Josh Miltier
HW1 - Stats

Theoretical Portion

1.
   a. The population to be concerned with is the properties of the Facebook of college students.
   b. The relationship between
   c. Some characteristics are age, gender, relationship status, and any other information given on their profile (assuming that the profile is public). The examples to pick would be which school the college student goes to, interest of study, and age.
   d. With an infinite amount of time and resources, I believe just about anything can be accomplished! For the matter of gathering characteristics based on the population interested in, it would be possible to gather ALL of the information.
   e. Of course part d is unrealistic, so obtaining a sample size of the population would be a good approach. It's important to assure a large enough sample to conclude plausible data for entire population.

2.
   a. The population to be working with are people in the United States over the of 65.
   b. The sample of the population to study will only be those that seek care from doctors (this may exclude the ER, since this isn't technically a doctor's office).
   c. The estimates will be based on the sample. Over a year time span and *all* doctor's offices would be a sufficient amount of data to generalize the entire US population over 65 which get the flu.

3.
   a. Mean, Median, Mode:
   - The mean is the sum of all data values (for a given set) divided by the total amount of data values. This is commonly referred to as an average.
   - The median is the middle value of the data set (once placed in order, either ascending or descending). If the total amount of data values is an even number, then the two middle numbers will be added then divided by two.
   - The mode of data is number of times the same value appears in the data set. These are tallied and compared against the other occurrences within the same data set.
   - Application would determine which of these to use and which not to use. If the data is skewed negatively, the average will seem higher than the median. The opposite will occur when there is a positive skew. Thus, non-symmetrical data will allow the median and mean to be equal to one another.

   b. Standard deviation and range
   - The standard deviation accounts for the spread of the data. That is, the average of how far the data is from the mean. A larger deviation would mean the data set has a greater value span difference versus a smaller deviation with a closely grouped data set. The determination of the latter depends on the expected values, so if the data seems "accurate" (as expected, even) then it's assumed to be a smaller deviation.
   - The range is simply the difference between the lowest and highest number in a data set.
   - It would be dangerous to report only the range of a data set because the standard deviation would be important for knowing how much each value varied from the mean, on average. The range could also assist the understanding of the standard deviation if the upper and lower values are outliers. If the standard deviation is low but the range is high, it's assumed that there are some value points which could have skewed the data. With any type of research, it's important to include as much information as possible. I believe both of these are pertinent.

Josh Miltier
HW1 - Stats

4. 1/50 people has disease D - test T claims 85% positive and 60% negative for disease D

a. The probability that the results will come back negative **

b. **

c. **

d. **

5.

a. To convert from frequency histogram to density histogram (on next page)

- Density = frequency / bin size (= 0.5)

BIN	0.0-0.5-	0.5-1.0-	1.0-1.5-	1.5-2.0-	2.0-2.5-	2.5-3.0-	3.0-3.5-	3.5-4.0-	4.0-4.5-	4.5-5.0-
DENSITY	28	6	12	176	230	10	6	10	6	2
BIN	5.0-5.5-	5.5-6.0-	6.0-6.5-	6.5-7.0-	7.0-7.5-	7.5-8.0-	8.0-8.5-	8.5-9.0-	9.0-9.5-	9.5-10.0-
DENSITY	4	4	2	2	0	0	0	0	2	0

b. Approximation of box plot are is the next page, attached. (Hand-written)

Josh Miltier
HW1 - Stats

Computational Portion

1.

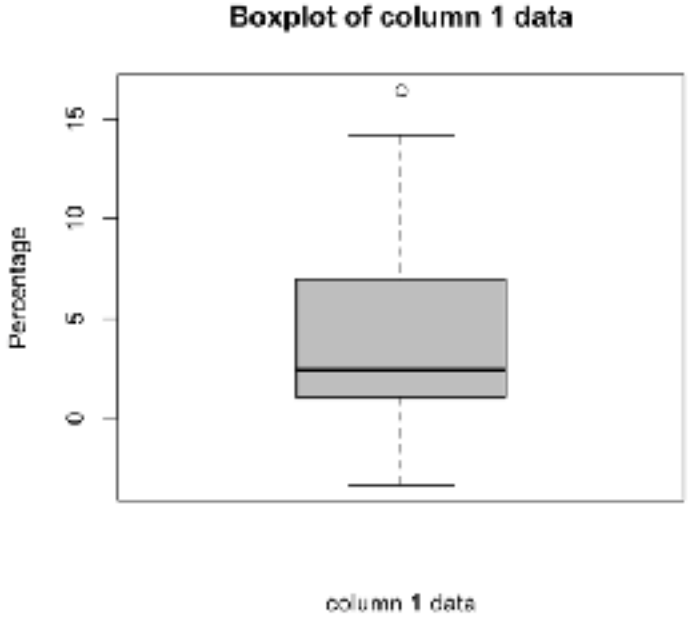
a.

