Problem Set 2

Applied Stats/Quant Methods 1

Name: Darragh McGee (18319331)

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.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

[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.

(a) Calculate the χ^2 test statistic by hand/manually (even better if you can do "by hand" in `R`).

Step 1: Assumptions

- The data consists of categorical variables (social class and type of police interaction)
- The data is from a random sample.
- Observations are independent (one observation does not influence another).

The two variables are independent if the conditional distributions across categories are identical.

Step 2: Setting Up Hypothesis

- **Null Hypothesis:** The relationship between social class and the type of police interaction is statistically independent.
- **Alternative Hypothesis:** The relationship between social class and the type of police interaction is statistically dependent.

Step 3: Calculate the Test Statistic

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Squared is equal to the sum of the squared difference between the observed frequency and expected frequency, divided by the expected frequency for each cell in the contingency table.

Create the Observed Frequency Table in R

```r
observed_frequency_table <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2,
                                   byrow = TRUE)
```

```r
rownames(observed_frequency_table) <- c("Upper class", "Lower class")
colnames(observed_frequency_table) <- c("Not Stopped", "Bribe Requested",
                                        "Stopped/Given Warning")
```

```r
observed_frequency_table
```

Formula to Calculate the Expected Frequency

$$\text{Expected frequency}_{ij} = \frac{\text{Row total}_i \times \text{Column total}_j}{\text{Grand total}}$$

Calculate the Row Totals

```
row_totals <- apply(observed_frequency_table, 1, sum)
row_totals
```

```
Upper class Lower class
27          15
```

Calculate the Column Totals

```
column_totals <- apply(observed_frequency_table, 2, sum)
column_totals
```

```
Not Stopped    Bribe Requested   Stopped/Given Warning
21             13                 8
```

Calculate the Grand Total

```
grand_total <- sum(observed_frequency_table)
grand_total
```

```
Grand Total = 42
```

Calculate the Expected Frequency

```
expected_frequency <- (row_totals / grand_total) %*% t(column_totals)
expected_frequency
```

```
     Not Stopped   Bribe Requested    Stopped/Given Warning
[1,] 13.5          8.357143           5.142857
[2,] 7.5           4.642857           2.857143
```

Calculate the Test Statistic using Chi-Squared Formula

```
chi_squared_statistic <- sum((observed_frequency_table - expected_frequency)^2
                            / expected_frequency)
chi_squared_statistic
```

```
Chi-Squared Statistic = 3.791168
```

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

Step 4: Calculate the p-value

Calculate the Degrees of Freedom

$$\text{Degrees of freedom} = (\text{Number of rows} - 1) \times (\text{Number of columns} - 1)$$

```r
degrees_of_freedom <- (2-1)*(3-1)
```

Chi-Squared p-value Formula in R

```r
p_value <- pchisq(chi_squared_statistic, df = degrees_of_freedom, lower.tail=
    FALSE)
p_value
```

```
p-value = 0.1502306
```

Step 5: Conclusion

- The p-value (0.1502306) is greater than the significance level of 0.1. Therefore, there is insufficient evidence to reject the null hypothesis that social class and the type of police interaction are statistically independent.
- In other words, we cannot conclude that there is a statistically significant relationship between social class and police interactions based on this data.

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

Standardised Residual Formula:

$$z = \frac{\text{Observed frequency} - \text{Expected frequency}}{\sqrt{\text{Expected frequency}}}$$

Standardised Residual refers to how far away each observation is from expectation.

Standardised Residual Calculation:

```
standardised_residual <- (observed_frequency_table - expected_frequency) /
                         (sqrt(expected_frequency))
standardised_residual
```

Table of Standardised Residual Output Summary:

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.1360828	-0.8153742	0.818923
Lower class	-0.1825742	1.0939393	-1.098701

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

- The expected frequency represents the values that would be expected in each cell of a contingency table if the two categorical variables were independent.
- Standardised residuals measure how much the observed frequencies deviate from the expected frequencies, helping to identify which cells contribute most to the Chi-squared statistic.
- Larger standardised residuals indicate greater deviations from expected frequencies, while smaller residuals suggest the observed and expected values are similar.
- In this analysis, the residuals for "Not Stopped" (e.g., 0.136 and -0.183) are small, indicating minimal deviations from expected values and contributing little to the Chi-squared statistic.
- For "Bribe Requested" (e.g., -0.815 and 1.094) and "Stopped or Given Warning" (e.g., 0.819 and -1.099), the residuals suggest moderate deviations from expected values.
- None of the standardised residuals provide substantial evidence of deviations from independence. As a result, these residuals support a higher p-value and reinforce the conclusion that there is insufficient evidence to reject the null hypothesis.

Question 2: Economics

Chattopadhyay and Duflo were interested in whether women promote different policies than men.[2] 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

[2] 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.

