Problem Set 2

Applied Stats/Quant Methods 1

Due: October 14, 2024

Name: Eimhin ONeill
Student Number: 20332107

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, I need to get the row, column and grand totals.

```
upper_class_total <- 14 + 6 + 7
lower_class_total <- 7 + 7 + 1
grand_total <- upper_class_total + lower_class_total

not_stopped_total <- 14 + 7
bribe_requested_total <- 6 + 7
stopped_warning_total <- 7 + 1
```

Then, calculate the observed and expected frequency of occurrence

```
# fe is (row total / grand total) multiplied by column total
uc_ns_fo <- 14
uc_ns_fe <- (upper_class_total / grand_total) * not_stopped_total

uc_br_fo <- 6
uc_br_fe <- (upper_class_total / grand_total) * bribe_requested_total

uc_sw_fo <- 7
uc_sw_fe <- (upper_class_total / grand_total) * stopped_warning_total

lc_ns_fo <- 7
lc_ns_fe <- (lower_class_total / grand_total) * not_stopped_total

lc_br_fo <- 7
lc_br_fe <- (lower_class_total / grand_total) * bribe_requested_total

lc_sw_fo <- 1
lc_sw_fe <- (lower_class_total / grand_total) * stopped_warning_total
```

Finally, I need to calculate X^2 using the formula...

```
X2 <- ((uc_ns_fo - uc_ns_fe)^2 / uc_ns_fe) + ((uc_br_fo - uc_br_fe)^2 /
    uc_br_fe) + ((uc_sw_fo - uc_sw_fe)^2 / uc_sw_fe) + ((lc_ns_fo - lc_ns_
    fe)^2 / lc_ns_fe) + ((lc_br_fo - lc_br_fe)^2 / lc_br_fe) + ((lc_sw_fo
    - lc_sw_fe)^2 / lc_sw_fe)
```

```
X2
# X2 = 3.791168
```

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

To calculate the p-value, I first need to get the degrees of freedom.

```
# degrees of freedom
# df - (rows - 1) (columns - 1)

df <- (2 - 1) * (3 - 1)
```

Then, to calculate the p-value, I will use the pchisq function

```
pchisq(X2, df, lower.tail = FALSE)

# p-value = 0.1502306
```

Since the calculated p-value is 0.1502306 and therefore, higher than our alpha of 0.1, we fail to reject the null hypothesis. I do not think there is enough evidence to conclude that there is a significant association between class and the likelihood of being solicited for a bribe.

[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	0.3220	-1.6420	1.5230
Lower class	-0.3220	1.6420	-1.5230

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

Standardised residuals aid us in understanding how far away each observed value is from what we expected. It's calculated using the formula: z = f(observed) minus f(expected) / standard error.

Additionally, they help us understand what the outliers in our data are. Normally, any figure above or below 3 is considered an outlier. So in respect to our results, we have no outliers and all the residuals are under an absolute value of 1.645 (or 90 per cent confidence interval)

Finally, standardised residuals are useful in identifying the direction of our results and what they mean. For example, positive residuals indicate to us that the observed count is higher than we expected. With upper class and bribe requested being positive, it suggests that upper class people are more likely to be solicited for a bribe than we expected.

A negative residual suggests the opposite. For example, with lower class and not stopped being negative, this suggest lower class people are less likely to not be stopped than we 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.

Null hypothesis (H0): reservation policy has no effect on the number of new/repaired water facilities (Beta of reservation policy = 0)

Alternate hypothesis (Ha): reservation policy does have an effect on the number of new/repaired water facilities (Beta of reservation policy is not = 0)

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

```
# b) run a bivariate regression
# have to use the variables of water and reserved for my regression

# fit the linear regression model
model <- lm(water~reserved, data=econ_data)

# t test for the slope
summary(model)
```

(c) Interpret the coefficient estimate for reservation policy.

```
# model summary

# lm(formula = water ~ reserved, data = econ_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
```

Reservation policy has a coefficient estimate of 9.252. This suggest that every additional seat reserved for women (or one unit increase in reservation policy), there is an increase of 9.252 new or repaired water facilities in the villages.

This indicates, to me, a positive relationship between reservation policy and water facilities. We can use the p-value generated in the summary to clarify this further.

The p-value of 0.0197 indicates statistically significance given likely and commonly used alphas of 0.05 or 0.1. In this case, if alpha was = 0.05 or 0.1, we could reject our null hypothesis of reservation policy not having an effect on water facilities.