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).

The χ^2 test result is 3.791168. Below is the R code:

```
1
2 ############### Question 1 ###############
3
4 ######## (a) ########
5
6 # Formulate the row of cross table
7 class_type <- c(rep("Upper class", 27), rep("Lower class", 15))
8 class_type <- factor(class_type)
9 # Formulate the columns of cross table
10 bribe_type <- c(rep("Not stopped", 14), rep("Bribe requested", 6),
11                  rep("Stopped warning", 7),
12                  rep("Not stopped", 7), rep("Bribe requested", 7),
13                  rep("Stopped warning", 1))
14 bribe_type <- as.factor(bribe_type)
15 # Check cross table
16 table(class_type, bribe_type)
17 # Calculate row and columns counts
18 row_1 <- 15 ; row_2 <- 27
19 col_1 <- 13 ; col_2 <- 21; col_3 <- 8
20 n <- 42
21 # Calculate expected frequencies
22 e11 <- (row_1 * col_1)/n; e12 <- (row_1 * col_2)/n; e13 <- (row_1 * col_
       3)/n
23 e21 <- (row_2 * col_1)/n; e22 <- (row_2 * col_2)/n; e23 <- (row_2 * col_
       3)/n
24 # Calculate Chi-squared Value
25 x11 <- (7 - e11)^2/e11; x12 <- (7 - e12)^2/e12; x13 <- (1 - e13)^2/e13
26 x21 <- (6 - e21)^2/e21; x22 <- (14 - e22)^2/e22; x23 <- (7 - e23)^2/e23
27 # Calculate chi-square value and print
28 x_square_value <- x11 + x12 + x13 + x21 + x22 + x23
```

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

Here is the null hypothesis and alternative hypothesis:
H0: These 2 variables in the population are not related and are independent of each others.
H1: These 2 variables in the population are related to each others.
Below is the R code:

```
1
2 ######## (b) ########
3
4 # Calculate degree of freedom
5 df <- (3 - 1) * (2 - 1)
6 # Calculate p value
7 p_value <- 1 - pchisq(x_square_value, df)
8 print(p_value)
9 # Set alpha
10 alpha <- 0.1
11 # Check chi-square value with 0.1 alpha
12 critical_chisq <- qchisq(1 - alpha, df)
13 print(critical_chisq)
14 # Compared with p value and alpha
15 if (p_value > alpha) {
16   print("Reject H1, Accept H0")
17 } else {
18   print("Reject H0, Accept H1")
19 }
20 # Another way: compared with critical level and chi-square results
21 if (x_square_value < critical_chisq) {
22   print("Reject H1, Accept H0")
23 } else {
24   print("Reject H0, Accept H1")
```

p value = 0.1502306 > 0.1 = alpha;
and $\chi^2_{0.1} = 3.791168 < 4.60517 = \chi^2$(K=3)

so we reject H1 and accept H0, we can think these 2 variables in the population are not related and are independent of each others.

[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	0.14	-0.82	0.82
Lower class	-0.18	1.09	-1.10

Below is the R code:

```
1
2 ######## (c) ########
3
4 # According to (a), we have calculated expected frequencies
5 # Now we can calcualte residual by (frequencies - expected values)
6 resi11 <- 7 - e11; resi12 <- 7 - e12; resi13 <- 1 - e13;
7 resi21 <- 6 - e21; resi22 <- 14 - e22; resi23 <- 7 - e23;
8 # Calculate standard residual by [residual / sqrt(expected value)]
9 s_resi11 <- resi11/sqrt(e11); s_resi12 <- resi12/sqrt(e12);
10 s_resi13 <- resi13/sqrt(e13)
11 s_resi21 <- resi21/sqrt(e21); s_resi22 <- resi22/sqrt(e22);
12 s_resi23 <- resi23/sqrt(e23)
13 # Print and check the results
```

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

Firstly, we can determine whether the difference between the observed value and the expected value in each cell is significant.

