Problem Set 2 Clare Zureich

Applied Stats/Quant Methods 1

Due: October 14, 2024

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).

```
The Chi-Square test statistic is 3.791
```

```r
#Create contingency table
bribery_table <- matrix(
                 c(14,7,6,7,7,1),
                 nrow = 2,
                 ncol = 3,
                 byrow = FALSE
)
rownames(bribery_table) = c("Upper Class",
                            "Lower Class")
colnames(bribery_table) = c("Not Stopped",
                            "Bribe Requested",
                            "Stopped/Given Warning")
#Hypotheses
#The null hypothesis is that the class of the driver and officers'
    response are statistically independent
#The alternative hypothesis is that the class of the driver and officers'
     response are statistically dependent

#Calculate a Test Statistic
row_total <- rowSums(bribery_table)
column_total <- colSums(bribery_table)
table_total <- sum(bribery_table)
df = (nrow(bribery_table)-1)*(ncol(bribery_table)-1)

#Expected table
expected_table <- matrix(0, nrow = 2, ncol = 3)
for (i in 1:nrow(bribery_table)) {
  for (j in 1:ncol(bribery_table)) {
    expected_table[i, j] <- (row_total[i] * column_total[j]) / table_
    total
  }
}
rownames(expected_table) <- rownames(bribery_table)
colnames(expected_table) <- colnames(bribery_table)

```

```r
#Test statistic
chi_square <- sum((bribery_table-expected_table)^2/expected_table)
chi_square
```

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

```
The p-value is 0.150. There is insufficient evidence to reject the
null hypothesis that the class of the driver and officers' responses
are statistically independent at a 90% confidence level, as the
p-value is greater than the .1 alpha.
```

```r
p_value = pchisq(chi_square, df = df, lower.tail = FALSE)
```

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

```r
adjusted_residuals <- matrix(0, nrow = 2, ncol = 3)
row_prop = row_total/table_total
col_prop = column_total/table_total
residuals <- matrix(0, nrow = 2, ncol = 3)

for (i in 1:nrow(bribery_table)) {
```

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

```
  for (j in 1:ncol(bribery_table)) {
    adjusted_residuals[i, j] <- (bribery_table[i, j] - expected_table[i,
    j]) /
      sqrt(expected_table[i, j] *
             (1-row_prop[i]) *
             (1-col_prop[j]))
  }
}
rownames(expected_table) <- rownames(bribery_table)
colnames(expected_table) <- colnames(bribery_table)
adjusted_residuals
```

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

Standardized residuals can help us interpret results by explaining which cells lead to the largest deviations of observed vs expectations if the null hypothesis (that the two variables are statistical independent) of a chi-squared test is rejected. The standardized residuals are the number of standard deviations of the sampling distribution from what we would expected under the null. They are helpful if we reject the null hypothesis by explaining which cells likely contributed to the rejection the most. In our example, the observed bribe requested response from the officers is the most standard devations away (1.642) from the expected value. The standardization takes into account amount of information.

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: The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages (Beta 1, the slope coefficient) = 0.
Alternative hypothesis: The reservation policy has an effect on the number of new or repaired drinking water facilities in the villages. (Beta 1, the slope coefficient) is not equal to 0.

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

```
str(women)
head(women)

regression_model <- lm(water ~ reserved, data = women)
summary(regression_model)
coefficients(regression_model)
```

Assumptions:
1) Linear relationship between the reservation policy and the number of new or repaired drinking water facilities.
2) The observations are independent
3) The data generation is randomized
4) Constant variance in the number of new and repaired drinking water facilities for all values of the reservation polcies

```
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 *
---
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
```

(c) Interpret the coefficient estimate for reservation policy.

There is sufficient evidence to reject the null hypotheses that there is no relationship between the reservation policy and the number of new/repaired drinking water #facilities, at a 95% confidence level, as the p-value of the coefficient is .0197. For each additional reservation policy, the number of new or repaired drinking water facilities will, on average, increase by 9.25.