# Problem Set 2

Jacqueline Bouvier 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 Multi-method 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`).

```
# Stating my Null and Alternate Hypothesis
# Ho - Police officers are more likely to stop you depending on if you
    seeminly
# look upper class H1 - Police officers do not care of social status

#Finding X^2 test statistic Formula X^2 = (fo - fe)^2/fe - the observed -
     expected squared
# and then divided by the expected

# First find total of columns and rows
# Not Stopped = 21 Bribe Requested =13 Stopped/given warning = 8
# Upper class = 27 Lower Class = 15
# Total records = 42
# Find percentage of upper and lower class
upper <- 27/42
lower <- 15/42
# We calculate the percentage of occurance out to be 64.29% for Upper
    class and
# Lower class to be 35.71%
# We will then take these percentages and calculate and expected value
    for each outcome

# Upper class row total of not stopped = 21 we multiply this by 64% to
    get expected
```

```
# to equal 13.50. Then we continue this with each column total for upper class by 64.29%
Up_notstop_expec <- 21*0.6429
Up_Bribe_expec <- 13*0.6429
Up_Stop_expec <- 8*0.6429
# We repeat this with the lower class and the percentage 35.71% and totals
Low_notstop_expec <- 21*0.3571
Low_Bribe_expec <- 13*0.3571
Low_Stop_expec <- 8*0.3571
# We now have our expected values for our formula -we will take each piece of data and
# put it into the formula - each observed and corresponding expected value will be squared
# and then divided by expected. We will then sum all the values together for the X^2 test statistic

Xsqrd <- (((14-Up_notstop_expec)^2/Up_notstop_expec) +
            (6 - Up_Bribe_expec)^2/ Up_Bribe_expec +
              (7 - Up_Stop_expec)^2/ Up_Stop_expec +
          (7 - Low_notstop_expec)^2/ Low_notstop_expec +
              (7 - Low_Bribe_expec)^2/ Low_Bribe_expec +
              ( 1 - Low_Stop_expec)^2/ Low_Stop_expec
)
# Our X^2 test statistic is 3.79
```

]

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

```
# We can now find our p - value
# First to find the degrees of freedom with formula
# df = (r-1)(c-1)
df <- (2-1)*(3-1)
# Our degrees of freedom is 2
# plug into our p-value for chisq formula pchisq( X^2, df=, lower.tail=F)
pvalue <- pchisq(3.79, df=2, lower.tail=FALSE)
# Our p-value is 0.15032
```

]

[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 | 0.321 | -1.642 | 1.523 |
| Lower class | -0.321 | 1.642 | -1.523 |

```r
# We would have to accept our Ho as we do not have enough statistical
    significance to reject
# Since we are accepting our Ho we now should calculate standardized
    residuals
# We can use the formula z = (fo - fe)/sqrt(fe(1-rowprop)(1-columnprop))

# Standardized residual for upper class not stopped
z_UpNS <- (14 - Up_notstop_expec)/sqrt((Up_notstop_expec*(1-(27/42))*
    (1-(21/42))))

# Standardized residual for upper class bribe requested
z_UpBr <- (6 - Up_Bribe_expec)/sqrt((Up_Bribe_expec*(1-(27/42))*(1-(13/
    42))))

# Standardized residual for upper class stopped
z_UpStop <- (7 - Up_Stop_expec)/sqrt((Up_Stop_expec*(1-(27/42))*(1-(8/42)
    )))


# Standardized residual for lower class not stopped
z_LowNS <- (7 - Low_notstop_expec)/sqrt((Low_notstop_expec*(1-(15/42))*
    (1-(21/42))))


# Standardized residual for lower class bribe requested
z_LowBr <- (7 - Low_Bribe_expec)/sqrt((Low_Bribe_expec*(1-(15/42))*(1-(13
    /42))))

# Standardized residual for lower class bribe requested
z_LowStop <- (1 - Low_Stop_expec)/sqrt((Low_Stop_expec*(1-(15/42))*(1-(8/
    42))))
```

]

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

```
#As we are accepting the Ho we would expect that little to none of our results would
# exceed 2 in absolute value meaning that it further that we should be accepting our Ho.
# Since the points given to us are not considered outliers. We can also assume
# since that the value is not more than 2 the cell has less observations than expected if the
# variables were truly independent.
```

]

## 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.

```
# Ho — The gender of a leader of a village had no impact on water facilities
# Ha— Female leaders are more likely to support water facilities therefore there would be more facilities
```

]

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

```
#Looking at the data
plot(data$female,data$water)
str(data)
mean_female <- mean(data$female)
mean_water <- mean(data$water)
sd_female <- sd(data$female)
sd_water <- sd(data$water)
# we run the regression
biV <- lm(data$water ~ data$female)
summary(biV)
# regression by hand

r <- cov(data$water, data$female)/ (sd_water * sd_female)
n <- dim(data)[1]
t_stat <- ((r*sqrt(n-2))/ (1-r^2))
p_value <- 2*pt(t_stat, n-2, lower.tail=FALSE)
```

]

(c) Interpret the coefficient estimate for reservation policy.

```
# The formula we would use for a bivarate regression test is y=Bo +BiX + E (y=14.813+7.864X+E) but with large data sets through r
# Running our data through the lm(Y ~ X ) function and looking at the information given we can assume
# that there is little statistical evidence that the gender of the political leader
# has anything to do with the number of water facilities/polices. We would therefor accept the Ho.
```

]