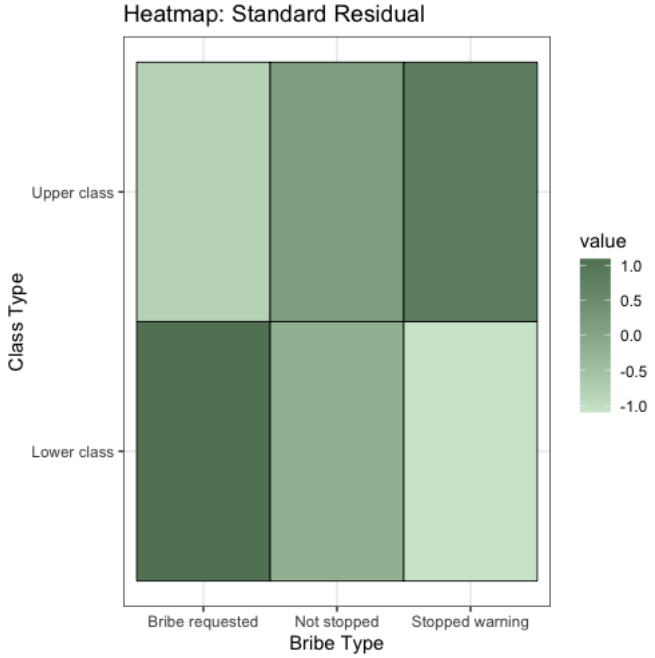
- Since no cell has a standardized residual greater than 1.61, we can assume that there is no difference between the observed and expected values in these cells.

Secondly, we can judge the relationship of observed and expected value based on the positive and negative conditions of the residual. If we assume each standard residuals is significant, then we can conclude that:

- For not stopped bribe categorical, upper class's observed value more than expected value; and lower class's observed value less than expected value.
- For bribe requested categorical, upper class's observed value less than expected value; and lower class's observed value more than expected value.
- For stopped or given warning categorical, upper class's observed value more than expected value, and lower class's observed value less than expected value.

Draw a heat map to visualize each cell's standard residual:

Figure 1: Heatmap: Standard Residual



Below is the R code:

```
1
2 ######## (d) ########
3 # Draw a heat map to visualize the differences between standard residual
4 # Make a matrix about standard residual
5 s_resi_mat <- matrix(
6   c(s_resi11, s_resi12, s_resi13, s_resi21, s_resi22, s_resi23),
7   nrow = 2, ncol = 3, byrow = TRUE)
8 # Give colnames and rownames in matrix
9 colnames(s_resi_mat) <- c("Bribe requested", "Not stopped", "Stopped
    warning")
10 rownames(s_resi_mat) <- c("Lower class", "Upper class")
11 # Transfer matrix into long data table to draw ggplot picture
12 s_resi_long <- melt(s_resi_mat)
13 print(s_resi_long)
14 # Use ggplot to draw heatmap
15 s_resi_heatmap <- ggplot(s_resi_long, aes(x=Var2, y=Var1)) +
16   geom_tile(aes(fill=value), color="black", size=0.3) +
17   scale_fill_gradient(low="#D0E7D2", high="#618264") +
18   labs(title="Heatmap: Standard Residual",
19        x="Bribe Type", y="Class Type") +
20   theme_bw()
```

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.

H_0: Whether there are female leaders in a region **is not related to** with the number of newly built drinking water facilities in the region.

H_1: Whether there are female leaders in a region **is related to** the number of newly built drinking water facilities in the region.

