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`).

Answers 1 (a)

Statistical independence: Two variables are statistically independent if the conditional distributions of the population are identical across categories.

H0: The variables are statistically independent
Ha: The variables are statistically dependent

We are going to calculate a test-statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution

f observed = O = observed frequency = the raw count

f expected = E = what we would expect for independent sample

If H0 is true, then we would expect f observed = f expected

Table 1: Observed Values

	Not Stopped	Bribe requested	Stopped/given warning	Total
Upper class	14	6	7	27
Lower class	7	7	1	15
Total	21	13	8	42

Table 2: Expected Values

To calculate Expected Values: (Raw Total/ Grand Total) * Column Total

	Not Stopped	Bribe requested	Stopped/given warning	Total
Upper class	13.5	8.357	5.142	27
Lower class	7.5	4.642	2.857	15
Total	21	13	8	42

Calculate the χ^2 Test Statistic by hand

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Inserting values into the formula and adding each total gives you:

0.0185 + 0.6647 + 0.671 + 0.0333 + 1.198 + 1.207

$\chi^2 =$ **3.7925**

Answers 1 (b)

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

Formula in R:

pchisq(χ^2, df = (rows-1)(columns-1), lower.tail=FALSE)

```
#OBSERVED VALUES
Not_Stopped <- c(14,7)
Bribed <- c(6,7)
Stopped_Given_Warning <- c(7,1)

df = data.frame (Not_Stopped, Bribed, Stopped_Given_Warning )
row.names(df) <- c("Upper Class", "Lower Class")

#EXPECTED VALUES
Not_Stopped_E <- c(13.5,7.5)
Bribed_E <- c(8.357,4.642)
Stopped_Given_Warning_E <- c(5.142, 2.857)

df_E = data.frame (Not_Stopped_E, Bribed_E, Stopped_Given_Warning_E )
row.names(df_E) <- c("Upper Class", "Lower Class")

chi_test_n <- chisq.test(df, df_E)
chi_test_n

 Chi_Result <- 3.7925

 pchisq(Chi_Result, df = 2, lower.tail=FALSE)

 P <- 0.1501306
```

The P-Value is greater than $\alpha = 0.1$. That being the case, we have evidence to accept the null hypothesis, and conclude that the two distributions are statistically independent in this instance, with a confidence level of $\alpha = 0.90$.

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

Answers 1 (c)

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

Formula used in R :

```
chi_test_n$stdres
```

Answers 1 (d)

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

Interpretation: The upper class were stopped less than the lower class, whilst a bribe was requested more from the lower class than the upper class. Those stopped or given a warning were more likely to be from the upper class.

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 1 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.

Figure 1: Names and description of variables from Chattopadhyay and Duflo (2004).

Name	Description
`GP`	An identifier for the Gram Panchayat (GP)
`village`	identifier for each village
`reserved`	binary variable indicating whether the GP was reserved for women leaders or not
`female`	binary variable indicating whether the GP had a female leader or not
`irrigation`	variable measuring the number of new or repaired irrigation facilities in the village since the reserve policy started
`water`	variable measuring the number of new or repaired drinking-water facilities in the village since the reserve policy started

[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.

Answers 2 (a)

The null hypothesis is that there is no association between the reservation policy in place in each GP for female leaders and the number of new or repaired drinking water facilities in the village. In that case the slope will = 0.

The alternative hypothesis states there is a correlation between the reservation policy in place for women leaders in a district (either positive or negative) and the water policies in place. In that case the slope will be greater or less than 0.

Answers 2 (b)

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

First, looking at the 108 GPs where the reservation policy is in place, we can see that there are female leaders in all of them. In the 214 unreserved GPs, only 16 GPs have female leaders. This appears to show a direct link between the reservation policy and number of female candidates.

```
length(which(data_India$reserved == 1))
#there arr 108 GPs with a reservation policy in place
length(which(data_India$reserved == 0))
#there are 214 GPs without a reservation policy in place

#all the reserved states have female candidates.
length(which(data_India$reserved == 1 & data_India$female == 1))

#another 16 out of 214 remaining states have female candidates
data_India$matched <- ifelse(data_India$reserved == data_India$female,"yes","no")
length(data_India$matched)

num_no <- length(which(data_India$matched == "no"))
num_no #there are 16 of the 214 unreserved states with female candidates.
```

That being the case, this analysis will run a regression analysis to look at the association between the GPs with female representation and the number of new or repaired water facilities.

```
data_female <- (data_India$female == 1)
data_male <- (data_India$female == 0)

model_water <- summary(lm(data_India$water~data_female))
print(model_water)


Call:
lm(formula = data_India$water ~ data_female)

Residuals:
```

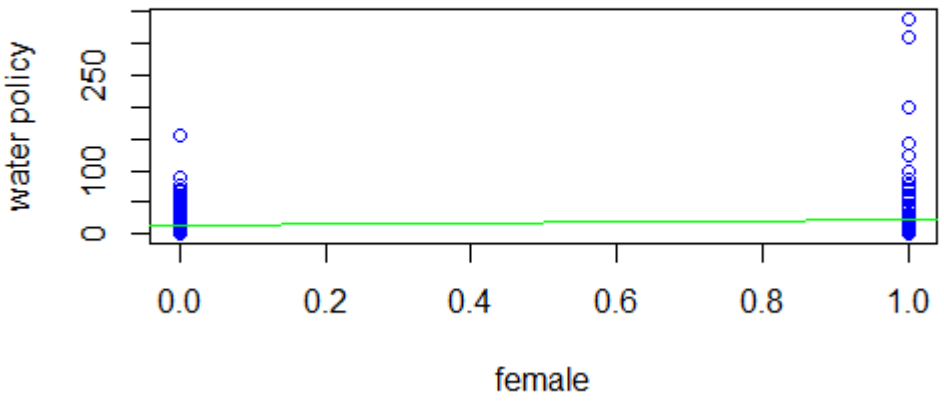
```
Min      1Q Median      3Q     Max
-22.68 -14.78  -7.81   2.29 317.32

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)       14.813      2.382   6.220 1.56e-09 ***
data_femaleTRUE    7.864      3.838   2.049   0.0413 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 320 degrees of freedom
Multiple R-squared:  0.01295,Adjusted R-squared:  0.009867
F-statistic: 4.199 on 1 and 320 DF,  p-value: 0.04126
```

We can also plot the association

```
plot(data_India$water ~ data_female, col = "blue")
abline(coef(lm(data_India$water~data_female)), col = "green")
abline(coef(lm(data_India$water~data_male)), col = "purple")
```



Answers 2 (c)

(c) Interpret the coefficient estimate for reservation policy.

Answer: The slope is equal to 7.864, indicating a correlation between female leaders and water policy.

The P-Value is less than 0.05, so the test can be considered significant, and we have evidence to reject the null hypothesis in this instance.