Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2023/Idi Amin Da Silva/ 23372225

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 Sunday October 15, 2023. 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.

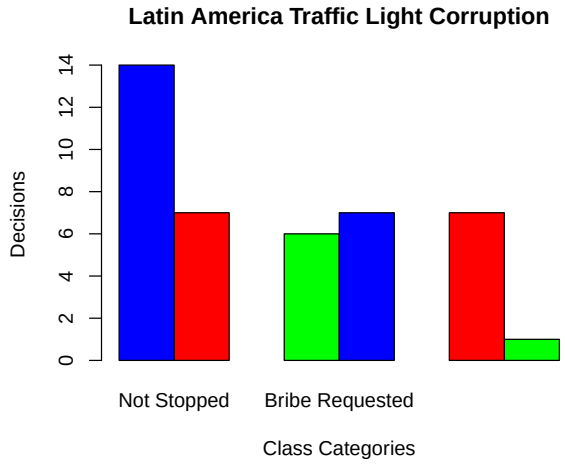

Figure 1: Enter Caption

	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).

1 Answers to Question 1

(1a) The chi-squared test statistic is given by:

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

$$= \frac{(14-13.50)^2}{13.50} + \frac{(6-8.36)^2}{8.36} + \frac{(7-5.14)^2}{5.14} + \frac{(7-7.50)^2}{7.50} + \frac{(7-4.46)^2}{4.46} + \frac{(1-2.86)^2}{2.86} = 3.80$$

```r
f_observed <- c(14, 6, 7, 7, 7, 1)
f_expected <- c(13.50, 8.36, 5.14, 7.50, 4.64, 2.86)
chi_squared <- sum((f_observed - f_expected)^2/(f_expected))
round(chi_squared, digits=2) # Answer: chi_squared = 3.80
##############################################################################
############################
## Extra calculations/works:
# Create a the given table in problem set2 Question 1 using R code
```

```r

traffic_light_corruption_vector <- as.table(rbind(c(14, 6, 7), c(7, 7, 1))) # table values in Rows and stored in a R object called traffic_light_corruption_vector
rownames(traffic_light_corruption_vector) <- c("Upper Class", "Lower Class") # labeling the rows
colnames(traffic_light_corruption_vector) <- c("Not Stopped", "Bride Requested", "Stopped/Given Warning") # Labeling the columns
traffic_light_corruption_vector # Display the table
###### Answer:
#             Not Stopped     Bride Requested     Stopped/Given Warning
#Upper Class          14                   6                         7

#Lower Class           7                   7                         1
####################
Chisq_test_Statistic <- chisq.test(traffic_light_corruption_vector) # perform the chi-squared test
Chisq_test_Statistic
####
#### Answer:
#   Pearson's Chi-squared test

#data:  traffic_light_corruption_vector
#X-squared = 3.7912, df = 2, p-value = 0.1502 # The chi-squared test statistics is the same as calculated
##############                                 # manually chi-sq=3.80, p-value=0.1502
#############################################################################
#(i) To recall the observed frequencies (fo):

Chisq_test_Statistic$observed
### Answer:
#              Not Stopped     Bride Requested       Stopped/Given Warning
#Upper Class          14                   6                         7
#Lower Class           7                   7                         1
#############
## (ii) To Calculate the expected frequencies (fe)

Chisq_test_Statistic$expected
## Answer:
#              Not Stopped      Bride Requested     Stopped/Given Warning
# Upper Class         13.5            8.357143                 5.142857
# Lower Class          7.5            4.642857                 2.857143
################
# (iii) Calculation of pearson residual: sum((fo - fe)^2/sqrt(fe))

Chisq_test_Statistic$residuals
###
#            Not Stopped   Bride Requested     Stopped/Given Warning
```

```
#Upper Class    0.1360828       -0.8153742              0.8189230
#Lower Class   -0.1825742        1.0939393             -1.0987005
##################

### Bar Plot - Using base R

color_names <- c("blue", "red", "green")
barplot(officers_bribe, beside=T, xlab="Class Categories", ylab="
    Decisions", main="Latin America Traffic Light Corruption", col=color_
    names)
legend(1, 2300, rownames(officers_bribe), cex=0.7, fill=color_names, bty=
    "n")
```

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

(1b) Before calculating the p-value I have decided first formulate the hypothesis testing:

H_0 : Upper Class is statistically independent from lower class

H_a : Upper Class is statistically dependent from lower class

The degree of freedom is given by:

$$df = (rows - 1)(columns - 1) = (2 - 1)(3 - 1) = 2$$

$$p - value = (3.80, df = 2, lower.tail = FALSE) = 0.1495686$$

Since the p-value = 0.1495686 is greater than $\alpha = 0.1$ we can draw the following conclusions:

(i) Since the (p-value = 0.1502) greater than (alpha=0.1), so there is no sufficient evidence to reject the the null hypothesis (i.e., I reject the alternative hypothesis in favor of the Null Hypothesis), the upper class is statistically independent of lower class.

