Problem Set 2

Applied Stats/Quant Methods 1

Due: October 16, 2022

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.

	Not Stopped	Bribe requested	Stopped/given warning
Upper class	14	6	7
Lower class	7	7	1

[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.

(a) *Calculate the χ^2 test statistic by hand/manually (even better if you can do "by hand" in R).*

H0: The variables are statistically independent.
H1: The variables are statistically dependent.

First, create a matrix of our data and name the rows and columns.

```
driver <- matrix(c(14,6,7,7,7,1),ncol=3,byrow=TRUE)
colnames(driver) <- c("Not Stopped","Bribe Requested","Stopped/Given
    Warning")
rownames(driver) <- c("Upper Class","Lower Class")
driver <- as.table(driver)
driver
```

```
            Not Stopped  Bribe Requested  Stopped/Given Warning
Upper Class          14                6                      7
Lower Class           7                7                      1
```

Calculate the table of expected values by multiplying the vectors of the margins and dividing by the total number of observations, using the "t" function take the transpose of the array.

```
expected <- as.array(margin.table(driver,1)) %*% t(as.array(margin.table(
    driver,2))) / margin.table(driver)
expected
```

```
            Not Stopped  Bribe Requested  Stopped/Given Warning
Upper Class        13.5         8.357143               5.142857
Lower Class         7.5         4.642857               2.857143
```

We need the square of the difference between the two tables divided by the expected values. The sum of all these values is the Chi-squared statistic:

```
chi <- sum((expected - as.array(driver))^2/expected)
chi
```

The χ^2 test statistic is:

```
[1] 3.791168
```

(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.

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

The degrees of freedom = 2

We can now find the p-value

```
p_value <- 1-pchisq(chi, df=2)
p_value
```

```
[1] 0.1502306
```

If a = 0.1, then a p-value of 0.1502306 (which is greater than 0.10) is considered not significant as p > 0.10, so we fail to reject the null hypothesis.

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

We calculate the standard residuals by subtracting the expected values from the observation and dividing by the square root of the expected values:

```
std_residuals <- (driver - expected) / sqrt(expected)
std_residuals
```

```
             Not Stopped         Bribe Requested         Stopped/Given Warning
Upper Class    0.1360828              -0.8153742             0.8189230
Lower Class   -0.1825742               1.0939393            -1.0987005
```

We can check if r agrees using the chisq.test() and str()function and calling the residuals

```
chi2 <- chisq.test(driver, correct=FALSE)
```

```
                        Pearson's Chi-squared test
data: driver
X-squared = 3.7912, df = 2, p-value = 0.1502
```

```
str(chi2)
chi2$residuals
```

[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	0.0.1360828	-0.1360828	0.8189230
Lower class	-0.1825742	1.0939393	-1.0987005

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

In this dataset, the standardized residuals indicate that more lower class drivers were approached for a bribe than expected (std res slightly greater than 1) , however fewer lower class drivers were stopped and given a warning than expected. A comparable number of upper and lower class drivers were not stopped at all when then numbers in each class are accounted for (std res close to 0).

Fewer bribes than expected were solicited from upper class individuals, however more than expected were stopped and given a warning. However, the absolute value of the standardized residuals in these cells were less than 1 and so not considered significant.

Overall the most significant contributor to the chi squared test were lower class drivers being asked for bribes and lower class drivers stopped and given a warning. However, the relatively small values for the standardized residuals support our rejection of the alternate hypothesis that these variables are dependent and do support the adoption of the null hypothesis.

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.

(a) *State a null and alternative (two-tailed) hypothesis.*

H0: The reservation policy has no effect on the number of new or repaired drinking water facilities in the villages.

H1: On average, the reservation policy has either a positive or negative effect 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.

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

```r
Reserved_Effect_Water <- lm(water ~ reserved, data = P2)
summary(Reserved_Effect_Water)
```

```
Call:
lm(formula = water ~ reserved, data = P2)

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
```

As the resulting p-value of 0.0197 for reserved is less than the usual significance level a = 0.05, we can conclude that our sample data provides enough evidence to reject the null hypothesis that the reservation policy has no effect on the number of new or repaired drinking water facilities on the village. On average, the reservation policy has either a positive or negative effect on the number of new or repaired drinking water facilities in the villages.

(c) *Interpret the coefficient estimate for reservation policy.*

Convert into a data matrix

```
matrix_coef <- summary(Reserved_Effect_Water)$coefficients
matrix_coef
```

Subset to extract only the coefficient estimates

```
my_estimates <- matrix_coef[ , 1]
my_estimates
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	14.738318	2.286300	6.446363	4.216474e-10
reserved	9.252423	3.947746	2.343723	1.970398e-02

As the table shows, the coefficient estimate of the reserved variable is 9.252423.
This positive coefficient indicates that as the value of the independent variable (reservation policy for women leaders) increases, the mean of the dependent variable (no. new or repaired drinking water facilities since the policy started) also tends to increase.

The number of new or repaired drinking water facilities, on average, in a village with a reservation policy is expected to be 9.25 times more compared to the number f new or repaired drinking water facilities in a village without a reservation policy.