2.

a. code:

line 1 dataset <- read.table("hw1data.txt", header = T)
line 2 boxplot(dataset[,1], col = "grey", ylab = "Percentage", xlab = "column 1 data",
line 2 main = "Boxplot of column 1 data")

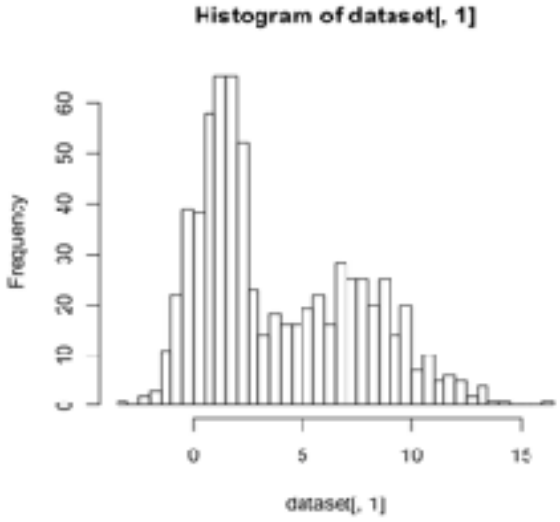
output:



b. code:

hist(dataset[,1], breaks=50) #first column of dataset, into histogram

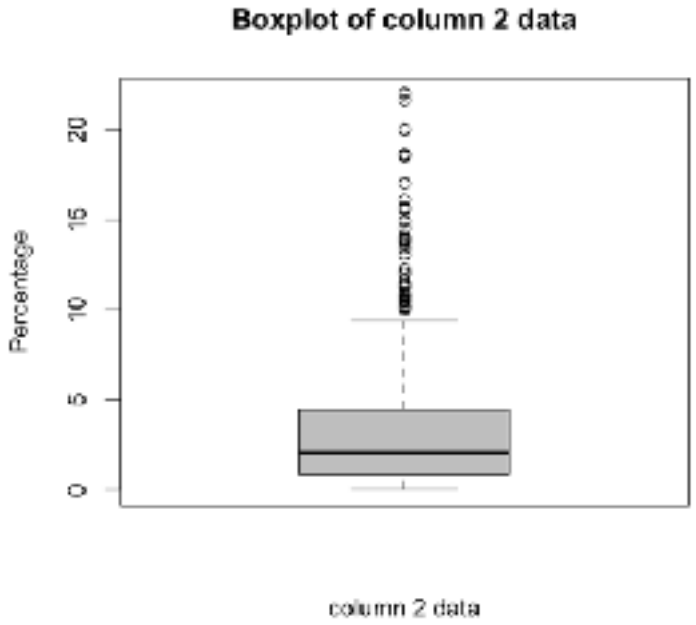
output:



c. code:

```
line 1   boxplot(dataset[,2], col = "grey", ylab = "Percentage", xlab = "column 2 data",
line 1             main = "Boxplot of column 2 data")
```

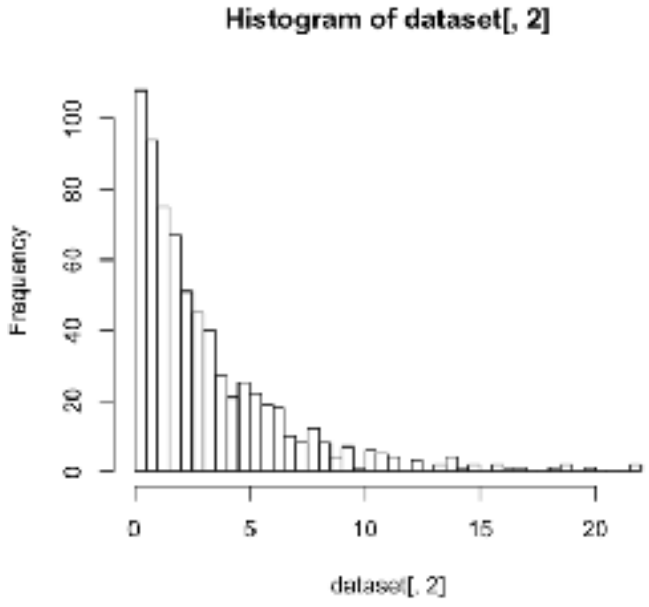
output:



code:

```
hist(dataset[,2], breaks=50)
```

output:



- For column 2 of the data, there are many more outliers in the box plot and the histogram shows more data on the left side (positively skewed). The box plot is only a summary so it isn't as precise in relaying the information as a histogram.

d. For this hw1 data, I believe that the histogram is a better representation of the data. With a small enough bin size, the information is better portrayed. However, I know with stock prices, the box plot is good for seeing the fluctuation in prices. I think once getting use to and increasing my readability of box plots, I may enjoy them more. But for the purposes of familiarity, histograms are better.

3.

