Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

Linette Lim

Question 1 (40 points): 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`).

1. Answer for Question 1a
   First, make sure global environment is empty, load packages such as Tidyverse, and set the working directory.
   Now create data in R using the following code:

```
A = matrix(

  # Taking sequence of elements
  c(14, 6, 7, 7, 7, 1),

  # No of rows
  nrow = 2,

  # No of columns
  ncol = 3,

  # By default matrices are in column-wise order
  # So this parameter decides how to arrange the matrix
  byrow = TRUE
)

```

```
# Naming rows
rownames(A) = c("Upper class", "Lower class")

# Naming columns
colnames(A) = c("Not stopped", "Bribe requested", "Stopped or warned")

cat("The 2x3 matrix:\n")
```

We get a matrix, A:

```
             Not stopped Bribe requested Stopped or warned
Upper class           14               6                 7
Lower class            7               7                 1
```

I manually calculated the row and column totals using these steps:

```
sum(A[1, ])
sum(A[2, ])
sum(A[ ,1])
sum(A[ ,2])
sum(A[ ,3])

sum.row <- sum(A[1, ]) + sum(A[2, ])
sum.column <- sum(A[ ,1]) + sum(A[ ,2]) + sum(A[ ,3])
```

Based on the row and column totals, I created a new matrix with the following code:

```
B = matrix(

  # Taking sequence of elements
  c(14, 6, 7, 27, 7, 7, 1, 15, 21, 13, 8, 42),

  # No of rows
  nrow = 3,

  # No of columns
  ncol = 4,

  # By default matrices are in column-wise order
  # So this parameter decides how to arrange the matrix
  byrow = TRUE
)

```

```r
# Naming rows
rownames(B) = c("Upper class", "Lower class", "Column total")

# Naming columns
colnames(B) = c("Not stopped", "Bribe requested", "Stopped or warned", "Row total")

cat("The 3x4 matrix:\n")
```

The new matrix, B, is as follows:

```
             Not stopped Bribe requested Stopped or warned Row total
Upper class           14               6                 7        27
Lower class            7               7                 1        15
Column total          21              13                 8        42
```

Now, I calculate the Fe, or Fexpected with the formula Fe = (Row total / grand total) * column total, and with that we can calculate the chi square statistic:

```r
c1 <- (21/42)*27
c2 <- (13/42)*27
c3 <- (8/42)*27
c4 <- (21/42)*15
c5 <- (13/42)*15
c6 <- (8/42)*15

chi.square <- ((14-c1)^2/c1) + ((6-c2)^2/c2) + ((7-c3)^2/c3) + ((7-c4)^2/c4) + ((7-c5)^2/c5) + ((1-c6)^2/c6)
```

The chi square is:

```
[1] 3.791168
```

To check if the answer 'by hand' is correct, we can run the chisq.test() function on A and we get chi square = 3.7912:

```
Pearson's Chi-squared test

data:  A
X-squared = 3.7912, df = 2, p-value = 0.1502
```

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

2. Answer for Question 1b
First, we calculate the degrees of freedom with the formula: df = (rows -1)(columns -1). We get df=2. We then use df=2 in the following formula to find the p-value:

```
p.value <- pchisq(chi.square, df = 2, lower.tail=FALSE)
```

We get:

```
[1] 0.1502306
```

To confirm if p-value is 0.15, I convert the matrix (A) created earlier into a table and run some code as follows. The summary() function reaffirms the chi sq statistic as 3.79 and the p-value as 0.15.

```
table <- as.table(A)
print(table)
summary(table)
```

The output is as follows:

```
> summary(table)
Number of cases in table: 42
Number of factors: 2
Test for independence of all factors:
Chisq = 3.791, df = 2, p-value = 0.1502
Chi-squared approximation may be incorrect
```

The p-value is 0.15. This is larger than the significance level alpha = 0.1, so we do not have enough evidence to reject the null hypothesis, that there is no relationship between the two variables, socioeconomic class and treatment by police officers.

