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).

```
table <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, ncol = 3, byrow = TRUE)
rownames(table) <- c("upper class", "lower class")
colnames(table) <- c("not stopped", "bribe request", "stopped/warning")

row_frequen <- rowSums(table)

col_frequen <- colSums(table)

grand_total <- sum(table)

expected_frequencies <- outer(row_frequen, col_frequen) / grand_total

chi_statis <- sum((table - expected_frequencies)^2 / expected_frequencies
    )
#chi statistic = 3.79
```

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

```
df <- (nrow(table) - 1) * (ncol(table)- 1)

p_val<-pchisq(chi_statis, df, lower.tail = FALSE)
result <- chisq.test(table)
print(result)
#alpha=0.1, p value=0.15

```

p value 0.15 > significant value alpha 0.1 in the contingency table, not enough evidence to reject the null hypothesis; Null Hypothesis (H0): There is no association between the officers' solicitation of a bribe and the drivers' class. In other words, there isn't enough evidence to conclude the variables are dependent or there is an association between a bribe and driver class.

[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.

```r
  standardized_residuals <- (table - expected_frequencies) / sqrt(
    expected_frequencies)
print(standardized_residuals)
sink("standardized_residuals_output.txt")
sink()
df_standardized_residuals <- as.data.frame(standardized_residuals)
df_standardized_residuals$class <- rownames(df_standardized_residuals)

install.packages("ggplot2")
library(ggplot2)
ggplot(df_standardized_residuals, aes(x = class, y = "bribe request")) +
  geom_point() +
  labs(x = "class", y = "Standardized Residual (bribe request)") +
  theme_minimal()

```

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	0.1360828	-0.8153742	0.818923
Lower class	-0.1825742	1.0939393	-1.098701

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

Positive values of standardized residuals indicate that the observed frequency of the cell is higher than expected; negative values indicate that the observed frequency is lower than expected. Larger absolute values of standardized residuals show stronger deviations in expectation of independence of the variable. for example: the frequency of giving warnings to lower-class drivers is significantly lower than expected. bribes requested to upper-class drivers is lower than expected, in contrast, bribes requested to lower-class driver is much higher than expected.

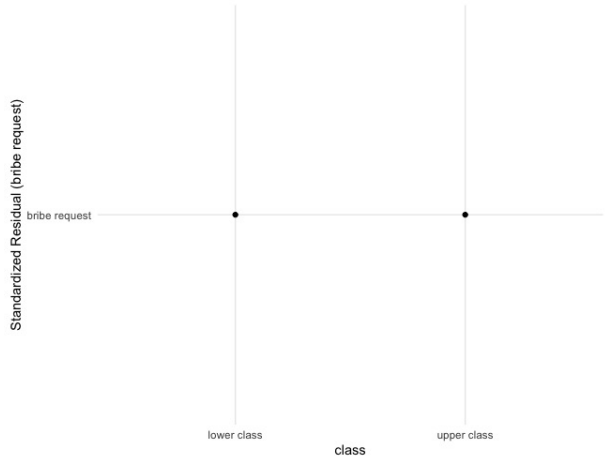

Figure 1: standardized residual$_b ribe$

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.

null hypothesis: The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages

Alternative Hypothesis (Ha): The reservation policy has two possible directions of effect on the number of new or repaired drinking water facilities in the villages

null hypothesis(H0): $\mu 1 = \mu 2$

Alternative Hypothesis (Ha): $\mu 1 \neq \mu 2$
$\mu 1$:population mean of number drinking water facilities with reservation policies
$\mu 2$: population mean of number drinking water facilities without reservation policies

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

```
install.packages("tidyverse")
library(tidyverse)
data<-read.csv("https://raw.githubusercontent.com/kosukeimai/qss/master/
    PREDICTION/women.csv")
summarise(data)
bivariate_model <- lm(water ~ reserved, data = data)
summary(bivariate_model)

ggplot(data, aes(x = reserved, y = water)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Reserved policy", y = "Water facility") +
  ggtitle("Scatterplot of Water facilityvs. Reserved policy") +
  theme_minimal()
```

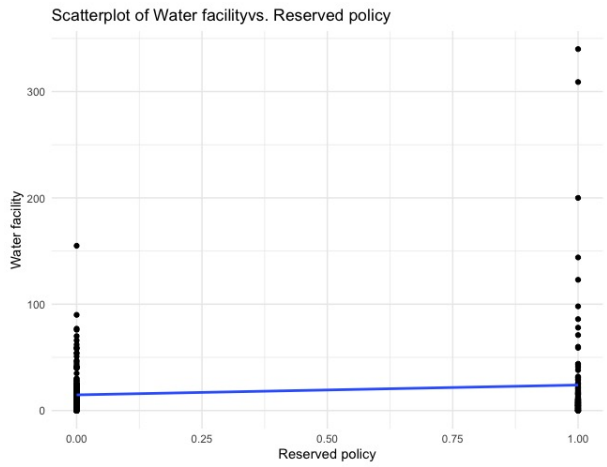

Figure 2: water facility vs policy

```
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
```

Figure 3: regression summary

(c) Interpret the coefficient estimate for reservation policy.

the estimated intercept is 14.738, which represents the value of Y(water facility) when X(reserved policy) is zero.
slope coefficient is 9.252, which represents the change in Y mean for a one-unit change in X.
A 1-unit increase in the "reservation policy " variable is associated with an estimated increase of 9.252 units in the mean number of Y (new or repaired drinking water facilities).
p-value is 0.0197, which is less than significance level of 0.05, indicating statistical significance, with evidence to reject the null hypothesis, which means the reservation policy affects the number of new or repaired drinking water facility