Problem Set 2

Ana Sofia Jesus 19327602

15/10/2023

Question 1

a)

To begin to answer the first question, I first inserted the contingency table with the data into R and printed it:

```
data_matrix <- matrix(c(14, 6, 7, 7, 7, 1),
              nrow = 2,
              byrow = TRUE,
              dimnames = list(c("Upper_class", "Lower_class"),
                              c("Not Stopped", "Bribe requested", "Stopped/given warning")))
data_matrix
```

I have obtained the following table:

Class	Not Stopped	Bribe Requested	Stopped/given warning
Upper Class	14	6	7
Lower Class	7	7	1

Next I moved on to calculating the expected frequency, based on the formula displayed during lectures:

$$\left(\frac{Row\,total}{Grand\,total}\right) \times Column\,total$$

The expected frequency is the count expected in a cell if the variables were independent
From left to right, here are my calculations by hand:

```
fe1=(27/42)*21
fe2=(27/42)*13
fe3=(27/42)*8
fe4=(15/42)*21
fe5=(15/42)*13
fe6=(15/42)*8
```

I have inserted the results, the expected frequencies, in the following table:

Class	Not Stopped	Bribe Requested	Stopped/given warning
Upper Class	13.5	8.36	5.14
Lower Class	7.5	4.64	2.86

Before calculating the chi-squared statistic, I have formulated my hypotheses:
H0: The variables social class and bribe solicitation are statistically independent.
H1: The variables social class and bribe solicitation are statistically dependent.

The Chi-Squared Test of independence makes the following assumptions:

- The two variables are categorical

- The sampling used is random

- Expected frequency exceeds 5 in each cell

The test statistic for independence summarises how close the expected frequencies fall to the observed frequencies. To calculate the chi-squared statistic I used the following formula:

$$\chi^2 = \sum \frac{(fo - fe)^2}{fe}$$

Here is a step-by-step demonstration of my calculations by hand, in R:

```
chi_squared= (14-13.5)^2/13.5 + (6-8.357143)^2/8.357143+(7-5.142857)^2/5.142857+
  +(7-7.5)^2/7.5+(7-4.642857)^2/4.642857+(1-2.857143)^2/2.857143

chi_squared=3.791169
```

I have also used the R function learned in tutorials to confirm this result:

```
chi_square_result <- chisq.test(data_matrix)

chi_square_statistic <- 3.7912
```

b)

Since the chi-square statistic alone does not allow to conclude if the difference between observed and expected frequencies is statistical significant, it is necessary to calculate the p-value, in order to determine the probability of obtaining this value for the chi-square statistic if H0 was true.
The significance level (alpha) is set: alpha - 0.1 The p-value is calculated from the chi-squared test statistic using the formula from lectures:

```
p_value <- 1 - pchisq(chi_square_statistic, df = 2, lower.tail=FALSE)
```

P-value: 0.1502282 is greater than alpha (0.1) thus, it is not possible to reject the null-hypothesis at alpha 0.1. There is not sufficient evidence to conclude that there is a relationship between social class and bribe solicitation from the above sample.
It is worth noting that the sample size for this experience is extremely small. The chi-squared test requires a large sample, the expected frequency should exceed 5 in each cell. Otherwise, the chi-squared distribution can only poorly approximate the actual distribution of the chi-squared statistic.

c)

The standardized residuals provide information regarding how far away is each observed value from the calculated expectation. I have used the formula provided in lectures:

$$z = \frac{f_{observed} - f_{expected}}{\sqrt{f_{expected} \cdot (1 - rowprop.) \cdot (1 - columnprop.)}}$$

Here are my calculations by hand:

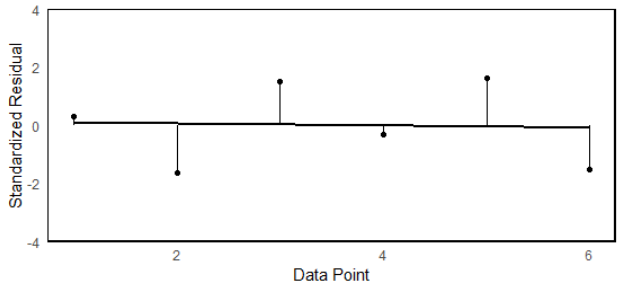
```
res1 - (14 - 13.5) / sqrt(13.5 * (1 - 27/42) * (1 - 21/42))
res2 - (6-8.357143)/ sqrt(8.357143 * (1-27/42)*(1-13/42))
res3 - (7-5.142857)/ sqrt(5.142857 * (1-27/42)*(1-8/42))
res4 - (7-7.5)/ sqrt(7.5 * (1-15/42)*(1-21/42))
res5 - (7-4.642857)/ sqrt(4.642857 * (1-15/42)*(1-13/42))
res6 - (1-2.857143)/ sqrt(2.857143 * (1-15/42)*(1-8/42))
```

The results from the above calculations are displayed in the following table:

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

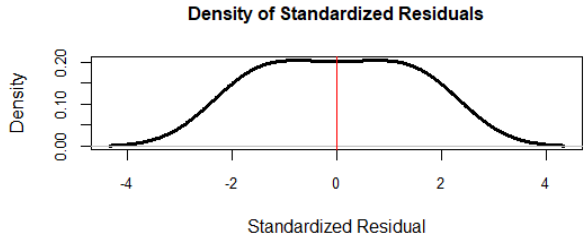
These residuals are also displayed in two graphs:
1) A scatterplot of residuals with a regression line using the following code:

```
ggplot(residuals_data, aes(x = seq_along(Residual), y = Residual)) +
geom_point() +
geom_smooth(method = "lm", se = FALSE, color = "black") +
geom_segment(aes(xend = seq_along(Residual), yend = mean(Residual)), linetype = "solid") +
labs(x = "Data Point", y = "Standardized Residual") +
ylim(min(residuals_data$Residual) - 2, max(residuals_data$Residual) + 2) +
theme_minimal() +
theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank(), panel.border
= element_rect(colour = "black", fill = NA, size = 1))
```



2) A density plot of the standardized residuals using the following code:

```
std_residuals <- c(0.3220306, -1.641957, 1.523026, -0.3220306, 1.641957, -1.523026)
plot(
density(std_residuals),
main = "Density of Standardized Residuals",
ylab = "Density",
xlab = "Standardized Residual",
cex.axis = 0.8,
cex.lab = 1,
cex.main = 1,
lwd = 3
  )
abline(v = 0, col = "red")
```



d)

The standardized residuals show how far away each point is from the predictive line. The further they are from the line of prediction the closer we are from being able to reject the null hypothesis, since this prediction is calculated according to the assumption of an absence of relation between the variables. A cell-by-cell comparison of observed and expected frequencies reveals the specifics of the association between the variables.

By visually inspecting the cells in the table, there are no standardized residuals values below -3 or above +3, thus variation could be random (Agresti and Finlay, 2009). There are no convincing evidence of a true effect in any of the cells. A large standardized residual provides evidence against independence. The observations closer to the predictions are those in the "Not Stopped" treatment, it has a magnitude of 0.32 which means the observations are 0.32 standard deviations away from the predictive line for that group. The other two treatments (Bribe Requested and Stopped/Given warning) 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, which probably is not enough to be statistically significant.

Question 2

a)

To assess the effect of the reservation policy on the number of new or repaired drinking water facilities in the villages, I stated the null and alternative hypotheses as follows:

- Null Hypothesis (H0): The reservation policy of a GP for women leaders has no effect on the number of new or repaired drinking water facilities in the villages.

