# Problem Set 2

Applied Stats/Quant Methods 1

Due: October 15, 2021

## 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 in .pdf form.
- This problem set is due before class on Friday October 15, 2021. No late assignments will be accepted.
- Total available points for this homework is 100.

## 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 $\chi^2$ test statistic by hand (even better if you can do ”by hand” in `R`).

```r
# identify observed frequency for each cell
f_o11 <- 14
f_o12 <- 6
f_o13 <- 7
f_o21 <- 7
f_o22 <- 7
f_o23 <- 1

# add observed frequencies to calculate row, column, and grand totals
row_1 <- f_o11 + f_o12 + f_o13
row_2 <- f_o21 + f_o22 + f_o23
col_1 <- f_o11 + f_o21
col_2 <- f_o12 + f_o22
col_3 <- f_o13 + f_o23
total <- row_1 + row_2

# calculate expected frequencies for each cell with formula:
# (row total / grand total) * column total
f_e11 <- (row_1 / total) * col_1
f_e12 <- (row_1 / total) * col_2
f_e13 <- (row_1 / total) * col_3
f_e21 <- (row_2 / total) * col_1
f_e22 <- (row_2 / total) * col_2
f_e23 <- (row_2 / total) * col_3

# calculate chi-square for each cell with formula:
# ((observed frequency - expected frequency)**2) / expected frequency
chsq_e11 <- ((f_o11 - f_e11)**2) / f_e11
chsq_e12 <- ((f_o12 - f_e12)**2) / f_e12
chsq_e13 <- ((f_o13 - f_e13)**2) / f_e13
chsq_e21 <- ((f_o21 - f_e21)**2) / f_e21
chsq_e22 <- ((f_o22 - f_e22)**2) / f_e22
chsq_e23 <- ((f_o23 - f_e23)**2) / f_e23

# add chi-square statistics of individual cells to equal the chi-square of the table
chsq_total <- chsq_e11 + chsq_e12 + chsq_e13 + chsq_e21 + chsq_e22 + chsq_e23

# print chi-square
print(chsq_total)
```

The chi-square is 3.791168.

(b) Now calculate the p-value from the test statistic you just created (in R).[2] What do you conclude if $\alpha = .1$?

```
# calculate degrees of freedom with formula:
# (rows-1)(columns - 1)
deg_fre <- (2-1) * (3-1)
print(deg_fre)
# degrees of freedom = 2

# calculate p-value of chi-square statistic
p_chsq_total <- pchisq(chsq_total, df = deg_fre, lower.tail=FALSE)
print(p_chsq_total)
```

The p-value is 0.1502306. We cannot reject the null hypothesis (that class of the drivers and bribe solicitation are statistically independent), because the p-value is greater than our alpha level of 0.1.

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

```
#create matrix
driver_matrix <- matrix(c(14, 6, 7, 7, 7, 1), nrow = 2, byrow = TRUE)

# run chi-square test
dm_chisq <- chisq.test(driver_matrix)

# call for residuals
dm_chisq$residuals
```

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

Note: I also attempted to calculate the residuals with a formula ”by hand,” but appear to have received incorrect results. Perhaps you can comment on what I did incorrectly below.

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

```
# calculate standardized residuals for each cell with formula:
# (observed frequency - expected frequency) / ((expected frequency * (1 - (row total / grand total))
# * (1 - (column total / grand total)))**.5)
z11 <- (f_o11 - f_e11) / ((f_e11 * (1 - (row_1 / total)) * (1 - (col_1 / total)))**.5) # = 0.322
z12 <- (f_o12 - f_e12) / ((f_e12 * (1 - (row_1 / total)) * (1 - (col_2 / total)))**.5) # = -1.642
z13 <- (f_o13 - f_e13) / ((f_e13 * (1 - (row_1 / total)) * (1 - (col_3 / total)))**.5) # = 1.523
z21 <- (f_o21 - f_e21) / ((f_e21 * (1 - (row_2 / total)) * (1 - (col_1 / total)))**.5) # = -0.322
z22 <- (f_o22 - f_e22) / ((f_e22 * (1 - (row_2 / total)) * (1 - (col_2 / total)))**.5) # = 1.642
z23 <- (f_o23 - f_e23) / ((f_e23 * (1 - (row_2 / total)) * (1 - (col_3 / total)))**.5) # = -1.523
```

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

The standardized residuals are useful in explaining where any deviations from independence might occur. This helps us to understand which values have high or low frequencies proportionally to their expected frequencies, and thus which values are contributing to the chi-square size. In this case, the overall chi-square is low, but much of the difference between observed and expected frequencies comes from the ”bribe requested” and ”stopped/given warning” values of the y variable. For these two values, the difference between the observed and expected frequencies are higher, whereas the standardized residuals for ”not stopped” are quite low. This indicates that the driver's class did not impact whether they were stopped by the police by very much (or at all), maybe because the police could not know the driver's supposed class from afar. Meanwhile, the higher residuals for ”bribe requested” and ”stopped/given warning” indicate that the driver's class may affect whether police requested a bribe. They requested bribes more often from the lower class driver, while more often only stoping or giving a warning to the upper class driver. However, overall, the chi-square is not large enough to reject the null hypothesis that the two variables are independent.

