Problem Set 2

Applied Stats/Quant Methods 1

Sania Suneeth

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

Table 1: Police encounters by class and type of action

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

- Assumption:
- Random assignment of the drivers were done as part of the experimental treatment in the study,so thereby ensuring random sampling of the police encounter too.this therefore fulfills the assumption of random sampling in the study
- with Random Sampling Independence of observation is achieved in the study, as each interaction of the driver with the police is independent of other interaction
- third, data is set to be categorical here the data represents the counts of police interactions 'Not stopped','bribe requested','Stopped/given warning', thus fulfilling the categorical data assumption
- **Hypothesis testing:**
  the Hypothesis for the chi square test of independence
- Null Hypothesis: There exists no association between the type of police interaction and the driver's class
- Alternative Hypothesis: There exists an association between the type of police interactions and the driver's class
- **Test Statistics**
  Here we are running a chi square test for independence

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

the observed data is given in the above table, we need to find the expected frequencies for the categorical data, therefore expected frequency has a formula of

$$\text{Expected Frequency} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$

the row total for upper class and lower class

```
upper_class_sum<- sum(14,6,7)
upper_class_sum
lower_class_sum<-7+7+1
lower_class_sum
row_total<- matrix(c(upper_class_sum,lower_class_sum))
row_total
rownames(row_total)<-c('upper class sum','lower class sum')
row_total
not_stopped_sum<-14+7
not_stopped_sum
bribe_requested_sum<-6+7
bribe_requested_sum
stopped_given_warning_sum<-7+1
col_total <- matrix(c(not_stopped_sum, bribe_requested_sum, stopped_
    given_warning_sum), ncol = 1)
col_total
rownames(col_total) <- c('Not Stopped', 'Bribe Requested', 'Stopped/
    Given Warning')
col_total
```

```
                [,1]
upper class sum   27
lower class sum   15
```

```
     [,1]
Not Stopped             21
Bribe Requested         13
Stopped/Given Warning    8
```

now the grand total of the rows or column

```
grand_total_sum<- upper_class_sum+lower_class_sum
grand_total_sum
```

```
> grand_total_sum
[1] 42
```

after we have calculated the necessary measures for expected frequency we now move to calculate the expected frequencies for each row and column

```
### expected data
upper_class_not_stopped_expected<-(upper_class_sum*not_stopped_sum)/
    grand_total_sum
upper_class_not_stopped_expected
upper_class_bribe_requested_expected<-(upper_class_sum*bribe_
    requested_sum)/grand_total_sum
upper_class_bribe_requested_expected
upper_class_stopped_expected<-(upper_class_sum*stopped_given_warning_
    sum)/grand_total_sum
```

```
upper_class_stopped_expected
lower_class_not_stopped_expected<-(lower_class_sum*not_stopped_sum)/
    grand_total_sum
lower_class_not_stopped_expected
lower_class_stopped_expected<-(lower_class_sum*stopped_given_warning_
    sum)/grand_total_sum
lower_class_stopped_expected
lower_class_bribe_requested_expected<-(lower_class_sum*bribe_
    requested_sum)/grand_total_sum
lower_class_bribe_requested_expected

### now forming the expected data table
expect<-(row_total/ grand_total_sum) %*% t(col_total)
expect
```

and thereby we can formulate the expected frequency table as

	Not Stopped	Bribe Requested	Stopped/Given Warning
upper class sum	13.5	8.357143	5.142857
lower class sum	7.5	4.642857	2.857143

now finding the chi square test

```
chi_sq_test<-(
  (14-upper_class_not_stopped_expected)^2/upper_class_not_stopped_
    expected+
    (6-upper_class_bribe_requested_expected)^2/upper_class_bribe_
    requested_expected+
    (7-upper_class_stopped_expected)^2/upper_class_stopped_expected+
    (7-lower_class_not_stopped_expected)^2/lower_class_not_stopped_
    expected+
    (7-lower_class_bribe_requested_expected)^2/lower_class_bribe_
    requested_expected+
    (1-lower_class_stopped_expected)^2/lower_class_stopped_expected
)
chi_sq_test
```

```
> chi_sq_test
[1] 3.79116
```

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = 0.1$? we first find the degrees of freedom using the formula

$$df = (r-1) \times (c-1)$$

where r is the number of rows and c is the number of columns

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

```r
observed_data <- matrix(c(14,6,7,7,7,1),nrow=2,byrow=TRUE)
rownames(observed_data)<-c('upper_class','lower_class')
colnames(observed_data)<-c('Not stopped','bribe requested','Stopped/given
    warning')
observed_data
rows = nrow(observed_data)
rows
col = ncol(observed_data)
df<- (1-rows)*(1-col)
df
```

```
 df
[1] 2
```

- **Finding the p-value**

```r
p_value<- 1-pchisq(chi_sq_test,df)
p_value
```

```
> p_value
[1] 0.1502306
```

the p value is of 0.15 which is greater than 0.1 significance level therefore we fail to reject the null hypothesis as there is no sufficient evidence against no association between the type of police interaction and the driver's class

