# 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 $\chi^2$ test statistic by hand/manually (even better if you can do "by hand" in R).

```r
#Question 1-a

# create the data frame
df_original <- data.frame(
  'Not Stopped' = c(14,7),
  'Bribe requested' = c(6,7),
  'Stopped/given warning' = c(7,1),
  row.names = c("Upper class","Lower class")
)
# set H0 and H1
cat("H0:There is no significant association between the variables")
cat("H1:There is a significant association between the variables")

#Calculate the Expected Frequencies:
result_1_1<-(27*21)/42
result_1_1
result_2_1<-(27*13)/42
result_2_1
result_3_1<-(27*8)/42
result_3_1
result_1_2<-(15*21)/42
result_1_2
result_2_2<-(15*13)/42
result_2_2
result_3_2<-(15*8)/42
result_3_2

#Calculate X2 test statistic
X2_result_1_1<-(14-result_1_1)^2/result_1_1
X2_result_1_1
X2_result_2_1<-(6-result_2_1)^2/result_2_1
X2_result_2_1
X2_result_3_1<-(7-result_3_1)^2/result_3_1
X2_result_3_1
X2_result_1_2<-(7-result_1_2)^2/result_1_2
X2_result_1_2
X2_result_2_2<-(7-result_2_2)^2/result_2_2
X2_result_2_2
X2_result_3_2<-(1-result_3_2)^2/result_3_2
```

```
X2_result_3_2
X2_result<-X2_result_1_1+X2_result_2_1+X2_result_3_1+X2_result_1_2+X2_
    result_2_2+X2_result_3_2
print(X2_result)
```

- Results: the square of X is 3.17.

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

```
#Question 1-b
#Calculate the df
df<-(2-1)*(3-1)
#Calculate p_value
p_value<-1-pchisq(X2_result,df)
print(p_value)
#Results
cat("because p value is 0.15 which is  higher than 0.1,
    we need to accept H0. There is no significant association between the
     variables.")
```

- Results: because p value is 0.15 which is higher than 0.1, we need to accept H0. There is no significant association between the variables.

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

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 0.322 | -1.644 | 1.523 |
| Lower class | -0.322 | 1.642 | -1.523 |

```
#Question 1-c
#Calculate standardized residuals
SR_result_1_1<-(14-result_1_1)/sqrt(result_1_1*(1-27/42)*(1-21/42))
SR_result_1_1
SR_result_2_1<-(6-result_2_1)/sqrt(result_2_1*(1-27/42)*(1-13/42))
SR_result_2_1
SR_result_3_1<-(7-result_3_1)/sqrt(result_3_1*(1-27/42)*(1-8/42))
SR_result_3_1
SR_result_1_2<-(7-result_1_2)/sqrt(result_1_2*(1-15/42)*(1-21/42))
SR_result_1_2
```

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

```
SR_result_2_2<-(7-result_2_2)/sqrt(result_2_2*(1-15/42)*(1-13/42))
SR_result_2_2
SR_result_3_2<-(1-result_3_2)/sqrt(result_3_2*(1-15/42)*(1-8/42))
SR_result_3_2
```

(d) How might the standardized residuals help you interpret the results?
- Results:
The standardized residual between "Upper class"and "Not Stopped" is 0.322. Because it is close to zero, it suggests that the observed and expected frequencies are close, and there may not be a strong association between "Upper class" and "Not Stopped".

The standardized residual between "Upper class"and "Bribe requested" is -1.644. Because it is negative, it suggests the observed frequency of "Upper class" individuals requesting a bribe is lower than what would be expected. It suggests that "Upper class" might not likely to request a bribe.

The standardized residual between "Upper class"and "Stopped/given warningd" is 1.523. Because it is a positive value.This suggests that the observed frequency of "upper-class" individuals being stopped is higher than expected. It suggests that "Upper class" might be likely to be stopped.

The standardized residual between "Lower class"and "Not Stopped" is -0.322. Because it is close to zero, it suggests that the observed and expected frequencies are close, and there may not be a strong association between "Lower class" and "Not Stopped".

The standardized residual between "Lower class"and "Bribe requested" is 1.642. Because it is a positive value.This suggests that the observed frequency of "Lower class" individuals being stopped is higher than expected. It suggests that "Lower class" might be likely to request a bribe.

The standardized residual between "Lower class"and "Stopped/given warning" is -1.532. Because it is negative, it suggests the observed frequency of "Lower class" individuals being stopped is lower than expected. It suggests that "Lower class" might not likely to be stopped.

