Problem Set 2/

Quant Methods 1/Due: October 14, 2024

Zengyuan Zhao / zhaoze@tcd.ie

Instructions

- Please show your work! You may lose points by simply writing in the answer. If the problem requires you to execute commands in `R`, please include the code you used to get your answers. Please also include the `.R` file that contains your code. If you are not sure if work needs to be shown for a particular problem, please ask.
- Your homework should be submitted electronically on GitHub.
- This problem set is due before 23:59 on Monday October 14, 2024. No late assignments will be accepted.

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.

[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.

	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).

```r
#Q1.1
#The X-squared statistic is calculated by subtracting the square of the expected value from the observed value and dividing by the expected value.
#The question gives a 2*3 contingency table.
resulting_data <- matrix(c(14, 6, 7,
                           7, 7, 1),
                         nrow = 2, byrow = TRUE,
                         dimnames = list(c("Upper class", "Lower class"),
                                         c("Not Stopped", "Bribe requested", "Stopped/given warning")))
# Calculate summary rows
row_sums <- rowSums(resulting_data)
col_sums <- colSums(resulting_data)
total_sum <- sum(resulting_data)
# Create a new matrix including raw data and summary rows
resulting_table <- rbind(resulting_data, c(col_sums, total_sum))
resulting_table <- cbind(resulting_table, c(row_sums, total_sum))
rownames(resulting_table) <- c("Upper class", "Lower class", "Total")
colnames(resulting_table) <- c("Not Stopped", "Bribe requested", "Stopped/given warning", "Total")
resulting_table
#Calculate expected frequency
Ex_Upper_NotStopped <- (27*21)/42
Ex_Upper_Bribe <- (13*27)/42
Ex_Upper_Stopped <- (8*27)/42
Ex_Lower_NotStopped <- (21*15)/42
Ex_Lower_Bribe <- (13*15)/42
Ex_Lower_Stopped <- (8*15)/42
#In order to be more intuitive, make an expected frequency table
Expected_table <- matrix(c(Ex_Upper_NotStopped, Ex_Upper_Bribe, Ex_Upper_Stopped,
                           Ex_Lower_NotStopped, Ex_Lower_Bribe, Ex_Lower_Stopped),
                         nrow = 2, byrow = TRUE,
                         dimnames = list(c("Upper class", "Lower class"),
                                         c("Not Stopped", "Bribe requested", "Stopped/given warning")))
Expected_table
#Calculate X^2
X_squared_statistic <- (14-13.5)^2/13.5 + (6-8.357143)^2/8.357143 +
```

```
  (7-5.142857)^2/5.142857 + (7-7.5)^2/7.5 +
  (7-4.642857)^2/4.642857 + (1-2.857143)^2/2.857143
X_squared_statistic
```

First, we need to calculate the row total, column total, and grand total, as shown in Figure 1

Figure 1: Summary table

```
> resulting_table
            Not Stopped Bribe requested Stopped/given warning Total
Upper class          14               6                    7    27
Lower class           7               7                    1    15
Total                21              13                    8    42
```

Next, we need to calculate the expected frequency, as shown in Figure 2

Figure 2: Expected Frequency table

```
> Expected_table
            Not Stopped Bribe requested Stopped/given warning
Upper class        13.5        8.357143                5.142857
Lower class         7.5        4.642857                2.857143
```

Finally, according to the X-squared statistical calculation formula, we get:

Figure 3: X-squared statistical calculation

```
> X_squared_statistic
[1] 3.791169
```

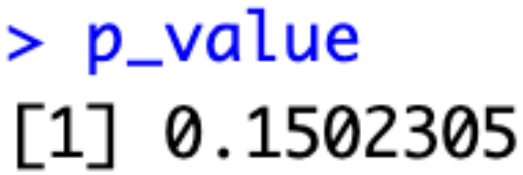
(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$?

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

```r
#Q1.2
#Calculate degrees of freedom
df <- (2-1)*(3-1)
df
# Calculate p_value
p_value <- pchisq(X_squared_statistic,df,lower.tail = FALSE)
p_value
```

After calculation, the p value is 0.1502305

Figure 4: p value



If a is 0.1, then the confidence interval is 90, and the p-value is greater than 0.1, so the null hypothesis cannot be rejected. Therefore, there is no significant correlation between the driver's class status and whether he is asked for a bribe.

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.3220306	-1.641957	1.523026
Lower class	-0.3220306	1.641957	-1.523026

```r
#1.3
#Formula for standardized residuals:(obeserved-expected)/sqrt(expected(1-
    row prop.)*(1-column prop.))
#row prop = row total/grand total
#column prop = column/grand total
sr_Upper_NotStopped <- (14-13.5)/sqrt(13.5*(1-27/42)*(1-21/42))
sr_Upper_Bribe <- (6-8.357143)/sqrt(8.357143*(1-27/42)*(1-13/42))
sr_Upper_Stopped <- (7-5.142857)/sqrt(5.142857*(1-27/42)*(1-8/42))
sr_Lower_NotStopped <- (7-7.5)/sqrt(7.5*(1-15/42)*(1-21/42))
sr_Lower_Bribe <- (7-4.642857)/sqrt(4.642857*(1-15/42)*(1-13/42))
sr_Lower_Stopped <- (1-2.857143)/sqrt(2.857143*(1-15/42)*(1-8/42))
```

```r
sr_table <- matrix(c(sr_Upper_NotStopped, sr_Upper_Bribe, sr_Upper_Stopped,
                     sr_Lower_NotStopped, sr_Lower_Bribe, sr_Lower_Stopped),
                         nrow = 2, byrow = TRUE,
                         dimnames = list(c("Upper class", "Lower class"),
                                         c("Not Stopped", "Bribe requested", "Stopped/given warning")))
sr_table
sr_Upper_NotStopped
sr_Upper_Bribe
sr_Upper_Stopped
sr_Lower_NotStopped
sr_Lower_Bribe
sr_Lower_Stopped
```

