Stat 104: Quantitative Methods
Homework 2 SOLUTIONS

EXERCISE 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.

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

Answer: Using the box plot rule and R,

```
> boxplot.stats(mydata$Tip)$out
[1] 10.00 10.00 10.41 10.00 10.49 15.00  9.76 10.00 10.00
```

There are 9 tip values that are considered outliers.

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

Answer: Using the box plot rule and R,

```
> boxplot.stats(tiper)$out
[1] 42.194093 31.786395 25.234815 24.703557  8.208955
[6] 40.962622  8.225108  6.666667
```

There are 8 tip percentage values that are considered outliers.

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

Answer:

```
> cor(mydata$Bill,mydata$Tip)
[1] 0.9150592
```

The correlation between dinner bill and top is 0.9151.

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.

Answer:

```
> cor(newdata$Bill,newdata$Tip)
[1] 0.9462058
```

The correlation is 0.9462. This number is not the same as from part (c) and is actually greater. This makes sense as removing outliers should strengthen the linear relationship between these two variables and increase correlation.

EXERCISE 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.

Answers may vary slightly.

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

Answer: I would use price each as the explanatory variable and number sold as the response variable. This is because I am interested in seeing how changing the price of a T-shirt affects sales of this T-shirt. I would expect to see a negative direction because as price increases, number sold decreases. The likely form of the scatterplot should be curved as I would expect that at lower prices consumers that can afford these t-shirts will be more sensitive to increases in t-shirt price which will result in more drastic reduction in sales. While at higher prices or luxury prices, consumers who can afford these t-shirts are less sensitive to increases and may buy the shirt regardless of price.

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

Answer: I would use house size as the explanatory variable and house price as the response variable. This is because I am interested in seeing how the price of a house is affected by the size of the house. I would expect to see a positive direction because as size increases, price should also increase. The likely form is linear as I would expect an increase in house size to result in about the same increase in house price regardless of whether we are looking at smaller or larger houses.

c. Economics: Interest rates, number of mortgage applications

Answer: I would use interest rates as the explanatory variable and number of mortgage applications as the response variable. This is because I am interested in seeing how changing interest rates affect the number of mortgage applications. I would expect to see a negative direction because as interest rates increase, number of mortgage applications should decrease. The likely form is curved as at lower interest rates, many individuals are interested in taking out mortgage applications to take advantage of a beneficial market so they are more sensitive to changes in interest rates. At higher interest rates, individuals who are interested in taking out mortgages are most likely doing it out of necessity. As a result, they will be less sensitive to changes in interest rates and may take out a mortgage application regardless of rate.

d. Employees: Salary, years of experience.

Answer: I would use years of experience as the explanatory variable and salary as the response variable. This is because I am interested in seeing how different years of experience affect salary amount. I would expect to see a positive direction because as years of experience increase, salary should also increase. The likely form is curved as after many years of experience, an individual may not be receiving job promotions frequently and you would expect their salary each year to plateau at a certain amount. Therefore after working for a long time, each additional year of experience will not increase salary as much as when an individual is initially starting out on the job.

EXERCISE 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.

Answer: A correlation of 0 implies that there is no linear relationship between two variables.

b. Load the blas data set into R and find the correlation of X and Y

Answer:

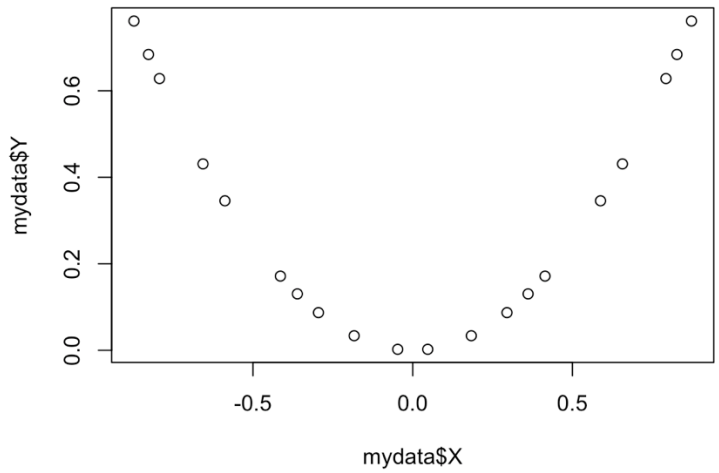
```
> cor(mydata$X,mydata$Y)
[1] 1.041004e-20
```

The correlation of X and Y is almost 0.

c. Plot the data-does it agree with your definition?

