Problem Set 2

Applied Stats/Quant Methods 1
Dan Zhang 23335541

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).

First,let's calculate the expected frequency $f_e = \frac{f_{\text{row total}} \times f_{\text{col total}}}{f_{\text{grand total}}}$. Then we can calculate the χ^2, which is $\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = 3.791168$

```
#create the resulting data table
resulting_data<-matrix(c(14,6,7,7,7,1),nrow = 2,byrow = TRUE)

#calculate the expected frequencies
  #calculate the sum of rows and columns of the data table
   total_samples<-sum(resulting_data)
   row_sums<-rowSums(resulting_data)
   col_sums<-colSums(resulting_data)
  #calculate the frequencies
   expected_freq<-outer(row_sums,col_sums)/total_samples

#calculate Chi square statistic
chi_square_statistic<-sum((resulting_data - expected_freq)^2/expected_
    freq)
```

Next, we need calculate the degree of freedom: $df = (nrow - 1) \times (ncol - 1) = 2$

```
#calculate degree of freedom
degree_of_freedom<-(nrow(resulting_data)-1)*(ncol(resulting_data)-1)
```

Last use R function to examine my results:

```
#examine the results
chisq_result<-chisq.test(resulting_data)
chisq_result
```

```
chisq_result
Pearson's Chi-squared testdata:
resulting_dataX-squared = 3.7912, df = 2, p-value = 0.1502
```

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

As we already calculated above, the p-value=0.1502, which is greater than $\alpha = 0.1$

So we can interpret that, Based on the chi-square test, at a 99% confidence level, we cannot reject the null hypothesis. Which indicates that we do not have enough evidence to support there is a statistically significant dependence between driver class and whether officers were more or less likely to solicit a bribe from drivers.

Also, we can use R function to calculate p-value:

```
alpha<-0.1
p_value<-pchisq(chi_square_statistic,degree_of_freedom,lower.tail = F)
p_value
```

[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.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class			
Lower class			

Calculate the residuals for each cell by using residual = $\frac{(f_o - f_e)}{\sqrt{f_e}}$

```
#calculate the standardized residuals for each cell
cell_residuals<-c()
cell_residual<-0
row_sums[[2]]
for (i in seq(1:nrow(resulting_data))){
  for (j in seq(1:ncol(resulting_data))){
    cell_residual<-(resulting_data[i,j]-expected_freq[i,j])/sqrt(expected
    _freq[i,j]*(1-row_sums[i])/sum(resulting_data)*(1-col_sums[j])/sum(
    resulting_data))
    cell_residuals<-c(cell_residuals,cell_residual)
  }
}

cell_residuals_table<-matrix(cell_residuals,nrow = 2,byrow = TRUE)
colnames(cell_residuals_table)<-c("Not Stopped","Bribe requested","
   Stopped/given warning")
rownames(cell_residuals_table)<-c("Upper class","Lower class")
```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.2506402	1.938782	2.549510
Lower class	-0.4582576	3.544769	-4.661392

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

A standardized residual with a larger absolute value indicates that a cell contributes more significantly to the chi-square statistic.In this sample, the observations of 'Not stopped' from two classed cannot give more significant contributions to the find whether there isa statistically significant dependence between driver class and police officer bribe.

Also, the sign of a standardized residual can help indicate the direction of the deviation of the observed frequency and the expected frequency. A positive value implies that the observed count is greater than expected, while a negative value suggests that the observed count is less than expected.

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.

H_0: The reservation policy does not have an effect on the number of new or repaired drinking water facilities in thevillages: β=0.

H_A: The reservation policy has an effect on the number of new or repaired drinking water facilities in thevillages: $\beta \neq 0$

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

```
#load the subset data from West Bengal
data<-read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")

#use bivariate regression
reserved_data<-as.factor(data$reserved)
bivariate_reg<-lm(data$water ~ reserved_data,data=data)
summary(bivariate_reg)
```

```
lm(formula = data$water ~ reserved_data, data = data)
Residuals:    Min      1Q      Median      3Q     Max
-23.991 -14.738  -7.865   2.262 316.009

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
 (Intercept)      14.738      2.286   6.446 4.22e-10 ***
 reserved_data1    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.

Based on the data, we can see from the bivariate regression results. When there is no reserved policy, the estimated mean value of new or repaired drinking water facilities is 14.738. When there is reserved policy, the estimated mean number of new or repired drinking water facilities in villages is 9.252 higher than when there is not.

In conclusion, the reservation of female politicians has a positive impact on the number of new or repqired drinking water facilities in villages, compared to the case where no female politicians are reserved. The p-value for this result is less than the significance level of 0.05, indicating that the effect is significant.