# Submission for Problem Set 2

Applied Stats/Quant Methods 1

Duc Minh, VU
TCD StudentID: 22996761 / UCD StudentID: 19211157

## Question 1: Political Science

The data on the police officer's response will first be recreated in R for analysis.

```r
crossroad_data = matrix(data = c(14, 7, 6, 7, 7, 1), nrow = 2, ncol = 3)
rownames(crossroad_data) <- c("Upper class", "Lower class")
colnames(crossroad_data) <- c("Not Stopped", "Bribe requested", "Stopped/given warning")
crossroad_data_tbl <- as.table(crossroad_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**

The expected frequency would first need to be calculated, which is the count expected in a cell if the variables were idependent. It equals the product of row and column totals for that cell, divded by the sum total.

$$f_e = \frac{\sum row * \sum column}{\sum N}$$

The row, column and sum total are calculated:

```r
total <- sum(crossroad_data_tbl) #total sample size
row_total <- rowSums(crossroad_data_tbl) #row totals
col_total <- colSums(crossroad_data_tbl) #column totals
```

which gives the results of:

```
> total
[1] 42
> row_total
Upper class Lower class
```

```
27          15
> col_total
Not Stopped        Bribe requested Stopped/given warning
21                     13                      8
```

Using these totals, the expected frequencies are then calculated by the above formula:

```r
#Calculate expected values based on the total
exp_up_clss <- row_total[1]*col_total/total
exp_lw_clss <- row_total[2]*col_total/total
expected_value <- as.table(rbind(exp_up_clss,exp_lw_clss))
```

which gives a table of expected frequency for each cell.

```
            Not Stopped Bribe requested Stopped/given warning
Upper class   13.500000        8.357143              5.142857
Lower class    7.500000        4.642857              2.857143
```

The test statistic is then calculated by using the expected and observed frequencies to summarizes how close the expected frequencies fall to the observed frequencies:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(14 - 13.5)^2}{13.5} + \frac{(7 - 7.5)^2}{7.5} + ...$$

```r
#Calculate test statistic
mnl_chi_test_stat <- sum((abs(crossroad_data_tbl - expected_value))^2/
    expected_value) #
```

```
> mnl_chi_test_stat
[1] 3.791168
```

which gives a $\chi^2$ test statistic of 3.7912. To cross check the result, the $\chi^2$ test function in R is also ran, which also gives the same result.

```r
auto_chi_test_stat <- chisq.test(crossroad_data_tbl)
```

```
> auto_chi_test_stat

Pearson's Chi-squared test

data:  crossroad_data_tbl
X-squared = 3.7912, df = 2, p-value = 0.1502
```

(b) **Calculate the p-value from the test statistic and make a conclusion for $\alpha = 0.1$**

Firstly, the degree of freedom will be calculated. For a 2x3 table, the degree of freedom is:

$$df = (r-1)*(c-1) = (2-1)*(3-1) = 2$$

Using R built in function, the p-value for the test statistic calculated from part (a) for df = 2 is:

```r
#Calculate degree of freedom
df = (nrow(crossroad_data_tbl) - 1) * (ncol(crossroad_data_tbl) - 1)

#Use R to find p-value
p_value <- pchisq(mnl_chi_test_stat, df, lower.tail = FALSE)
```

```
> df
[1] 2
> p_value
[1] 0.1502306
```

which gives the p-value of 0.1502. Since our p-value is larger than $\alpha$, we don't have enough evidence to reject the null hypothesis that the police officers' class and their bribe behaviour are independent.

(c) **Calculate standardized residuals**

The ***standardized residual*** for a cell is:

$$z = \frac{f_o - f_e}{se}$$

As the obersved and expected frequency are known, the standard error ($se$) of $f_o$-$f_e$ (presuming the null hypothesis is true), is calculated as:

$$se = \sqrt{f_e * (1 - row\ porportion) * (1 - column\ proportion)}$$

```r
#Calculate standard error for each expected value in each cell
se_1_1 <- expected_value[1]*(1 - row_total[1]/total)*(1 - col_total[1]/total) #row 1, col 1
se_1_2 <- expected_value[2]*(1 - row_total[2]/total)*(1 - col_total[1]/total) #row 2, col 1
se_2_1 <- expected_value[1,2]*(1 - row_total[1]/total)*(1 - col_total[2]/total) #row 1, col 2
se_2_2 <- expected_value[2,2]*(1 - row_total[2]/total)*(1 - col_total[2]/total) #row 2, col 2
se_3_1 <- expected_value[1,3]*(1 - row_total[1]/total)*(1 - col_total[3]/total) #row 1, col 3
se_3_2 <- expected_value[2,3]*(1 - row_total[2]/total)*(1 - col_total[3]/total) #row 2, col 3
```

```
#Create a table for these SE and taking their square roots
se <- matrix(data = c(se_1_1,se_1_2,se_2_1,se_2_2,se_3_1,se_3_2),nrow = 2,
    ncol = 3)
rownames(se) <- c("Upper class", "Lower class")
colnames(se) <- c("Not Stopped", "Bribe requested", "Stopped/given warning
  ")
se <- as.table(sqrt(se))
```

The calculated standard errors for each cell is:

```
> se
Not Stopped Bribe requested Stopped/given warning
Upper class    1.552648        1.435570              1.219377
Lower class    1.552648        1.435570              1.219377
```

The standardized residual is then computed using the above formula to give the results of:

```
std_res <- (crossroad_data_tbl - expected_value)/se
```

```
> std_res
Not Stopped Bribe requested Stopped/given warning
Upper class   0.3220306      -1.6419565             1.5230259
Lower class  -0.3220306       1.6419565            -1.5230259
```

Upon cross checking against the outputs generated by the built-in function in R, these ”manual” calculated standardized residuals also give the same results.