## Question 2 (20 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"). 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.

H-null: The reservation policy does not affect the number of new/repaired drinking water facilities in the villages (the two variables are statistically independent).

H-alt: The reservation policy affects the number of new/repaired drinking water facilities in the villages (the two variables are statistically dependent).

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

```r
# run a bivariate regression of y (number of new/repaired drinking water facilities) on x
# (dichotomous presence of reservation policy or not)
women_water <- lm(water ~ reserved, data=women)
print(women_water)
```

The y-intercept is 14.738 and the coefficient for the reserved variable is 9.252.

(c) Interpret the coefficient estimate for reservation policy.

The coefficient estimate of 9.252 indicates that villages with the reservation policy (x = 1) will have 9.252 more new/repaired drinking water facilities than those without the reservation policy (x = 0). Another way to describe this is to say that as x

[3] Chattopadhyay and Duflo. (2004). “Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica.* 72 (5), 1409-1443.

increases by one, y increases by 9.252. However, since x is dichotomous, it can only increase/decrease by one.

## Question 3 (40 points): Biology

There is a physiological cost of reproduction for fruit flies, such that it reduces the lifespan of female fruit flies. Is there a similar cost to male fruit flies? This dataset contains observations from five groups of 25 male fruit flies. The experiment tests if increased reproduction reduces longevity for male fruit flies. The five groups are: males forced to live alone, males assigned to live with one or eight newly pregnant females (non-receptive females), and males assigned to live with one or eight virgin females (interested females). The name of the data set is `fruitfly.csv`.[4]

| | |
|---|---|
| `No` | serial number (1-25) within each group of 25 |
| `type` | Type of experimental assignment<br>1 = no females<br>2 = 1 newly pregnant female<br>3 = 8 newly pregnant females<br>4 = 1 virgin female<br>5 = 8 virgin females |
| `lifespan` | lifespan (days) |
| `thorax` | length of thorax (mm) |
| `sleep` | percentage of each day spent sleeping |

1. Import the data set and obtain summary statistiscs and examine the distribution of the overall lifespan of the fruitflies.

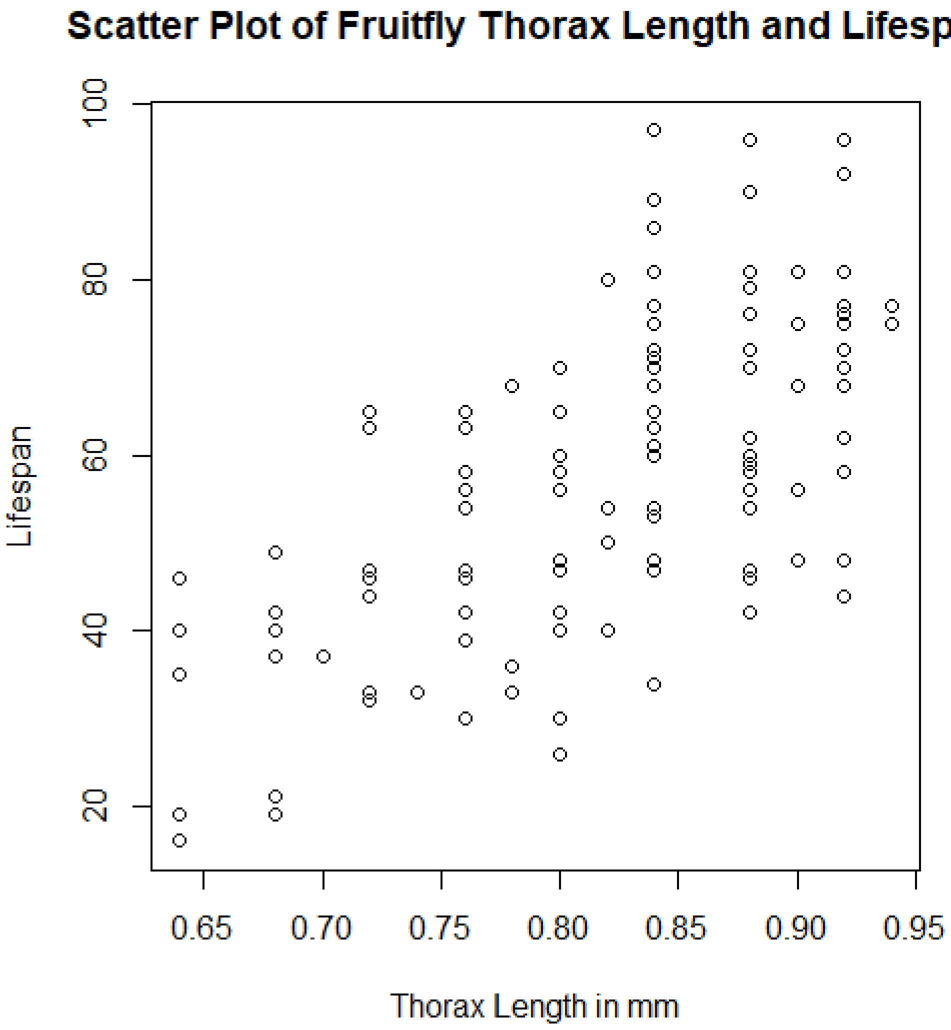
```r
# import data
fruitfly <- read.csv("https://raw.githubusercontent.com/ASDS-TCD/StatsI_Fall2021/main/datasets/fruitfly.csv")

# look at structure, which is data.frame
str(fruitfly)

# print first six observations
head(fruitfly)

# print summary statistics
summary(fruitfly)
```

