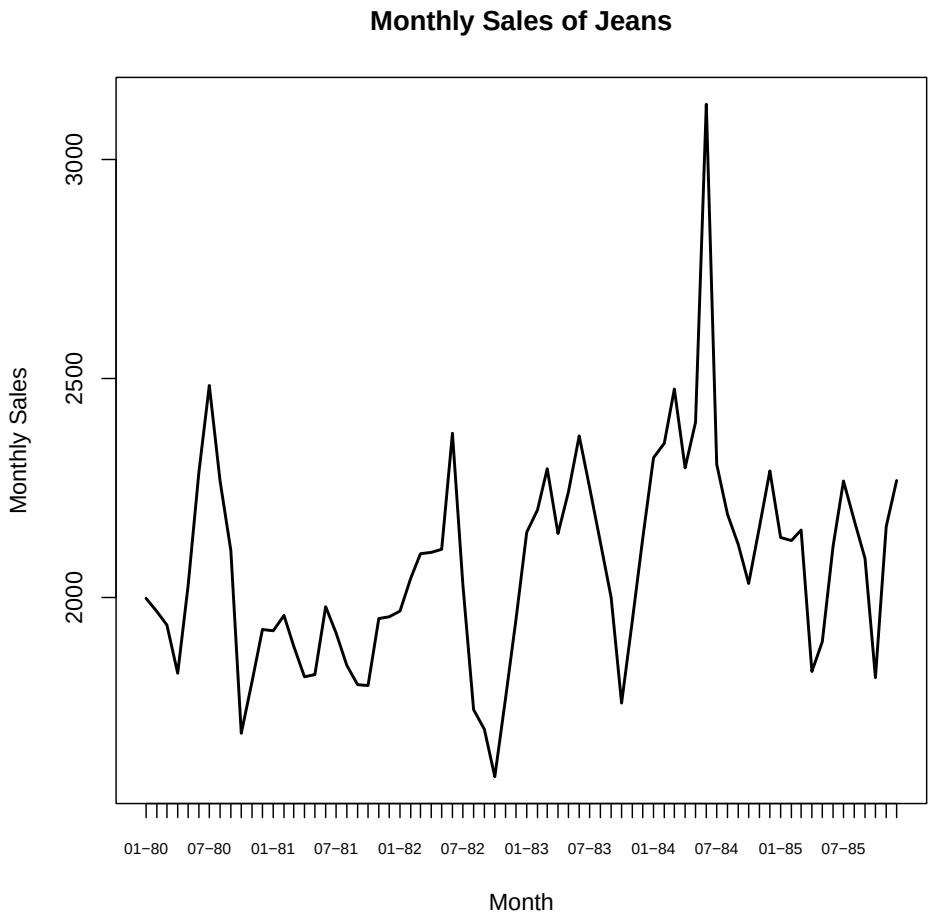
   The `Year` and `Month` variables are in integer format in separate columns. We can convert them into one `Date` column.

```
jeans$Date <- as_date(unlist(mapply(function(y, m){
  d = make_date(y,m)
  return(d)
```

```
  },jeans$Year, jeans$Month, SIMPLIFY = FALSE)))

class(jeans$Date)

## [1] "Date"
```

(c) Make a time series plot (or run chart) of the jean sales for all 72 months with the dates correctly labelled on the x-axis (which will require the variable that you created in the prior part). Comment on any trends or outliers.

```
plot(jeans$Date, jeans$Sales,
     type = 'l', lwd = 2, xaxt = "n",
     main = "Monthly Sales of Jeans", xlab = "Month",
     ylab = "Monthly Sales")
axis(1, jeans$Date, format(jeans$Date, "%m-%y"), cex.axis = .7)
```