- Alternative Hypothesis (H1): The reservation policy of a GP for women leaders has an effect on the number of new or repaired drinking water facilities in the villages. This is expressed as a two-tailed hypothesis: since there may be either an increase or a decrease in the association.

b)

Before using the variable representing the reservation of seats for women, I will verify that the reservation policy was followed, if villages did not obey to this policy I would not be estimating the intended effect of having women politicians on policy outcomes. To calculate the proportion of female politicians in reserved GP seats vs. unreserved seats I used the following code:

```
proportion_reserved <- mean(data$female[data$reserved == 1])
proportion_unreserved <- mean(data$female[data$reserved == 0])
```

I have obtained the following results:

Proportion of female politicians in reserved GP seats: 1

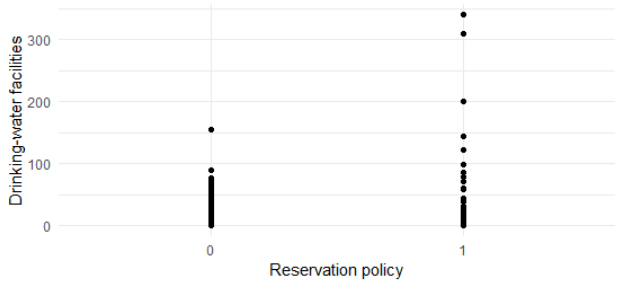
Proportion of female politicians in unreserved GP seats: 0.07476636

These results allow me to proceed with my analysis since the variable "reserve" represents well the intended factor, in other words, the reservation policy was followed in every GP.

I started by creating a plot to visualise the data using the following code:

```
ggplot(data, aes(x = factor(reserved), y = water)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE) +
  labs(x = "Reservation policy", y = "Drinking-water facilities") +
  theme_minimal()
```

From this I have obtained the following graph:



The regression equation for this model is:

$$Y = \alpha + \beta X$$

Y= Drinking water facilities (continuous variable)
X= Reservation policy (binary variable)

The remaining values will be known after the statistic.
The linear regression model assumes that the relationship between my response variable - drinking water facilities (x) and the mean of my predictor - reservation policy (y) follows a straight-line relationship. It also assumes the data is randomized, that observations are independent and that the population is normally distributed and has a constant variance. This model also assumes that the error is normally distributed and that there is no bias; no auto correlation are that X values are measured without error. It assumes that no causal variables were left out and that no non-causal variables were included.

Next, I ran a bivariate regression model to assess the relationship between the two variables: "water" (the number of new or repaired drinking water facilities) and "female" (a binary variable indicating whether the council head is female or not). I used the following code drawn from the lectures:

```
model <- lm(water ~ reserved, data = data)
summary(model)
```

I have obtained the following results:

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

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
```

$$\beta = 9.252$$

$$\alpha = 14.738$$

In summary, this regression analysis indicates that the GP reservation policy is statistically significant in explaining the number of new or repaired drinking water facilities. When the GP is reserved for females, there is on average an estimated increase of 9.252 units of new or repaired drinking water facilities, compared to when the council head is male. The relationship is significant but explains only a small portion of the variance in the outcome variable since there is low R-squared of 0.01688. Residual Standard Error: The residual standard error is 33.45, which indicates the typical error in predicting the number of drinking water facilities based on the model. Smaller values would indicate a better fit.

c)

The coefficient estimate for "reserved" is 9.252. The "reserved" coefficient (9.252) represents the estimated change in the output variable "water" (the number of new or repaired drinking water facilities) when a GP is reserved for women leaders compared to when it is not reserved. In this context, it means that reserving the GP is associated with an estimated increase of on average 9.252 units of new or repaired drinking water facilities compared to when GP is not reserved for women. The coefficient is positive, which suggests that the reservation policy is associated with more drinking water facilities. This coefficient is statistically significant, as indicated by the asterisk (*), with a p-value of 0.0197. This means that the relationship between reservation policies and the number of drinking water facilities is unlikely to be due to chance.