Problem Set 2

Athena Rodrigues

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

Null Hypothesis: There is no relation between a driver's socioeconomic class and bribe solicitation, the variables are statistically independent. Alternative Hypothesis: A driver's socioeconomic class impacts whether they are solicited for a bribe, the variables are statistically dependent.

(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 3.80. This was calculated by hand and through R using the Frequency table and code below.

Figure 1: Frequency Table

	Not Stopped	Bribe Requested	Stopped/Given Warning	Total
Upper Class	obs: 14 | exp: 13.5	obs: 6 | exp: 8.36	obs: 7 | exp: 5.14	27
Lower Class	obs: 7 | exp: 7.5	obs: 7 | exp: 4.64	obs: 1 | exp: 2.86	15
Total	21	13	8	42

```r
# Setting up Matrix
observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow=TRUE)
rownames(observed) <- c("Upper Class", "Lower Class")
colnames(observed) <- c("Not Stopped", "Bribe Requested", "Stopped/Given Warning")
row_total <- rowSums(observed)
column_total <- colSums(observed)
overall <- sum(observed)
#expected frequency
expected <- round((outer(row_total, column_total) / overall),2)
expected
#overall table
full_table <- matrix(paste("obs:",observed, "| exp:", round(expected, 2)), nrow = nrow(observed),ncol = ncol(observed))
rownames(full_table) <- rownames(observed)
colnames(full_table) <- colnames(observed)
full_table <- cbind(full_table, row_total = paste(row_total))
full_table <- rbind(full_table, c(column_total = paste(column_total), Total = paste(overall)))
rownames(full_table)[nrow(full_table)] <- "Total"
colnames(full_table)[ncol(full_table)] <- "Total"
#part a
chi_square_stat <- round(sum((observed - expected)^2 / expected), 2)
```

(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 of 0.15 is found from the test-stastic.
This allows the conclusion to be made that there is sufficient
evidence, a p-value of 0.15 is greater than confidence level of 0.1,
to fail to reject the null hypothesis that there is no relation
between socioeconomic class and solicitation of bribes.
Code:

```
chi_pvalue <- pchisq(chi_square_stat, df = (nrow(observed)-1)*(ncol(
    observed)-1), 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	-1.644	1.526
Lower class	-0.322	1.644	-1.525

```
standardized_residuals <- round(((observed - expected) / (sqrt(expected *
    (1-(row_total/overall)) %*% t(1-(column_total/overall))))),3)
```

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

Standardized residuals help in understanding the patterns of association
between cells and is calculated using the residuals and table proportions.
This comparison of the model helps describe
the pattern of association among the contingency table.
A large standard residual, one who's absolute value is
greater than 2, provides evidence to likeley reject the null hypothesis
while standardized residuals below -3 or above 3 provide convincing
evidence of a cells true effect.

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

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.

Null-Hypothesis: The reservation policy does not impact
the number of new or repaired drinking water facilities in villages.
Alternative Hypothesis: The number of
new or repaired drinking water facilities in villages
are impacted by the reservation policy.

Ho: b = 0 Ha: b $\neq$ 0

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

Figure 3: Bivariate Regression

```
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
```

```r
bireg <- lm(water~reserved, data = women)
summary(bireg)
```

(c) Interpret the coefficient estimate for reservation policy.

The bivariate regression shows a coefficient estimate of 9.252
and a p-value of 0.0197. According to sufficient evidence found
from the regression summary, the p-value of 0.0197 being lower
than the significance value of 0.05, we reject the null
hypothesis that reservation policies do not impact the amount
of new or repaired drinking facilities in villages.
This regression interprets that at the 95% confidence level,
villages with spots reserved for women leaders
have about 9 more new or repaired drinking water facilities
than villages without the policy.