Homework 1
STAT 104 - Introduction to Quantitative Methods for Economics

1) For the following surveys, discuss any problems you think exist and suggest how to fix the issues.

a) A retail store manager wants to conduct a study regarding the shopping habits of his customers. He selects the first 60 customers who enter his store on a Saturday morning.

➔ Problem:

This survey represents a **sampling bias.** First 60 customers on a Saturday morning is a representation of only a portion of the population.

Suggested fix:

If he wants a true sample of the population, he should sample customers at **random times of the day every day of the week or on random days**.

b) The village of Oak Lawn wishes to conduct a study regarding the income level of households within the village. The village manager selects 10 homes in the southwest corner of the village and sends an interviewer to the homes to determine household income.

➔ Problem:

This survey represents a **sampling bias**.

It could also potentially have **too small a sample size**.

Suggested fix:

The homes should be **randomly selected across the entire village** and not just a particular area.

The **number of homes should also be large enough** to account for the entire population.

c) An antigun advocate wants to estimate the percentage of people who favor stricter gun laws. He conducts a nationwide survey of 1,203 randomly selected adults 18 years old and older. The interviewer asks the respondents, "Do you favor harsher penalties for individuals who sell guns illegally?"

➔ Problem:

This survey represents **wording-deliberate bias** as well as **wording-unintentional bias**. **Harsher penalties is subjective** and could bring out different interpretations, hence inducing wording-unintentional bias. **Illegally** has negative connotations and induces a wording-deliberate bias.

Suggested Fix:

"Do you favor stricter gun laws?"

Homework 1
STAT 104 - Introduction to Quantitative Methods for Economics

2) A bank with branches in a large metropolitan area is considering opening its offices on Saturday, but it is uncertain whether customers will prefer (1) having walk-in hours on Saturday or (2) having extended branch hours during the week. Listed below are some of the ideas proposed for gathering data. For each, indicate what (if any) biases (problems) might result.

a) Put a big ad in the newspaper asking people to log their opinions on the bank's Web site.

➔ Problems:

1. Selection bias:

   This suggestion relies on newspaper readers being a large representation of the population.

2. Voluntary response bias:

   Only people with strong opinions would log them on the web site. This might not represent the entire population.

b) Randomly select one of the branches and contact every customer at that bank by phone.

➔ Problem:

Selection bias:

Customers of one branch are not representation of customers of all the branches.

c) Send a survey to every customer's home, and ask the customers to fill it out and return it.

➔ Problem:

Voluntary response bias:

Only people with strong opinions would return the survey. This might not represent the entire population.

d) Randomly select 20 customers from each branch. Send each a survey, and follow up with a phone call if he or she does not return the survey within a week.

➔ No problem as long as 20 customers is a large enough sample for the population.

3) Suppose you are back in high school and the campaign manager for your friend who is running for senior class president. You would like to know what proportion of students would vote for her if the election was held today. The class is too big to ask everyone (314 students). Comment on whether or not each of the following sampling procedures should be used. Explain why or why not.

a) Poll everyone in your friend’s math class.

➔ This procedure should not be used because it has a **sampling bias**. Just the math class is not representation of the entire population.

b) Assign every student in the senior class a number from 1 to 314. Then, use a random number generator to select 30 students to poll.

➔ This could be used but the **sample size** is a matter of concern. The **confidence level will not be high** for a sample that is less than 10% the population.

c) Ask every student who is going through the lunch line in the cafeteria who they will vote for.

➔ This procedure could have a sampling bias depending on what percentage of the population goes through the lunch line. If the percentage is very high then yes, this procedure should be used. If not, it possesses a risk of being biased towards a certain section of the population.

4) R Practice, Part 1. In R, read in the results of a small survey done by visitors to a regional mall. This is done with the following command in the R command window
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/smallsurvey.csv")
You can see the data with the command View(mydata)

a) How many rows of data are in this data set? (the nrow(mydata) command could be useful here but remember the first row has the variables names).

➔ There are **31 rows** of data in the data set. 1 row for the 10 variables and 30 rows for the observations of the variables.

```
> nrow(mydata)
  [1] 30
```

b) How variables are in this data set? (the ncol(mydata)command could be useful here).

➔ There are 10 variables in this data set.

```
> ncol(mydata)
[1] 10
```

c) How many categorical variables are in this data set?

