PS2

Clémence Pellissier

```
{r setup, include=FALSE} knitr::opts_chunk$set(echo = FALSE)
```

Instuctions

• 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.

Problem 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.

Problem 1 Sloved Manually (using the follosing steps) :

Question a : Calculate the χ2 test statistic by hand/manually

First I set Up the contingency Table :

Class	Not Stopped	Bribe Requested	Stopped/Given Warning	Row Total
Upper Class	14	6	7	27
Lower Class	7	7	1	15
Column Total	21	13	8	42

Second, I calculate the Chi-Square Test :

The goal of this test is asserting whether there is a significant association between he class of driver and the outcome of the police encounter (e.g.,Not topped, Bribe Requested, Stopped/Given Warning).:

I use the following formula : $\chi 2 = \Sigma (O-E)2/E$ Where: O = observed frequency E = expected frequency

Third, I calculate χ2 for each cell, and I I have the following results :

For Upper Class - Not Stopped => 0.0185

For Upper Class - Bribe Requested => 0.666

For Upper Class - Stopped/Given Warning => 0.675

For Lower Class - Not Stopped => 0.0333

For Lower Class - Bribe Requested => 0.330

For Lower Class - Stopped/Given Warning => 2.12

Fourth, I sum them all up to get the total Chi-square statistic:

$\chi 2 = 0.0185 + 0.666 + 0.675 + 0.0333 + 0.330 + 2.12 = 3.84$

The result is 3.84

Fifth, I calculate the degrees of freedom, using the following formula :

$df = (\text{number of rows} - 1) \times (\text{number of columns} - 1)$ Here I Have : $df = (2-1) \times (3-1) = 1 \times 2 = 2$

Sixth, I compare this with the critical value :

For alpha = 0.05 and 2 degrees of freedom, the critical value of $\chi 2$ from the Chi-square distribution table is : 5.991.

Interpretation :

My calculated $\chi 2$ =3.84 This is less than 5.991. In other words, I fail to reject the null hypothesis. This means there is no significant association between driver class and police outcome at the 5% significance level.

Question b : Calculate thestandardized residuals for each cell (manually)

I calculate the standardized residuals, using the following formula :

Standardized Residual= $(O-E)/\sqrt{E}$

Where: O is the observed frequency for a cell E is the expected frequency for a cell (calculated earlier)

First, I recalculate the observed and expected frequencies. I have the following table :

Class	Not Stopped	Bribe Requested	Stopped/Given Warning
Upper Class (ob)	14	6	7
Upper Class (ex)	13,5	8,36	5,14
Lower Class (ob)	7	7	1
Lower Class (ex)	7,5	5,64	3,86

Second, I calculate the standardized residuals :

To do so, I apply the formula (e.g.$(O-E)/\sqrt{E}$) for each cell. I get the following results :

For Upper Class - Not Stopped => 0.136

For Upper Class - Bribe Requested => −0.816

For Upper Class - Stopped/Given Warning => 0.820

For Lower Class - Not Stopped => −0.183

For Lower Class - Bribe Requested => 0.573

For Lower Class - Stopped/Given Warning => −1.456

I observe the following residual table :

Class	Not Stopped	Bribe Requested	Stopped/Given Warning
Upper Class	0.136	-0.816	0.820
Lower Class	-0.183	0.573	-1.456

Interpertation

I know that :

1a) A positive residual indicates that the observed frequency is higher than the expected frequency.

1b) A negative residual indicates that that the observed frequency is lower than the expected frequency.

I also know that :

2a) A value closer to 0 indicates that he observed frequencies are close to the expected frequencies

2b) A larger standardized residual shows greater deviations

Here, I have the following (positive/negative) :

1a) Positive :

Upper Class - No Stopped => positive residual

Upper Class - Stopped/Given Warning => positive residual

Lower Class - Bribe Requested => positive residual

In these three cases : the observed frequency is higher than the expected frequency.

1b) Negative :

Upper Class - Bribe Requested => negative residual

Lower Class - No Stopped => negative residual

Lower Class - Stopped/Given Warning => negative residual

In these three cases : the observed frequency is lower than the expected frequency.

2a) Close to 0 (2 to -2):

Upper Class - No Stopped => close to 0

Upper Class - Bribe Requested => close to 0

Upper Class - Stopped/Given Warning => close to 0

Lower Class - No Stopped => close to 0

Lower Class - Bribe Requested => close to 0

Lower Class - Stopped/Given Warning => close to 0

In every cases : the deviation between the observed frequency and the expected frequency is low/small (i.e. between 2 and -2)

2b) Large :

N/A : no value greater than 2 or -2

Question c : How might the standardized residuals help you interpret the results?

First, as indicated above, the standard residual gives two types of information :

1 - Magnitude (large or close to 0):

The absolute value of the standardized residual helps you identify which cells deviate the most from what would be expected if there were no association between the variables (i.e., if the null hypothesis were true). In other words :

=> Residual close to 0 means the observed values are very close to the expected values for that cell, indicating that the cell is not contributing much to the overall X2 statistic => Large residual (either positive or negative) indicates cells where there is a significant difference between observed and expected frequencies.

