# Assignment 4

*Hongyu He (2632195) & Bruno Hoevelaken (2645065)*

*Group CS 6*

## & Theoratical exercises

### Exercise 4.1 (Section 9.2, E6)

**1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: $\rho = 0$;
- $H_a$: $\rho \neq 0$; (original claim: there is a linear correlation between the durations of eruptions and the heights of the eruptions)

**2) the significance level;**

- $\alpha = 10\%$

**3) the test statistic and its distribution under the null hypothesis;**

- The test statistic $T_\rho$ has a t-distribution with (n - 2) degrees of freedom;

**4) the observed value of the test statistic (the observed score);**

- r = 0.091548
- n = 40

$$T_\rho = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.091548}{\sqrt{\frac{1-0.091548^2}{40-2}}} \approx 0.567$$

**5) the P-value or the critical region;**

- `df` = 40 - 2 = 38
- According to *Table 3*, the two-tailed critical values are

$$+t_{38,\ 0.05} = +2.024 \ and \ -t_{38,\ 0.05} = -2.024$$

**6) whether or not the null hypothesis is rejected and why.**

Since the test statistic lies between the two critical values:

$$-2.024 < T_\rho < +2.024$$

Therefore,

- we fail to reject the null hypothesis $H_0$;
- there is sufficient evidence to support the claim that there is a linear correlation between the durations of eruptions and the heights of the eruptions.

## Exercise 4.2 (Section 10.2, E8)

**1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: the frequency counts agree with the claimed that the results fit an uniform distribution; (original claim)
- $H_a$: the frequency counts do not agree with the claimed that the results fit a uniform distribution.

**2) the significance level;**

- $\alpha = 10\%$

**3) the test statistic and its distribution under the null hypothesis;**

- When all $E_i$ are at least 5, the test statistic $\chi^2$ has approximately a chi-square distribution with (k-1) degrees of freedom under H~0 where k represents the number of different categories.

**4) the observed value of the test statistic (the observed score);**

(1) the **claimed uniform distribution** is as follows:

| Tire | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Expected Distribution | 10% | 10% | 10% | 10% |

(2) we obtain the following **expected frequencies**:

| Tire | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Expected frequency | 10 | 10 | 10 | 10 |

(3) we **observed** the following frequencies:

| Tire | Left Front | Right Front | Left Rear | Right Rear |
|---|---|---|---|---|
| Observed frequency | 11 | 15 | 8 | 6 |

Since all $E_i$ are at least 5, the requirements are met, so the test statistic is

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - E_i)^2}{E_i} = \frac{(11-10)^2}{10} + \frac{(15-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(6-10)^2}{10}$$
$$= 4.6$$

**5) the P-value or the critical region;**

- k = 4 - 1 = 3

- According to *Table 4*, the critical value is

$$\chi^2_{k-1,\ \alpha} = \chi^2_{3,\ 0.10} = 6.251$$

**6) whether or not the null hypothesis is rejected and why.**

Since,

$$\chi^2 = 4.6 < 6.251 = \chi^2_{3,\ 0.10}$$

Therefore,

- we fail to reject the null hypothesis $H_0$;
- there is no sufficient evidence to warrant the rejection of the claim that the results fit a uniform distribution.

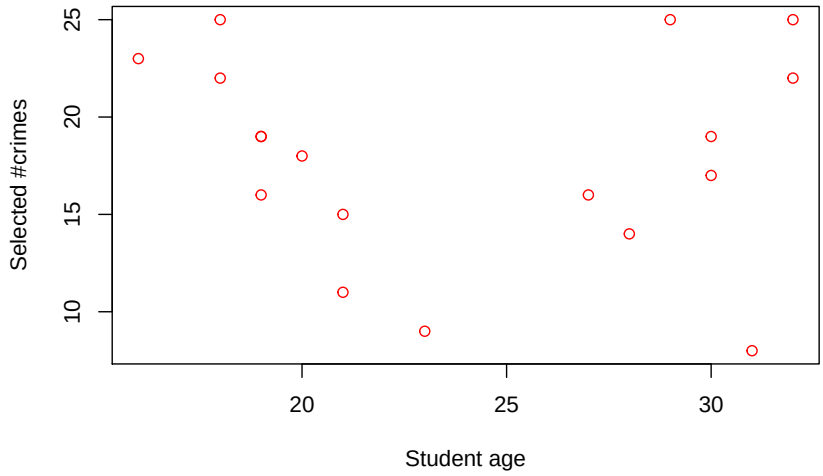
**7) What does the result suggest about the ability of the four students to select the same tire when they really didn't have a flat?**

Based on the test conclusions, the choices that are made by those students tend to form an uniformed distribution.