And in math language, we can also make the statement like this:

H_0: $\beta = 0$
H_1: $\beta \neq 0$

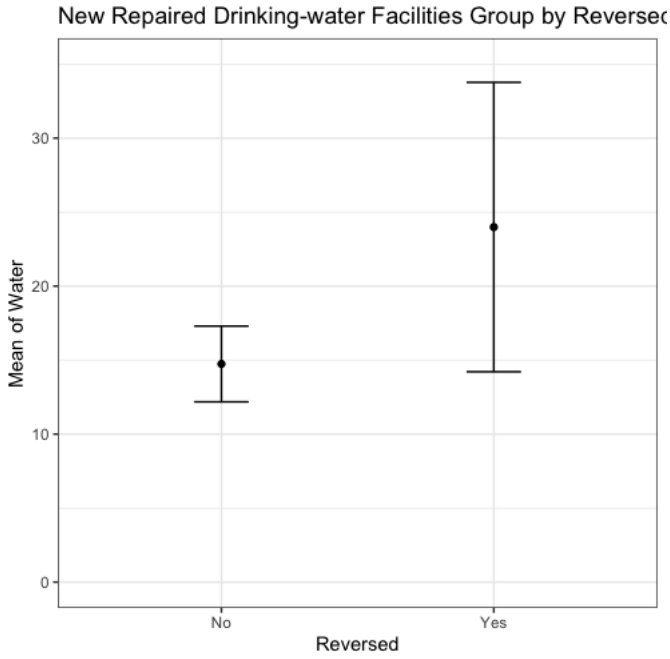
β_0 means the slop in regression model $y = \beta x + \alpha + \varepsilon$

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

Put "water" as dependent variable and binary variable "reversed" as independent variable. And because binary variable "reversed" is already a dummy variable, so we don't need to operate anything on it.

Before we start regression model, we can draw a point plot to see the mean value of newly repaired drinking water facilities group by reversed.

Figure 3: Point: Water Group by Reversed



Below is the R code of the plot:

```
1
2 ######## (b) ########
3
4 # Import data from website link
5 df <- readr::read_csv(
6   "https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/
      women.csv")
7 # Set regression's dependent variable and independent variable
8 reg_y <- df$water; reg_x <- df$reserved
9 # Because the independent variable is categorical,
10 # so draw a bar plot instad of scatter
11 # Draw a scatter to see differences between these two variables
12
```

```r
# Create a data frame about mean and upper, lower ci
reg_mean_ci <- aggregate(reg_y ~ reg_x, data=df, FUN=function(x){
  mea_y <- mean(x)
  ci <- t.test(x)$conf.int
  return <- (c(mea_y, ci[1], ci[2]))
})
# Create x and y for point plot
sca_x <- c("0", "1"); sca_y <- c(14.74, 23.99)
# Create lowerci and upperci for point plot
lower_ci <- reg_mean_ci$reg_y[,2]; upper_ci <- reg_mean_ci$reg_y[,3]
reg_sca <- ggplot(reg_mean_ci, aes(x=sca_x, y=sca_y)) +
  geom_point() +
  geom_errorbar(aes(ymin=lower_ci, ymax=upper_ci), width=0.2) +
  labs(x="Reversed", y="Mean of Water",
       title="New Repaired Drinking-water Facilities Group by Reserved")
    +
  scale_x_discrete(labels=c("No", "Yes")) +
  ylim(0,35) +
  theme_bw()
```

Use R to get the data frame and create regression model:

```r
# Create a regression model
reg_mod <- lm(reg_y ~ reg_x, data=df)
summary(reg_mod)
```

And we can get the regression model table from R:

Table 1: Abstract of Regression Model

	Estimate	Std. Error	t value	Pr($>\|t\|$)
(Intercept)	14.738	2.286	6.446	4.22e-10 ***
Reversed	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

Because the slop's p value is $0.0197 < 0.05$, so it is significant.

And this suggests we can reject H_0 and accept H_1, which means $\beta \neq 0$.

(c) Interpret the coefficient estimate for reservation policy.

According to (b), the regression model is:

$y = 9.252\ x + 14.738$

According the regression model, because we pass the F test and the slop is significant, so we can make this statement:

In comparison to non-female leaders GP, the female leaders GP's newly repair drinking water facilities will expect to increase 9.252.