➔ There are **4** categorical variables in this data set. They are:

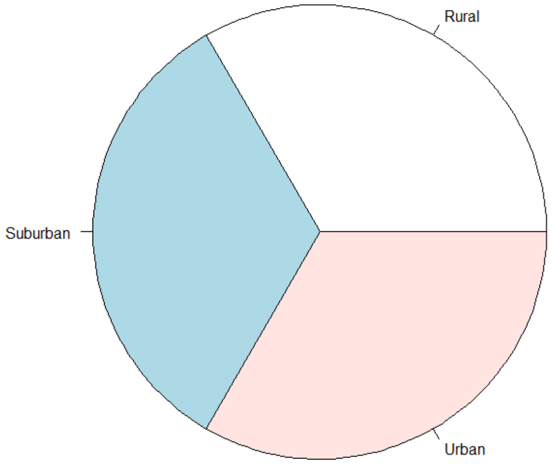
Gender
Residence
Political Party
Job Happy

A categorical variable is a variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some qualitative property

```
  > sapply(mydata, class)
```

This does not return 'Job Happy' as a factor with fixed levels but on further examination it can be seen that it is a factor with 11 levels ranging from 0 to 10.

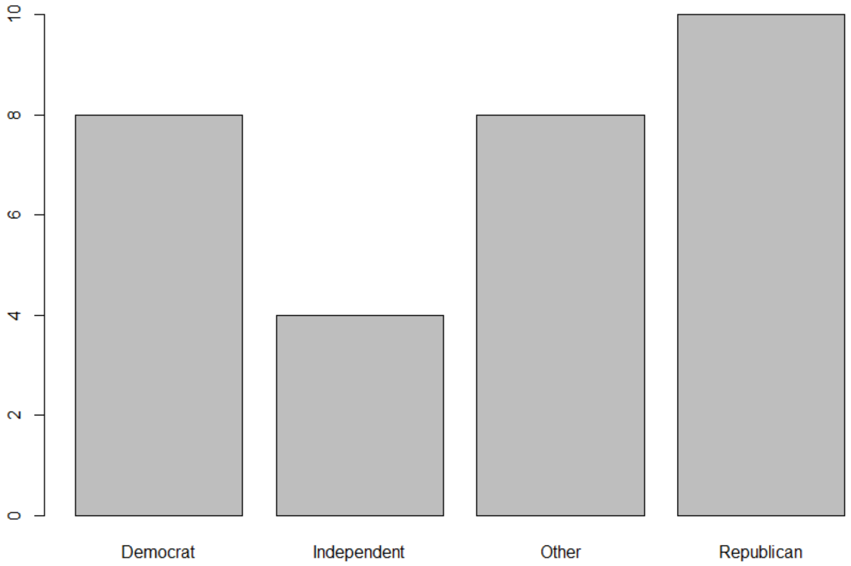
d) One way to examine categorical variables is with a pie chart. Produce a pie chart of where people live (the *residence* variable) by using the following command. Comment on the graph:
pie(table(mydata$residence))



➔ The graph tells us that residence is a **categorical variable with 3 levels**: Suburban, Rural and Urban. These are the three categories in which the people can be grouped into with respect to the Residence variable.

e) Another way to examine categorical variables is with a bar chart. Produce a bar chart of political affiliation (the *politicalparty* variable) by using the following command. Comment on the graph-why can’t we use a histogram for this variable?

```
barplot(table(mydata$politicalparty))
```



➔ The bar chart tells us that residence is a **categorical variable with 4 levels**: Democrat, Republican, Independent and Other. These are the four categories in which the people can be grouped into with respect to the Political Party variable.

➔ Histograms plot **quantitative data** with ranges of the data grouped into bins or intervals. If we want to plot a histogram, we would need to plot the frequency of the levels of the factor. Bar charts on the other hand are used to plot **categorical data**.

f) Find the average of the income variable.

➔ The average of the income variable is **45.4**

```
> mean(mydata$income)
[1] 45.4
```

g) We can subset data in different ways. We could create a new data set just for all the females respondents by creating femdata=subset(mydata,gender=="F"). As another example, one could create a new data set for those people that have income over 50 with the command newdata=subset(mydata,income>50).