# & R-exercises

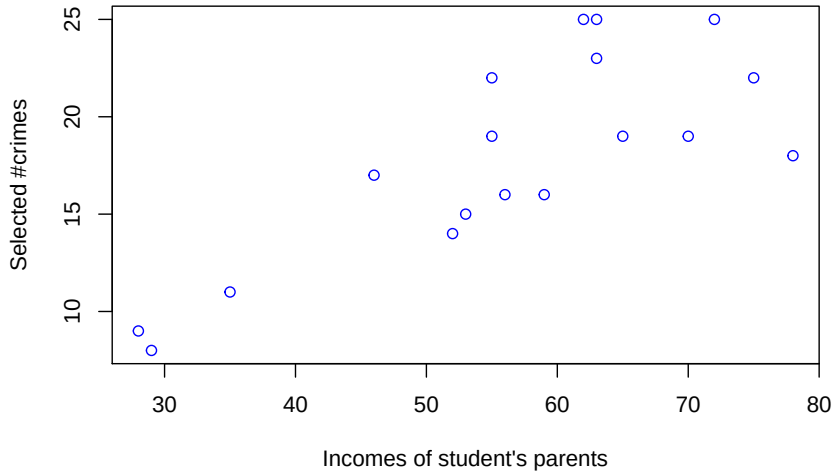
## Exercise 4.3

**a)**




As the above scatter plot shows, the spots spread all over the space and do not illustrate a linear-shaped trend whatsoever. Furthermore, since the linear correlation coefficient is around -0.071, which is quite close to 0 and therefore demonstrates a weak linear relationship between them. Thus, we think there is no linear correlation between these two variables `age` and `crimes`.

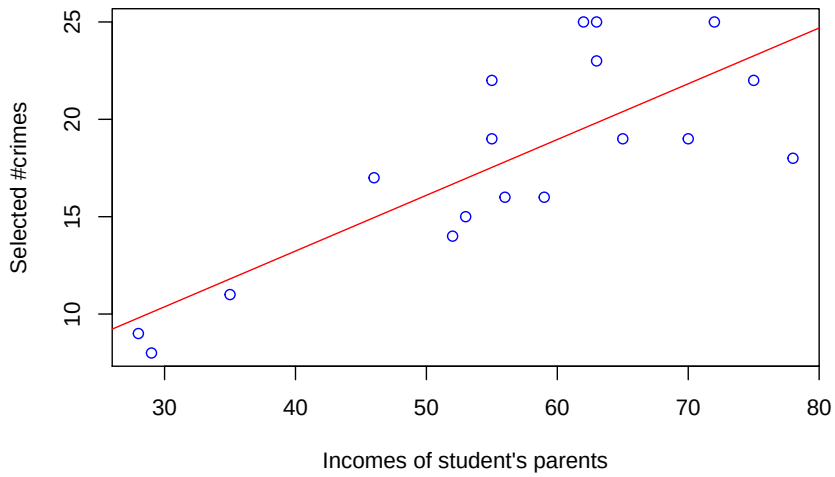
**b)**




The above scatter plot shows an approximate positive linear correlation between these two variables and furthermore, their linear correlation coefficient 0.792 is very close to 1 which suggests a strong linear association between the `income` and `crimes`. Thus, we there there should be a linear correlation between them.

**c)**

- intercept: 1.781
- slop: 0.286



**d)**

**d1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: $\rho = 0$; (original claim: there is no linear relationship between the two variables `income` and `crimes`)
- $H_a$: $\rho \neq 0$;

**d2) the significance level;**

- $\alpha = 5\%$

**d3) the test statistic and its distribution under the null hypothesis;**

- The test statistic $T_\rho$ has a t-distribution with (n - 2) degrees of freedom (n is 18 in this case).

