PS2 Response

Applied Stats/Quant Methods 1

Zhexuan Yin

Instructions

This is my PS2 responses in `R` *and* `Latex`*.*

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).

According to the Week3 Slides, we can know that

$$H_0 : \text{The variables are statistically independent}$$
$$H_a : \text{The variables are statistically dependent}$$

We are going to calculate a test statistic (the χ^2 statistic) that is distributed according to the χ^2 distribution.

$$f_{\text{observed}} = f_o = \text{observed frequency} = \text{the raw count (NOT THE \%)}$$

$$f_{\text{expected}} = f_e = \text{what we would expect for independent samples} = \frac{\text{Row total}}{\text{Grand total}} \times \text{Column total}$$

If H_0 is true, then we would expect:

$$f_{\text{observed}} = f_{\text{expected}} = \frac{\text{Row total}}{\text{Grand total}} \times \text{Column total}$$

And

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

So "By hand" in R

```r
#(a)
# Create the observed contingency table
observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# Label the rows and columns
rownames(observed) <- c("Upper Class", "Lower Class")
colnames(observed) <- c("Not Stopped", "Bribe requested", "Stopped/given
    warning")

```

```r
# Print the observed data
print(observed)

# Row and column totals
row_totals <- rowSums(observed)
col_totals <- colSums(observed)
grand_total <- sum(observed)

row_totals
col_totals
grand_total
```

$> row_{totals}$
Upper Class Lower Class
27 15

$> col_{totals}$
Not Stopped Bribe requested Stopped/given warning
21 13 8

$> grand_{total}$
[1]42

```r
# Expected frequencies
expected <- outer(row_totals, col_totals) / grand_total
expected
```

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

```r
# Chi-square test statistic
chi_square_stat <- sum((observed - expected)^2 / expected)

# Print the results of the test
chi_square_test
```

Then

Pearson's Chi-squared test

data: observed
X-squared = 3.7912, df = 2, p-value = 0.1502

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

According to the Week3 slides, we can know that how to calculate the p-value in R.

Conditions:

$$\text{Frequency} \geq 5 \text{ for all cells}$$

Degrees of freedom:

$$\text{df} = (\text{rows} - 1)(\text{columns} - 1)$$

In R:

```
    pchisq($\chi$^2, df = (rows - 1)(columns - 1), lower.tail = FALSE)

```

So

```
#(b)
# Degrees of freedom = (number of rows - 1) * (number of columns - 1)
df <- (2 - 1) * (3 - 1)

# Chi-square test statistic
chi_square_stat <- 3.7912

# Calculate p-value
p_value <- pchisq(chi_square_stat, df, lower.tail = FALSE)
p_value
```

[1] 0.1502282

If $\alpha = 0.1$ and the p-value is less than 0.1, we reject the null hypothesis. If the p-value is greater than 0.1, we fail to reject the null hypothesis.

[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	0.1336	-0.816	0.821
Lower class	-0.182	1.096	-1.1

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

Standardized residuals help to identify which cells contribute most to the χ^2 statistic and where the discrepancies between observed and expected values lie. Residuals above 2 or below -2 typically indicate a significant difference between observed and expected values.

The residuals for Upper Class, Bribe requested (-0.816) and Lower Class, Bribe requested (1.096) suggest some difference from expected behavior, but they are not extreme.

The largest residuals occur in the Lower Class, Stopped/given warning (-1.1) category, indicating that lower-class drivers were stopped less than expected.

These residuals suggest that class may have some influence on police response, especially in the case of being stopped or given warning, though the differences are not very large.

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.

Null Hypothesis (H_0):
The reservation policy for female GP leaders has no effect on the number of new or repaired drinking water facilities in the village.

$$H_0 : \beta_1 = 0$$

Alternative Hypothesis (H_A):
The reservation policy for female GP leaders has a significant effect (positive or negative) on the number of new or repaired drinking water facilities in the village.

$$H_A : \beta_1 \neq 0$$

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

```r
#(b)
# Load necessary libraries
library(tidyverse)

# Load the dataset
url <- "https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv"
data <- read.csv(url)

# View the first few rows of the data
head(data)

# Run a bivariate regression
# Outcome variable: water
# Predictor variable: reserved
model <- lm(water ~ reserved, data = data)

# Print the summary of the regression results
```

```
summary(model)
```

(c) Interpret the coefficient estimate for reservation policy.
The results of the bivariate regression

Call:
lm(formula = water reserved, data = 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 ***
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

Interpret the coefficient estimate

Intercept: The intercept represents the expected number of new or repaired drinking water facilities when the GP leader is not female. In this case, the estimate is 14.738, meaning that on average, villages with male GP leaders have about 14.73 new or repaired drinking water facilities.

Reserved: The estimated coefficient for the reserved variable is 9.252. This means that, on average, villages with female GP leaders have 9.252 more new or repaired drinking water facilities than villages with male GP leaders.

Statistical significance: The p-value for the reserved variable is 0.016, which is less than the conventional significance level of 0.05. This means the effect of having a female GP leader (due to the reservation policy) on the number of drinking water facilities is statistically significant.

So, based on the results of the regression:
We reject the null hypothesis and conclude that the reservation policy for female GP leaders does have a significant positive effect on the number of new or repaired drinking water facilities.