Problem Set 2

Robert Baker - Applied Stats/Quant Methods 1

15th October, 2023

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 Sunday October 15, 2023. 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.

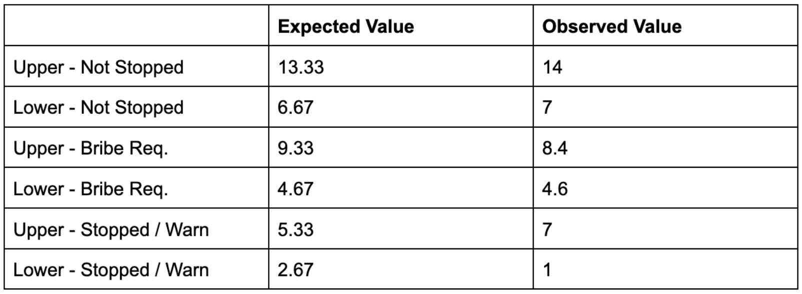
Figure 1: Traffic Experiment

	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_data <- matrix(c(14, 7, 6, 7, 7, 1), nrow=2, ncol=3)
observed_data

# Calculating row totals
row_totals <- rowSums(observed_data)

# Calculating column totals
col_totals <- colSums(observed_data)

# Calculating total sample size
total_sample_size <- sum(observed_data)

# Checking
total_sample_size

# Initialising empty matrix for 'expected' values
expected_data <- matrix(0, nrow = 2, ncol = 3)

# Calculating expected values using a nested loop
for (i in 1:2) {
  for (j in 1:3) {
    expected_data[i, j] <- (row_totals[i] * col_totals[j]) / total_sample_size
  }
}

expected_data

# checking this is the same as total observed values
sum(expected_data)

# Calculating chi-squared contributions for each cell
chi_squared_contributions <- ((observed_data - expected_data) ^ 2) / expected_data
chi_squared_contributions
```

```r

# Calculating total chi-squared test statistic
chi_squared_statistic <- sum(chi_squared_contributions)
chi_squared_statistic

# Using R function to check chi-squared
chi <- chisq.test(observed_data)
chi # these match

# Calculating Degrees of Freedom
df <- (nrow(observed_data) - 1) * (ncol(observed_data) - 1)
df
```



	Expected Value	Observed Value
Upper - Not Stopped	13.33	14
Lower - Not Stopped	6.67	7
Upper - Bribe Req.	9.33	8.4
Lower - Bribe Req.	4.67	4.6
Upper - Stopped / Warn	5.33	7
Lower - Stopped / Warn	2.67	1

Interpretation of results:
The χ^2 test statistic = Sum of ((observed - expected) squared / expected) = Individual Scores: 0.03333333 + 0.5833333 + 0.52083330.06666667 + 1.1666667 + 1.0416667 = **3.4125**

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

```r
# P Value

# Calculating P-value using the chi-squared distribution
p_value <- 1 - pchisq(chi_squared_statistic, df)
p_value
```

Interpretation of results:
DF was 2 P-Value = **0.185** (checked using pchisq() function in R). The P-Value is the probability that the test statistic (in this case the chi sq test) equals or is greater than the observed value. The smaller the P-Value the stronger the evidence is against Ho. Alpha / Our significance level was given as 0.1 (i.e Confidence level = 90 per cent). As P-Value of 0.185 is greater than 0.1 we won't reject Ho. .

[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	0.46	-3.03	2.17
Lower class	-0.55	1.92	2.86

```r
# Calculating standardised residuals for each cell
standardised_residuals <- (observed_data - expected_data) / sqrt(expected_data * (1 - rowSums(expected_data) / total) * (colSums(expected_data) / total))

round(standardised_residuals,2)

sum(standardised_residuals)
```

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

The larger the residual the more different it is to what you would ‘expect’. Larger residuals suggest something ‘significant’ may be happening in the data. We can observe there might be a correlation between being lower class and a bribe being solicited - whereas an Upper Class driver may be more likely to get away with a warning. In terms of interpreting the results, it is also worth bearing in mind the small sample size.

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 **??** 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.

[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.

The null hypothesis (H_o) = The reservation policy had no effect on the number of new or repaired water facilities.

The alternate hypothesis (H_a) = The reservation policy had an effect on the number of new or repaired water facilities .

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

Residuals: Min -23.991 1Q -14.738 Median -7.865 3Q 2.262 Max 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 * —

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
# Performing bivariate regression
regression <- lm(water ~ reserved, data=data)
regression

summary(regression_model)
```

(c) Interpret the coefficient estimate for reservation policy.

P-Value of 0.02 is statistically significant - if Significance Level was set at 0.05. If this was alpha than we may have evidence to reject Ho. The predictor variable (reserved seats for women) is positive which suggests there may be a positive correlation between the predictor variable and the dependent variables (new / repaired water facilities).

The coefficient of reserved seats for women politicians is 9.2 which suggests on average villages with reserved seats for women politicians have c. 9.2 more water facilities than villages without reserved seats for women politicians. It suggests there is a meaningful relationship between the 2 variables according to a linear regression model. It's also worth noting that the R Squared value looks weak (0.016) which suggests other variables are also at play in determining the number of new or repaired water facilities.