**d4) the observed value of the test statistic (the observed score);**

- r = 0.792
- n = 18

$$T_\rho = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.792}{\sqrt{\frac{1-0.792^2}{18-2}}} \approx 5.189 \ (Tech: \ 5.181)$$

**d5) the P-value or the critical region;**

- P-value = 9.097 x $10^{-5}$

*(Our TA Luminita agrees that we don't need to round the `p-value` calculated by R)*

**d6) whether or not the null hypothesis is rejected and why.**

Since the P-value is significantly smaller than $\alpha$,

therefore,

- we fail to reject the null hypothesis $H_0$;
- there is sufficient evidence to warrant the rejection of the claim that there is no linear relationship between the two variables `income` and `crimes`.

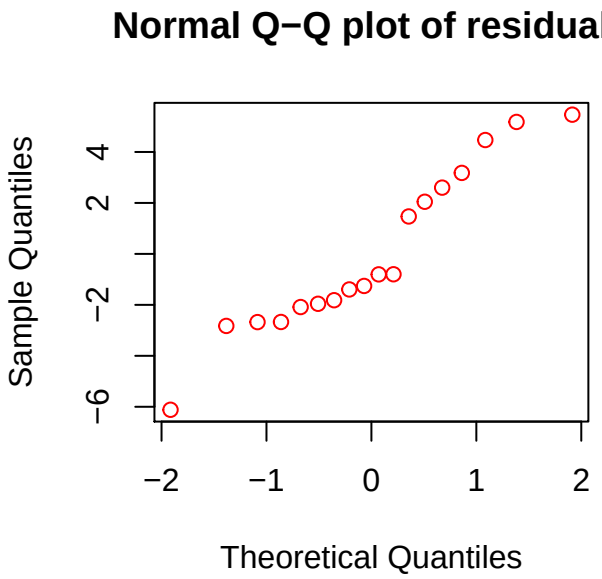
**e)**

**Requirements for testing linearity:**

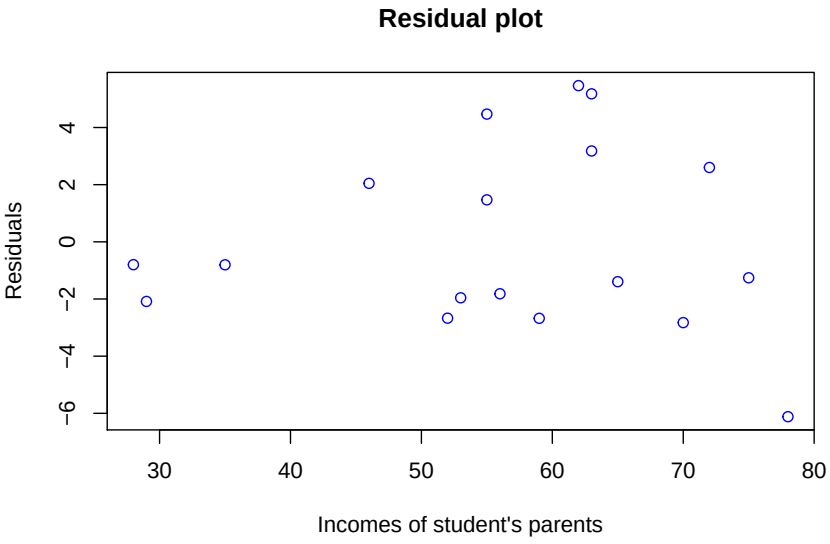
1. Two samples should be independent;
2. The residual should be normally distributed;
3. The standard deviation should be fixed;

**Requirement checks:**

- **R1 check**: According to our lecture, this requirement is difficult to check and can be assumed to be true;
- **R2 check**: Based on the normal QQ plot of residuals shown below, both tails of the distribution of the residuals are lighter than the standard normal distribution (excluding the outlier that lies in the left-bottom corner). Thus, we suppose the second requirement is not quite met;




- **R3 check**: As the residual plot below illustrates, there is no obvious pattern in the residuals and no increase/decrease of residuals with the x variables (income), we, therefore, believe the third requirement is met.



