Claire Ott
20302999
Applied Statistics I
Problem set 2

Question 1

a) Calculating chi-squared in R

```r
## Question 1
## part a
## creating a matrix for the data
crossroads <- matrix(c(14,6,7,7,7,1), nrow=2, byrow=TRUE)
## getting expected frequencies from row/column/grand totals
row_totals <- rowSums(crossroads)
col_totals <- colSums(crossroads)
grand_total <- sum(crossroads)
expectedf <- outer(row_totals, col_totals)/grand_total
print(expectedf)
## calculate chi-squared
chi_squared <- sum((crossroads-expectedf)^2/expectedf)
print(chi_squared)
```

i) Chi squared value was found to be 3.791168

ii) Tried calculating by hand too just to be sure:

iii)

Problem Set 2

→ from Week 3 slides

1. (a) calculating χ^2 statistic

	Not stopped	Bribe	stopped	
upper	14	6	7	[27]
lower	7	7	1	[15]
	[21]	[13]	[8]	

total observances = 42

calculate expected frequencies:

(upper, Not stopped): $\frac{27\times 21}{42} = 13.5 = E$

(lower, Not stopped): $\frac{15\times 21}{42} = 7.5 = E$

(upper, Bribe): $\frac{27\times 13}{42} = 8.357 = E$

(lower, Bribe): $\frac{15\times 13}{42} = 4.643 = E$

(upper, Stopped): $\frac{27\times 8}{42} = 5.143 = E$

(lower, stopped): $\frac{15\times 8}{42} = 2.857 = E$

$$\chi^2 = \frac{(14-13.5)^2}{13.5} + \frac{(7-7.5)^2}{7.5} + \frac{(\cancel{8.357}\ 6-8.357)^2}{8.357} + \frac{(7-4.643)^2}{4.643} + \frac{(7-5.143)^2}{5.143} + \frac{(1-2.857)^2}{2.857}$$

$$\chi^2 = \frac{0.25}{13.5} + \frac{0.25}{7.5} + \frac{5.556}{8.357} + \frac{5.555}{4.643} + \frac{3.448}{5.143} + \frac{3.448}{2.857}$$

$$= 0.0185 + 0.0333 + 0.665 + 1.196 + 0.670 + 1.207 = 3.7898 = \chi^2$$

b) Calculating the p-value in R

i)

```r
## part b
## getting degrees of freedom
df <- (nrow(crossroads)-1)*(ncol(crossroads)-1)
print(df)
## getting p-value
p_value <- 1-pchisq(chi_squared,df)
print(p_value)
```

ii) The degrees of freedom calculated = 2 and the p-value = 0.1502306

iii) Because in this case the p-value is greater than the significance level, a=0.1, we fail to reject the null hypothesis as there isn't sufficient evidence to conclude that there's a statistically significant association between socioeconomic status and bribe solicitation.

c) Calculating standardised residuals in R

```
## part c
## getting standardized residuals
standard_res <- (crossroads-expectedf)/sqrt(expectedf)
print(standard_res)
```

i) I

ii)

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

d) A standardised residual of approximately 0 would suggest that the observed and expected frequencies are similar whereas residuals of greater than 2 or less than -2 would indicate some significant deviation from the observed/expected values, meaning that that specific element would contribute more to the chi-squared value calculated. However, as there are no residuals greater than 2 or less than -2, in this case I think it is reasonable to expect that the values further from 0, such as 1.09 (bribe requested, lower class) and -1.09 (stopped, lower class) contributed a bit more to the chi-squared values.

Question 2

a) Stating the hypotheses:

i) H0 (null hypothesis): Having a female leader has no effect on the number of new/repaired drinking water facilities in a village.

ii) H1 (alternative hypothesis): Having a female leader leads to a greater number of new/repaired drinking water facilities in a village.