(ii) On the other hand if I calculate the critical value of chi-squared with two degree of freedom by using the quantille, i.e., the syntax is given by qchisq(alpha, df, lower.tail=FALSE) that implies qchisq(0.1, df=2, lower.tail=FALSE)= 4.60517, we can reject the alternative hypothesis in favor of null hypothesis since the critical value fall in region of non-rejection

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

(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

(1c) The standardized residual for each cell are given by:

$$standardizedResiduals = \frac{(f_o - f_e)}{se}$$

Where $se = \sqrt{(1 - rowprop)(1 - colprop)}$

z_{ij} where i= rows and j=columns, then

$$z_{11} = \frac{(14 - 13.50)}{\sqrt{(13.50)(1 - \frac{27}{42})(1 - \frac{21}{42})}} = 0.322$$

;

$$z_{12} = \frac{(6 - 8.36)}{\sqrt{(8.36)(1 - \frac{27}{42})(1 - \frac{13}{42})}} = -1.644$$

$$z_{13} = \frac{(7 - 5.42)}{\sqrt{(5.42)(1 - \frac{27}{42})(1 - \frac{8}{42})}} = 1.262$$

$$z_{21} = \frac{(7 - 7.50)}{\sqrt{(7.50)(1 - \frac{15}{42})(1 - \frac{21}{42})}} = -0.322$$

$$z_{22} = \frac{(7 - 4.64)}{\sqrt{(4.64)(1 - \frac{15}{42})(1 - \frac{13}{42})}} = 1.644$$

$$z_{23} = \frac{(1 - 2.86)}{\sqrt{(2.86)(1 - \frac{15}{42})(1 - \frac{8}{42})}} = -1.525$$

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

(1d) After running the R code in the question (1c) I have found that the rows cells for upper class and lower class are symmetric for each columns cells and the sum of each columns cells is equal to zero then I can state that the standardized residual show that the Null hypothesis true and the observed frequencies e close to the expected frequencies.

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 2 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 2: 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. (2a) State the Hypothesis test

$$H_0 : \mu = 0$$

or

μ : The reservation policy has no effect on the number of new or repair water facilities in the village

$$H_a : \mu \neq 0$$

or

μ : The reservation policy has effect on the number of new or repair water facilities in the village.

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

```
# (2b)

fitted_model_women <- lm(water ~ reserved, data=women)
summary(fitted_model_women)
### Answer:
#Call:
# lm(formula = water ~ reserved, data = women)

#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 *
#  ---


#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
#######
# The predicted line or the best fitting line or the regression line is
    given by:

# water_i= 14.738 +  9.252* reserved_1; Note that the reserved_i is a
    binary variable/dummy variable can can take two values 0
# and 1.
```

(c) Interpret the coefficient estimate for reservation policy.

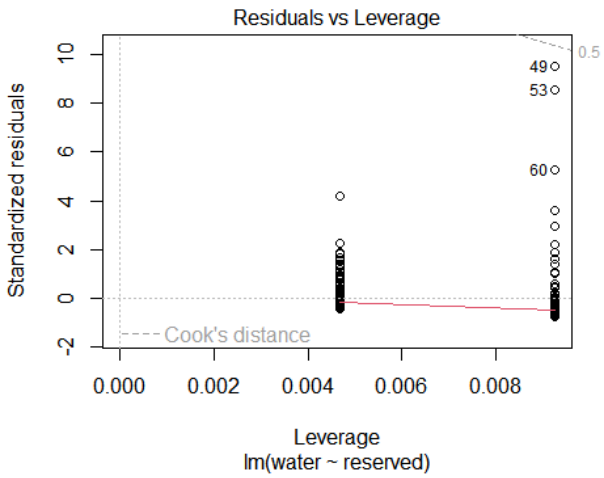

Figure 3: Enter Caption

The predicted best fitting regression line is given by water_i= 14.738 + 9.252* reserved_1, the intercept water_i=14.738 represent the average value when reserved_i=0.
For each increase in one unit in reserved_i=1 , the value of water_i is expected to increase by average of 9.252.