PS02 Assignment

Philip Kruger
18328699
11/10/2022

Question 1

Part (a)

First the values from the table provided were entered into R in the form of a matrix x. Following this the expected value of each entry in the matrix is calculated using the following the following r code:

```
x_e <- matrix(data = NA,ncol = 3, nrow =2) #creating empty matrix to put the expected values in
#x_e
for(i in 1:nrow(x)){
    for(j in 1:ncol(x)){
        x_e[i,j] <- sum(x[i,])*sum(x[,j])/sum(x) #going through all the slots in the matrix and
        #putting in the expected value
    }}
```

which yields the following table of values:

Class	Not Stopped	Bribe Requested	Stopped/given warning
Upper Class	13.5	8.36	5.14
Lower Class	7.5	4.64	2.85

Following this, another matrix was created containing the contribution of each term to the χ^2 test statistic, this was done with the following code and the table is as follows:

```
x_chi <- matrix(data = NA,ncol = 3, nrow =2) #creating empty matrix to put the chi squared values in
for(i in 1:nrow(x_chi)){
    for(j in 1:ncol(x_chi)){
        x_chi[i,j] <- (x[i,j]-x_e[i,j])**2/x_e[i,j] #going through all the slots in the matrix and
        #putting in the chi squared contribution of that slot
    }}
```

Class	Not Stopped	Bribe Requested	Stopped/given warning
Upper Class	0.02	0.66	0.67
Lower Class	0.03	1.2	1.2

The terms in this table were then summed with the following code:

```
chi_squared <- sum(x_chi) #summing terms in x_chi to get chi squared term
chi_squared
```

Leading to the following result:

$$\chi^2 = 3.791168$$

Part (b)

To calculate the p-value we can simply apply the code from lecture 3:

```
pchisq(chi_squared, df = (ncol(x)-1)*(nrow(x)-1),  lower.tail=FALSE) #gets p-value from chi
#squared test statistic
```

This yields the result:

$$p - value = 0.1502306$$

for $\alpha = 0.1$ we have insufficient evidence to exclude that the variables are statistically independent. Since the p-value is greater than our α.

Part (c)

The standardized residuals of each entry was gotten using the following code:

```
x_sr <-  matrix(data = NA,ncol = 3, nrow =2) # creating empty matrix to put the standardized
#residuals in
for(i in 1:nrow(x_sr)){
    for(j in 1:ncol(x_sr)){
        x_sr[i,j] <- x[i,j]-x_e[i,j]/(x_e[i,j]*(1-sum(x[i,])/sum(x))*(1-sum(x[,j])/sum(x))) #going
        #through all the slots in the matrix and putting in the expected value
    }}
```

This resulted in the following values:

Class	Not Stopped	Bribe Requested	Stopped/given warning
Upper Class	0.32	-1.64	1.52
Lower Class	-0.32	1.64	-1.52

Part (d)

A standardized residual is the raw residual divided by an estimate of the standard deviation of the residuals. So it shows how far away each point is from the prediction line relative to how far away other points are from the predictive line. From the fact that the magnitude of the numbers in each column of the standardized residual table are the same, we know that the predictive line goes right through the average of the two terms in the each column. By the magnitude of the number we can also know how different our observation are from the prediction.

For Not Stopped the predictive line is very close to the observations, it has a magnitude of 0.32 which means the points are 0.32 standard deviations away from the predictive line for that group. The other two categories have a standardized residual magnitude of 1.64 and 1.52 respectively. This means that they fall reasonably far away from the expected frequency however they would still fall within 2 standard deviations. From this we can interpret that for Bribe Requested and Stopped/Given warning, there was a noticeable difference in frequency and predictive frequency, however probably not enough to be statistically significant.

