# 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.

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.

| | 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).

First I calculate the Expected frequencies if the two variables were independent, using the formula ((row total * column total)/grand total)

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 13.5 | 8.36 | 5.14 |
| Lower class | 7.5 | 4.64 | 2.86 |

Next, I calculate the difference between the Actual and expected values for each, square the difference in each case, and divide the squared difference of each by the expected value:

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 0.02 | 0.66 | 0.67 |
| Lower class | 0.03 | 1.2 | 1.21 |

I sum the resulting figures = 3.79. This is the Chi-squared statistic.

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

I calculate the degree of freedom:
(df) = (rows - 1) x (columns - 1) = 2 x 1 = 2.
Next, I calculate the p-value using the chi-squared statistic and the degrees of freedom

```
pchisq(3.79, df=2, lower.tail=FALSE)
# P = 0.1503183
```

Taking our significance level as 0.1, because p > 0.1, we can conclude that we do not have sufficient evidence to reject the null hypothesis, that class and police officers' reaction to an illegal left turn are independent.

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

The standardized residual for each value is found by dividing the difference between the actual and expected values for each by the square root of the expected values.

| | 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?

The standardized residuals suggest that, whilst there may not be a very significant difference between Upper and Lower class people in whether they are pulled over by the police or not, if they are stopped, there may be a more significant difference in how they are treated (i.e. whether a bribe is requested or whether they are given a warning)

Repeating the chi-squared test, but excluding the cases where drivers were not pulled over confirms this to be the case.
P < 0.1, and so we can say that we have sufficient evidence to reject the null-hypothesis, with a 90% confidence level that there is no association between drivers' class and their treatment by the police IF they are pulled over.

```
pchisq(3.59, df=1, lower.tail=FALSE)
# P = 0.05812824
```

# Question 2 (40 points): Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men. 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).

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

The Null Hypothesis (H0) is that a villlage having a female village council head is not associated with or is negatively associated with the number of new or repaired drinking water facilities in villages.

The alternative hypothesis (HA) is that a village having a female village council head is positively associated with the number of new or repaired drinking water facilities in villages.

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

First, I control for the effect of "confounding problems" (i.e. districts choosing female leaders are likely to systematically differ in other respects too) by removing from the sample the villages where female village leaders were not randomly selected.
Then I run my regression analysis.

```
random_sample_villages <- subset(Villages, female==0 | reserved==1)
female_random_lm <- lm(water ~ female, data = random_sample_villages)
summary(female_random_lm)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   14.813      2.429   6.099 3.24e-09 ***
female         9.178      4.088   2.245   0.0255 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.18 on 304 degrees of freedom
Multiple R-squared:  0.01631,Adjusted R-squared:  0.01307
F-statistic: 5.039 on 1 and 304 DF,  p-value: 0.0255
```

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate for my regression suggests that having a female village council leader is associated with an increase in the number of new or repaired drinking water facilities, by 9.178 units, with a standard error of 4.088 units.
This result is significant at the 0.05 level (p<0.05). Therefore, we would expect that if we repeated this analysis with a different sample, in more than 95% of cases having a female village leader would be associated with a larger number of new and repaired drinking water facilities.
However, the adjusted R-squared figure is 0.013, meaning that only 1.3% of the overall variance in the number of new and repaired drinking water facilities is explained by the explanatory variable of female village council leaders.
In comparison to this, including irrigation as an explanatory variable produces a model with an adjusted R-Squared of 0.1868, therefore explaining around 18.7% of the variation in the sample.

```r
female_random_with_irrigation_lm <- lm(water ~ female + irrigation, data = random_sample_villages)
summary(female_random_with_irrigation_lm)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   9.7224     2.2921   4.242 2.95e-05 ***
female        9.7608     3.7118   2.630  0.00898 **
irrigation    1.4933     0.1839   8.120 1.18e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 31.02 on 303 degrees of freedom
Multiple R-squared:  0.1921,Adjusted R-squared:  0.1868
F-statistic: 36.03 on 2 and 303 DF,  p-value: 9.191e-15
```

The effect of female village council leadership doesn't decrease or disappear when irrigation is included as an explanatory variable (in fact it increases slightly with $p<0.01$), suggesting that they may play some independent causal role, but are not as important an explanatory variable in explaining differential patterns of new and repaired drinking water facilities as are other "water-related" projects occurring such as irrigation projects.