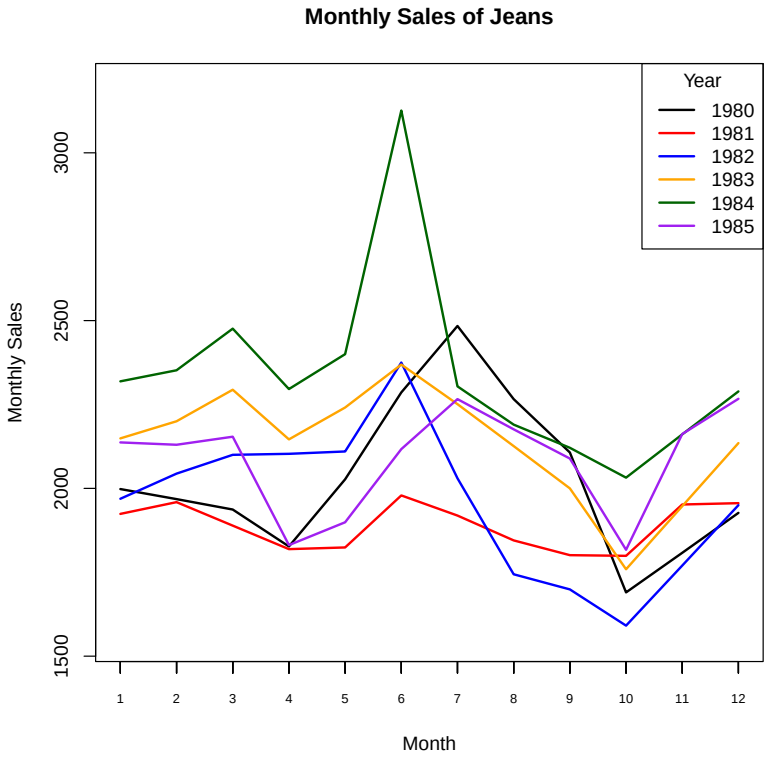


It seems that June 1984, there was a huge spike in jeans sales. This is either a true sales spike or a data entry typo. This appears to be an outlier. There seems to be seasonality, although this plot makes it difficult to truly compare each annual sales pattern.

(d) Plot separate run charts for different years on the same set of axes. Use a different color for each year using a sensible color spectrum. Does this display make it easier

to detect trends and outliers?

```
y1 <- jeans[jeans$Date >= "1980-01-01" & jeans$Date <= "1980-12-31", ]
y2 <- jeans[jeans$Date >= "1981-01-01" & jeans$Date <= "1981-12-31", ]
y3 <- jeans[jeans$Date >= "1982-01-01" & jeans$Date <= "1982-12-31", ]
y4 <- jeans[jeans$Date >= "1983-01-01" & jeans$Date <= "1983-12-31", ]
y5 <- jeans[jeans$Date >= "1984-01-01" & jeans$Date <= "1984-12-31", ]
y6 <- jeans[jeans$Date >= "1985-01-01" & jeans$Date <= "1985-12-31", ]


plot(y1$Month, y1$Sales,
     type = 'l', lwd = 2, xaxt = "n", ylim = c(1550, 3200),
     main = "Monthly Sales of Jeans", xlab = "Month",
     ylab = "Monthly Sales", col = "black")
lines(y2$Month, y2$Sales,
     type = 'l', lwd = 2, xaxt = "n", col = 'red')
lines(y3$Month, y3$Sales,
      type = 'l', lwd = 2, xaxt = "n", col = 'blue')
lines(y4$Month, y4$Sales,
      type = 'l', lwd = 2, xaxt = "n", col = 'orange')
lines(y5$Month, y5$Sales,
      type = 'l', lwd = 2, xaxt = "n", col = 'darkgreen')
lines(y6$Month, y6$Sales,
      type = 'l', lwd = 2, xaxt = "n", col = 'purple')
axis(1, jeans$Month, cex.axis = .7)
legend("topright",
       legend = c("1980", "1981", "1982", "1983", "1984", "1985"),
       lwd = 2,
       col = c("black", "red", "blue", "orange", "darkgreen", "purple"),
       title = "Year")
```



This display makes it easier to detect outliers, because you can observe seasonal trends more easily.