Answer:

```
> plot(mydata$X,mydata$Y)
```



Yes, the data does agree with my definition. These variables are in a quadratic relationship and not a linear relationship.

EXERCISE 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.

Answers may vary.

Answer: Even though there is an interesting positive correlation between the strength of the U.S. economy and the height of women's hemlines, it is important to note that correlation does not imply causation. It is hard to identify whether fashion trends, disposable income, or cultural context, etc. truly causes increases in hemlines. As statisticians, we can only point out interesting associations and should not extend these associations into conclusions regarding causation.

EXERCISE 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.

a. What is the correlation between these two variables?

Answer:

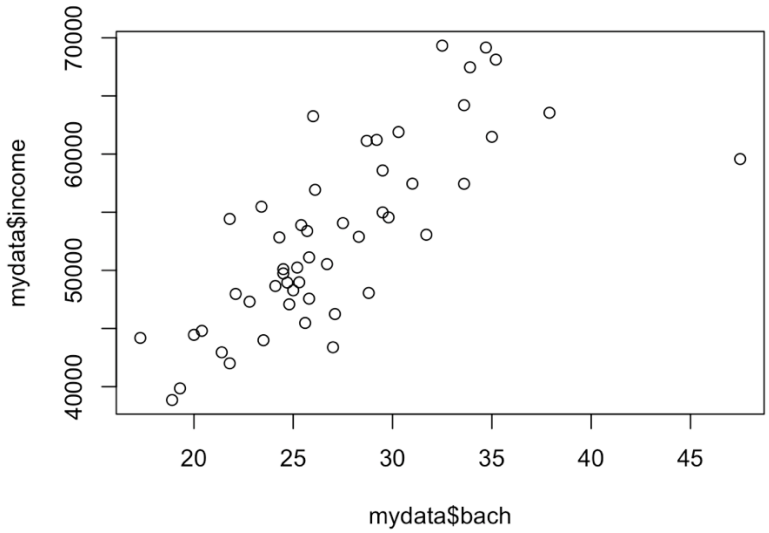
```
> cor(mydata$bach,mydata$income)
[1] 0.754167
```

The correlation between the percentage of residents over the age of 25 who have at least a bachelor's degree and median salary is 0.7542.

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?

Answer:

```
> plot(mydata$bach,mydata$income)
```



The outlier data point belongs to the District of Columbia. Residents of this area are lawmakers who are often very educated but earn public servant salaries. This may result in 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?

Answer:

```
> newdata = subset(mydata,mydata$bach<45)
> dim(mydata)
[1] 51  3
> dim(newdata)
[1] 50  3
> cor(newdata$bach,newdata$income)
[1] 0.8205775
```

With the outlier removed, the correlation is now greater. This illustrates that correlation is affected by outliers. This makes sense as removing outliers can strengthen the linear relationship between two variables and increase correlation.

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

Answer:

Blank 1

$$r_{23} = \frac{s_{23}}{s_2 s_3} = \frac{(1.579785)}{(4.40)(0.85)} = 0.4224$$

Blank 2

$$cov(x_1, x_1) = var(x_1) = sd(x_1)^2 = (2949.50)^2 = 8699550$$

EXERCISE 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.

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.

Answer: MRK represents Merck & Co., Inc. with a Beta of 0.94. YUM represents YUM! Brands, Inc. with a Beta of 0.69. NKE represents Nike, Inc. with a Beta of 0.41.

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?

Answer:

```
> lm(mrkret~spyret)

Call:
lm(formula = mrkret ~ spyret)

Coefficients:
(Intercept)       spyret
  -0.001555     0.935385

> lm(yumret~spyret)

Call:
lm(formula = yumret ~ spyret)

Coefficients:
(Intercept)       spyret
   0.009039     0.664544
```

```
> lm(nkeret~spyret)

Call:
lm(formula = nkeret ~ spyret)

Coefficients:
(Intercept)       spyret
   0.007185     0.425569
```

The calculated beta values for MRK, YUM, and NKE are 0.9354, 0.6645, 0.4256, respectively. These values are very similar to those reported by Yahoo! Finance with MRK and YUM being slightly lower and NKE being slightly higher when compared to Yahoo!.