Compare the average income and standard deviation of income for men and women.

```
> femaleData=subset(mydata,gender == "F")
> maleData=subset(mydata,gender == "M")
>
> femaleIncome = femaleData$income
> maleIncome = maleData$income
>
> femaleIncomeAverage = mean(femaleIncome)
> maleIncomeAverage = mean(maleIncome)
>
> maleIncomeAverage
[1] 53.4
> femaleIncomeAverage
[1] 37.4
```

➔ **Average male income (53.4) is greater than average female income (37.4)**

```
> maleIncomeSd
[1] 15.54624
> femaleIncomeSd
[1] 12.0226
```

➔ **Standard deviation of income for men(15.546) is greater than standard deviation of income for women(12.023)**

h) The variable jobhappy measures on a 1-10 scale how happy someone is with their job. Compare the average income for someone with a jobhappy rating of 8 or more versus the average income of someone with a jobhappy rating of 3 or less. What do you find?

➔ **The average income for someone with a jobhappy rating of 8 or more(36.143) is lower than the average income of someone with a jobhappy rating of 3 or less(50.923).**

```
> above8jobhappy = subset(mydata,jobhappy >="8")
> above8jobhappy_income = above8jobhappy$income
> above8jobhappy_income_avg = mean(above8jobhappy_income)
>
> below3jobhappy = subset(mydata,jobhappy <="3")
> below3jobhappy_income = below3jobhappy$income
> below3jobhappy_income_avg = mean(below3jobhappy_income)
>
> above8jobhappy_income_avg
[1] 36.14286
> below3jobhappy_income_avg
[1] 50.92308
```

5) R practice, part 2. This question uses an old data set on cars from Consumer Reports. To load the data into R enter the following command in R's command line:
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/cars10.csv")
To see what is in this data set, you can enter the R command View(mydata).

a) Calculate the mean price of the automobiles in the data set.

➔ **The mean price of the automobiles in the data set is 6165.27**

```
> prices = mydata$price
> meanPrice = mean(prices)
> meanPrice
[1] 6165.257
```

b) Calculate the median price of the automobiles in the data set.

➔ **The median price of the automobiles in the data set is 5006.5**

```
> prices = mydata$price
> medianPrice = median(prices)
> medianPrice
[1] 5006.5
```

c) What does the difference between the mean and median price indicate about the shape of the distribution for the price?

➔ The difference between the mean and median price indicates that the distribution for the price is **skewed to the left.** That implies that there are more data points less than the mean than there are greater than the mean.

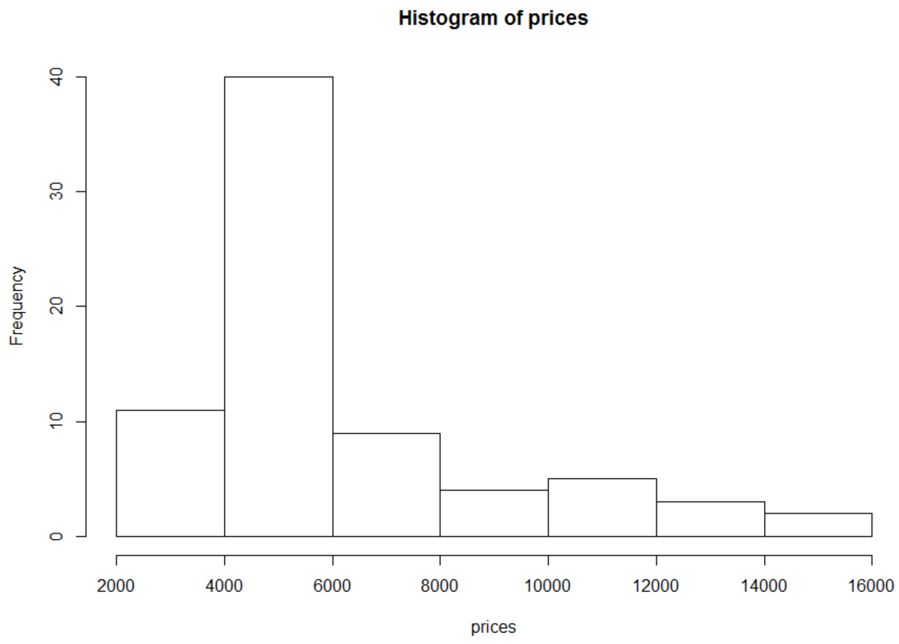
d) Calculate the mean price of automobiles separately for the domestic and foreign cars and compare the results.