(e) What is evident from the run charts that is missed by the histogram?

Because the run charts are a time series visualization, we can observe seasonal trends. This is something that the histogram misses.

(f) Which of these displays would be most useful for forecasting future sales?

The final run chart from part (d) would probably provide the most useful visual on how to model the forecast this time series data.

6. The "powerball.csv" file from the class website contains 214 sets of 5 numbers drawn in the Colorado Powerball game. This game is played by selecting 5 numbers without replacement from the integers between 1 and 59, inclusive. The final "powerball" number is selected from the integers between 1 and 39, inclusive, and is not included in the dataset. Imagine that a television news reporter has called you up and asked you to apply your statistical skills to answer the following question, "What 5 numbers are chosen the most frequently?" This is a simple question, but the reporter just wants to tell people to choose those 5 most frequently selected numbers. Answer the following:

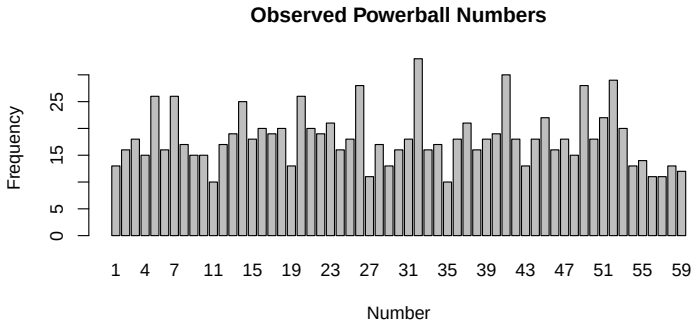
   (a) What are the 5 most frequently chosen numbers? We can find the five most drawn numbers by executing the following `R` code.

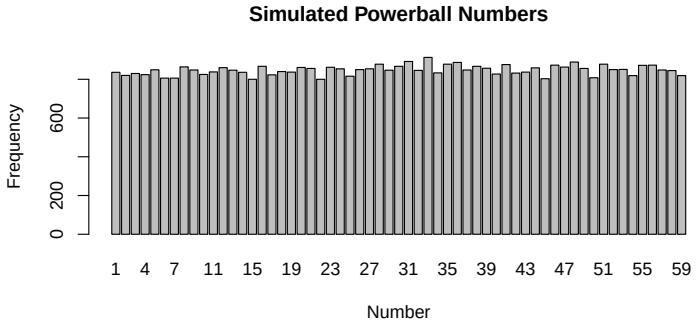
```
powerball <- read.csv("powerball.csv", header = FALSE)
powerball_vec <- c(powerball[[1]], powerball[[2]],
                   powerball[[3]], powerball[[4]],
                   powerball[[5]])

obs_vals <- table(powerball_vec)

sort(obs_vals, decreasing = TRUE)[1:5]

## powerball_vec
## 32 41 52 26 49
## 33 30 29 28 28
```

   It appears the 5 most common numbers are 32, 41, 52, 26, and 49 respectively.

   (b) Simulate 10,000 sets of 5 numbers at a time in the way that is described for the game. Compare the quantiles of these simulated values with the quantiles of the observed values. Comment on what you see.

```
set.seed(613)

results <- replicate(10000, sample(1:59, 5, replace = FALSE)) |>
  table()

par(mfrow = c(2,1))
barplot(obs_vals,
        main = "Observed Powerball Numbers",
        xlab = "Number",
        ylab = "Frequency")
barplot(results,
        main = "Simulated Powerball Numbers",
        xlab = "Number",
        ylab = "Frequency")
```






```
# Quantiles simulated data
replicate(10000, sample(1:59, 5, replace = FALSE)) |> quantile()

##   0%  25%  50%  75% 100%
##    1   15   30   45   59

# Quantiles for real data
powerball_vec |> quantile()

##   0%  25%  50%  75% 100%
##    1   16   30   44   59
```

	0%	25%	50%	75%	100%
Observed Powerball Quantiles	1	15	30	45	59
Simulated Powerball Quantiles	1	15	30	44	59

Table 2: Table of observed versus simulated quantiles for Powerball data.