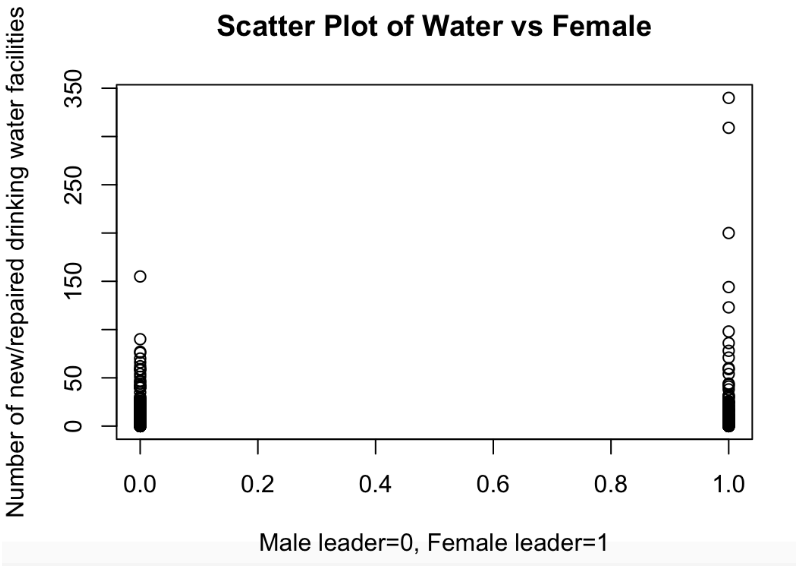
b) Running a bivariate regression in R

i) The two variables I want to investigate are “female” (which represents whether or not a village has a woman leader) and “water” (represents the number of new/repaired drinking water facilities in a village) however because “female” is a binary variable, I found the regression a bit more complicated at first.

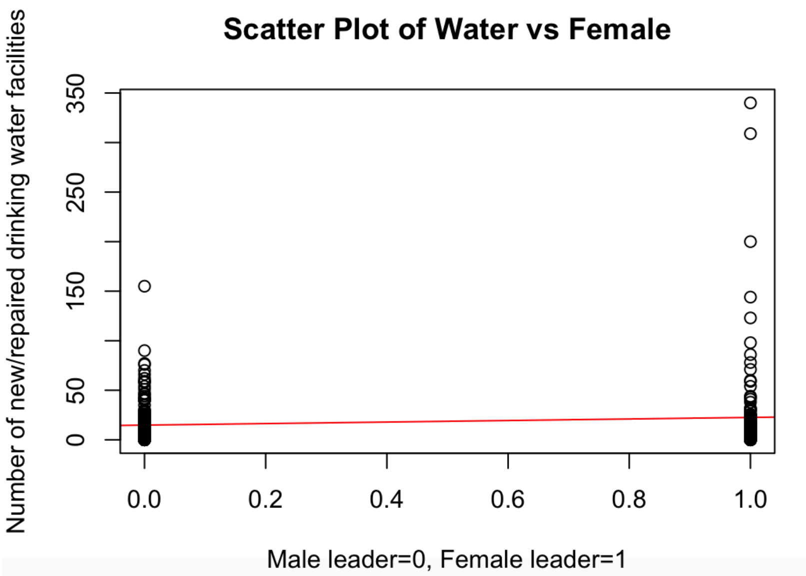
```
## trying to do a regression
water_female_model <- lm(women$water~women$female, data=women)
water_female_model
summary(water_female_model)
print(water_female_model$coefficients)
```

ii)

iii) I started by plotting the two variables:



iv) Then added in the regression line (red):



v)

c) It does seem like in general, the number of new/repaired drinking water facilities is higher when there is a female leader. However, I want to confirm this so I looked at the model summary in R:

```
Call:
lm(formula = women$water ~ women$female, data = women)

Residuals:
   Min     1Q Median     3Q    Max 
-22.68 -14.78  -7.81   2.29 317.32 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)    14.813      2.382   6.220 1.56e-09 ***
women$female    7.864      3.838   2.049   0.0413 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.51 on 320 degrees of freedom
Multiple R-squared:  0.01295,	Adjusted R-squared:  0.009867 
F-statistic: 4.199 on 1 and 320 DF,  p-value: 0.04126
```

i)

ii) When female=0, or there is a male leader, the average number of new/repaired drinking water facilities is 14.813 (the intercept), and when female=1, there’s a female leader, the average number of new/repaired drinking water facilities is 7.864 (women$female coefficient) higher than if the leader was male. Since the p-value is 0.04126, these results are significant for a=0.05, thus; we can reject the null hypothesis and conclude that having a female leader leads to a higher number of new/repaired drinking water facilities.