➔ **The mean price for domestic automobiles(6072.423) is less than the mean price for foreign cars(6384.682)**

```
> domesticAuto = subset(mydata, foreign == "Domestic")
> foreignAuto = subset(mydata, foreign == "Foreign")
> domesticAutoPrices = domesticAuto$price
> foreignAutoPrices = foreignAuto$price
> meanPrice_domesticAuto = mean(domesticAutoPrices)
> meanPrice_foreignAuto = mean(foreignAutoPrices)
> meanPrice_domesticAuto
```

```
[1] 6072.423
> meanPrice_foreignAuto
[1] 6384.682
```

e) Make a histogram of the price of cars. What shape does the histogram take? (Is it symmetric? Skewed?)



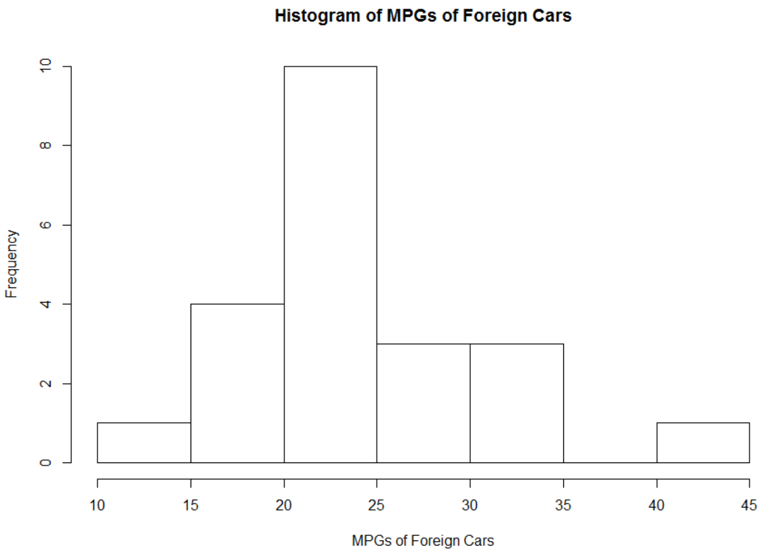
➔ **The histogram is bell shaped but not symmetric and skewed to the left.**

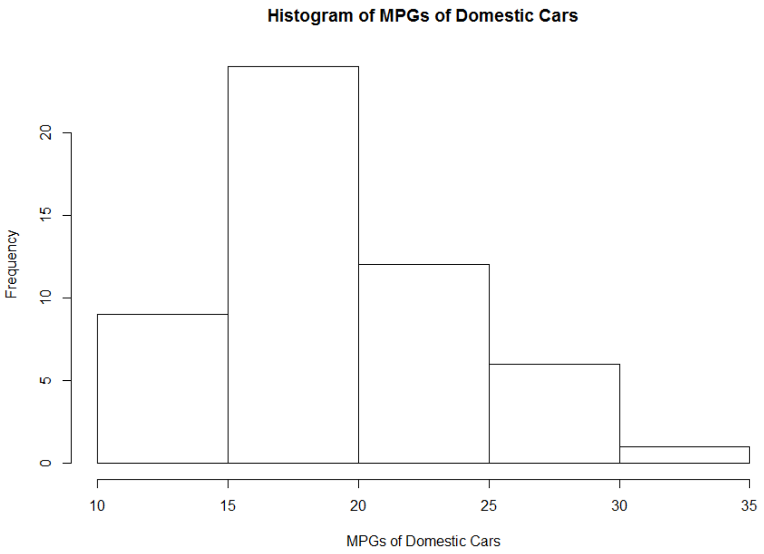
f) Discuss the difference in distributions of mpg for foreign and domestic cars. [do this by comparing means, medians and histograms).

➔ Mean MPG for foreign cars: 24.773
Mean MPG for domestic cars: 19.827
Median MPG for foreign cars: 24.5
Median MPG for domestic cars: 19

The distribution for both foreign and domestic cars will be left skewed since the mean is greater than the median. The lower difference between mean and median in foreign cars can be seen in the form of lower variance and hence lower deviation in the distribution as compared to domestic cars.

