Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2023

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Sunday October 15, 2023. No late assignments will be accepted.

Question 1: Political Science

The following table was created using the data from a study run in a major Latin American city.[1] As part of the experimental treatment in the study, one employee of the research team was chosen to make illegal left turns across traffic to draw the attention of the police officers on shift. Two employee drivers were upper class, two were lower class drivers, and the identity of the driver was randomly assigned per encounter. The researchers were interested in whether officers were more or less likely to solicit a bribe from drivers depending on their class (officers use phrases like, "We can solve this the easy way" to draw a bribe). The table below shows the resulting data.

[1] Fried, Lagunes, and Venkataramani (2010). "Corruption and Inequality at the Crossroad: A Multimethod Study of Bribery and Discrimination in Latin America. *Latin American Research Review.* 45 (1): 76-97.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

(a) Calculate the χ^2 test statistic by hand/manually (even better if you can do ”by hand” in `R`).

For this first question I was not certin weather or not to include the ”not stopped” columb. ”Not Stopped” does not seem very relevent to the research question. However, in part c we are asked to calculate the standard residuals for that the ”not stopped” values so I was not sure. By the time I noticed this it was too late to email and ask so I decided to preform the calculations of all the data.

The total sum is 42, this will be used in every formula as the denomiator.
I will first find the expected value for each observation if the variables are independent. After that, I will plug the values into the formula which is the sum of all (observed values - expected values)$^2/expectedvalues$.

```
    (27/42)*21 = 13.5   #expected value
    (14-13.5)^2/13.5
    #0.01851852

    (15/42)*21 = 7.5 #expected value
    (7-7.5)^2/7.5
    #0.03333333

    (27/42)*13
    (6-8.36)^2/8.36
    # 0.6662201

    (15/42)*13
    (7-4.64)^2/4.64
    # 1.200345

    (27/42)*8
    (7-5.14)^2/5.14
    #0.6730739

    (15/42)*8
    (1-2.86)^2/2.86
    #1.20965
```

```

  #Now to add them all together and get the Chi Squared statistic

  0.01851852 + 0.03333333 +  0.6662201 + 1.200345 + 0.6730739 + 1.20965
  # My  Chi Squared Statistic 3.80

```

There is also is the quick way to do it in R, which I will use to check my answers. First I create a matrix with the values, then use the chisqu.test() funciton. Since I rounded when prefroming calcualtions by hand, the returned value is slightly different when I run the function. 3.7912 is the more precise value which I will use from here on out.

```

  #MY CODE:
  mat <- matrix(c(14, 7, 6, 7, 7, 1), nrow = 2, ncol = 3)
  chimat <- chisq.test(mat)
  print(chi_mat)

```

(b) Now calculate the p-value from the test statistic you just created (in `R`).[2] What do you conclude if $\alpha = 0.1$?

To calculate the P-Value, the funciton requires the following arguments. Chi Squared Statistic, degrees of freedom, and logial argument. Degrees of freedom = (number of rows - 1)(number of columns - 1) = 2. Whenever we conduct a Chi Squared Test we will only look at the upper tail (there is no lower tail), hence the third argument.

```
#MY CODE:
  pchisq(3.7912, df = 2, lower.tail = FALSE)

```

The P-Value is 0.1502282

Since the P-Value of 0.15 is greater than $\alpha = 0.1$ we fail to reject the Null Hypothesis that driver class and solicitation for bribes are indipendent.

We can not support the alternate hypothesis that there is a realtionship between driver class and soliciation for bribes.

[2] Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

(c) Calculate the standardized residuals for each cell and put them in the table below.

The formula for standard residuals is (observed - expected) / sqrt(expected). The following R functions returns standard residuals, I used the "chimat" matrix which I created in part a of the asssignment. Then I rounded the standard residuals to the third decimal point for convience.

```
    standard_residuals <- round(chimat$stdres,3)
    print(standard_residuals)

```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

(d) How might the standardized residuals help you interpret the results?

You can use the standard residuals to identify at the variable that are further from or closer to being independent, smaller values are closer to being independent. You can also see if your observed value is greater than or less than what we would expect.