[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			
Lower class			

3. Answer for Question 1c
   To access the standardized residuals, you can assign the output of your chisq.test() to an object, and print out using object$stdres:

```r
model <- chisq.test(table)
model$stdres
```

We get:

```
> model$stdres
            Not stopped Bribe requested Stopped or warned
Upper class   0.3220306      -1.6419565         1.5230259
Lower class  -0.3220306       1.6419565        -1.5230259
```

It is also possible to calculate by hand in r. First, we calculate the row and column proportions:

```r
c1.rowprop <- 27/42
c1.colprop <- 21/42
c2.rowprop <- 27/42
c2.colprop <- 13/42
c3.rowprop <- 27/42
c3.colprop <- 8/42
c4.rowprop <- 15/42
c4.colprop <- 21/42
c5.rowprop <- 15/42
c5.colprop <- 13/42
c6.rowprop <- 15/42
c6.colprop <- 8/42
```

Next, we can find out the standardized residuals using the following formula:

```
z1 <- (14-c1)/sqrt((c1*(1-c1.rowprop)*(1-c1.colprop)))
z2 <- (6-c2)/sqrt((c2*(1-c2.rowprop)*(1-c2.colprop)))
z3 <- (7-c3)/sqrt((c3*(1-c3.rowprop)*(1-c3.colprop)))
z4 <- (7-c4)/sqrt((c4*(1-c4.rowprop)*(1-c4.colprop)))
z5 <- (7-c5)/sqrt((c5*(1-c5.rowprop)*(1-c5.colprop)))
z6 <- (1-c6)/sqrt((c6*(1-c6.rowprop)*(1-c6.colprop)))
```

Finally using the same formula earlier for creating matrix A and B, we can arrange the standardized residuals for each cell in another matrix, C. You can see we have arrived at the same standardized residuals through calculation by hand in R.

```
> print(C)
            Not stopped Bribe requested Stopped or warned
Upper class   0.3220306       -1.641957          1.523026
Lower class  -0.3220306        1.641957         -1.523026
```

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

4. Answer for Question 1d
   First, we create a dataframe so that it is easier to manipulate the data with ggplot.

```
X <- c("upper", "upper", "upper", "lower", "lower", "lower")
Y1 <- c(14, 6, 7, 7, 7, 1)
Y2 <- c("not.stopped", "bribe.requested", "stopped.warned","not.stopped",
    "bribe.requested", "stopped.warned")
```

```r

dataframe <- data.frame(X, Y1, Y2)
names(dataframe) <- c('Class', 'Cases', 'Treatment')
print(dataframe)
```

We get:

```
> print(dataframe)
  Class Cases       Treatment
1 upper    14     not.stopped
2 upper     6 bribe.requested
3 upper     7  stopped.warned
4 lower     7     not.stopped
5 lower     7 bribe.requested
6 lower     1  stopped.warned
```

Next, we apply the ggplot() function.

```r
ggplot(dataframe, aes(x=X, y=Y1, colour=Y2, shape=Y2)) +
  geom_point()  +
  theme_bw() +
  theme(axis.title = element_text(size=20),
        axis.text = element_text(size=15),
        legend.title = element_text(size=17),
        legend.text = element_text(size=15))
```

Based on plot 1, we see that both classes have a similar likelihood of being targeted for bribes. It also looks like (a) the upper class are more likely to be not stopped when flouting traffic rules. However, it seems puzzling that at the same time, (b) a member of the upper class is also more likely to be stopped or warned than someone deemed lower class. Since the interpretation is unclear, let us try using standardized residuals to interpret the results. We use the following formula:

```r
install.packages("corrplot")
library(corrplot)
corrplot(model$stdres, is.cor = FALSE)
```

Conclusion: The positive residuals are in blue, signifying a positive association between the corresponding row and column variables. In the plot, we can see a strong

Figure 1: Relationship between Class and Police treatment.