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

standardised residual is calculated by the formula

$$\text{Standardized Residual} = \frac{O - E}{\sqrt{E \cdot (1 - row_proportion) \cdot (1 - col_proportion)}}$$

where O is the observed value E is the expected value now calculating the standardised residuals now calculating the standardised residual using R

```
### standardised residual
observed_data <- matrix(c(14,6,7,7,7,1),nrow=2,byrow=TRUE)
rownames(observed_data)<-c('upper_class','lower_class')
colnames(observed_data)<-c('Not stopped','bribe requested','Stopped/given
    warning')
observed_data
addmargins(observed_data)
expect<-(row_final/ grand_total_sum) %*% t(col_final)
expect
row_proportion1<-upper_class_sum/grand_total_sum
row_proportion1
row_proportion2<-lower_class_sum/grand_total_sum
row_proportion2
col_proportion1<-not_stopped_sum/grand_total_sum
col_proportion1
col_proportion2<-bribe_requested_sum/grand_total_sum
col_proportion3<-stopped_given_warning_sum/grand_total_sum
col_proportion3

standard_residual_upperclass_notstopped<- (14-upper_class_not_stopped_
    expected)/sqrt(upper_class_not_stopped_expected*(1-row_total_
    proportion_1)*(1-col_total_proportion_1))
standard_residual_upperclass_notstopped
standard_residual_upperclass_briberequested<- (6-upper_class_bribe_
    requested_expected)/sqrt(upper_class_bribe_requested_expected*(1-row_
    proportion1)*(1-col_proportion2))
standard_residual_upperclass_briberequested
standard_residuals_upperclass_stopped<- (7-upper_class_stopped_expected)/
    sqrt(upper_class_stopped_expected*(1-row_proportion1)*(1-col_
    proportion3))
standard_residuals_upperclass_stopped
standard_residuals_lowerclass_notstopped<-(7-lower_class_not_stopped_
    expected)/sqrt(lower_class_not_stopped_expected*(1-row_proportion2)*
    (1-col_proportion1))
standard_residuals_lowerclass_notstopped
standard_residual_lowerclass_bribe<-(7-lower_class_bribe_requested_
    expected)/sqrt(lower_class_bribe_requested_expected*(1-row_proportion2
    )*(1-col_proportion2))
standard_residual_lowerclass_bribe
```

```
standard_residual_lowerclass_stopped<-(1-lower_class_stopped_expected)/
    sqrt(lower_class_stopped_expected*(1-row_proportion2)*(1-col_
    proportion3))
standard_residual_lowerclass_stopped
```

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

Table 2: Standardized Residuals for Police Encounters by Class

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

- standerdised residuals help in understanding the variances of the observed data with that of the expected data,
- it also tells us how significant each cell is in contributing to the chi square value when the null hypothesis is true stating that the officers likelihood to solicit a bribe is independent of the driver's class
- Also the magnitude of the residuals closer to zero suggest least deviation of the observed value from the expected value and supports the null hypothesis,and a positive residual suggest the observed frequency is greater than the expected, while a negative residual suggest that the observed frequency is less than expected
- here the residual values are between the range of -2 and +2, and the largest residual value in the table is evident amongst lower class drivers who were requested for Bribe, where the observed frequency was much greater than the expected count. Overall as most of residuals are between -2 and +2 we can assume that there is no significant deviation from the expected frequencies in each cells and therefore we fail to reject the null hypothesis, that there is no association between driver's class and police interaction

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.

(a) State a null and alternative (two-tailed) hypothesis.
H0: the reservation policy has no significant effect on the number of new or repaired drinking water facilities in the village ($\beta = 0$)
H1: the reservation policy has a significant effect on the number of new or repaired drinking water facilities in the village ($\beta \neq 0$)

Run a bivariate regression to test this hypothesis include your code.

now running the bi variate model

```
## Running the bivariate regression model
URL<- 'https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/
    women.csv'
df2<-read_csv(URL)
View(df2)
head(df2)
model <-lm(water~reserved,data=df2)
summary(model)
```

```
Call:
lm(formula = water ~ reserved, data = df2)

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
```

p-value
The p-value for our slope associated with the reservation policy is 0.0197.
Conclusion
Since this p-value is less than our significance level of 0.05, we conclude that the slope is statistically significant. Therefore, we have sufficient evidence to reject the null hypothesis, which states that there is no relationship between the reservation policy and the number of new or fixed water facilities. This suggests that changes in the reservation policy are

associated with changes in the number of water facilities, indicating a significant relationship between these variables.

Interpret the coefficient estimate for reservation policy

- the linear regression model indicates that when the reservation policy is at zero the estimated number of new or fixed water facilities is 14.73, moreover, the number of fixed or new water facilities tends to increase by 9.25 for every one unit change in the reservation policy.
- also the p-value for the slope(reservation policy) is 0.01 which is less than the significance level of0.05 therefore we can reject the null hypothesis that there exists no relationship between reservation policy and number of water facilities, and the slope estimate of 9.252 did not occur by random chance
- Additionally, the p-value for the slope (reservation policy) is 0.01, which is less than the significance level of 0.05. Therefore, we can reject the null hypothesis that there is no relationship between the reservation policy and the number of water facilities. This suggests that the slope estimate of 9.25 is statistically significant and did not occur by random chance.
- also only 33.45percent of variance in the number of water facilities can be explained by the reservation policy