# Problem Set 2

Applied Stats/Quant Methods 1
Yucheng Wang/Student ID: 23367784

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
# Create the observed frequency table
observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)
chisq.test(observed)

# Calculate row totals
row_totals <- rowSums(observed)

# Calculate column totals
col_totals <- colSums(observed)

# Calculate the overall total
total <- sum(observed)

# Calculate the expected frequency table
expected <- outer(row_totals, col_totals) / total

# Calculate chi-squared values
chi_square_values <- (observed - expected)^2 / expected

# Calculate the chi-squared test statistic
chi_square_statistic <- sum(chi_square_values)
cat("Chi-Squared Test Statistic: ", chi_square_statistic, "\n")
```

1. Chi-Squared Test Statistic: 3.791168

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

```r
# Degrees of freedom
df <- (nrow(observed) - 1) * (ncol(observed) - 1)

p_value <- 1 - pchisq(chi_square_statistic, df)

alpha <- 0.1
cat("P-Value: ", p_value, "\n")
if (p_value <= alpha) {
  cat("Reject the null hypothesis. There is evidence of an association
   between variables.\n")
} else {
  cat("Fail to reject the null hypothesis. There is no strong evidence of
    an association between variables.\n")
}
```

2. Fail to reject the null hypothesis. There is no strong evidence of an association between variables.

[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 | | | |

```r
# Calculate the standardized residuals for each cell
standardized_residuals <- (observed - expected) / sqrt(expected)

result_table <- matrix(NA, nrow = 2, ncol = 3)
colnames(result_table) <- c("Not Stopped", "Bribe requested", "Stopped/
    given warning")
rownames(result_table) <- c("Upper class", "Lower class")
result_table <- format(standardized_residuals, digits = 2)
rownames(result_table) <- c("Upper class", "Lower class")

result_table
```

**result**

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 0.14 | -0.82 | 0.82 |
| Lower class | -0.18 | 1.09 | -1.10 |

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

- For the cell corresponding to "Upper class" and "Not Stopped," the standardized residual is 0.14. Since it is close to zero, it suggests that the observed and expected frequencies for this cell are reasonably close, and there may not be a strong association between being in the upper class and not getting stopped.
- For the cell corresponding to "Upper class" and "Bribe requested," the standardized residual is -0.82. This negative value indicates that the observed frequency of upper-class individuals requesting a bribe is lower than expected. It suggests that being in the upper class might be associated with a lower likelihood of requesting a bribe.
- For the cell corresponding to "Upper class" and "Stopped/given warning," the standardized residual is 0.82. This positive value suggests that the observed frequency of upper-class individuals being stopped or given a warning is higher than expected. It implies that being in the upper class might be associated with a higher likelihood of being stopped or given a warning.
- For the cell corresponding to "Lower class" and "Not Stopped," the standardized residual is -0.18, which is close to zero. It suggests that there may not be a strong association between being in the lower class and not getting stopped.
- For the cell corresponding to "Lower class" and "Bribe requested," the standardized residual is 1.09, indicating that the observed frequency of lower-class individuals requesting a bribe is higher than expected. It suggests that being in the lower class might be associated with a higher likelihood of requesting a bribe.
- For the cell corresponding to "Lower class" and "Stopped/given warning," the standardized residual is -1.10, indicating that the observed frequency of lower-class individuals being stopped or given a warning is lower than expected. It implies that being in the lower class might be associated with a lower likelihood of being stopped or given a warning.

# 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.

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

- Null Hypothesis (): The reservation policy for women in village council heads has no effect on the number of new or repaired drinking water facilities in villages. Mathematically, this can be stated as: , where is the coefficient of the "reserved" variable in the regression model.
- Alternative Hypothesis (): The reservation policy for women in village council heads has an effect on the number of new or repaired drinking water facilities in villages. Mathematically, this can be stated as: . This is a two-tailed alternative hypothesis because we are testing if the coefficient is significantly different from zero in either direction.

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

```
model <- lm(water ~ reserved, data = data)
summary(model)
```

```
##
## Call:
## lm(formula = water ~ reserved, 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       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 for the "reserved" variable is 9.252, with a standard error of 3.948 and a associated p-value of 0.0197. This coefficient represents the change in the number of new or repaired drinking water facilities in villages associated with a one-unit change in the "reserved" variable, which indicates whether the village council head position is reserved for women.
  - 1.Coefficient Value: The coefficient estimate of 9.252 is positive, indicating that when the village council head position is reserved for women (i.e., "reserved" = 1), there is, on average, an increase of approximately 9.252 in the number of new or repaired drinking water facilities compared to when there is no reservation policy (i.e., "reserved" = 0).
  - 2.Statistical Significance: The coefficient is statistically significant with a p-value of 0.0197, which is less than the typical significance level of 0.05. This suggests that the presence of a reservation policy for women in village council heads has a significant effect on the number of new or repaired drinking water facilities in villages.