Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 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.

	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).

In order to perform the χ^2 test, I assume a) both variables are categorical, and b) both are nominal (i.e. not ranked, in contrast to ordinal which means ranked/ordered).

The χ^2 test compares the observed distribution of variable observations with the expected distribution IF the predictor variable has no relationship with the effect/outcome variable (i.e. the null hypothesis). A statistically significant difference gives evidence to reject the null hypothesis i.e. that the variables are independent.

First, calculate the totals for observations (rows).

```
upper_class_observed <- 14 + 6 + 7
lower_class_observed <- 7 + 7 + 1
grand_total_class <- upper_class_observed + lower_class_observed
```

Second, calculate the totals for outcome variables (columns).

```
not_stopped <- 14 + 7
bribe_requested <- 6 + 7
stopped_or_warning <- 7 + 1
```

Third, calculate expected observations.

```
upper_class_expected_not_stopped <- (upper_class_observed/grand_total_
    class)*not_stopped
lower_class_expected_not_stopped <- (lower_class_observed/grand_total_
    class)*not_stopped

upper_class_expected_bribe_requested <- (upper_class_observed/grand_total
    _class)*bribe_requested
lower_class_expected_bribe_requested <- (lower_class_observed/grand_total
    _class)*bribe_requested

upper_class_expected_stopped_or_warning <- (upper_class_observed/grand_
    total_class)*stopped_or_warning
lower_class_expected_stopped_or_warning <- (lower_class_observed/grand_
    total_class)*stopped_or_warning
```

Fourth, calculate the sum of the squared differences over expected values to find the χ^2 test. Note that for negative numbers, I had to swap the sqrt() function to avoid NaN warnings.

```
chi_squared_statistic <- sum(
  sqrt(14 - upper_class_expected_not_stopped)/upper_class_expected_not_stopped,
  (7 - lower_class_expected_not_stopped)^2/lower_class_expected_not_stopped,
  (6 - upper_class_expected_bribe_requested)^2/upper_class_expected_bribe_requested,
  sqrt(7 - lower_class_expected_bribe_requested)/lower_class_expected_bribe_requested,
  sqrt(7 - upper_class_expected_stopped_or_warning)/upper_class_expected_stopped_or_warning,
  (1 - lower_class_expected_stopped_or_warning)^2/lower_class_expected_stopped_or_warning
)
chi_squared_statistic
```

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

Calculate the degrees of freedom, which is the product of the number of rows (minus 1) and number of columns (minus 1). I choose the upper tail as it is a chi-squared hypothesis test.

```
df <- (2-1)*(3-1)
p <- pchisq(chi_squared_statistic, df = df, lower.tail = FALSE)
p
```

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

The standarised residual is the difference between the observed and the expected counts, divided by the square root of expected count. Note I have rounded each value to 2 decimal places.

```
sr_upper_class_not_stopped <- round((14 - upper_class_expected_not_stopped) / sqrt(upper_class_expected_not_stopped), 2)
sr_lower_class_not_stopped <- round((7 - lower_class_expected_not_stopped) / sqrt(lower_class_expected_not_stopped), 2)

sr_upper_class_bribe_requested <- round((6 - upper_class_expected_bribe_requested) / sqrt(upper_class_expected_bribe_requested), 2)
sr_lower_class_bribe_requested <- round((7 - lower_class_expected_bribe_requested) / sqrt(lower_class_expected_bribe_requested), 2)

sr_upper_class_stopped_or_warning <- round((7 - upper_class_expected_stopped_or_warning) / sqrt(upper_class_expected_stopped_or_warning), 2)
sr_lower_class_stopped_or_warning <- round((1 - lower_class_expected_stopped_or_warning) / sqrt(lower_class_expected_stopped_or_warning), 2)
```

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

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.14	-0.82	0.82
Lower class	-0.18	1.09	-1.1

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

Standardised residuals measure the strength of the difference between observed and expected values. I note two details: the positive and negative values largely cancel each other out, and also each value is relatively small (less than ± 1.1).

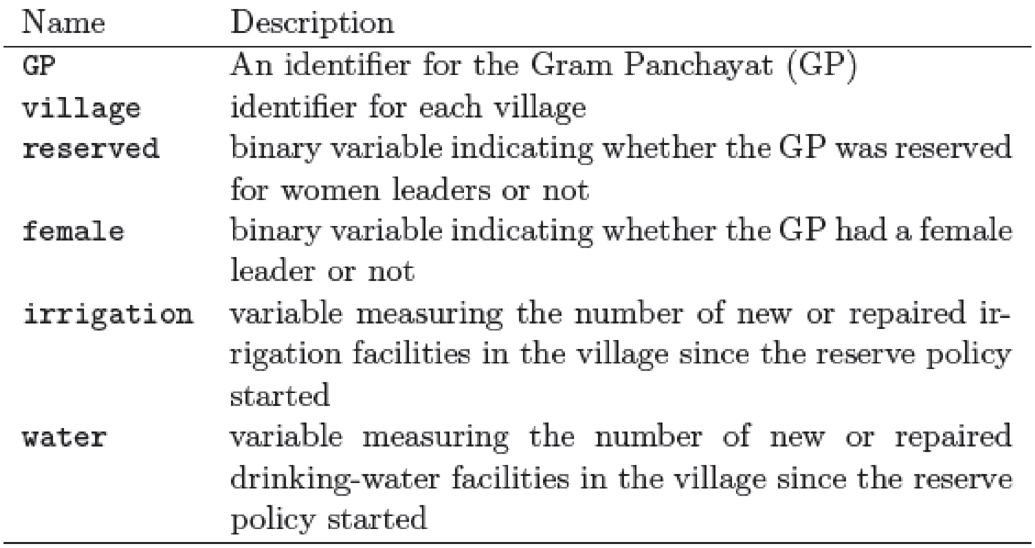
These both indicate that the differences are so minor that it would require a large n-size for such results to be statistically significant.

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.

H_0: The number of new or repaired drinking water facilities is identical in villages with a reservation policy and without a reservation policy: $water_1 = water_0$

H_a: The number of new or repaired drinking water facilities is different in villages with a reservation policy and without a reservation policy: $water_1 \neq water_0$

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

I extract the data.

```
WB_df <- read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv")
```

I create a linear regression model with the presence of a reservation policy (0 denotes no policy, 1 denotes policy present)

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

I then print this for input into LaTeX.

```
stargazer(model)
```

Table 1:

	Dependent variable:
	water
reserved	9.252**
	(3.948)
Constant	14.738***
	(2.286)
Observations	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493** (df = 1; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01

(c) Interpret the coefficient estimate for reservation policy.

The linear regression indicates an average increase of 9.252 drinking water facilities for villages which have a reservation policy compared to villages which don't.

Further, $p < 0.05$, indicating this result is statistically significant.

In other words, I have evidence to reject the null hypothesis (the number of new and repaired facilities in villages with and without a reservation policy is the same) and I have evidence supporting the alternative hypothesis (the number of facilities in villages with the reservation policy is different to the number in villages without the reservation policy).