Problem Set 2

Ella Karagulyan

Applied Stats/Quant Methods 1

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

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

To test the association between driver's class and the actions of the police officers, we need to calculate the χ^2 statistic using the formula below:

$$\chi^2 = \sum \frac{(f_o - f_i)^2}{f_i}$$

To do this, first, we input the experiment results table into R.

[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.

```
# Inputting the observed data
observed_table <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# Assign row and column labels
rownames(observed_table) <- c(”Upper class”, ”Lower class”)
colnames(observed_table) <- c(”Not stopped”, ”Bribe requested”, ”Stopped/given warning”)

# Convert to a table with observed frequencies
observed_table <- as.table(observed_table)
observed_table
```

Now that we have our observed frequencies (f_o), we manually calculate the expected frequencies (f_e) using the following formula:

$$f_e = \frac{\text{row total}}{\text{grand total}} \times \text{column total}$$

```
# Calculate the sums
row_totals <- rowSums(observed_table)
column_totals <- colSums(observed_table)
grand_total <- sum(observed_table)
row_totals
column_totals
grand_total

# Generate the table with expected frequencies
expected_table <- outer(row_totals, column_totals) / grand_total
expected_table
```

The table below shows the expected frequencies.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	13.5	8.4	5.1
Lower class	7.5	4.6	2.9

Lastly, we calculate the χ^2 statistic.

```
# Calculating the Chi-sqr statistic
chi_sqr_table <- (observed_table - expected_table)^2 / expected_table
chi_sqr_table
chi_sqr <- sum(chi_sqr_table)
chi_sqr
```

The Chi-sqr statistic of $\chi^2 = 3.79$.

(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 test the significance of the χ^2 statistic, we need to calculate the p-value with the degrees of freedom.

```r
#Calculating the p-value and degrees of freedom
p_value <- pchisq(chi_sqr, df = df, lower.tail = FALSE)
round(p_value, 2)
df <- (nrow(observed_table) -1) * (ncol(observed_table) - 1)
df
```

The p-value of the χ^2 statistic 3.79 (df = 2) equals to 0.15. With the confidence level of 10% (alpha=0.1), we fail to reject the null hypothesis that there is no association between employee driver's class and the actions of the police officers.

To check our calculations, we also verified the results using the built-in chisqr.test function in R.

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

Finally, we calculate the adjusted standardized residuals with the formula below. We opt for the adjusted standardized residuals, rather than standardized residuals, taking into the account the small sample size of some cells (frequency > 5).

$$z = \frac{f_{\text{observed}} - f_{\text{expected}}}{\sqrt{f_{\text{expected}} \cdot (1 - \text{row prop.}) \cdot (1 - \text{column prop.})}}$$

```r
# Calculating Standardized residuals
prop_rows <- prop.table(observed_table, margin = 1)
prop_cols <- prop.table(observed_table, margin = 2)
prop_rows
prop_cols

residuals <- (observed_table - expected_table) / sqrt(expected_table * (1
    - prop_rows) * (1 - prop_cols))
round(residuals, 2)
```

The calculated adjusted standardized residuals are presented in the below table.

[2] Remember frequency should be > 5 for all cells, but let's calculate the p-value here anyway.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.34	-1.26	2.69
Lower class	-0.31	2.20	-1.22

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

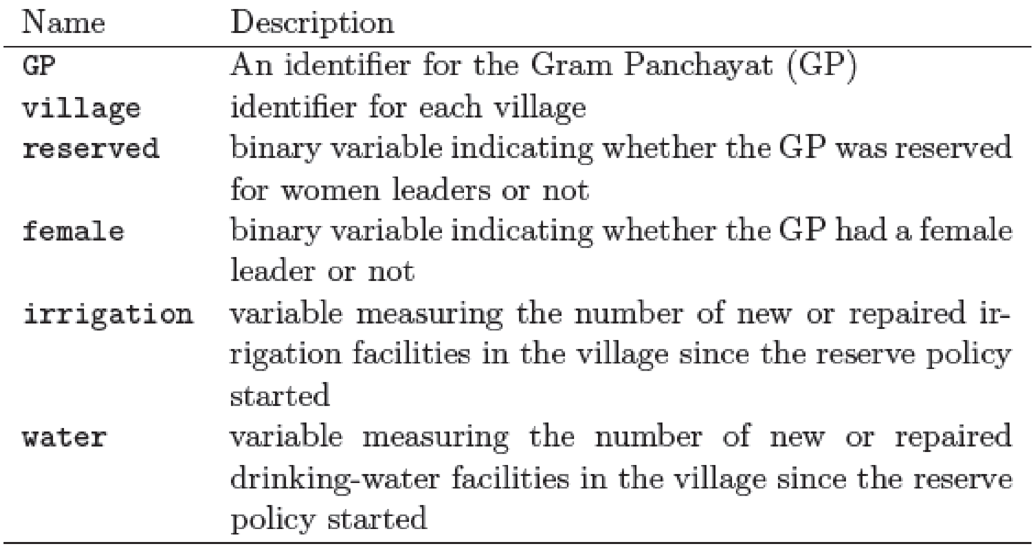
Standardized residuals show the extent each observed value in the contingency table differs from the expected value. They help identify which cells contribute the most to the overall χ^2 statistic. In case of the association between the driver's class and actions of the police officer, we can state the following:
- Upper class and "Stopped/given warning": the residual of 2.69 is large (more that 1.96, 5% confidence level) and positive indicating that more upper-class drivers were stopped and given a warning than expected.
- Lower class and "Bribe requested": the residual of 2.2 is positive and exceeds the threshold of 1.96, indicating that more lower-class individuals were asked for a bribe than 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.

To estimate the effect of reservation policy on the number of new or repaired drinking-water facilities, first, we input the dataset into R and explore the variables.

```r
#Loading the data and exploring
df <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
summary(df)
table(df$reserved)
table(df$female)
mean(df$water)
tapply(df$water, df$female, mean)
tapply(df$water, df$reserved, mean)
female_reserved_table <- table(df$female, df$reserved)
female_reserved_table
```

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

H_0: The number of new or repaired drinking-water facilities in the reserved villages is the same as in the unreserved ones. Hence, the reservation policy has no effect on the number of new or repaired drinking water facilities.

H_1: The number of new or repaired drinking-water facilities in the reserved villages is not the same as in unreserved ones. Hence, the reservation policy has an effect (positive or negative) on the number of new or repaired drinking water facilities.

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

The bivariate regression will test whether the policy of reservation of female village councils has an effect on the number of new or repaired drinking-water facilities in the village.

$$Y = \beta_0 + \beta \cdot (\text{Reserved})$$

where:
Y - number of new or repaired drinking-water facilities
Reserved - the village councils have been randomly reserved for women

```r
# Linear regression model
lm(df$water~df$reserved)

# Saving the model as an object
model <- lm(df$water~df$reserved)
```

```
# Creating a table of the results
install.packages(”stargazer”)
library(stargazer)
stargazer(model, type = ”latex”, title = ”Table: Linear Regression
    Results”, out = ”regression_table.tex”)
```

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

The below table contains the results of the regression model. The coefficient for Reserved is significant at the 5% level. The villages with reserved female councils on average have by 9.252 more new or repaired drinking-water facilities compared to unreserved villages. We, hence, have enough evidence to reject the null hypothesis that the reservation policy has no effect on the number of new or repaired drinking-water facilities.

Table 1: Linear Regression Results

	Dependent variable:
	water
reserved	9.252^{**}
	(3.948)
Constant	14.738^{***}
	(2.286)
Observations	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493^{**} (df = 1; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01