Step 1: Assumptions about the Data

- Linear relationship: There is a linear relationship between the explanatory and response variables.
- Independence: The observations are independent of each other.
- Normally Distributed Errors: For any given value of the explanatory variable, the errors (residuals) are assumed to follow a normal distribution.
- Constant variance (Homoscedasticity): The variance of the errors is constant across all values of the independent variable.

Step 2: Setting Up Hypothesis

- **Null Hypothesis:** The policy of reserving village council head positions for women does not affect the number of new or repaired drinking-water facilities in the village. ($\beta = 0$)
- **Alternative Hypothesis:** The policy of reserving village council head positions for women does affect the number of new or repaired drinking-water facilities in the village. ($\beta \neq 0$)

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

Load Relevant Data:

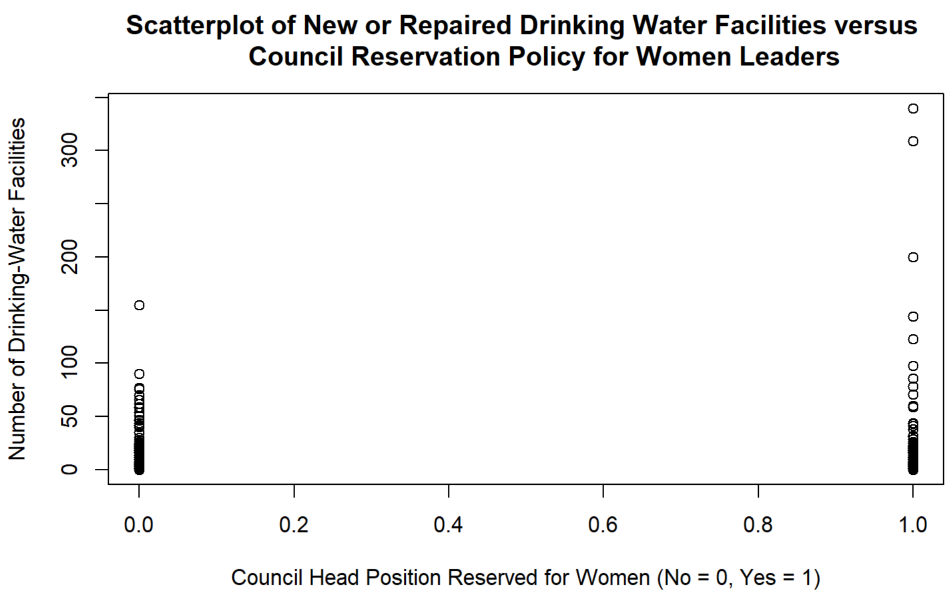
```r
women_policies_data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
str(women_policies_data)
```

Operationalise the Relevant Variables:

```r
Y <- women_policies_data$water
X <- women_policies_data$reserved
```

Create a Scatterplot to Visualise the Relationship:

```r
png("Figure_2_1.png", width = 1500, height = 950, res = 200)
# Produce Scatterplot to Visualize Data.
plot(X, Y,
     xlab = "Council Head Position Reserved for Women (No = 0, Yes = 1)",
     ylab = "Number of Drinking-Water Facilities",
     main = "Scatterplot of New or Repaired Drinking Water Facilities versus
     Council Reservation Policy for Women Leaders")
dev.off()
```



Run Bivariate Regression Analysis

```r
model <- lm(Y ~ X)
model
summary(model)
```

Bivariate Regression Analysis Output:

```
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 ***
X              9.252         3.948    2.344      0.0197 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Step 3: Calculate Test Statistic
t-statistic available from R Summary Model

```
t-value = 2.344
```

Step 4: Calculate p-value
p-value available from R Summary Model

```
p-value = 0.0197
```

Step 5: Conclusion

- In a bivariate regression analysis, the slope represents both the strength and direction of the relationship between two variables (an explanatory and response variable). Specifically, in this analysis, the slope illustrates how the policy of reserving village council head positions for women impacts the number of new or repaired drinking-water facilities in the village.
- If the slope is statistically significantly different from 0, it indicates a relationship between the two variables. In this case, the p-value is 0.0197, which is statistically significant at the 95 percent confidence level ($\alpha < 0.05$).
- Therefore, we reject the null hypothesis that the policy of reserving village council head positions for women has no effect on the number of new or repaired drinking-water facilities.
- There is sufficient evidence to conclude that the reservation policy has an impact on the number of new or repaired drinking-water facilities in the village.

(c) Interpret the coefficient estimate for reservation policy.

- The coefficient for the reservation policy is 9.252, which represents the slope of the relationship between the reservation policy and the number of new or repaired drinking-water facilities in the village. This coefficient explains how a one-unit change in the explanatory variable (reservation policy) affects the response variable (number of drinking-water facilities).
- The coefficient is positive meaning that the reservation policy is associated with an increase in the number of new or repaired drinking-water facilities.
- The reservation policy is a binary variable (coded as 0 for no reservation policy and 1 for reservation policy). This indicates that, villages that have implemented the reservation policy tend to have approximately 9.252 more new or repaired drinking-water facilities on average when compared to villages that have no reservation policy.