1 PS02 Answers

Aryan Goyal
Student number: 18306046

2 Question 1

2.1 Answer 1a)

First,I check the assumptions that we have two categorical variables and if the expected frequency is greater than/equal to 5 in all cells. In this case, the expected frequency is not greater than 5 but we do the chi-square test anyways
The null hypothesis is that the variables of class and police behavior are statistically independent
The alternative hypothesis is that the variables of class and police behavior are statistically dependent

In order to calculate the chi-square test statistic "by hand" in R:
I first create a table with the given information on our variables:

```
Not_Stopped <- c(14,7)
Bribe_Requested <- c(6,7)
Stopped_Given_Warning <- c(7,1)

PolSci <- data.frame(
Not_Stopped,Bribe_Requested, Stopped_Given_Warning,
row.names = c('Upper Class','Lower class')
)
str(PolSci)
```

Table 1: Observed values

	Not_Stopped	Bribe_Requested	Stopped_Given_Warning
Upper Class	14	6	7
Lower class	7	7	1

Next, I calculate the expected frequency for each row and column
The formula to calculate expected frequencies:
((Row total)*(Column Total))/Total sample size

```
(27*21)/42 # = 13.50
(15*21)/42 # = 7.50
(27*13)/42 # = 8.36
(15*13)/42 # = 4.64
```

```
(27*8)/42 # = 5.14
(15*8)/42 # = 2.86
```

Table 2: Expected values

	Not_Stopped	Bribe_Requested	Stopped_Given_Warning
Upper Class	13.500	8.357	5.143
Lower class	7.500	4.643	2.857

Now, I have all the required values to calculate the chi-square test statistic value.

```
The formula is ((observed frequency-expected frequency)^2)/expected frequency

ChiSquareStatistic <- (((14-13.5)^2)/13.5) + (((6-8.36)^2)/8.36) +
(((7-5.14)^2)/5.14) +  (((7-7.5)^2)/7.5) +
(((7-4.64)^2)/4.64) + (((1-2.86)^2)/2.86)
```

From this, I got the chi-square test statistic value = 3.8
I double check our "by hand" chi square test using the chisq.test function in R:

```
    chisq.test(PolSci)

chi_testR <- chisq.test(PolSci) #Saving chi-square test in an object
```

Through this, I get chi squared value = 3.79

2.2 Answer 1b)

Next, I calculate the p-value from the test statistic. I use the pchisq function in R:

```
degrees of freedom=(rows-1)*(columns-1)
Hence, df=(2-1)(3-1)=1*2=2
pvalue <- pchisq(3.80,df=2, lower.tail=FALSE)
```

From this, I find that the p-value = 0.15 In this case, the p-value is 0.15 and above the alpha level= 0.1 and hence, we do not have enough evidence to reject the null hypothesis that the variables of class and police behavior are statistically independent.

If the p-value was below the alpha level = 0.1, we could conclude that we have enough evidence to reject the null hypothesis, and hence have support for the alternative hypothesis that the variables of class and police behavior are statistically dependent

2.3 Answer 1c)

Finally, to calculate the standardized residuals for each cell:
The formula for each cell is (observed frequency - expected frequency)/se, where standard error= sqrt(expected frequency(1-row proportion)(1-column proportion))

First, I do this by hand:

```
    ##For Cell 1:
numerator1 <- 14-13.5 #Numerator
denominator1 <- sqrt((13.5*(1-0.64)*(1-0.5)))
cell1 <- numerator1/denominator1

##For Cell2:
numerator2 <- 6-8.36
denominator2 <- sqrt((8.36*(1-0.64)*(1-0.31)))
cell2 <- numerator2/denominator2
##For Cell3:
numerator3 <- 7-5.14
denominator3 <- sqrt((5.14*(1-0.64)*(1-0.19)))
cell3 <- numerator3/denominator3
##For Cell4:
numerator4 <- 7-7.5
denominator4 <- sqrt((7.5*(1-0.36)*(1-0.5)))
cell4 <- numerator4/denominator4
##For Cell5:
numerator5 <- 7-4.64
denominator5 <- sqrt((4.64*(1-0.36)*(1-0.31)))
cell5 <- numerator5/denominator5
##For Cell6:
numerator6 <- 1-2.86
denominator6 <- sqrt((2.86*(1-0.36)*(1-0.19)))
cell6 <- numerator6/denominator6
```

To check these values, I extract the standardized residuals that R's chisq.test function calculates:

```
    ls(chi_testR)
##storing the standardized residuals in an object
StdRes <- chi_testR$stdres

#Using the stargazer function, to get code for
making a table in Latex
stargazer(StdRes)
```

From the calculated standardized residuals, I can make this table:

Table 3: Standardized Residuals

	Not_Stopped	Bribe_Requested	Stopped_Given_Warning
Upper Class	0.322	-1.642	1.523
Lower class	-0.322	1.642	-1.523

2.4 Answer 1d)

Standardized residuals are useful to describe the pattern of association among the cells. A large standardized residual (greater than 2 in absolute value) provides evidence against independence in that cell. Since none of the the standardized residuals exceed 2 in absolute value, the cells do not have more observations than we would expect if the variables were truly independent.

The value for Bribe requested for upper class individuals is -1.642 which suggests that this was less what the hypothesis of independence predicts, however not enough to be statistically significant.

Similarly, for upper class individuals who were stopped/given warning, a value of 1.523 suggests that it was more than the hypothesis of independence predicts, but not enough to be statistically significant.

3 Question 2

First I input the dataset for this question:

```
Economics <- read.csv("https://raw.githubusercontent.com/
kosukeimai/qss/master/PREDICTION/women.csv")
```

3.1 Answer 2a)

The null hypothesis (H0): There is no significant relationship between the reservation policy and the number of new or repaired drinking water facilities in the village.
The alternative hypothesis (Ha): There is a significant relationship between the reservation policy and the number of new or repaired drinking water facilities in the village.

3.2 Answer 2b)

I conduct a bivariate regression in R using the "lm" function:

```
reg <- lm(Economics$water ~ Economics$reserved, Economics)
reg
summary(reg)
```

This is the table showing the results

Table 4: Bivariate Regression

	Dependent variable:
	water
Constant	14.738***
	(2.286)
reserved	9.252**
	(3.948)
Observations	322
R^2	0.017
Adjusted R^2	0.014
Residual Std. Error	33.446 (df = 320)
F Statistic	5.493** (df = 1; 320)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
# using the confint function to get confidence intervals
for our estimates
confint(reg)
```

Table 5: Confidence intervals of y-intercept and slope

	2.5 %	97.5 %
(Intercept)	10.240	19.236
reserved	1.486	17.019

3.3 Answer 2c)

Finally, I interpret the coefficient estimate for reservation policy:

The value of the slope suggests that villages with a female representative have 9.25 more new/repaired drinking water facilities on average with a 95 percent confidence interval of [1.49, 17.02]. The confidence interval suggests that with a female representative, the number of new/drinking water facilities increases by as few as 1.49 or as many as 17.02. At alpha level = 0.05, we can reject the null hypothesis and conclude that the slope is statistically differentiable from 0.

In addition, the y-intercept suggests that having a non-female representative means that the number of drinking water facilities are 14.738 in the village with a 95 percent confidence interval of [10.240, 19.236]. At alpha level = 0.01, we can reject the null hypothesis that the y-intercept = 0.