Homework 5

Stat 345 - Spring 2020

Name: ____

Problem 1

X and Y are jointly continuous with joint pdf

$$f(x,y) = 2, \quad x > 0, y > 0, x + y \leq 1$$

and 0 otherwise.

a) Find marginal pdf's of X and of Y.

Note that since both variables are positive and $x + y \leq 1$, it follows that $0 \leq x, y \leq 1$

$$f(x) = \int_0^{1-x} 2dy = 2 - 2x, \quad 0 \leq x \leq 1$$

$$f(y) = \int_0^{1-y} 2dx = 2 - 2y, \quad 0 \leq y \leq 1$$

b) Find covariance Cov(X,Y).

The expected values are

$$E(X) = \int_0^1 2x(x-1)dx = \frac{1}{3}$$

$$E(Y) = \int_0^1 2y(y-1)dy = \frac{1}{3}$$

$$E(XY) = \int_0^1 \int_0^{1-x} 2xydydx = \frac{1}{12}$$

The covariance is

$$Cov(X,Y) = E(XY) - E(X)E(Y) = \frac{1}{12} - \frac{1}{3}\frac{1}{3} = -\frac{1}{36}$$

c) Find correlation Corr(X,Y). What you can say about the relationship between X and Y?

$$E(X^2) = \int_0^1 2x^2(x-1)dx = \frac{1}{6} = E(Y^2)$$

$$Var(X) = Var(Y) = \frac{1}{6} - \frac{1}{3}\frac{1}{3} = \frac{1}{18}$$

The correlation is

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} = \frac{-1/36}{1/18} = -0.5$$

X and Y have moderate strong negative linear relationship.

Problem 2

The number, Y, of spam messages sent to a server in a day has a Poisson distribution with parameter $\lambda = 24$. Each spam message independently has a probability $p = 0.25$ of not being detected by the spam filter. Let X denote the number of spam massages getting through the filter.

a) How are you going to find the distribution of X?

$$X|Y \sim Binomial(n = Y, p = 0.25)$$
$$Y \sim Poisson(\lambda = 24)$$

$X|Y$ has Binomial distribution with probability of success of 0.25 (spam message is not detected) and size Y (the number of spam messages).

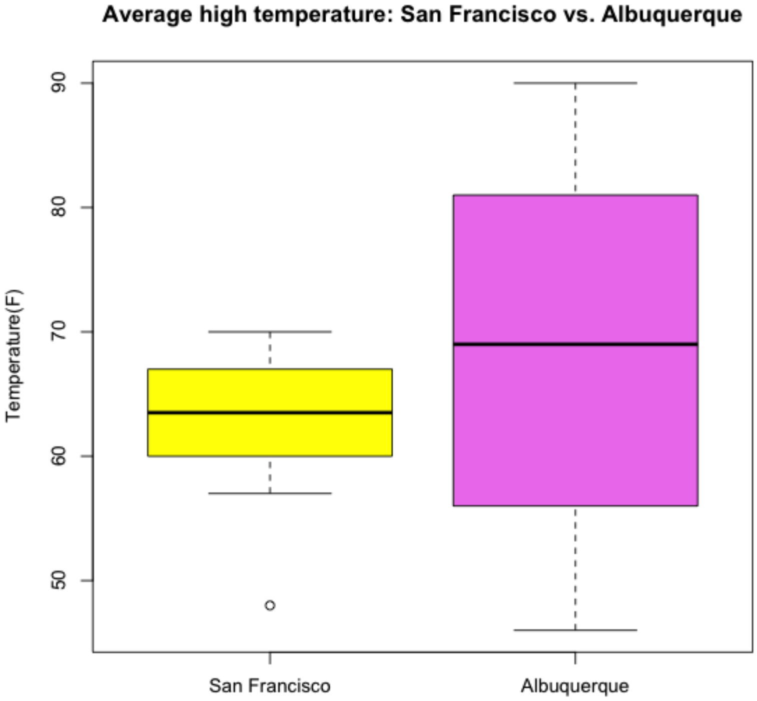
b) Calculate the expected daily number of spam messages which get into the server.

$$E(X) = E[E(X|Y)] = E(np) = E(0.25Y) = 0.25E(Y) = 0.25\lambda = 0.25(24) = 6$$

On average, six spam messages will get into the server through the filter.