Question 2
Part (a)

H_0 = Having a Reservation policy requiring a female head of GP is not correlated with the number of new or repaired drinking water facilities

H_a = Having a Reservation policy requiring a female head of GP is correlated with the number of new or repaired drinking water facilities

Part (b)

When tackling this bivariate regression there is a major decision to be made with respect to wrangling the data. We care about the effect having a female reservation policy on number of water projects in a village. However, a key assumption is that observations are independent. However, with the data we have, they are clearly not independent, since villages in the same Gram Panchayat (GP) have the same leader. Thus the degrees of freedom would be more accurately represented with the following adjustments:
We first need to change the two village entries into one with the entries in the water column added together. Looking at the data we can see that the GP number is increasing as you go down the dataset, furthermore all even numbered columns have the village number 1. To test that this is the case for the whole database the following 2 functions are used:(its not really necessary for a dataset with around 300 values as you can just eyeball it to see the formatting is all identical, however this makes the code generalizable for larger datasets).

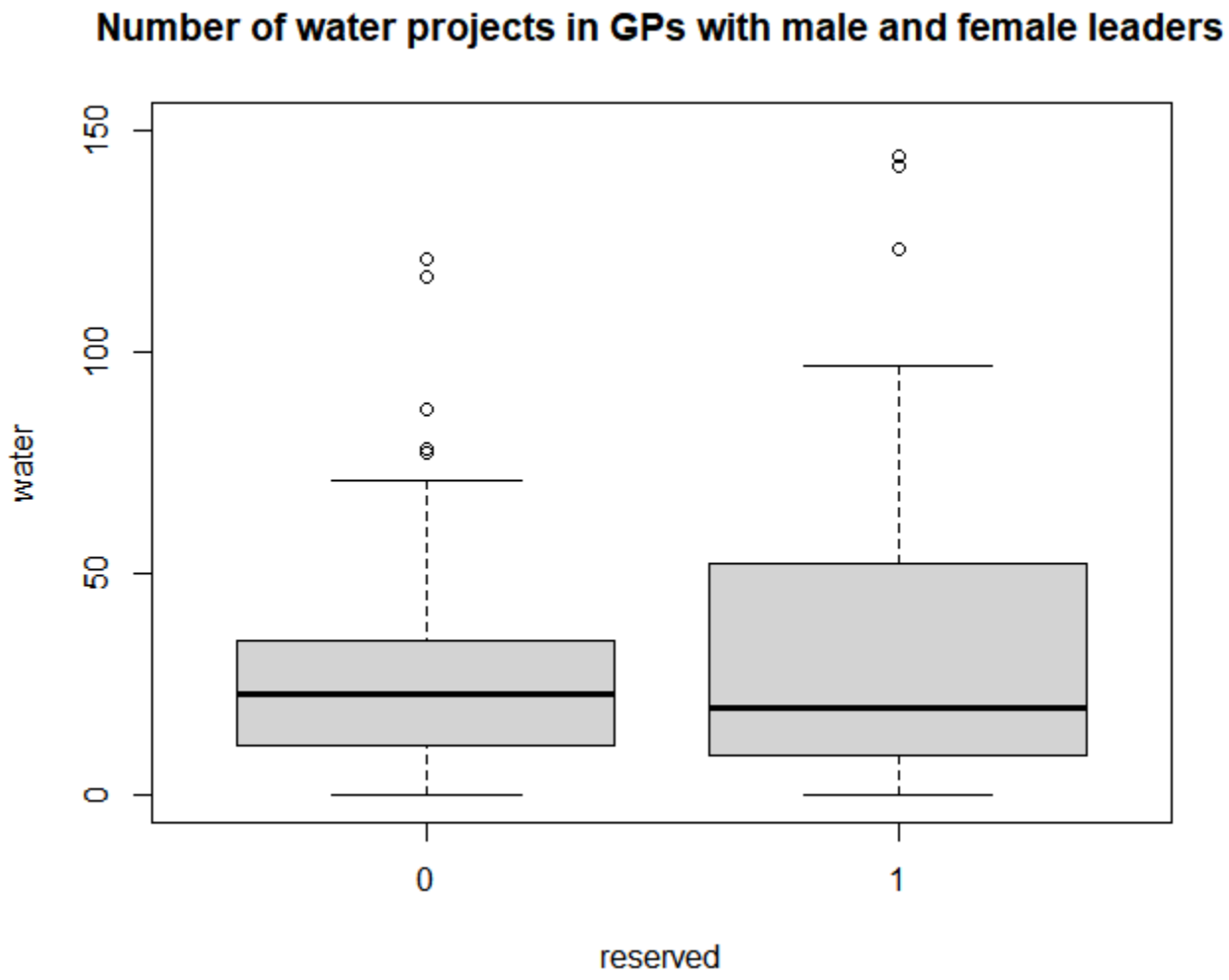
```
#This function checks that the GP number is always increasing down the column.
#i.e. that no entries are out of order.
increase.function <- function(a){
    for(i in 2:length(a[,1])){
        if(a[i,1] < a[(i-1),1]){print("Error in line ")
            print(i)}
        else{}
}}

#This function checks that the second entry of each GP is in an even numbered column
two_gp.function <- function(a){
    for(i in 2:length(a[,1])){
        if(i %% 2 == 0 & a[(i-1),1] != a[i,1]){print("Error in line ")
            print(i)}
        else{}
}}
```