Figure 5: standardized residuals

```
> sr_table
            Not Stopped Bribe requested Stopped/given warning
Upper class   0.3220306       -1.641957                1.523026
Lower class  -0.3220306        1.641957               -1.523026
> sr_Upper_NotStopped
[1] 0.3220306
> sr_Upper_Bribe
[1] -1.641957
> sr_Upper_Stopped
[1] 1.523026
> sr_Lower_NotStopped
[1] -0.3220306
> sr_Lower_Bribe
[1] 1.641957
> sr_Lower_Stopped
[1] -1.523026
```

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

(1)The value of the standardized residual for the upper class who has not been stopped is positive and relatively small (0.3220306), less than 1 and close to 0, indicating that this observation is slightly higher than the predicted value of the model, but the degree of deviation is not large.
(2)The standardized residual value for the upper class who has been asked for a bribe is negative and relatively large (-1.641957), with an absolute value greater than 1 and less than 2, indicating that this observation is significantly lower than the predicted value of the model, and the degree of deviation is large, but it is still within a reasonable range.
(3)The value of the upper class who has been stopped or warned is positive and relatively large (1.523026), greater than 1 but less than 2, indicating that this observation is significantly higher than the predicted value of the model, although the degree of deviation is large, but it is still within a reasonable range.
Lower Class
(1)The value of the standardized residual for the lower class who has not been stopped is negative and relatively small (-0.3220306), with an absolute value less than 1 and close to 0, indicating that this observation is slightly lower than the predicted value of the model, but the degree of deviation is not large.
(2)The standardized residual value of the lower class being asked for bribes is positive and relatively large (1.641957), greater than 1 and less than 2, indicating that this observation is significantly higher than the model's predicted value, and the degree of deviation is large, but it is still within a reasonable range.
(3)The standardized residual value of the lower class being stopped or warned is negative and relatively large (-1.523026), and the absolute value is greater than 1 but less than 2, indicating that this observation is significantly lower than the model's predicted value. Although the degree of deviation is large, it is still within a reasonable range.

Conclusion
The standardized residual values of the upper and lower classes being treated differently are all between [-2, 2], all within the normal range, and there are no abnormal values.The results passed the standardized residual test, and the hypothesis test model met the requirements.

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 6 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 6: 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.

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

Null Hypothesis: The reservation policy for women as village government leaders has no effect on the number of new or repaired drinking water facilities in the villages.

Alternative Hypothesis: The reservation policy for women as village government leaders has positive/negative effect on the number of new or repaired drinking water facilities in the villages.

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

```
#read data
women_data <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
women_data
#inspect data
inspect_null <- sum(is.na(women_data))
inspect
str(women_data)
class(women_data)
#bivariate regression
model1 <- lm(women_data$irrigation ~ women_data$reserved,data = women_data)
summary(model1)
model2 <- lm(women_data$water ~ women_data$reserved,data = women_data)
summary(model2)
```

By observing the data and combining it with the variable explanation table, we can see that the values of the "women" variable and the "reserved" variable are the same. Therefore, I decided to use the "reserved" variable as the independent variable and "irrigation" and "water" as the dependent variables to explore their regression relationship.

Figure 7: The binary regression bewteen irrigation and reserved

```
Call:
lm(formula = women_data$irrigation ~ women_data$reserved, data = women_data)

Residuals:
   Min     1Q Median     3Q    Max
-3.388 -3.388 -3.019 -1.019 86.612

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)           3.3879     0.6498   5.214 3.33e-07 ***
women_data$reserved  -0.3693     1.1220  -0.329    0.742
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.506 on 320 degrees of freedom
Multiple R-squared:  0.0003385, Adjusted R-squared:  -0.002785
F-statistic: 0.1084 on 1 and 320 DF,  p-value: 0.7422
```

From the above figure, we can see that the p-value of the regression analysis of "irrigation" and "reserved" is 0.7422, which is much larger than 0.1, and the null hypothesis cannot be rejected. Therefore, there is no significant correlation between the implementation of the reservation policy and the number of irrigation facilities.

Next, I will calculate the regression relationship between the variables water and reserved.

Figure 8: The binary regression bewteen water and reserved

```
Call:
lm(formula = women_data$water ~ women_data$reserved, data = women_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 ***
women_data$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
```

From the above figure, we can see that the p-value of the regression analysis of "water" and "reserved" is 0.0197, which is less than 0.05, and the null hypothesis is rejected. Therefore, there is a significant correlation between the implementation of the reservation policy and the number of drinking water facilities.

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate of the regression model of water and reserved is 9.252, which is a positive number and shows a positive correlation. This usually means that if the independent variable reserved increases by one unit, the dependent variable water may increase by 9.252 units. Under this premise, it is possible that for every leadership position reserved for a woman, 9.252 drinking water facilities will be added.
Assuming that the variables are linear and considering that the intercept is 14.738, we can get the relationship between reserved and water:
Y(water)=9.252*X(reserved)+14.738