Problem 3

These boxplots compare the average high temperatures of San Francisco (x) to those in Albuquerque (y) during one year.



a) Which of the following statements are true? Select all.

- Half of SF temperatures are below 70 F
- **The ABQ IQR is around 23 F**
- About 10% of temperatures are below 65 F
- **A quarter of SF temperatures are above 67 F**
- **Half of ABQ temperatures are below 69 F**
- A quarter of ABQ temperatures are above 55 F
- A quarter of ABQ temperatures are below 69 F
- **Mean and median of ABQ temperatures are very close to each other**
- The spread of SF temperatures is greater than the spread of ABQ temperatures
- The median of SF temperatures is greater than the median of ABQ temperatures
- The third quartile of SF temperatures is less than the first quartile of ABQ temperatures

- **The first quartile of SF temperatures is less than the third quartile of ABQ temperatures**

a) Can you identify any outliers from the plot? Identify quartiles from the plot and use the $1.5IQR$ rule to check suspected values.

For San Francisco $47F$ looks like an outlier. Approximately $IQR = Q_3 - Q_1 = 67 - 60 = 7F$. Since $47 < 60 - 7 = 53$, $47F$ is the outlier.

Problem 4

Online R: https://rdrr.io/snippets/. If you will use R attach your histograms at the end of this homework.

The scores of two exams are given:

Exam 1: 42, 41, 76, 48, 48, 59, 58, 49, 56, 88, 83, 49, 78, 47, 61, 64, 54, 76, 91, 63, 51, 95, 64, 88, 47

Exam 2: 58, 100, 50, 62, 68, 67, 69, 78, 79, 72, 75, 83, 81, 75, 89, 87, 91, 97, 59

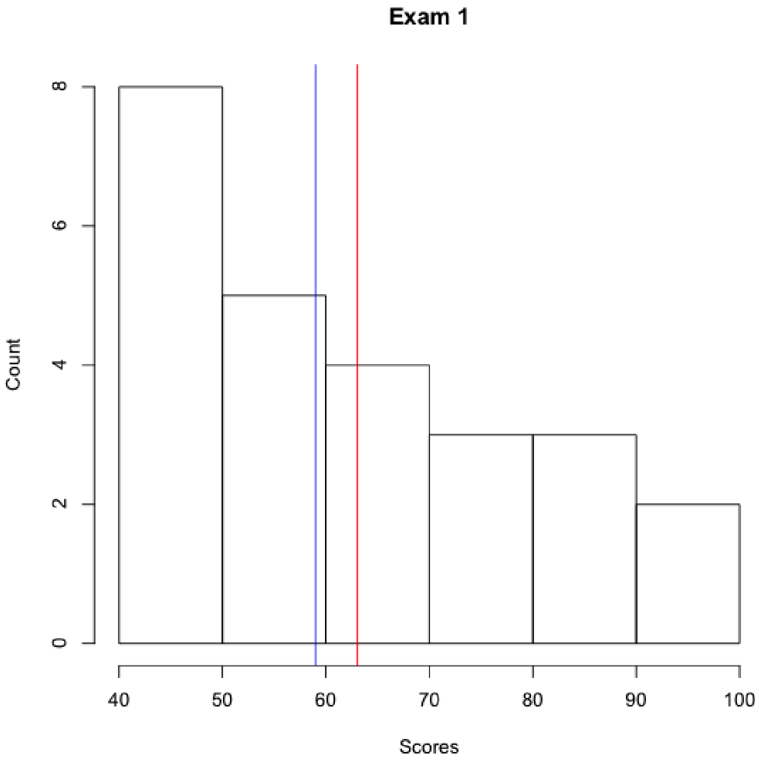
a) Make a histogram (by hand or use *hist(x)* in R) of the test scores on Exam 1 and Exam 2.

