Stat 104: Quantitative Methods
Homework 2: Due Monday, September 18

Homework policy: Homework is due by 8:00am(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 s p a c e 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) This problem is to simply get you more comfortable with R. We have data on the amount of the dinner bill and the resulting tip from a local restaurant. Read the data in as follows:

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat104/RestaruantTips.csv")
> names(mydata)
[1] "Bill"   "Tip"    "Credit" "Guests" "Day"    "Server" "PctTip"
```

Let's first build a variable that has the tip percentage:

```
> tiper=100*mydata$Tip/mydata$Bill
> summary(tiper)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.667  14.286  16.204  16.620  18.192  42.194
```

The median tip is a bit above 15% and more than 25% of the people tip more than 15%. Someone tipped an amazing 42%. [my bad-I just realized there is a variable in this data set called PctTip which is the same thing I just defined. Oh well.]

How many people tipped above 40%? Looks like two people:

```
> sum(tiper>40)
[1] 2
```

Suppose we want to remove these big tippers from our data set. One way to do so is to make a new data set with these 2 points removed:

```
> dim(mydata)
[1] 157   7
> newdata=subset(mydata,tiper<40)
> dim(newdata)
[1] 155   7
> newtiper=100*newdata$Tip/newdata$Bill
> summary(newtiper)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  6.667  14.276  16.069  16.297  18.025  31.786
```

The code above shows we started with 157 rows of data, and when we delete the two largest tippers our new data set has 155 rows of data. Note that we have to create a new variable for tip percentage for the new data set, and this new variable has a max less than 40.

Part a: Using the box plot rule, how many Tip values are considered outliers (use the original data set).

Part b: Using the box plot rule, how many tiper (tip percentage) values are considered outliers (use the original data set).

Part c: Using the original data set, what is the correlation between dinner bill and tip?

Part d: Using the data set with the two largest tip percentages removed, what is the correlation between dinner bill and tip? Is this number the same as from part (c)? Explain.

2) Suppose you were to collect data for each pair of variables. You want to make a scatterplot. Which variable would you use as the explanatory variable and which as the response variable? Why? What would you expect to see in the scatterplot? Discuss the likely direction and form.

   a) T- shirts at a store: price each, number sold.

   b) Real estate: house price, house size (square footage).

   c) Economics: Interest rates, number of mortgage applications.

   d) Employees: Salary, years of experience.

3) This question moves us in the direction of understanding that just because two variables are uncorrelated does not mean they are independent.

   a) Explain in words what a correlation of 0 implies.
   b) Load the blas data set into R and find the correlation of X and Y
      ```
      mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/blas.csv")
      ```
   c) Plot the data-does it agree with your definition?

4) It has been noted that there is a positive correlation between the U.S. economy and the height of women's hemlines (distance from the floor of the bottom of a skirt or dress) with shorter skirts corresponding to economic growth and lower hemlines to periods of economic recession. Comment on the conclusion that economic factors cause hemlines to rise and fall. (for historical references see for example http://www.edelmanfinancial.com/education-center/articles/t/the-relationship-between-hemlines-and-the-stock-market).

5) We have state by state data (plus Washington, DC) on percentage of residents over the age of 25 who have at least a bachelor's degree and median salary. Load this data into R with the command
   ```
   mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/bach.csv")
   ```

   a) What is the correlation between these two variables?
   b) Produce a scatter plot of the data with percentage with bachelor's degree on the X axis. Notice the outlier? Who does that point belong to? Can you think of any reasons why this location might have a high percentage of residents with a bachelor's degree but a lower than expected median income?
   c) Remove the outlier point found in (b) and recalculate the correlation. How do the two correlation values compare? What does this illustrate about correlation?

6) Fill in the blanks (show your work).