The only difference between the simulated and real powerball quantiles is the 75% quantile. The observed powerball 75% quantile is 45, whereas the simulated one is 44.

(c) If the public is directed to choose the 5 numbers from part (a), can they expect to win a lot of cash? (If all 5 numbers selected are a match to those drawn, you win $200,000.) Explain using your answer from part (b) or any other numerical and graphical summaries that you choose.

```
# Observed values
sort(obs_vals, decreasing = TRUE)[1:5]
```

```
## powerball_vec
## 32 41 52 26 49
## 33 30 29 28 28

# Simulated values
sort(results, decreasing = TRUE)[1:5]

##
##  33  31  48  36  28
## 913 892 889 887 878
```

Refer to the histograms shown in part (b). There are $n_1 = 1065$ observed values for the first (top) histogram, while there are $n_2 = 50000$ simulated values created in the same experimental fashion as the observed data. Although the first histogram might not strike someone as being shaped like a uniformly distributed random variable, the simulation seems to lend to the idea that each factor in the random variable of a number being drawn is equally likely (for each draw that is, obviously the first and the third draw will have different probabilities for respective numbers). Because the sample data and simulation data were created in exactly the same manner, the argument could be made that as $n \to \infty$, $X =$ the set of five numbers drawn, converges to a Uniform distribution. Hence, the person who selects the five numbers found in part (a) will not be more likely than any other person to win the grand prize.

7. The idea of a normal plot can be extended to other distributions. Let $X_1, X_2, \ldots, X_n$ be a random sample. To check if a specified (continuous) distribution with cdf F fits this sample, we need to plot the theoretical quantiles against the sample quantiles. The order statistic, $X_{(i)}$, is the $\left(\frac{i-0.5}{n}\right)th$ sample quantile, and $F^{-1}\left(\frac{i-0.5}{n}\right)$ is the corresponding theoretical quantile. Apply this method to check the exponentiality of a dataset.

   (a) For the exponential distribution with cdf $F(x) = 1-\exp(-\lambda x)$, show that the $\left(\frac{i-0.5}{n}\right)th$ theoretical quantile is given by

$$F^{-1}\left(\frac{i-0.5}{n}\right) = \frac{1}{\lambda}\ln\left(\frac{n}{n-i+0.5}\right).$$

   We begin by finding the inverse CDF for an exponential random variable,

$$\begin{aligned} F(x) &= 1-\exp(-\lambda(x)) \\ 1-F(x) &= \exp(-\lambda(x)) \\ \ln(1-F(x)) &= -\lambda x \\ -\frac{1}{\lambda}\ln(1-F(x)) &= x \\ \implies F^{-1}(x) &= -\frac{1}{\lambda}\ln(1-x) \end{aligned}$$

   So then we would substitute $x = \left(\frac{i-0.5}{n}\right)$

$$\begin{aligned} F^{-1}\left(\frac{i-0.5}{n}\right) &= -\frac{1}{\lambda}\ln\left(1-\left(\frac{i-0.5}{n}\right)\right) \\ &= -\frac{1}{\lambda}\ln\left(\frac{n-i+0.5}{n}\right) \\ &= \frac{1}{\lambda}\ln\left(\frac{n}{n-i+0.5}\right) \end{aligned}$$