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.

Answer:

```
> sd(mrkret)
[1] 0.04447468
> sd(yumret)
[1] 0.05111473
> sd(nkeret)
[1] 0.05408185
```

From lowest to highest standard deviation: MRK, YUM, NKE.

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.]

From lowest to highest Beta value: NKE, YUM, MRK. No, this order is not the same as when I ranked them on their standard deviations. In fact, it is the complete opposite order.

EXERCISE 8: We have data on frozen pizza sales (in pounds) and average price ($/unit) from Dallas Texas for 39 recent weeks.

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.

Answer:

```
> fit = lm(mydata$sales~mydata$price)
> coef(fit)
 (Intercept) mydata$price
   141865.53    -24369.49
```

$$sales = 141865.53 - 24369.49 * price$$

b. What does the slope mean in this context?

Answer: For every increase of $1 in average price, the number of pounds of frozen pizza sold decreases by an average of 24369.49.

c. What does the y-intercept mean in this context? Is it meaningful?

Answer: The y-intercept means that when the average price of pizza is $0/unit, there will be 141865.53 pounds of frozen pizza sold. This is not meaningful as reasonably pizza will never be priced at $0/unit. Even if it was priced at $0/unit, we would expect infinite pounds of pizza to be sold as the pizza would be free.

d. What do you predict the sales to be if the average price charged was $3.50 for a pizza?

Answer:

$$sales = 141865.53 - 24369.49 * (3.50) = 56572.31$$

If the average price charged was $3.50 for a pizza, we would expect 56572.31 pounds of pizza to be sold.

e. If the sales for a price of $3.50 turned out to be 60,000 pounds, what would the residual be?

Answer:

$$residual = (60000) - (56572.31) = 3427.69$$

The residual is 3427.69 pounds of pizza.

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

Answer:

```
> b =
cor(mydata$sales,mydata$price)*(sd(mydata$sales))/(sd(mydata$pri
ce))
> b
[1] -24369.49
```

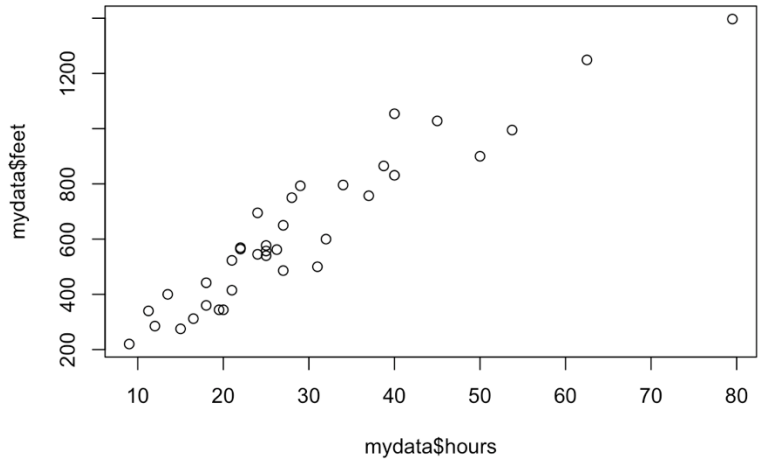
As you can see, the slope coefficient calculated through R and through the equation are the same.

EXERCISE 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.

a. Create a scatter diagram of the data.

Answer:

```
> plot(mydata$hours,mydata$feet)
```



b. Fit a least squares regression line to this data and interpret the slope.

Answer:

```
> lm(mydata$hours~mydata$feet)

Call:
lm(formula = mydata$hours ~ mydata$feet)

Coefficients:
(Intercept)  mydata$feet
   -2.36966      0.05008
```

For every additional cubic foot moved, the number of labor hours increases by an average of 0.05008 hours.

c. Predict the labor hours for a 500 cubic feet move using the estimated regression equation developed in part (b).

Answer:

$$hours = -2.36966 + 0.05008 * cubicfeet = -2.36966 + 0.05008(500) = 22.670$$

We would expect a 500 cubic feet move to result in 22.670 labor hours.

EXERCISE 10: A fair six-sided die is rolled.

a. What are the possible outcomes of this event?

Answer: The possible outcomes are {1,2,3,4,5,6}

b. Calculate the probability of rolling a prime number.