```
> describe(cbind(x1,x2,x3),skew=FALSE)
   vars  n    mean      sd    min   max   range     se
x1    1 74 6165.26 2949.50 3291.0 15906 12615.0 342.87
x2    2 74   39.65    4.40   31.0    51    20.0   0.51
x3    3 74    2.99    0.85    1.5     5     3.5   0.10


> cor(cbind(x1,x2,x3))
          x1        x2        x3
x1 1.0000000 0.3096174 0.1145056
x2 0.3096174 1.0000000 0.4244646
x3 0.1145056  blank 1  1.0000000


> cov(cbind(x1,x2,x3))
            x1          x2          x3
x1    Blank 2   4017.557201 285.7209367
x2    4017.5572   19.354313   1.5797853
x3     285.7209    1.579785   0.7157071
```

Blank 1 = _____

Blank 2 = _____

7) We are going to work with stock data for the companies MRK, YUM and NKE. We are going to work with monthly returns of these stocks.
This problem requires use of the R package called `quantmod`. To install this package enter the command

```
install.packages("quantmod")
```

To use the package enter the command

```
library(quantmod)
```

To create the monthly returns run the following command in R

```
source("http://people.fas.harvard.edu/~mparzen/stat104/getstockdata1.R")

This will create four new variables, mrkret,yumret,nkeret and spyret [the
index return] which are monthly returns.
```

a) What company does each symbol represent? Go to finance.yahoo.com to find out. While you're on the yahoo finance page, also write down the Beta yahoo has for each stock.

b) Find the Beta for each stock. That is run a regression of each stock return as the Y variable and index returns as the X variable. Beta is the slope from this regression. How do your calculated betas compare to the reported Betas from yahoo finance?

c) What is the standard deviation for the returns of the stocks (just do `sd(mkret)` for example) ? Rank them from lowest to highest standard deviation.

d) Rank the stocks based on their Beta values (smallest to larges). Is the order the same as if you ranked them on their standard deviations from smallest to largest? [there are many risk measures wall street uses so no reason why one ranking is the same as another.]

8) We have data on frozen pizza sales (in pounds) and average price ($/unit) from Dallas Texas for 39 recent weeks. Load the class survey data into R using the command

```
mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/pizzasales1.csv")
```

a) Using price as the explanatory variable and sales as the response variable, run a regression and write down the linear equation relating sales to price from the output.
b) What does the slope mean in this context?
c) What does the y-intercept mean in this context? Is it meaningful?
d) What do you predict the sales to be if the average price charged was $3.50 for a pizza?
e) If the sales for a price of $3.50 turned out to be 60,000 pounds, what would the residual be?

f) Show that the slope coefficient for the regression model can also be calculated using the equation $b_1 = r\frac{s_y}{s_x}$.

9) The owner of a moving company typically has his most experienced manager predict the total number of labor hours that will be required to complete an upcoming move. This approach has proved useful in the past, but the owner has the business objective of developing a more accurate method of predicting labor hours (Y). In a preliminary effort to provide a more accurate method, the owner has decided to use the number of cubic feet moved as the independent variable (X) and has collected data for 36 moves in which the origin and destination were within the borough of Manhattan in New York City and in which the travel time was an insignificant portion of the hours worked. The data may be loaded into R as follows

mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/moving.csv")

Use R to answer the questions below.

a) Create a scatter diagram of the data.
b) Fit a least squares regression line to this data and interpret the slope.
c) Predict the labor hours for a 500 cubic feet move using the estimated regression equation developed in part (b).

10) A fair six-sided die is rolled.
a) What are the possible outcomes of this event?
b) Calculate the probability of rolling a prime number.
c) Calculate the probability of rolling an even number.
d) What is the probability of rolling a number greater than seven?

11) The probability that a driver is speeding on a stretch of road is 0.27. What is the probability that a driver is not speeding?

12) A department store manager has monitored the numbers of complaints received per week about poor service. The probabilities for numbers of complaints in a week, established by this review, are shown in the table. Let *A* be the event "There will be at least one complaint in a week," and B the event "There will be less than 10 complaints in a week."

NUMBER OF COMPLAINTS	0	1-3	4-6	7-9	10-12	More than 12
PROBABILITY	.15	.29	.16	?	.14	.06

a) Find the value of ?
b) Find the probability of *A*.
c) Find the probability of *B*.
d) Find the probability of the complement of *A*.
e) Find the probability of *A* or *B*.
f) Find the probability of *A* and *B*.

13) Answer the following questions using the following joint probability table

	No wind	Some wind	Strong wind	Storm
No rain	0.1	0.2	0.05	0.01
Light rain	0.05	0.1	0.15	0.04
Heavy rain	0.05	0.1	0.1	0.05

a) Find the marginal probability P(light rain).
b) Find the marginal probability P(strong wind).
c) Find the conditional probability P(heavy rain | strong wind) .
d) Find the conditional probability P(some wind | light rain) .

14) Read the pdf document on the website entitled Birthday Problems. Then answer the following question (question 3 on page 199 of the document):

A small class contains 6 students. What is the chance that at least two have the same *birthmonth*?

15) In this question we work through some basic R commands for simulating rolling a 6 sided die.

The main command we need to know for this is `sample`. It has two arguments and two options. Imagine drawing names from a hat. The command sample just picks a certain number of names from the hat. You have to tell R what hat to pick from and how many to pick. R can sample with and without replacement. With replacement means we pick a name out of the basket and put it back. Without replacement means we pick a name out of the basket and we don't put it back.

So, say we want R to roll a six-sided die once. We are going to use a new format for writing a sequence of numbers, X:Y. This is just shorthand for make a list of numbers from 1 to 6 increasing by 1 each step. Take a look at this command:

```
> 1:6
[1] 1 2 3 4 5 6
```

So, now we tell R to choose 1 number from the numbers 1:6 and spit it back to us. 1:6 is the "hat" and 1 is the number of picks.

```
> sample(1:6,1)
[1] 3
```

If we want R to roll a die 10 times, we need to tell R to do it with replacement. With replacement means pick a number from 1-6 from the hat. Put the number back in the hat, and pick again. This means our rolls are independent of one another.

```
> sample(1:6,10,replace=TRUE)
 [1] 5 4 6 1 5 2 2 6 1 3
```

Below, I'm going to make a new command. Don't worry about the syntax, unless you want to know how functions are defined in R. The result of running this code is that R has a new command called `Roll1Die()`. You pick the number of rolls and put that as the argument to your command.

```
Roll1Die = function(n) sample(1:6, n, rep = T)
```

As an example, the following code will generate 10 random dice rolls:

```
> Roll1Die(10)
 [1] 6 1 4 3 4 5 2 1 5 5
```

Let's do an experiment of rolling two dice 100 times each and taking their sum.

```
> die1=Roll1Die(100)
> die2=Roll1Die(100)
> diesum=die1+die2
> prop.table(table(diesum))
```

```
diesum
    2    3    4    5    6    7    8    9   10   11
 0.02 0.09 0.09 0.16 0.11 0.20 0.12 0.12 0.07 0.02
```

These numbers represent the probabilities of the different dice sums when rolling two dice. If you ran the code above you probably will get different results since it is randomly generated data.

Question a: There is something unusual about the probability table above-what is it?

We can find the probability of rolling a seven by the following R command

```
> sum(diesum==7)/100
[1] 0.2
```

Question b:
Using http://www.mathcelebrity.com/2dice.php?gl=1&pl=7&opdice=1&rolist=+&dby=&ndby=&montect=+, what is the probability that the sum of two dice equals 7? Is our simulated example close?

Question c: Increase the number of dice rolls to 10000 each time. What is the new simulated probability that the sum equals 7?

Question d: Using 10000 rolls for each time, what is the simulated probability that the value of dice 1 equals the value of dice 2? This can be done in R using the command `sum(die1==die2)`. What is the true probability of the dice equaling each other? You can find this from the weblink above.