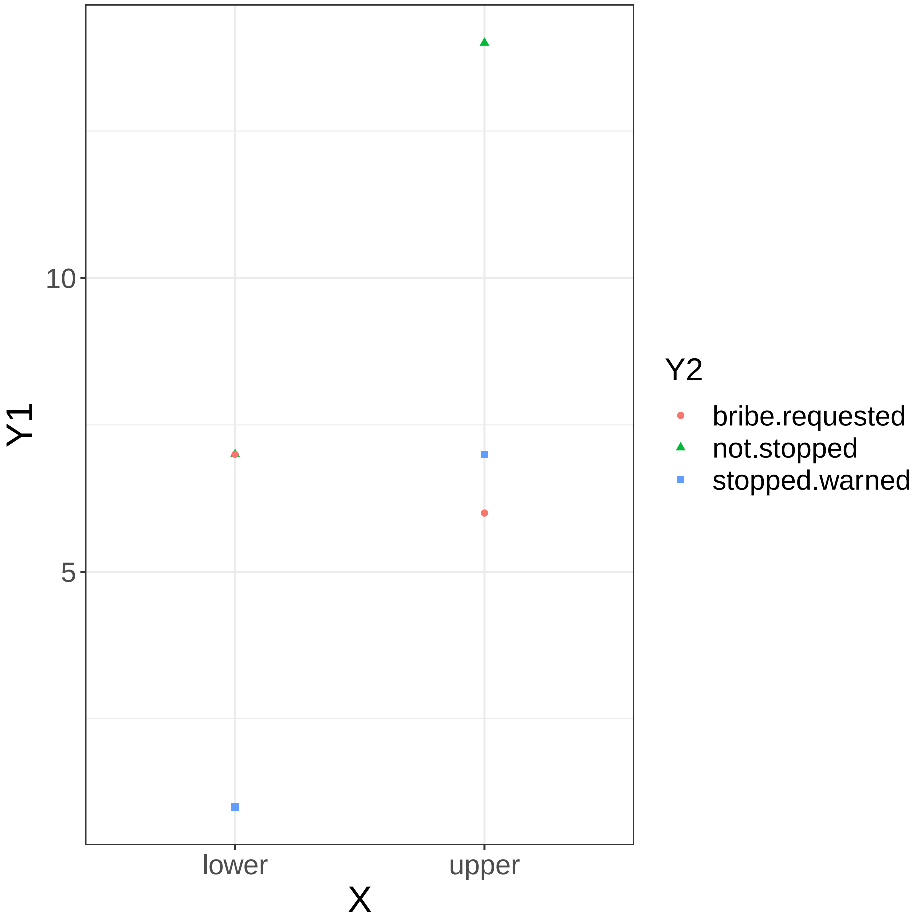
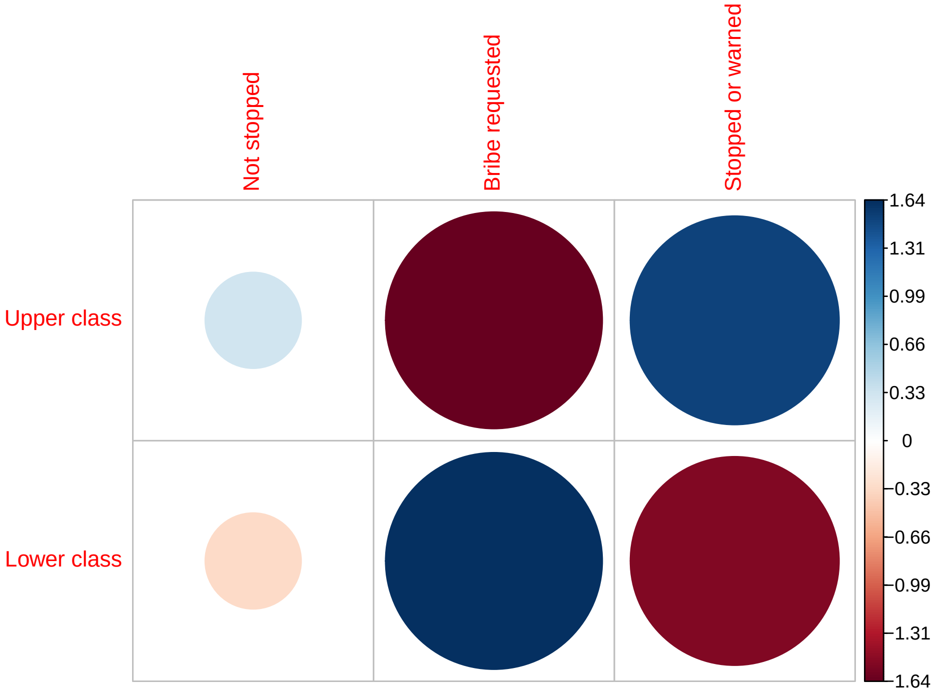