Homework 1
STAT 104 - Introduction to Quantitative Methods for Economics



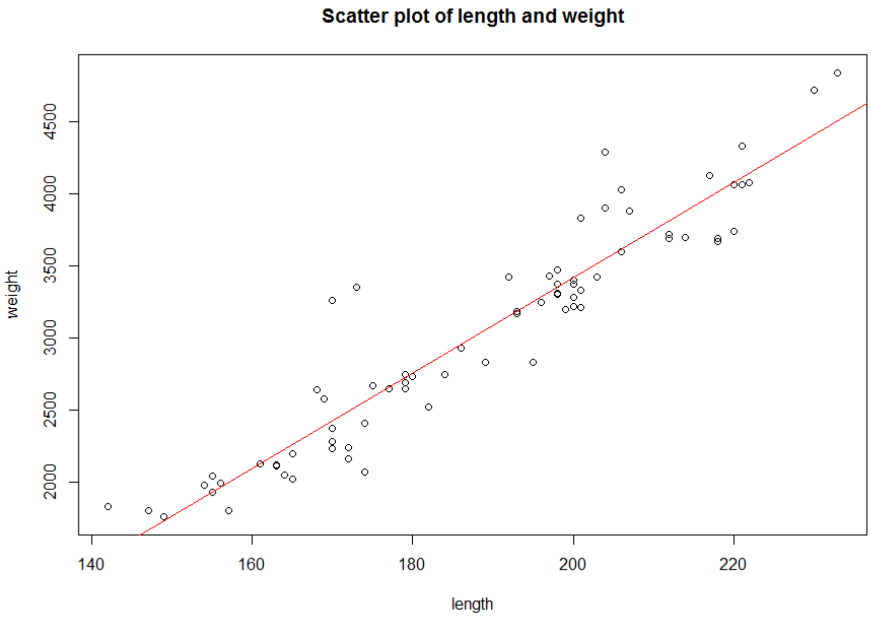


```
> foreignCars = subset(mydata, foreign == "Foreign")
> domesticCars = subset(mydata, foreign == "Domestic")
>
> mpg_foreignCars = foreignCars$mpg
> mpg_domesticCars = domesticCars$mpg
>
> meanMpg_foreignCars = mean(mpg_foreignCars)
> meanMpg_domesticCars = mean(mpg_domesticCars)
> meanMpg_foreignCars
[1] 24.77273
> meanMpg_domesticCars
[1] 19.82692
```

Karan A. Bhandarkar

```
>
> medianMpg_foreignCars = median(mpg_foreignCars)
> medianMpg_domesticCars = median(mpg_domesticCars)
> medianMpg_foreignCars
[1] 24.5
> medianMpg_domesticCars
[1] 19
>
> hist(mpg_foreignCars, main = "Histogram of MPGs of Foreign Cars", xlab = "
MPGs of Foreign Cars")
> hist(mpg_domesticCars, main = "Histogram of MPGs of Domestic Cars", xlab =
"MPGs of Domestic Cars")
```

g) Make a scatter plot of the variables weight and length. Does there appear to be any association between the variables?



➔ **There appears to be a directly proportional linear association between the variables.**

```
> weight = mydata$weight
> length = mydata$length
>
> plot(length, weight, main = "Scatter plot of length and weight")
> abline(lm(weight ~ length), col = "Red")
```

6) R practice, part 3. For this question we will use the following data set.
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/stat111survey.csv")
Create the following variable (which is number of texts students send per day)
texts=mydata$texts

a) Using the mean command, find the mean number of texts. Uh oh you should get a weird response- what is it?

➔ **We get a missing value represented by the symbol NA**

```
> # Using the mean command, find the mean number of texts.
> meanTexts = mean(mydata$texts)
> meanTexts
[1] NA
```

b) Use the command length(texts)to find how many data points are in the variable texts.

➔ **There are 107 data points in the variable texts.**

```
> # Use the command length(texts)to find how many data points are in the vari
able texts.
> lengthTexts = length(mydata$texts)
> lengthTexts
[1] 107
```

c) Use the command describe(texts)to get the summer statistics. How does the n from this output compare to what you found in (b)?