In this case the standard residuals mirror each other with the first row being the same as the lower but with the opposite sign (negative or positive). As one category of one goes up the other goes down, this suggests a negitive association between the variables. However, we were not able to reject the Null that the variables are indipendent so this does not mean much.

Question 2: Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[3] Answering this question with observational data is pretty difficult due to potential confounding problems (e.g. the districts that choose female politicians are likely to systematically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, $\frac{1}{3}$ of village council heads have been randomly reserved for women. A subset of the data from West Bengal can be found at the following link: `https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv`

Each observation in the data set represents a village and there are two villages associated with one GP (i.e. a level of government is called "GP"). Figure **??** below shows the names and descriptions of the variables in the dataset. The authors hypothesize that female politicians are more likely to support policies female voters want. Researchers found that more women complain about the quality of drinking water than men. You need to estimate the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages.

For some reason I could not get the graphic to show in my Tex.pdf so deleted it.

[3] Chattopadhyay and Duflo. (2004). “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica.* 72 (5), 1409-1443.

(a) State a null and alternative (two-tailed) hypothesis.

Null Hypothesis: The reservation policy will not have an effect on the number of new or repaired water treatment facilities in the village.
Alternate Hypothesis: The reservation policy will have an effect on the number of new or repaired water treatment facilities in the village.

(b) Run a bivariate regression to test this hypothesis in R (include your code!).

MY CODE:

```r
# I load in the data and store it in "data"
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
#Take an initial look at the data and plot the variables of interest, to make it easier to view
View(data)
plot(data$reserved,data$water)
#I subset the two variables we are interested in to make it easier to look at.
relevant_data <- data[c("reserved","water")]
print(relevant_data)
#Now I use the lm() and summary() functions to get my p-score
sum <- summary(lm(data$water~data$reserved))
print(sum)sum$coefficients

```

P-Value is 0.0197
The y intercept is 14.738318
Beta is 9.252423

We can also use these two lines of code to find the t-statistic and critical value respectiley. In the second line of code I use 1 - .05/2 to signify a two tailed test at 95 percent confidence. I use the degrees of freedom which was an output from the first line of code.

```r
cor.test(data$water, data$reserved)
critical_value <- qt(1-.05/2,df=320)

```

The t-statistic is 2.3437 and the critical value is 1.967405.

We can determine if we reject the Null either comparing the p-vaue to alpha or the t-statistic to the critical value, both will result in the same outcome.

We have sufficient evidence to reject the Null Hypothesis at 5 percent significance level because p-value less than .05 and/or the t-statistic is greater than the critical value. Therefore, we support that alternate hypothesis that the seat resererervation policy effects the number of new or repaired water treatment facilities.

(c) Interpret the coefficient estimate for reservation policy.

The Beta X coefficient represents an estimate of how much one incremental change in X will affect Y. In this case, since X is binary, one change in x (from 0 to 1) means that the reservation policy is in place. Beta shows us that on average we estimate the implementation of the reservation policy is correlated with approximately 9.25 new or renovated water treatment facilities. Similarly, alpha, the Y-intercept, represents the average amount of new or repaired water treatment facilities when the reservation policy is not in place.

The formula Y = 14.73 + 9.252 (x) helps us understand this. When X=0 (no reservation policy), 9.252 * 0 = 0. We are left only with the Y-intercept. Inversely, when X = 1, (reservation policy implemented), 9.252 * 1 = 9.252 and 14.73 + 9.252 = 23.982 which is the estimated average amount of new or repaired water treatment facilities when the reservation policy is in place.

As we can see, the y intercept (14.73) represents the estimated average number of new or repaired drinking water facilites when the reservation policy is not in place.
x (9.252) represents the estimated average difference of new or repaired drinking water facilities when the reservation policy is in effect vs when it is not.

Finally (x*1) + y intercept = estimated number of new or repaired drinking water faciliteis when the reservation policy is in place (23.982).

The correlation coefficient of the sample is 0.1299079. displaying that the correlation is predicted to not be relitivley weak. A correlation further from zero and cloer to 1 or -1 would be signify a stronger relationship. None the less, the correlation coefficient is above zero so we do have positive cerrolation. (I got the correlation coefficient from the qt() code shown above).