Figure 2: Relationship between Class and Police treatment.



positive association between being lower class and being asked to pay a bribe. There is also a strong positive association between being upper class and being stopped or warned. This seems to suggest that the lower class are often stymied by 'informal' ways of doing things, possibly due to a combination of low social capital and lack of literacy in the law, while the upper class are somewhat protected by and benefit from formal institutions like the law and the police force.

The negative residuals are in red. This implies a negative association between the corresponding row and column variables. From the plot, we see the upper class is less likely (or 'not associated' with) to be asked to pay a bribe, while the lower class is less likely to be stopped or warned. Taken together, this suggests that there are different traffic offense resolution methods for the upper class and the lower class. The lower class is more usually let go with a bribe, while the upper class is more likely to be formally warned or sanctioned.

Question 2 (40 points): 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.

1. Answer for Question 2a
   First, we clear the global environment, load packages, set the working directory, and import the data.

```
rm(list=ls())
library(tidyverse)
setwd(”/Users/linettelim/Documents/GitHub/StatsI_Fall2022/problemSets/PS02/”)
data <- read.csv (”https://raw.githubusercontent.com/kosukeimai/qss/master/PREDICTION/women.csv”)
```

Let us find the difference in means for the number of drinking water facilities between the villages with treatment (reservation for women) and without treatment using this formula:

```
mean(data$water[data$reserved == 1]) - mean(data$water[data$reserved == 0])
```

We get:

```
[1] 9.252423
```

Thi suggests that the reservation policy increased the number of drinking water facilities in a GP on average by about 9 (new or repaired). Therefore, the null hypothesis is that there is zero treatment effect, i.e.:

```
# H0 : B1 = 0
# Ha: B1 =/= 0
```

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

2. Answer for Question 2b
   We run a bivariate regression using the lm() function:

```r
lm(water ~ reserved, data = data)
fit.data <- lm(water ~ reserved, data = data)
summary(fit.data)
```

We get:

```
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
```

Here, we see that the p-value = 0.0197. This is less than the typical threshold of alpha=0.05, so we reject the null hypothesis that treatment effect (of reservation policy on new or repaired drinking water facilities) is zero.

(c) Interpret the coefficient estimate for reservation policy.

3. Answer for Question 2c
From the output above, we can see that the coefficient estimate for reservation policy is 9.252, the standard error is 3.948, and the t-statistic for the estimated slope coefficient is 2.344. The slope coefficient is equal to the difference-in-means estimator earlier calculated (9.252423). When the explanatory variable is binary, the estimated average treatment effect equals the estimated slope coefficient. We can run the confint() function to check:

```
confint(fit.data)
```

We get:

```
> confint(fit.data)
                2.5 %   97.5 %
(Intercept) 10.240240 19.23640
reserved     1.485608 17.01924
```

The result tells us that having reservations for women is estimated to increase the number of drinking water facilities by 9.25 facilities with a 95 percent confidence interval of [1.49, 17.02].