➔ **The n from this output, i.e. 91, is less than the number of data points found in (b), i.e. 107**

```
> # Use the command describe(texts)to get the summer statistics
> describe(mydata$texts)
   vars  n  mean    sd median trimmed   mad min max range skew kurtosis   se
X1    1 91 39.24 48.06     20   29.34 22.24 0.5 300 299.5 2.88    10.14 5.04
```

d) Do the command sum(is.na(texts))which counts the number of values that are missing. How many values are missing? Does this agree with (b) and (c)?

➔ **16 values are missing and this does agree with (b) and (c) (107-91 = 16)**

```
> # Do the command sum(is.na(texts))which counts the number of values that ar
e missing.
> sum(is.na(mydata$texts))
[1] 16
```

e) Create a new variable texts.comp= texts[complete.cases(texts)]. This removes all the missing data.

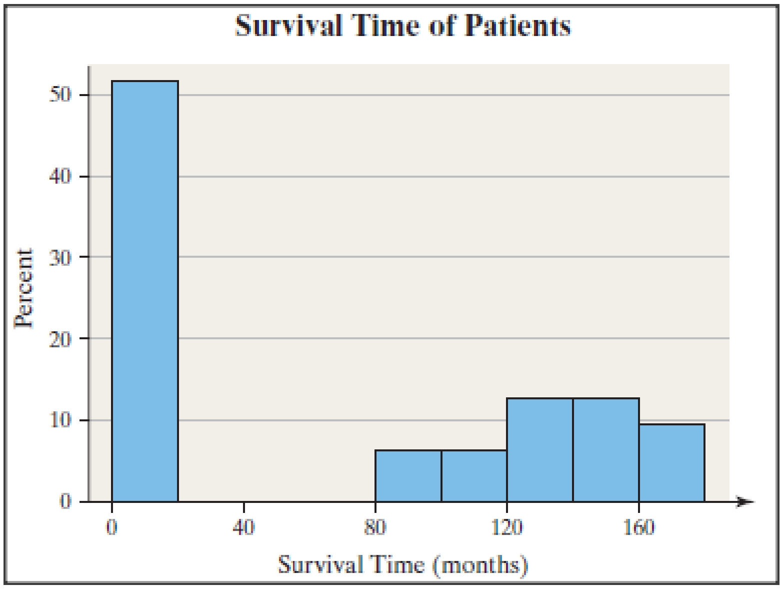
```
> # Create a new variable texts.comp= texts[complete.cases(texts)].
> texts.comp = mydata$texts[complete.cases(mydata$texts)]
```

f) Using the boxplot outlier rule, how many outliers does the data set texts.comp have?

➔ **Using the boxplot outliet rule, there are 5 outliers in the texts.comp data set**

```
> # Using the boxplot outlier rule, how many outliers does the data set texts
.comp have
> length(boxplot.stats(texts.comp)$out)
[1] 5
```

7) Unfortunately, a friend of yours has been diagnosed with cancer. You obtain a histogram of the survival time (in months) of patients diagnosed with this form of cancer as shown in the figure below. The median survival time for individuals with this form of cancer is 11 months, while the mean survival time is 69 months. What words of encouragement should you share with your friend from a statistical point of view?



➔ Median is an abstraction that should not be relied upon. The 11 months median is not much to be fixated on. The variations are the hard realities. The mean survival time of 69 months was sending a message. Almost 50% of the patients survived more than 80 months and almost 25% over 140 months. There are a lot of factors that the histogram does not bring out like the age, health, time of detection of the disease, mode of treatment, etc. There is every reason to believe you will come out in the higher than mean half of the data set.

8) When my friend Seth transferred from Harvard to Yale, many of his friends remarked that the average student IQ increased at both places. Is this possible and if so, how? Briefly explain.

➔ The average of a data set changes with the values in the data set. The given outcome is possible if Seth's IQ was below the mean IQ at Harvard and above the mean IQ at Yale. Adding a value to the data set that is above the mean, moves the mean up. Removing a value from below the mean, moves the mean up as well. Hence, taking Seth's IQ out of the data set at Harvard would move the mean up and adding his IQ to the Yale data set would move the mean up as well.

9) Suppose the diameters of a sample of new tires coming off one production line turned out to have a standard deviation of 0. Would the manufacturer be happy or unhappy, assuming the average diameter was correct? Explain.

