Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

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 16, 2022. No late assignments will be accepted.
- Total available points for this homework is 80.

Question 1 (40 points): 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).

The first step in calculating the χ^2 test statistic is to calculate the expected frequency (Fe) for each cell of the original table. The equation to calculate Fe in each cell is Fe = (row total/grand total)*(column total). I added margins to the original table (entered into a csv file) so that I could see the row, column, and grand totals. This expanded table of observed frequencies with row/column/grand totals is included below:

Table of Observed Values with Row, Column, and Grand totals

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

Then, I calculated Fe for each cell using the equation:
Fe = (row total/grand total)*(column total)

The table with calculated Fe in each cell is below:

Table of Expected Values: Calculated for Question 1(a)

	Not Stopped	Bribe requested	Stopped/given warning	
Upper class	13.50	8.36	5.14	
Lower class	7.50	4.64	2.86	

Now that I have the Fe for each cell, I can calculate the χ^2 test statistic. The equation for this is:

χ^2 test statistic = $\sum$ ((Fo-Fe)^2/Fe)

I manually calculated this equation (on paper), by summing the result of ((Fo-Fe)^2/Fe) for each cell of the table, and got the result:

χ^2 test statistic = 3.8269

Note: I calculated the test statistic using values for each cell that were rounded to 4 decimal places, so the resultant test statistic might be slightly off from what I would have calculated had I done so without these rounded values. The difference should not be substantial however, and I would expect the test statistic calculated without rounded values to equal 3.8 as well.

(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 from my test statistic, I need to know my degrees of freedom.

I calculate my degrees of freedom as follows:
df = (number of rows - 1)(number of columns - 1)
df = (2-1)(3-1)
df = (1)(2)
df = 2.

Next, I plug the degrees of freedom (2) and the χ^2 test statistic (3.8269) into the pchisq() function in R:

```
pchisq(3.8269, df = 2, lower.tail = FALSE)
```

Which returns a p-value from the test statistic of: 0.1475704.

Because the p-value is greater than $\alpha = 0.1$ (p-value > 0.1), we fail to reject the null Hypothesis (that the observations for upper and lower classes are statistically independent).

[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.

I now calculate the standardized residual for each cell in the table. The equation I use to calculate each cell's standardized residual is: z = (Fo - Fe)/se.
se is the standard error of Fo-Fe, presuming that H0 is true, and the equation I use to calculate the standard residual for each cell is:
z = ((Fo-Fe) / $\sqrt{Fe(1 - rowproportion)(1 - columnproportion)}$), or in even more detail:
z = ((Fo-Fe) / $\sqrt{Fe(1 - (RowTotal/Total))(1 - (ColumnTotal/Total))}$).

I calculate this for each cell and get the following standardized residuals:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220	-1.6437	1.5039
Lower class	-0.3220	1.6619	-1.5246

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

The standardized residual allows us to better see if the difference between the observed frequency and expected frequency is due to more than chance. The standardized residual of a cell is the number of standard errors (se) that Fo-Fe falls from the value we expect when H0 is true.

The higher (in absolute terms) a standardized residual is, is an indication of greater chance of association between the two variables (or in other words, stronger chance that the variables are not independent statistically). Generally, if this absolute value of standardized residual in a cell is over 3, we can assume the two variables are statistically associated. Becuase all absolute values of standardized residuals that we found in this case are less than 2, we cannot reject the null hypothesis that the variables are statistically independent for any cell.

However, because the absolute values of the standardized residuals for Bribe requested and class, and Stopped/warned and class are closer to 2 than for not stopped and class, there is a stronger chance of association between class and whether a bribe was requested, and between class and whether the individual is warned/stopped, than there is between class and not stopped. In other words, we can say with greater confidence that class and not being stopped are statistically independent, and say

with less confidence that class and bribed or class and stopped/given warning are statistically independent (though for all variables, we fail to reject claim that they are statistically independent of one another).

Question 2 (40 points): 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.

First, I loaded the csv data into R , saving it as the object "women", with the code:

```
women <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
```

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

Null Hypothesis (H0) = There is no association between whether a village had the reservation policy (the reserved variable) and the number of new or repaired drinking water facilities (water variable) in the village.

Alternative Hypothesis (Ha) = The number of new or repaired drinking water facilities in a village is correlated with whether or not they had the reservation policy (there is a relationship) between these two variables.

Note: I am using the reserved variable not the female variable, because the question is asking me to estimate the effect that the reservation policy (not explicitly whether a village has a woman leader) has on the water variable.

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

I wanted to test the strength of the relationship between the reserved policy (my x, or Explanatory Variable) and the water variable (my y, or Response Variable), so I ran the code below to test my hypothesis using bivariate regression:

```
lreg_water_reserved <- lm(women$water~women$reserved)
summary(lreg_water_reserved)
```

The summary output of this bivariate regression test was:

```
Residuals:
 Min      1Q     Median      3Q     Max
 -23.991 -14.738  -7.865   2.262 316.009

Coefficients
              Est   Std.Error t value Pr(>|t|)
(Intercept)    14.738  2.286  6.446   4.22e-10***
women$reserved 9.252   3.948  2.344   0.0197*

---Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 33.45 on 320 degrees of freedom
Multiple R-squared:  0.01688,Adjusted R-squared:  0.0138
F-statistic: 5.493 on 1 and 320 DF,  p-value: 0.0197
```

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate is an estimate change in mean of the response variable (in our case, water), that we are estimated to see with a one unit change in the explanatory variable (reservation).

As seen in the above section, the coefficient estimate for reservation policy is 9.252.

We therefore interpret the coefficient estimate as: "we estimate that on average, villages with the reservation policy will have 9.252 more new or repaired drinking-water facilities than villages without the reserve policy have, since the reserve policy started."

We also interpret the p-value to test our hypothesis and add to our interpretation of the coefficient estimate. The p-value is the probability of observing this coefficient estimate or one that is more extreme if the null hypothesis is actually correct. A low p value indicates that we can reject our null hypothesis. The p-value for this regression test is 0.0197. This is a low p-value (lower at least then 0.05, which is often used as the "benchmark"). Thus, this bivariate regression output leads us to reject our null hypothesis, that there is no relationship / association between the two variables (reservation policy and water). Using these statistics, we predict villages with the policy to have introduced more (on average) new / repaired drinking water facilities than those without the policy.