2 - Positive or Negative variation :

=> Positive standardized residual means the observed frequency is higher than expected.
=> Negative standardized residual means the observed frequency is lower than expected.

Second, interpretation and application to the data :

1 - Understanding the significance of the deviation

=> Between -2 and 2: variation is generally considered not significant, meaning the cell's observed and expected frequencies are close.

=> Greater than 2 or less than -2: this indicates significant deviations and suggest that the cell contributes a lot to the overall X2

2 - Aplying to the data

Class	Not Stopped	Bribe Requested	Stopped/Given Warning
Upper Class	0.136	-0.816	0.820
Lower Class	-0.183	0.573	-1.456

3 - Conclusion(s)

Upper Class - No Stopped :

The standardized residual is 0.136. This means the observed frequency of No Stopped for upper-class drivers is slightly higher than expected, but it's not significantly different because the residual is still within the range of – 2 to 2.

Upper Class - Bribe Requested :

The standardized residual is -0.816. This means the observed frequency of bribe requests for upper-class drivers is slightly lower than expected, but it's not significantly different because the residual is still within the range of – 2 to 2.

Upper Class - Stopped/Given Warning :

The standardized residual is 0.820. This means the observed frequency ofStopped/Given Warning for upper-class drivers is slightly higher than expected, but it's not significantly different because the residual is still within the range of – 2 to 2.

Lower Class - No Stopped :

The standardized residual is -0.183 . This means the observed frequency of no Stopped for lower Class drivers is slightly lower than expected, but it's not significantly different because the residual is still within the range of – 2 to 2.

Lower Class - Bribe Requested :

The standardized residual is 0.573 . This means the observed frequency of bribe Requested for lower Class drivers is slightly higher than expected, but it's not significantly different because the residual is still within the range of – 2 to 2.

Lower Class - Stopped/Given Warning :

The residual is -1.456, which indicates that fewer lower-class drivers were given warnings than expected. While it's a noticeable deviation, it's not extreme enough to be statistically significant.

All Cells :

None of the residual exceeds 2 or falls below -2. This means that no individual cell shows a significant deviation from what would be expected under the null hypothesis. Therefore, no particular cell is contributing heavily to the overall $\chi 2$ statistic, aligning with the fact that the overall test did not reject the null hypothesis.

Problem 1 Solved Using R Studio (coding)

I start the X2 test (i.e. Chi-square test). Once again, the goal of this test is asserting whether there is a significant association between he class of driver and the outcome of the police encounter (e.g.,Not Stopped, Bribe Requested, Stopped/Given Warning)

I start by generating the :

1) The given table

```{r cars}

experiment <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE) rownames(experiment) <- c("Upper class", "Lower class") colnames(experiment) <- c("Not Stopped", "Bribe requested", "Stopped/given warning") experiment


### 2) I calculate the expected frequencies

```{r carsb}
row_total <- rowSums(experiment)
col_total <- colSums(experiment)
total <- sum(experiment)
f <- outer(row_total, col_total) / total
f
```

Having this, I also calculate :

3) the chi-squared

```{r carsc}

chi_squared <- sum((experiment - f)^2 / f) chi_squared


#### 4) The degree of freedom

```{r carsd}

pchi <- (nrow(experiment) - 1) * (ncol(experiment) - 1)
pchi
```

5) the p-value

```{r carse}

p_value <- pchisq(chi_squared, pchi, lower.tail = FALSE) p_value


#### 6) the standardised residuals
```

```{r carsf}

stand_residuals <- (experiment - f) / sqrt(f)
stand_residuals
```

Problem 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 system- atically differ in other aspects too). Hence, they exploit a randomized policy experiment in India, where since the mid-1990s, 13 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.

Question a : State a null and alternative (two-tailed) hypothesis

I formulate the hypotheses

=> Null Hypothesis (H_0): The reservation policy has no effect on the number of new or repaired drinking water facilities. The mean number of water facilities in villages with a reserved female GP head is equal to that in villages without a reserved female GP head.

=> Alternative Hypothesis (H_1): The reservation policy has an effect on the number of new or repaired drinking water facilities. The mean number of water facilities in villages with a reserved female GP head is not equal to that in villages without a reserved female GP head.

I start by loading the dataset from the provided URL

```{r c} # Set up libraries options(repos = c(CRAN = "https://cloud.r-project.org")) library(tidyverse)
```

Load the dataset

url <- “https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv” data <- read.csv(url)

I start inspecting the data

```{r d}

head(data)
```

I summarize of the dataset and check the missing values

```{r e} summary(data)

colSums(is.na(data))
```

I Check the structure of the dataset and view the first few rows of the dataset

```{r f}

str(data)

head(data)
```

I calculate the means for each group

```{r g} means <- data %>%  group_by(reserved) %>% summarise(mean_water = mean(water, na.rm = TRUE)) print(means)
```

I run a two-sample t-test

```{r h} t_test <- t.test(water ~ reserved, data = data) print(t_test)
```

Question b : Run a bivariate regression to test this hypothesis in R

```{r i}
library(tidyverse)
url <-
"https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"
data <- read.csv(url)
model <- lm(water ~ reserved, data = data)
summary(model)
```