- ➔ Assuming the average diameter was correct, the **manufacturer would be happy**. Standard deviation is used to measure the displacement of the data set from the mean. A standard deviation of 0 means that there is no displacement and all the values in the data set have the same value i.e. the mean. The manufacturer would be happy because this means that **the tires are all of the same size and there is no error in the manufacturing process.**

10) Use this data set for the following question {10,20,30,40,50}. Feel free to use R for this problem. You can define this data set in R with the command `x=c(10,20,30,40,50)`.

a) Find the standard deviation and mean.

```
> x = c(10,20,30,40,50)
> sd(x)
[1] 15.81139
> mean(x)
[1] 30
```

b) Add 5 to each value, and then find the standard deviation and mean.

```
> x1 <- x+5
> sd(x1)
[1] 15.81139
> mean(x1)
[1] 35
```

c) Subtract 5 from each value and find the standard deviation and mean.

```
> x2 <- x-5
> sd(x2)
[1] 15.81139
> mean(x2)
[1] 25
```

d) Multiply each value by 5 and find the standard deviation and mean.

```
> x3 <- x*5
> sd(x3)
[1] 79.05694
> mean(x3)
[1] 150
```

e) Divide each value by 5 and find the standard deviation and mean.

```
> x4 <- x/5
> sd(x4)
[1] 3.162278
> mean(x4)
[1] 6
```

11) A company has 30 employees, including a director. The lowest salary among the 30 employees is $22,000. The director's salary is $180,000, which is more than twice as much as anyone else's salary. Decide for each of the following statements about the 30 salaries whether it is true, false, or you cannot tell *on the basis of the information at hand*.

a) The average salary is below $60,000. - **Can't tell**

We need to calculate the max possible average salary and the min possible average salary.

Total max salary is when 1 person makes the min 22000, the director makes 180000 and the remaining 28 make 90000 i.e. less than half the director's salary

```
> totalMaxSalary = 22000+180000+28*90000
> maxAverageSalary = totalMaxSalary/30
> maxAverageSalary
[1] 90733.33
```

Total min salary is when 29 people makes the min 22000 and the director makes 180000

```
> totalMinSalary = 22000+180000+28*22000
> minAverageSalary = totalMinSalary/30
> minAverageSalary
[1] 27266.67
```

Based on the information at hand, the average salary is between 27266.67 and 90733.33

b) The median salary is below $60,000. – **Can't tell**

Based on the min and max examples considered in (a), the median can range from 22000 to 90000

c) If all salaries are increased by $1,000, that adds $1,000 to the average. **- True**

Irrespective of the values, given the **linear transformation property** of the mean of a data set, if all salaries are increased by $1,000, that adds $1,000 to the average.

d) If the director's salary is doubled, and all other salaries remain the same, that increases the average salary. - **True**

The director's salary is more than the mean of the data set. **If a value to the right of the mean of the data set is increased, the mean increases.**

e) If the director's salary is doubled, and all other salaries remain the same, that increases the median salary. - **False**

The director's salary is an outlier and the highest value in the data set. **Further increasing the largest value in the data set increases the mean, but not the median of the data set as the middle value still remains unchanged.**

f) The standard deviation of the salaries is larger than $180,000. – **False**

The range of the data set (180000-22000) is 158000. **The standard deviation can never be more than the range of the data set.**

12) In this problem we will look at the sexual partner dataset mentioned in class. Load it into R using the command
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/sexpart.csv")
sexpart=mydata$x

a) Compare the standard deviation and IQR as measures of spread on the full data set. Which measure do you think is more appropriate to describe the spread in the data set?

```
> # Compare the standard deviation and IQR as measures of spread on the full data s
et
> sd(sexpart)
[1] 585.1631
> IQR(sexpart)
[1] 5
> # What are the outliers
> boxplot.stats(sexpart)$out
 [1]  150   40   19  150   30   19   30   18 6000   15   45   15
```

The IQR is more appropriate to describe the spread in data because there are clearly **outliers influencing the standard deviation**.

b) Compare which points are flagged as outliers using the two methods discussed in class (Z score and boxplot method).

```
> # Calculate Z score of the data set
> Z <- abs((sexpart - mean(sexpart))/sd(sexpart))
> sexpartWithZ = data.frame(sexpart,Z)
> # Retrieve outliers
> zOutliers = sexpartWithZ[sexpartWithZ$Z > 2, ]
> zOutliers$sexpart
[1] 6000

> # What are the outliers using boxplot method
> boxplot.stats(sexpart)$out
 [1]  150   40   19  150   30   19   30   18 6000   15   45   15
```

c) Remove the outliers flagged using the boxplot method. Recalculate the IQR and standard deviation of this smaller dataset. Are the values closer to each other now?