The data set has 125 observations of five variables. It is a data frame. ”thorax” is composed of numbers and ”lifespan” is composed of integers. The range of the lifespan is from 16-97 days. The first quartile is 46, median is 58, and third quartile is 70. The mean (57.44) is slightly lower than the median.

[4] Partridge and Farquhar (1981).“Sexual Activity and the Lifespan of Male Fruitflies”. *Nature*. 294, 580-581.

2. Plot `lifespan` vs `thorax`. Does it look like there is a linear relationship? Provide the plot. What is the correlation coefficient between these two variables?

```r
# make scatter plot of lifespan vs thorax
plot(fruitfly$thorax, fruitfly$lifespan,  # specify variables
     main = ”Scatter Plot of Fruitfly Thorax Length and Lifespan”,  # main title
     xlab = ”Thorax Length in mm”,  # x axis title
     ylab = ”Lifespan”)  # y axis title
```



The relationship appears relatively linear (and positive), with perhaps a slight upward curviture.

```r
# calculate correlation coefficient
cor(fruitfly$thorax, fruitfly$lifespan)
```

The correlation coefficient between these two variables is 0.6365.

3. Regress `lifespan` on `thorax`. Interpret the slope of the fitted model.

```r
# regress thorax on lifespan
lm(lifespan ~ thorax, data = fruitfly)
```

The intercept coefficient (alpha) is -61.05, and the coefficient (slope) for thorax (beta) is 144.33. This means that for an increase in thorax length of one mm, the fruitfly lifespan increases by 144.33 days. However, since the scales are smaller than this, a better way to explain the relationship is as follows: for an increase in thorax length of .01 mm, the fruitfly lifespan increases by 1.4433 days.

4. Test for a significant linear relationship between `lifespan` and `thorax`. Provide and interpret your results of your test.

```r
# regress thorax on lifespan
ff_reg <- lm(lifespan ~ thorax, data = fruitfly)

# call summary of regression, which includes p-values
summary(ff_reg)
# p-value of thorax coefficient = 1.5e-15
```

The p-value is very near 0 and is statistically significant even with a significance level of less than .001. We can reject the null hypothesis of no relationship between the reservation policy and new/repaired drinking water facilities in the villages.

5. Provide the 90% confidence interval for the slope of the fitted model.

- Use the formula of confidence interval.

```r
# To use the confidence interval formula, first I need to specify the
# coefficient, degrees of freedom, t-score, and standard error.

# make object with thorax coefficient from ff_reg summary
thorax_coef <- 144.33

# calculate degrees of freedom: n - # of estimated coefficients)
deg_fre2 <- 125 - 2
# deg_fre2 = 123

# calculate t-score with 123 degrees of freedom
```

```
t <- qt(.95, df = deg_fre2)
# t = 1.657336

# use se from ff_reg summary
st_er <- 15.77

# calculate 90% confidence interval with formula:
# coefficient + or - (t-score * standard error)
CI_90_lower <- thorax_coef - (t * st_er)
CI_90_upper <- thorax_coef + (t * st_er)
CI_90 <- c(CI_90_lower, CI_90_upper)
print(CI_90)
```

The confidence interval according to this formula is [118.1938, 170.4662].

- Use the function `confint()` in `R` .

```
# calculate 90% confidence interval with confint()
confint(ff_reg, level = .90)
```

The confidence interval according to the confint() function is [118.1962, 170.4700].