Answer: The prime numbers within the possible outcomes are 2,3, and 5. Therefore, the probability of rolling a prime number is ½.

c. Calculate the probability of rolling an even number.

Answer: The even numbers within the possible outcomes are 2,4, and 6. Therefore, the probability of rolling a prime number is ½.

d. What is the probability of rolling a number greater than seven?

Answer: There is no number greater than seven in the possible outcomes. Therefore, the probability of rolling a number greater than seven is 0.

EXERCISE 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?

Answer: Since these are complementary events, the probability that the driver is not speeding is $1 - 0.27 = 0.73$.

EXERCISE 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 ?

Answer:

$$0.15 + 0.29 + 0.16 + ? + 0.14 + 0.06 = 1$$

$$? = 0.2$$

b. Find the probability of A.

Answer: The probability that there will be at least one complaint in a week is the same as 1 – the probability that there will be only 0 complaints in a week. Therefore,

$$P(A) = 1 - 0.15 = 0.85$$

c. Find the probability of B.

Answer: The probability that there will be less than 10 complaints in a week is 1 – the probability that there will be 10-12 or more than 12 complaints in a week. Therefore,

$$P(B) = 1 - 0.14 - 0.06 = 0.8$$

d. Find the probability of the complement of A.

Answer: $P(A^c) = 1 - P(A) = 1 - 0.85 = 0.15$

e. Find the probability of A or B.

Answer: There will always be at least one complaint in a week or less than 10 complaints in a week. Therefore,

$$P(A \text{ or } B) = 1.$$

f. Find the probability of A and B.

Answer: There will be at least one complaint in a week and less than 10 complaints in a week when there are 1-3, 4-6, or 7-9 complaints in a week. Therefore,

$$P(A \text{ and } B) = 0.29 + 0.16 + 0.2 = 0.65.$$

EXERCISE 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).

Answer: $P(light\ rain) = 0.05 + 0.1 + 0.15 + 0.04 = 0.34.$

b. Find the marginal probability P(strong wind).

Answer: $P(strong\ wind) = 0.05 + 0.15 + 0.1 = 0.3.$

c. Find the conditional probability P(heavy rain | strong wind).

Answer: $P(heavy\ rain \mid strong\ wind) = \frac{P(heavy\ rain\ and\ strong\ wind)}{P(strong\ wind)} = \frac{0.1}{0.3} = \frac{1}{3}$

d. Find the conditional probability P(some wind | light rain).

Answer: $P(some\ winds \mid light\ rain) = \frac{P(some\ wind\ and\ light\ rain)}{P(light\ rain)} = \frac{0.1}{0.34} = 0.2941$

EXERCISE 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?

Answer: Let A be the event that at least two have the same birthmonth. Then A' is the event that no students have the same birthmonth. Since these events are complements,

$$\boldsymbol{P(A) = 1 - P(A')}$$

It is much easier to calculate P(A'). The probability that no students have the same birthmonth is simply the probability that student 2 doesn't have the same birthmonth as student 1, student 3 doesn't have the same birthmonth as student 1 and 2, and so on. Therefore,

$$\boldsymbol{P(A') = 1 * \frac{11}{12} * \frac{10}{12} * \frac{9}{12} * \frac{8}{12} * \frac{7}{12} = 0.2228}$$

Plugging this in,

$$\boldsymbol{P(A) = 1 - (0.2228) = 0.7772}$$

The chance that at least two have the same birthmonth is 0.7772.

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

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

Answers may vary.

Answer: It is unusual that there is no simulated case where both players rolled a 6 and the sum of the rolls equals 12. We expect to roll a sum of 12 once every 36 rounds, so it is unusual that we have not rolled a sum of 12 in 100 rounds.

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

Answer: The probability that the sum of two dice equals 7 is 1/6. Our simulated example got 0.2 which is pretty close to 1/6 or 0.166667.

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

Answers will vary slightly.

Answer:

```
> die1=Roll1Die(10000)
> die2=Roll1Die(10000)
> diesum=die1+die2
> sum(diesum==7)/10000
[1] .1672
```

The new simulated probability that the sum equals 7 is 0.1672.

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.

Answers may vary slightly.

Answer:

```
> die1=Roll1Die(10000)
> die2=Roll1Die(10000)
> sum(die1==die2)/10000
[1] 0.1676
```

Our simulated probability is 0.1676. The true probability is 1/6 or 0.1667.