```
chisq.test(crossroad_data_tbl)$stdres
```

```
> chisq.test(crossroad_data_tbl)$stdres
Not Stopped Bribe requested Stopped/given warning
Upper class   0.3220306      -1.6419565             1.5230259
Lower class  -0.3220306       1.6419565            -1.5230259
```

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

```
#Constructing a table with the original observations along with the
  residuals
obs_with_res = matrix((paste(crossroad_data_tbl, "(", round(std_res,3),")"
  )), nrow = 2, ncol = 3)
rownames(obs_with_res) <- c("Upper class", "Lower class")
colnames(obs_with_res) <- c("Not Stopped", "Bribe requested", "Stopped/
  given warning")
obs_with_res_tbl <- as.table(obs_with_res)
```

```
> obs_with_res_tbl
Not Stopped  Bribe requested Stopped/given warning
Upper class 14 ( 0.322 ) 6 ( -1.642 )    7 ( 1.523 )
Lower class 7 ( -0.322 ) 7 ( 1.642 )     1 ( -1.523 )
```

Although the p-value can provide us with evidence to make conclusion on whether the variables are dependennt/associated or not, it doesn't tell us anything about the nature or strength of the association if there is any. It also doesn't indicate whether all cells or just one or two cells that deviate greatly from independence.

The standardized residuals help providing us with these information, as they acts like a z-scores to show if a residual for each cell is large enough to indicate a deviation from independence. The standardized residuals follow a standard normal distribution, in which they fluctuate around the mean of 0 and with a standard deviation of 1. According to the empirical rule, 95% of the observations fall within two standard deviations from the mean.

From our results, the standardized residuals for all cells are smaller than two, which doesn't indicate deviance from independence. None of the cell are more or less than what we would expected if the variables were truly independent. This helps us to be more confident about the conclusion from part (c), but we would still need the $\chi^2$ test to make conclusions.

## Question 2: Economics

The subset of the data on female politicians will first be imported into R for analysis.

```r
###Import data:
female_policy_data <- read.csv("https://raw.githubusercontent.com/kosukeimai/
    qss/master/PREDICTION/women.csv")
```

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

Our concern is the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages. Therefore, the hypotheses of interest are:

• The null hypothesis: having the village council heads reserved for woman **doesn't change** the number of new or repaired drinking water facilities in villages.

$$\mathbf{H_0\!: \; \beta = 0}$$

• The alternative hypothesis: The reservation policy **has an effect** on the number of new or repaired drinking water facilities in villages

$$\mathbf{H_1\!: \; \beta \neq 0}$$

(b) **Run a bivariate regression to test (a) hypothesis**

For our analysis, the dependent or outcome variable is the change in the number of

drinking water facilities in a village which is the ”water” variable in the dataset. The independent or explanatory variable is whether if the reservation policy is in employed or in other words if there are reserved seats for woman leaders in a village, which is the ”reserved” variable in the dataset.
Bivariate regression involves fitting a linear straight-line equation $\hat{y} = \alpha + x.\beta$ to the observed data. The least square estimates of $\alpha$ and $\beta$ are the best estimates for this equation, as it has the minimum value for the residual sum of squares which is the difference between what is expected and what is observed. The formulas to calculate these estimates are as followed:

$$\beta = \frac{\sum(x-\bar{x}).(y-\bar{y})}{\sum(x-\bar{x})^2} \quad and \quad \alpha = \bar{y} - \beta.\bar{x}$$

Using these formulas, the value for $\alpha$ and $\beta$ are calculated

```
# Manual calculation of alpha and beta:
beta = with(female_policy_data,sum((water - mean(water))*(reserved - mean(reserved)))/sum((reserved - mean(reserved))^2))
alpha = with(female_policy_data, mean(water) - beta*mean(reserved))
```

```
> beta
[1] 9.252423
> alpha
[1] 14.73832
```

which give the result of $\alpha = 14.74$ and $\beta = 9.252$. The regression function is then $\hat{y} = 14.74 + \text{x}.9.252$. The built-in R function for linear regression is also ran, which provides the same results for the estimates.

```
# A regression modelling with water as the independent/outcome variable and indication of seats reserved for woman
female_policy_model <- lm(data = female_policy_data, water ~ reserved)
summary(female_policy_model)
```

```
> summary(female_policy_model)

Call:
lm(formula = water ~ reserved, data = female_policy_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
```

(c) **Interpret the coefficient estimate for reservation policy**

Based on the regression analysis, the estimated coefficient for reservation policy is 9.252. So, there is a positive relation between reservation policy and the number of new/repaired drinking-water facilities. For every increase in reserved seats for female politicians, the number of new/repaired drinking-water facilities increase by 9.252, or in other words, having a seat reserved for a woman in the village heads increase the number of drinking-water facilities by 9.252 units. Since p-value is below $\alpha = 0.05$ threshold $(0.0197 < 0.05)$, there is sufficient statistical evidence that the estimated coefficient is different from 0 at the 95% confidence level.