# 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).

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

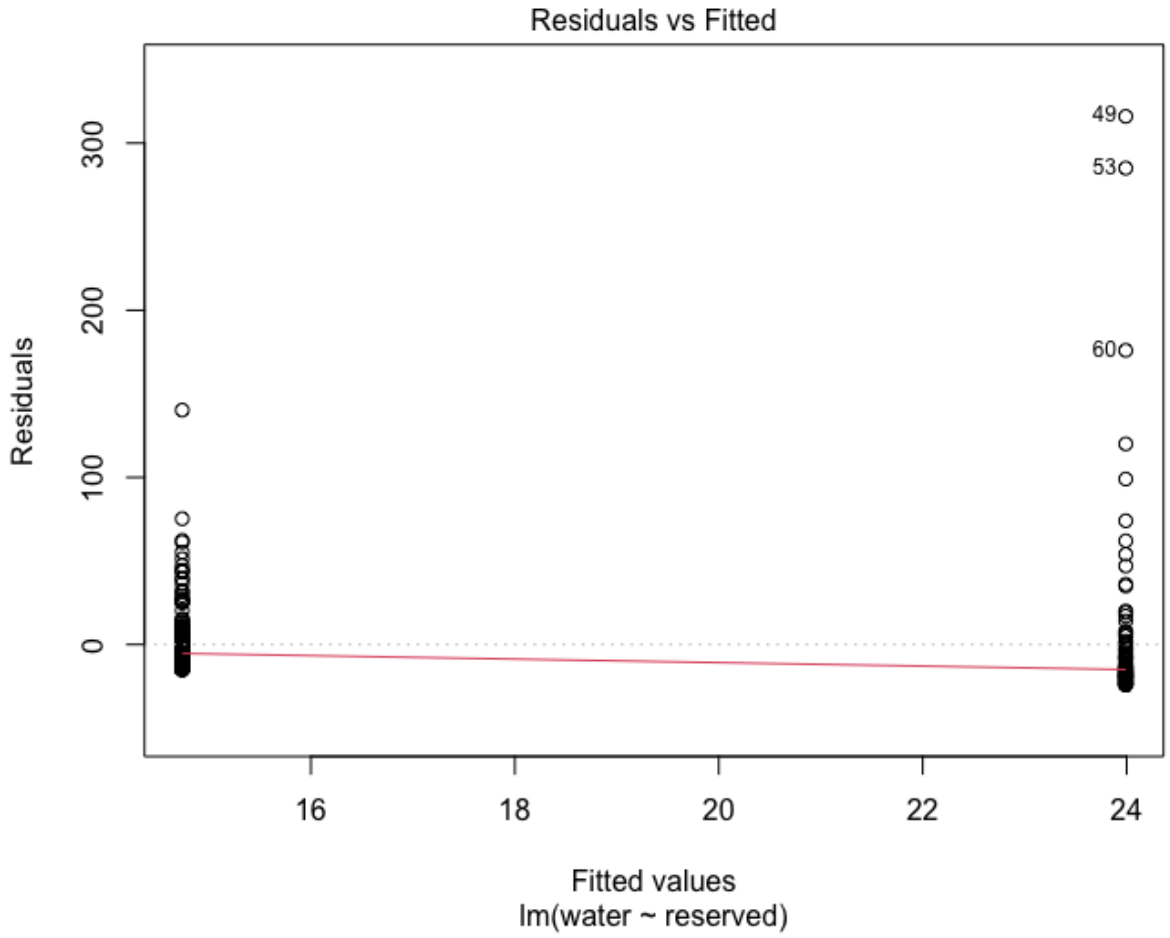
(a) State a null and alternative (two-tailed) hypothesis.
Null Hypothesis (H0): The reservation policy for women leaders in village has no effect on the number of new or repaired drinking water facilities in the villages.

Alternative Hypothesis (H1): The reservation policy for women leaders has a significant effect on the number of new or repaired drinking water facilities in the villages.

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

```r
#Question 2-b

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

# Check the structure of the dataset
str(dataset)

# Run the bivariate regression
model_bivariate_regression <- lm(water ~ reserved, data=dataset)
model_bivariate_regression

# Summarize the regression results
summary(model_bivariate_regression)
```

```
Call:
lm(formula = water ~ reserved, data = dataset)

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       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.
- Results:
Coefficient Value: The positive coefficient estimate of 9.252 suggests that when the village council head's position is reserved for women, there is an increase by 9.252 on the number of new or repaired drinking water facilities compaired to situations with no reservation policy.

Statistical Significance: The p-value of 0.0197 suggests statistical significance which is smaller than threshold of 0.05. This suggests that the implementation of a reservation policy has a significant impact on the number of new or repaired drinking water facilities.