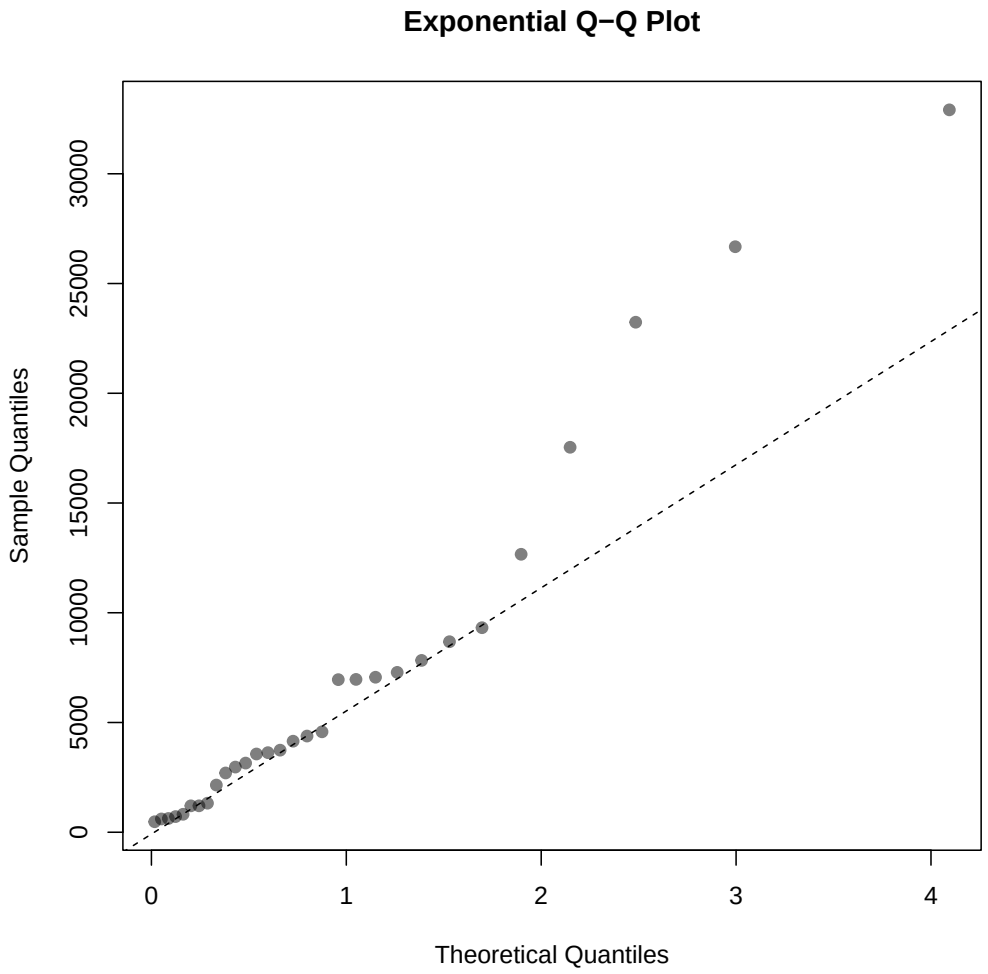
   (b) Explain why, although λ is not known, it suffices to use $\lambda = 1$ when making the plot. The corresponding quantiles are called the *unit exponential scores.*

   In this case, $\frac{1}{\lambda}$ is a scaling parameter. By scaling up, the shape of the function upholds, but the numerical values of the quantiles stay the same. Hence, $\lambda = 1$ will suffice.

   (c) Make an exponential plot for the cost of hospitalization data from Table 4.5 (p. 122) in the textbook for the control group. Does the exponential distribution fit the data well?

```
control <- c(478, 605, 626, 714, 818, 1203, 1204,
             1323, 2150,2700, 2969, 3151, 3565,
             3626, 3739, 4148, 4382, 4576, 6953,
             6963, 7062, 7284, 7829, 8681, 9319,
             12664, 17539, 23237, 26677, 32913)

n <- length(control)
```

```r
theoretical <- log(n/(n-seq(1,30,1) + 0.5))

plot(theoretical, control, col = alpha("black", 0.5), pch = 19,
     main="Exponential Q-Q Plot", xlab = "Theoretical Quantiles",
     ylab = "Sample Quantiles")
qqline(control, distribution = qexp, col = "black", lty = 2)
```



The exponential seems to fit the data well for smaller values, but toward the tail the distribution poorly fits the data, with more heaviness than an exponential distribution traditionally has.