Problem 2

Yajie Dong 23339234

September 14, 2023

1 Question 1: Political Science

1.1 (a) Calculate the chi-Square test "by hand" in R.

Step 1: Create the observed data matrix and display the matrix with totals

```
observed_data <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow =
   TRUE)
rownames(observed_data) <- c("Upper class", "Lower class")
colnames(observed_data) <- c("Not Stopped", "Bribe requested", "
   Stopped/Given Warning")
observed_data_with_totals <- addmargins(observed_data)
print(observed_data_with_totals)
```

```
              Not Stopped   Bribe requested   Stopped/Given Warning   Sum
Upper class            14                 6                       7    27
Lower class             7                 7                       1    15
Sum                    21                13                       8    42
```

Step 2: Calculate expected frequencies

```
total <- sum(observed_data)
rows <- nrow(observed_data)
columns <- ncol(observed_data)
```

```
expected_data <- outer(rowSums(observed_data), colSums(observed_data)) / total
print(expected_data)
```

```
            Not Stopped   Bribe requested    Stopped/Given Warning
Upper class        13.5          8.357143                 5.142857
Lower class         7.5          4.642857                 2.857143
```

Step 3: Calculate chi-squared test statistic

```
chi_squared <- sum((observed_data - expected_data)^2 / expected_data)
print(chi_squared)
```

$$\chi^2 \approx 3.79$$

1.2 (b) Calculate the P-value.

```
df <- (rows - 1) * (columns - 1)
p_value <- pchisq(chi_squared, df = df, lower.tail = FALSE)
print(p_value)
```

$$P\text{-value} \approx 0.15$$

Null Hypothesis H_0: The variables are independent

Alternative Hypothesis H_a: The variables are dependent

Since P-value $= 0.15 > \alpha = 0.1$, we fail to reject the null hypothesis.

Based on the given P-value and α, the null hypothesis cannot be rejected, meaning there isn't sufficient evidence to suggest that the two variables are dependent.

1.3 (c) Calculate the standardized residual.

$$z = \frac{f_o - f_e}{\sqrt{f_e \times \left(1 - \frac{N_r}{N}\right) \times \left(1 - \frac{N_c}{N}\right)}} \tag{1}$$

```r
standardized_residuals <- matrix(0, nrow = nrow(observed_data), ncol = ncol(observed_data))
row_totals <- rowSums(observed_data)
col_totals <- colSums(observed_data)
grand_total <- sum(observed_data)

for (i in 1:nrow(observed_data)) {
  for (j in 1:ncol(observed_data)) {
    fo <- observed_data[i, j]
    fe <- expected_data[i, j]
    Nr <- row_totals[i]
    Nc <- col_totals[j]
    N <- grand_total
    standardized_residuals[i, j] <- (fo - fe) / sqrt(fe * (1 - Nr / N) * (1 - Nc / N))
  }
}
rownames(standardized_residuals) <- c("Upper class", "Lower class")
colnames(standardized_residuals) <- c("Not Stopped", "Bribe requested", "Stopped/Given Warning")
print(standardized_residuals)
```

Table 1: Standardized Residuals

	Not Stopped	Bribe requested	Stopped/Given Warning
Upper class	0.32	-1.64	1.52
Lower class	-0.32	1.64	-1.52

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

Interpretation with $\alpha = 0.1$

1. Upper class, Not Stopped (0.32) - Within ± 1.645: Not statistically significant at the 0.1 alpha level. The observed frequency is close to what would be expected under independence between social class and police action.

2. Upper class, Bribe requested (-1.64) - Almost at -1.645: This result is borderline significant at the 0.1 alpha level. It suggests that fewer people from the upper class were asked for bribes than would be expected under independence.

3. Upper class, Stopped/Given Warning (1.52) - Within ± 1.645: Not statistically significant at the 0.1 alpha level. More people from the upper class were stopped or given warnings than would be expected under independence, but the result is not highly significant.

4. Lower class, Not Stopped (-0.32) - Within ± 1.645: Not statistically significant at the 0.1 alpha level. The observed frequency aligns closely with what would be expected if social class and police action were independent.

5. Lower class, Bribe requested (1.64) - Almost at 1.645: Borderline significant at the 0.1 alpha level. More people from the lower class were asked for bribes than would be expected under independence. This might warrant further investigation.

6. Lower class, Stopped/Given Warning (-1.5) - Within ± 1.645: Not statistically significant at the 0.1 alpha level. Although fewer people from the lower class were stopped or given warnings than would be expected under independence, the result isn't considered highly significant.

In summary, using an alpha level of 0.1 and a corresponding Z-score cut-off of ± 1.645 alters the threshold for what is considered statistically significant. In this question, none of the cells show standardized residuals outside of ± 1.645, suggesting that, at the 0.1 significance level, **there isn't strong evidence to reject the null hypothesis** of independence for any of these categories. However, the **"Upper class, Bribe requested" and "Lower class, Bribe requested" cells come close** to the threshold, indicating that these might be areas to explore further.

2 Question 2:Economics

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

- Null Hypothesis (H_0): There is no effect of the reservation policy on the number of new or repaired drinking water facilities.

$$H_0 : \beta_1 = 0$$

- Alternative Hypothesis (H_a): There is an effect of the reservation policy on the number of new or repaired drinking water facilities.

$$H_a : \beta_1 \neq 0$$

2.2 (b) Run a bivariate regression to test this hypothesis in R.

```
data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
names(data)
model <- lm(water ~ reserved, data=data)
summary(model)
```

Table 2: Linear Model Summary: Effect of Reservation Policy(x) on Water Facilities(y)

Variable	Estimate	Std. Error	t-value	p-value
(Intercept)	14.738	2.286	6.446	4.22×10^{-10}
Reserved	9.252	3.948	2.344	0.0197

Additional Summary Statistics:

Residual standard error: 33.45

Degrees of Freedom: 320

Multiple R^2: 0.01688

Adjusted R^2: 0.0138

F-statistic: 5.493

p-value: 0.0197

$$y = 14.738 + 9.252x$$

y: This is the dependent variable, representing the number of new or repaired drinking water facilities in a village.

x: This is the independent variable, representing the status of the reservation policy.

2.3 (C) Interpret the coefficient estimate for reservation policy.

```
confint(model)
```

```
                2.5 %   97.5 %
(Intercept) 10.240240 19.23640
reserved     1.485608 17.01924
```

Confidence Interval for β_1: 95% CI: $[1.4856, 17.0192]$

p**-value: 0.0197** $< \alpha = 0.05$

Therefore,at $\alpha = 0.05$,reject the null hypothesis. The CI does not include zero which confirms effect of the reservation policy on the number of new or repaired drinking water facilities is statistically significant.

Practical Significance: The estimate for β_1 is 9.252. This indicates that for every unit increase in "reserved," the number of new or repaired drinking water facilities increases by approximately 9.252, assuming all other variables are constant.The interval suggests that, with 95% confidence, a unit increase in "reserved" could result in an increase in the number of new or repaired drinking water facilities between approximately 1.49 and 17.02, assuming all other variables are constant.

Direction: The sign of the coefficient is positive, indicating that as the reservation for women increases, the number of new or repaired drinking water facilities also increases.The CI is entirely above zero, confirming that the effect is positive.

Effect Size: The magnitude of the effect is moderate but not very strong, as indicated by the coefficient 9.252.