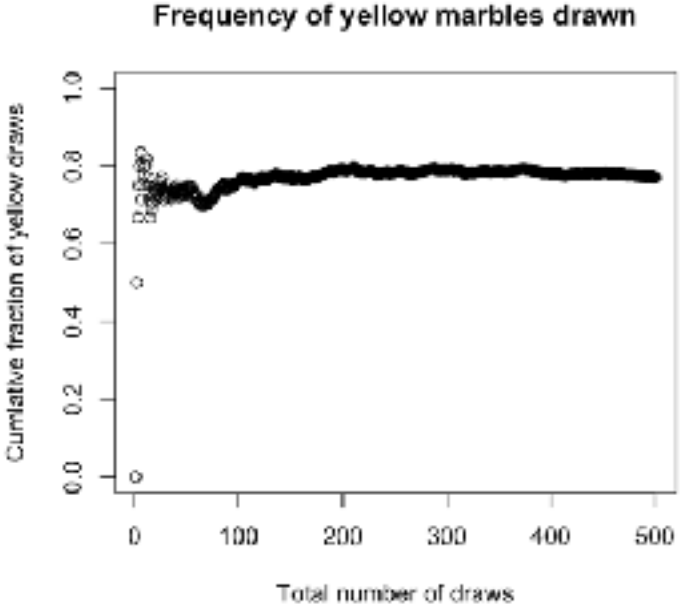
a. code:

```
yellow.draw=NULL
sum.yellow.draw=0

for (i in 1:500){
        #'1' = yellow marble drawn, '0' = yellow marble not drawn
        yellowDraws<-sample(0:1, size=1, replace=TRUE, prob=c(.26,.74))
        sum.yellow.draw=sum.yellow.draw+yellowDraws
        yellow.draw=c(yellow.draw, sum.yellow.draw/i)
}

##plot results
plot(yellow.draw, ylim=c(0,1), ylab="Cumlative fraction of yellow draws",
xlab="Total number of draws", main="Frequency of yellow marbles
drawn")
```

output:



- The number fluctuation occurs due to the smaller sample data, the running average is less certain. With more data, the overall mean evens out to be a more realistic data set.

b. The ideal draws to be certain appears to be around 200. Though the more the data obtained will give better results and more accurate information.

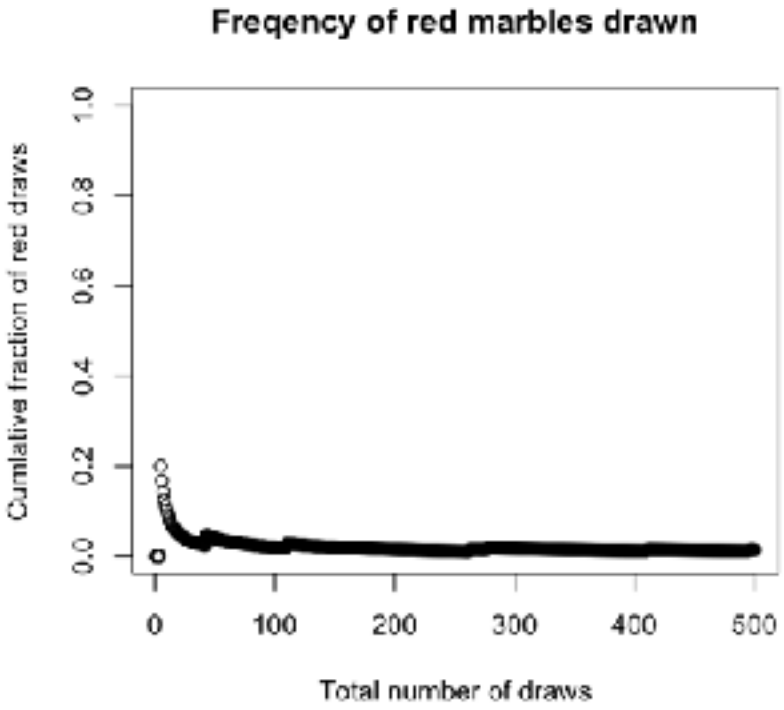
c. code:

```
red.draw=NULL
sum.red.draw=0

for (i in 1:500){
        #'1' = red marble drawn, '0' = red marble not drawn
        redDraws<-sample(0:1, size=1, replace=TRUE, prob=c(.99,.01))
        sum.red.draw=sum.red.draw+redDraws
        red.draw=c(red.draw, sum.red.draw/i)
}

##plot results
plot(red.draw, ylim=c(0,1), ylab="Cumlative fraction of red draws", xlab="Total
        number of draws", main="Freqency of red marbles drawn")
```

output:



- The probability seems to stabilize quicker than the yellow marble, since it's more unlikely to be drawn than the yellow marble. Their percentage differences are just enough to link closer probabilities to 50% will vary more than those that vary further away from it. Though with a large enough sample, the data should equalize and remain steady.