- ➔ **Yes, the values are now closer to each other.**

```
> # Remove the outliers flagged using the boxplot method.
> outliers = boxplot.stats(sexpart)$out
> cleanSexpart = setdiff(sexpart,outliers)
> # Recalculate the IQR and standard deviation of this smaller dataset.
> IQR(cleanSexpart)
[1] 6.5
> sd(cleanSexpart)
[1] 4.1833
```

13) A mutual fund has a mean rate of return of about 12.3%, with a standard deviation of 15.7%.

For any set of data and for any number k, greater than one, the proportion of the data that lies within k standard deviations of the mean is at least:

$$1-\frac{1}{k^2}$$

a) According to Chebyshev's Inequality, at least 75% of returns will be between what values?

For this scenario, $1-\frac{1}{k^2}=75\%$ or $1-\frac{1}{k^2}=\frac{3}{4}$ $\therefore k=2$

75% of returns lies within 2 standard deviations i.e. -19.4% and 43.3%

b) According to Chebyshev's Inequality, at least 88.9% of returns will be between what two values?

For this scenario, $1-\frac{1}{k^2}=88.9\%$ or $1-\frac{1}{k^2}=\frac{889}{100}$ $\therefore k=3$

88.9% of returns lies within 3 standard deviations i.e. -34.8% and 59.4%

c) Should an investor be surprised if she has a negative rate of return? Why?

No, a negative rate of return should not surprise an investor because returns as low as -3.7% are within one standard deviation (as low as -19.4% are within two standard deviation and as low as -34.8% are within 3 standard deviations)

d) If we were going to use the Empirical Rule, what would we need to assume about the returns?

If we were going to use the Empirical Rule, we would need to assume that the **data is mound shaped**.

14) Suppose $x_1 = 2$, $x_2 = -1$, $x_3 = 0$. Find $2 + \sum_{i=1}^{3} 5x_i$ and $\frac{1}{\sum_{i=1}^{3} x_i^2}$

$$\mathbf{2 + \sum_{i=1}^{3} 5x_i = 2 + (5*x_1 + 5*x_2 + 5*x_3) = 7}$$

$$\frac{1}{\sum_{i=1}^{3} x_i^2} = \frac{1}{x_1^2 + x_2^2 + x_3^2} = \frac{1}{5} = 0.2$$

15) $\bar{x} = 11$ and $y_i = 2x_i - 5$. Find the (numerical) value of $\bar{y}$

y is a dataset obtained by performing a linear transformation on x i.e. y = a + b*x
where a = -5 and b = 2

Deducing from the effects of linear transformation,
Mean $_{new}$ = a + b*mean
$\therefore \bar{y} = a + b*\bar{x}$
$\therefore \bar{y} = -5 + 2 * 11$
$\therefore \bar{y} = 17$

16) We have a data set that explores airline on time performance of domestic flights operated by large air carriers. The information was compiled from the Bureau of Transportation Statistics. We will only be analyzing the data from randomly selected flights from November 2008 which is in the data set airline2008NovS.csv. The variable names and definitions are listed in another file on the course web site.
You can read the dataset into R as follows
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/Airline2008NovS.csv")

a) Which day of the week has the most flights? Use the following R command to help answer the question: `table(mydata$DayOfWeek)`

➔ **The 7th day of the week has the most flights.**

```
> table(mydata$DayOfWeek)

   1    2    3    4    5    6    7 
1089 1056 1060 1431 1626 1436 2299 
```

b) How many unique carriers are in this data set?

➔ **There are 19 unique carriers in this data set**

```
> length(unique(mydata$UniqueCarrier))
[1] 19
```

c) How many flights in this data set had a zero minute weather delay?

➔ **9562 flights in this data set had a zero minute weather delay**

```
> nrow(mydata[mydata$WeatherDelay < 1,])
[1] 9562
```

d) Which is larger, the median departure delay or the median arrival delay?

➔ **The median arrival delay is larger than the median departure delay**

```
> median(mydata$DepDelay)
[1] 30
> median(mydata$ArrDelay)
[1] 34
```