# Problem Set 2

Owen Eglinton

October 14, 2024

## Question 1: Political Science

| | 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 set the data up as a matrix in R and generated an object for the number of tests done:

```
data_1 <- matrix(c(14, 7, 6, 7, 7, 1), nrow=2)
total <- sum(data_1)
```

- Then I generated a matrix of the expected frequencies, using the formula $\frac{ColumnSum \times RowSum}{Total}$:

```
exp <- matrix(c(1:6), nrow=2)

for (i in 1:2) {
  for (j in 1:3) {
    exp[i, j] <- sum(data_1[i,])*sum(data_1[,j])/total
  }
}
```

- Finally, I generated a matrix of the difference between each observed frequency and the expected, and used this to generate the $\chi^2$ test statistic using the formula $\chi^2 = \Sigma \frac{Difference^2}{ExpectedFrequency}$:

```
diff <- data_1 - exp
chi2 <- sum(diff^2/exp)
```

- This returned a $\chi^2$ test statistic of ∼3.7912.

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

- First, I generated an object for the degrees of freedom:

```r
df_1 <- (nrow(data_1)-1)*(ncol(data_1)-1)
```

- Then I checked the P-value using the `pchisq` command:

```r
pchisq(chi2, df_1, lower.tail = FALSE)
```

- This returned a P-value of $\sim$0.1502.
- Since this is greater than $\alpha = 0.1$, we fail to reject the null hypothesis, namely that the variables of class and road traffic police interaction are independent.

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

- I generated a matrix in R to contain the standardised residuals, calculated using the formula $\frac{Difference}{\sqrt{ExpectedFrequency \times (1 - RowProportion) \times (1 - ColumnProportion)}}$:

```r
std_res <- matrix(c(1:6), nrow=2)

for (i in 1:2) {
  for (j in 1:3) {
    std_res[i, j] <- diff[i, j]/sqrt(exp[i, j])
  }
}
```

- This output the following:

| | Not Stopped | Bribe requested | Stopped/given warning |
|---|---|---|---|
| Upper class | 0.1361 | -0.8154 | 0.8189 |
| Lower class | -0.1826 | 1.0939 | -1.0987 |

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

- Since none of the standardised residuals exceed 2 in absolute value, their usefulness as statistically significant indicators is limited. If we are to draw anything, we can see that people of lower class were somewhat more likely to have a bribe requested of them than expected, with the latter being true for people of an upper class. Neither group seems significantly more or less likely to be stopped or not.

## Question 2: Economics

My first step was to import the data into R:

```
data_2 <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv", header=TRUE)
```

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

- $H_0$: The reservation policy has no effect on the number of new or repaired drinking-water facilities.
- $H_A$: The reservation policy has an effect on the number of new or repaired drinking-water facilities.

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

- I ran the following code in R:

```
summary(lm(water~reserved, data = data_2))
```

- This gave the following 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 ***
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.

- This coefficient estimate can be interpreted as follows: "The reservation policy is associated with a 9.252 unit increase in the number of new or repaired drinking-water facilities, significant at the 5% level".