when these functions are called there are no prints in the console thus the structure for the whole dataset is the same as the first few lines. Therefore, the following two functions are effective in summing the water projects from the same GP, entering that value into the first row of that GP and deleting the second row from each GP.

```
#this function is used sum the water projects from each GP and entering this value into the
#first village from that GP.
total_water.function <- function(a){
    for(i in 1:length(a[,6])){
        if(i %% 2 == 1){
            a[i,6] <- a[i,6]+a[(i+1),6]}
        else{}
    }
    return(a)
}
```

```
# This line deletes the second village row which does not contain the total water
#projects for this GP.
remove_even.function <- function(a){
    remove_vector <- c() #no idea why i have to do it this way, removing lines in the if
    #loop gets the incorrect answer
    for(i in 1:length(a[,6])){
        if(i %% 2 == 0){
            remove_vector <- append(remove_vector, i)
        }
        else{}
    }
    a <- a[-c(remove_vector),]
    return(a)
}
```

Now we have two functions which edit the data such that each row represents one leader where the corresponding entry in the water column is the total number of new and repaired drinking water projects in the two village they are in charge of. I then graphed this data in a boxplot to give an idea what the data looks like:



**some higher outliers in the female group are excluded as to make the graph clearer*

From this plot we can see the mean number of projects in each GP seems similar GPs with and without the reservation policy. However, having a reservation results in a larger standard deviation and the data is skewed left.(which has a longer tail to the right).

The following code (adapted from lecture 4) is then used to find the correlation co-efficient and to test our null hypothesis:

```
cor(remove_even.function(total_water.function(dat))[,3],remove_even.function(
total_water.function(dat))[,6], method = "pearson") #gets the correlation coefficient
#between female and water

cor.test(remove_even.function(total_water.function(dat))[,3],remove_even.function(
total_water.function(dat))[,6]) #does the hypothesis test that female and water are correlated
```

This yields a correlation coefficient of 0.1785658 . Thus there is a very weak positive correlation. The p-value of 0.02343 is also obtained, this passes under the $\alpha = 0.05$ requirement. Thus we can exclude the null hypothesis that the correlation is zero. We also get the 95 percent confidence intervals:

Lower Interval	0.0246
Upper Interval	0.3243

These do not cross the null, which verifies our earlier exclusion of the null hypothesis.

If the whole dataset is taken without concern for if the observations are independent we get:

```
t = 2.3437, df = 320, p-value = 0.0197
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.02090616 0.23585751
sample estimates:
cor
0.1299079
```

So all conclusions from the independent dataset apply.

Part (c)

Using the following code:

```
cor(remove_even.function(total_water.function(dat))[,3],remove_even.function(
total_water.function(dat))[,6], method = "pearson") #gets correlation coefficent
#between reservation policy and water
```

This yielded a correlation coefficient of 0.1785658. Thus, there is a very weak positive correlation. However due to the degrees of freedom this is statistically significant. This indicates that there is a positive correlation between reservation for women and the building/repair of drinking water projects. i.e. more places reserved for women = more drinking water projects.