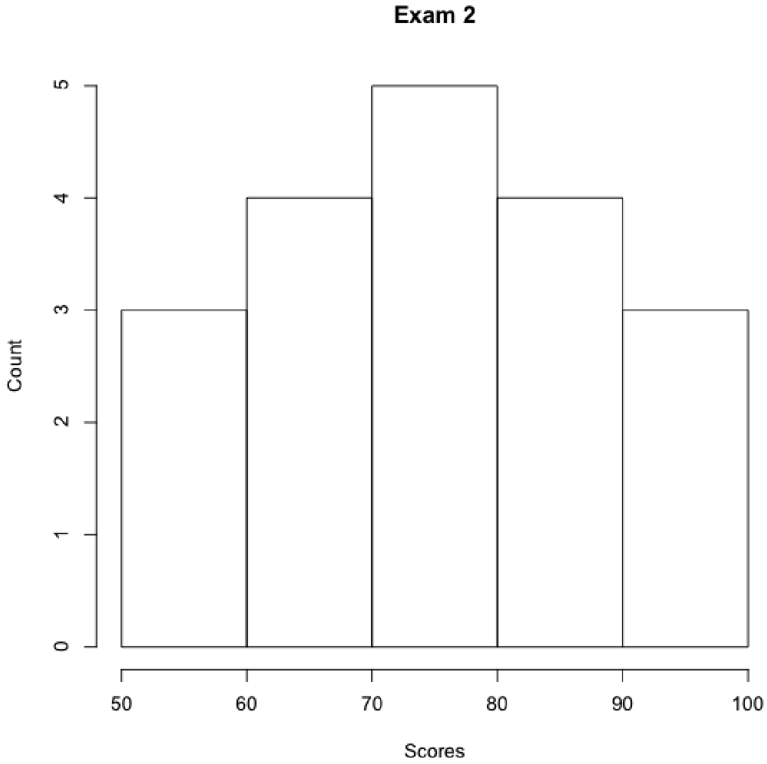
```
ex1 = c(42, 41, 76, 48, 48, 59, 58, 49, 56, 88, 83, 49, 78, 47, 61, 64, 54, 76, 91, 63, 5
hist(ex1, xlab ="Scores", ylab="Count", main = "Exam 1")
abline(v=mean(ex1), col="red")
abline(v=median(ex1), col="blue")

ex2=c(58, 100, 50, 62, 68, 67, 69, 78, 79, 72, 75, 83, 81, 75, 89, 87, 91, 97, 59)
hist(ex2, xlab ="Scores", ylab="Count", main = "Exam 2")
```

b) Calculate mean and median for both exams. You can calculate this by hand or use R. For Exam 2 scores, also calculate three quartiles by hand.

```
mean(ex1)
[1] 63.04
median(ex2)
[1] 75
mean(ex2)
[1] 75.78947
median(ex2)
[1] 75
```

Sort scores for Exam 2 in increasing order ($sort(ex2)$ in R). There are $length(ex2) = 19$ observations:





50 58 59 62 **67** 68 69 72 75 **M=75** 78 79 81 83 **87** 89 91 97 100

The quartiles Q_1, Q_2, Q_3 divide the data into four equal sized parts.

c) Was one test more reasonable than the other? Which measure of center is better to use for each exam?

The second exam was more reasonable. The scores are approximately normally distributed whereas the number of students with higher scores of exam 1 are decreasing. It is better to use median as the measure of center for the first exam because distribution is skewed, median is more resistant than mean in this case. The mean and median for the second exam are almost the same since the distribution is approximately normal.