6. Use the `predict()` function in `R` to (1) predict an individual fruitfly's lifespan when `thorax`=0.8 and (2) the average `lifespan` of fruitflies when `thorax`=0.8 by the fitted model. This requires that you compute prediction and confidence intervals. What are the expected values of lifespan? What are the prediction and confidence intervals around the expected values?

```
# set x value at 0.8
new_DF <- data.frame(thorax = 0.8)

# predict individual fruitfly's lifespan when thorax = 0.8
ind_response <- predict(ff_reg, newdata = new_DF, interval = "prediction"
    , level = 0.95)
print(ind_response)
```

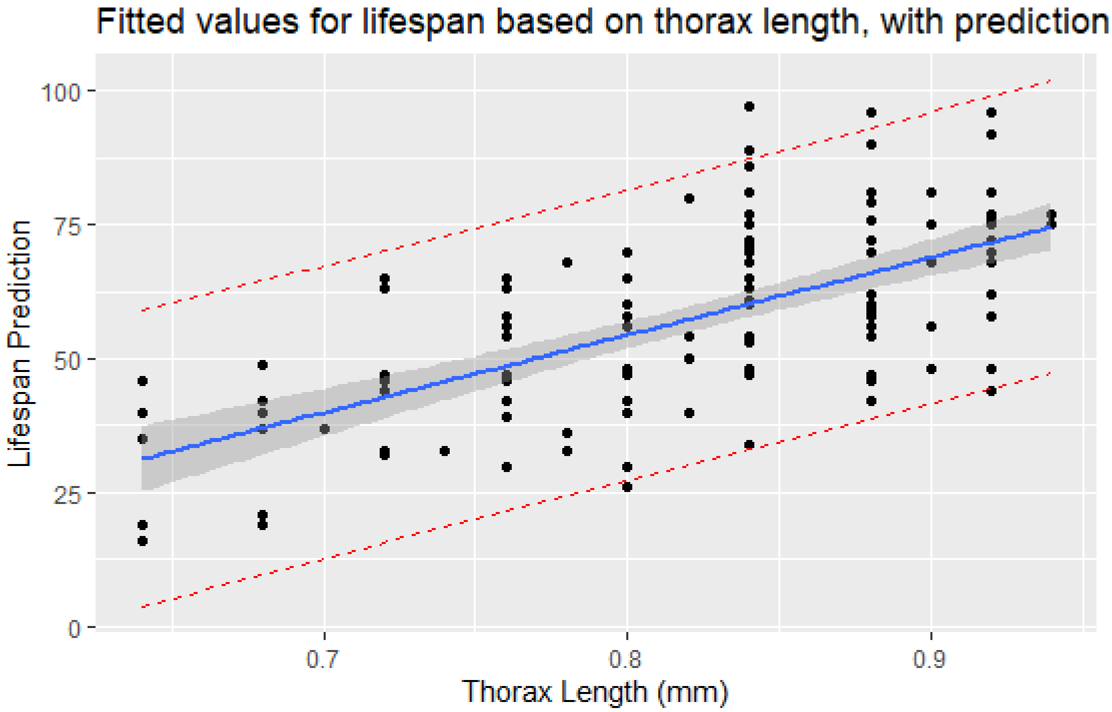
This prediction is 54.41478 days, with a prediction interval of [27.37542, 81.45414].

```
# predict average fruitfly's lifespan when thorax = 0.8
avg_response <- predict(ff_reg, newdata = new_DF, interval = "confidence"
    , level = 0.95)
print(avg_response)
```

This prediction is also 54.41478 days, with a much smaller interval of [51.91932, 56.91024].

7. For a sequence of `thorax` values, draw a plot with their fitted values for `lifespan`, as well as the prediction intervals and confidence intervals.

```r
# create linear model
model <- lm(lifespan ~ thorax, data = fruitfly)

# create predictions
pred <- predict(model, interval = "prediction")

# specify data
mydata <- cbind(fruitfly2, pred)

# create regression line
p <- ggplot(mydata, aes(thorax, lifespan)) +
  geom_point() +
  ggtitle("Fitted values for lifespan based on thorax length, with prediction and confidence intervals") +
  labs(y="Lifespan Prediction", x = "Thorax Length (mm)") +
  stat_smooth(method = lm) # add confidence interval

# create prediction intervals
p + geom_line(aes(y = lwr), colour = "red", linetype = "dashed")+
  geom_line(aes(y = upr), colour = "red", linetype = "dashed")

# Source for graph: http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/
```



The plot above shows the fitted values for lifespan as predicted by thorax length (blue line). It also has the confidence intervals (gray shading) and prediction intervals (dotted red lines). Just to provide additional information, I kept the scatterplot points as well.