Stat 104: Quantitative Methods
Homework 1: Due Monday, September 11

Homework policy: This homework is due by 8:0am (EST) on the due date. Homework is to be handed in via the course website in pdf format. You do not need to type the homework; there are many ways (scanner in the library or phone apps) to convert written homework into a pdf file. Ask the teaching staff if you need assistance.

Late homework will not be accepted. You are encouraged to discuss homework problems with other students (and with the instructor and TFs, of course), but you must write your final answer in your own words. Solutions prepared “in committee” or by copying someone else’s paper are not acceptable.

- Please submit your homework in **pdf format**; this can be done in Word, or OpenOffice or via cellphone apps that will scan and turn into pdf.

- Please make your homework solutions legible by **bolding** or using circles to identify your solution.

- Since we are not printing out anything, use lots of space for your solutions, and put each answer on a different page if it makes the solution easier to read.

- Please make sure your submitted solutions are in numerical order [problem 1, problem 2 and so on].

- Please keep your computer output to a minimum and focus on the required answer. The easiest way to put your computer output into your homework is to cut and paste it into a Word file and use the font “courier new”.

- Please keep in mind the course rules on Academic Honesty and Collaboration

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.

   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.

   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?"

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.
   b) Randomly select one of the branches and contact every customer at that bank by phone.
   c) Send a survey to every customer's home, and ask the customers to fill it out and return it.
   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.

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

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).

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

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

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))`

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))`

f) Find the average of the income variable.

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.

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?

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.
b) Calculate the median price of the automobiles in the data set.
c) What does the difference between the mean and median price indicate about the shape of the distribution for the price?
d) Calculate the mean price of automobiles separately for the domestic and foreign cars and compare the results.
e) Make a histogram of the price of cars. What shape does the histogram take? (Is it symmetric? Skewed?)
f) Discuss the difference in distributions of mpg for foreign and domestic cars. [do this by comparing means, medians and histograms).
g) Make a scatter plot of the variables weight and length. Does there appear to be any association between the variables?

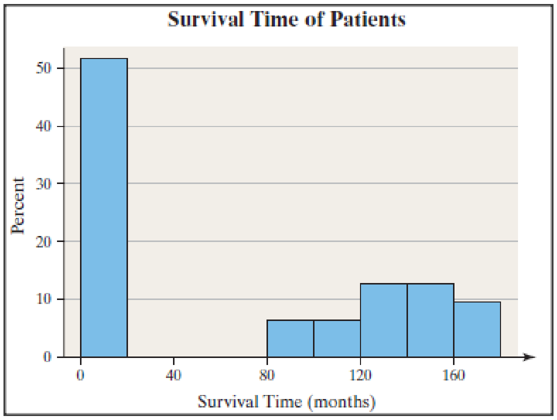
6) R practice, part 3. For this question we will use the following data set.

```
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?
b) Use the command `length(texts)` to find how many data points are in the variable texts.
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)?
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)?
e) Create a new variable `texts.comp = texts[complete.cases(texts)].` This removes all the missing data.
f) Using the boxplot outlier rule, how many outliers does the data set texts.comp have?

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? [It also recommended you read the essay "the median isn't the message" found on the course web site.]



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.

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.

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.
   b) Add 5 to each value, and then find the standard deviation and mean.
   c) Subtract 5 from each value and find the standard deviation and mean.
   d) Multiply each value by 5 and find the standard deviation and mean.
   e) Divide each value by 5 and find the standard deviation and mean.
   f) Generalize the results of parts b through e.

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. True Can't Tell False

b) The median salary is below $60,000. True Can't Tell False

c) If all salaries are increased by $1,000, that adds $1,000 to the average. True Can't Tell False

d) If the director's salary is doubled, and all other salaries remain the same, that increases the average salary. True Can't Tell False

e) If the director's salary is doubled, and all other salaries remain the same, that increases the median salary. True Can't Tell False

f) The standard deviation of the salaries is larger than $180,000. True Can't Tell False

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

13) A mutual fund has a mean rate of return of about 12.3%, with a standard deviation of 15.7%.
a) According to Chebyshev's Inequality, at least 75% of returns will be between what values?
b) According to Chebyshev's Inequality, at least 88.9% of returns will be between what two values?
c) Should an investor be surprised if she has a negative rate of return? Why?
d) If we were going to use the Empirical Rule, what would we need to assume about the returns?

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

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

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)`
b) How many unique carriers are in this data set?
c) How many flights in this data set had a zero minute weather delay?
d) Which is larger, the median departure delay or the median arrival delay?