Problem Set 2

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 Multi-method 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).

```r

observed <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)
colnames(observed) <- c("Not_Stopped", "Bribe_requested", "Stopped_warning")
rownames(observed) <- c("Upper_class", "Lower_class")

print("Observed counts:")
print(observed)


row_totals <- rowSums(observed)
col_totals <- colSums(observed)
grand_total <- sum(observed)


print("Row totals:")
print(row_totals)
print("Column totals:")
print(col_totals)
print("Grand total:")
print(grand_total)


expected <- matrix(0, nrow = 2, ncol = 3)
for (i in 1:2) {
  for (j in 1:3) {
    expected[i, j] <- (row_totals[i] * col_totals[j]) / grand_total
  }
}


print("Expected counts:")
print(expected)

chi_squared_values <- (observed - expected)^2 / expected
print("Chi-squared components for each cell:")
print(chi_squared_values)


chi_squared_stat <- sum(chi_squared_values)
print("Chi-squared statistic:")
```

```
print(chi_squared_stat)
```

First the expected counts are calculated with the row and columns totals. Next the individual chi-square statistics are computed for each cell. Finally these individual chi-square statistics are summarised to result at the final chi-square statistic of 3.791168.

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

```

p_value <- pchisq(chi_squared_stat, df = df, lower.tail = FALSE)
print("P-value:")
print(p_value)


standardized_residuals <- (observed - expected) / sqrt(expected)
print("Standardized Residuals for each cell:")
```

At a significance level of a=0.1, we obtained a chi-squared statistic of 3.791 with 2 degrees of freedom and a corresponding p-value of 0.1502. Since the p-value is greater than the significance threshold of 0.1, we fail to reject the null hypothesis. This means that the data does not provide sufficient evidence to suggest a statistically significant association between social class (upper or lower) and type of police interaction (whether individuals were not stopped, had a bribe request, or were stopped and given a warning). The differences observed between the upper and lower class in this table are not significant at the 10% level and could reasonably be attributed to random variation rather than a true underlying relationship between class and police treatment.

[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	0.1361	-0.8155	0.8191
Lower class	-0.1826	1.0943	-1.0982

```
standardized_residuals <- (observed - expected) / sqrt(expected)
print("Standardized Residuals for each cell:")
print(standardized_residuals)
```

The standardized residuals are calculated by subtracting the expected from the observed values and dividing this by the squared root of the expected values.

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

Standardized residuals help interpret the results by highlighting which specific cells in the contingency table contribute most to the overall chi-squared statistic. While the chi-squared test evaluates whether an association exists between variables (e.g., social class and police interaction), it does not specify which categories drive this association or lack thereof. Standardized residuals measure the difference between observed and expected counts in each cell, scaled by the expected count's standard deviation. Positive residuals indicate overrepresentation, while negative residuals suggest underrepresentation. In this case, the residual for Lower Class and Bribe Requested is positive (1.0943), showing bribe requests occur more often than expected, while the residual for Lower Class and Stopped/Given Warning is negative (-1.0982), indicating warnings are less frequent for the lower class than expected. These residuals help identify patterns that may not be apparent from the overall chi-squared statistic, allowing for a clearer understanding of the relationships between variables.

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 for female leaders has no effect on the number of new or repaired drinking water facilities in the villages.

HA: The reservation policy for female leaders has an effect on the number of new or repaired drinking water facilities in the villages.

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

```

url <- "https://raw.githubusercontent.com/kosukeimai/qss/master/
    PREDICTION/women.csv"
data <- read_csv(url)

head(data)

model <- lm(water ~ reserved, data = data)

summary(model)
```

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

(c) Interpret the coefficient estimate for reservation policy.

The coefficient for the reserved variable is 9.252, indicating that villages reserved for female leaders had, on average, 9.252 more new or repaired drinking water facilities than non-reserved villages. This suggests the reservation policy significantly increased water-related infrastructure. The p-value of 0.0197 is below the 0.05 threshold, meaning the effect is statistically significant, and we can reject the null hypothesis. The t-value of 2.344 further supports this significance, and despite a standard error of 3.948, the effect remains robust. The intercept of 14.738 shows the average number of water facilities in non-reserved villages, and an R-squared of 0.01688 indicates the policy explains 1.69% of the variance. Although small, this is still a measurable effect. In conclusion, the reservation policy significantly increased drinking water facilities, reinforcing the idea that female leaders prioritize water improvements.