## Exercise 4.4

### a)

Using the formula $E_i = n \cdot p_i$, the expected frequency of German men's wins against Italy under the assumption that the teacher's guess was true is calculated as follows:

| | Wins | Draws | Defeats |
|---|---|---|---|
| Expected requency | 10.5 | 14 | 10.5 |

### b)

**d0) Chi-test requirement check:**

Before conducting the chi-test, we checked that all the expected values are greater than 5, so the requirement is met.

**b1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: the frequency counts agree with the teacher's claimed distribution (i.e. p(Win) = 30%, p(Draw) = 30%, p(Defeat) = 30%)
- $H_a$: the frequency counts do not agree with the claimed distribution.

**b2) the significance level;**

- $\alpha = 5\%$

**b3) the test statistic and its distribution under the null hypothesis;**

- When all $E_i$ are at least 5, the test statistic $\chi^2$ has approximately a chi-square distribution with (k-1) degrees of freedom under H~0;

**b4) the observed value of the test statistic (the observed score);**

$$\chi^2 = \sum_{i=1}^{4} \frac{(o_i - E_i)^2}{E_i} = \frac{(8-10.5)^2}{10.5} + \frac{(12-14)^2}{14} + \frac{(15-10.5)^2}{10.5}$$
$$= 1.619$$

**b5) the P-value or the critical region;**

The critical value = $\chi^2_{k-1,\alpha} = \chi^2_{3-1,0.05} = 5.991$

**b6) whether or not the null hypothesis is rejected and why.**

Since

$$\chi^2 = 1.619 < \chi^2_{k-1,\alpha} = 5.991$$

therefore,

- we fail to reject the null hypothesis $H_0$;
- there is no sufficient evidence to warrant the rejection of the claim that the frequency counts agree with the teacher's claimed distribution (i.e. p(Win) = 30%, p(Draw) = 30%, p(Defeat) = 30%).

**c)**

**i) We believe that we should use a test of homogeneity to test the teacher's second claim.**

This is because what we are concerned is whether or not two populations have the same proportions of some characteristic, rather than if two categorical variables of the same population are dependent from each other.

**ii) Hypothesises**

- $H_0$: the German men and women have the same proportion of winning or losing against or tie with the Italians;
- $H_a$: the German men and women do not have the same proportion of winning or losing against or tie with the Italians.

**d)**

**d0) Chi-test requirement check;**

Before conducting the chi-test, we use R program to calculate the **expected frequencies table** of the constructed `result` matrix. Because all the data are greater than 5, the requirement is met.

**d1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: the German men and women have the same proportion of winning or losing against or tie with the Italians;
- $H_a$: the German men and women do not have the same proportion of winning or losing against or tie with the Italians.

**d2) the significance level;**

- $\alpha = 5\%$

**d3) the test statistic and its distribution under the null hypothesis;**

- When all $E_i$ are at least 5, the test statistic $\chi^2$ has approximately a chi-square distribution with (k-1) degrees of freedom under H~0;

**d4) the observed value of the test statistic (the observed score);**

- test statistic = 9.171

**d5) the P-value or the critical region;**

- The critical value = $\chi^2_{k-1,\alpha} = \chi^2_{3-1,0.05} = 5.991$
- p-value = 0.0102

**d6) whether or not the null hypothesis is rejected and why.**

Since

$$\chi^2 = 9.171 > \chi^2_{k-1,\alpha} = 5.991 \; and \; \alpha = 0.05 > p - value = 0.0102$$

therefore,

- we reject the null hypothesis $H_0$;
- there is sufficient evidence to warrant the rejection of the claim that the German men and women have the same proportion of winning or losing against or tie with the Italians.

**e)**

```
p(men wim) = p(women win) = 16/28 = 4/7
```

The expected number of wins of the German men = 35 * p(men win) = 20.

**f)**

**f1) the hypotheses in terms of the population parameter of interest;**

- $H_0$: the proportion of winning or losing against or tie with the Italians of the German men and women are independent;
- $H_a$: the German men are *more likely* to lose against/tie with Italy than the German women. (the original, directed claim)

**f2) the significance level;**

- $\alpha = 1\%$

**f3) the test statistic and its distribution under the null hypothesis;**

The test statistic is under a known discrete distribution: a hypergeometric distribution with parameters m = 35, k = 39 and N = 63.

**f4) the observed value of the test statistic (the observed score);**

- test statistic = 9.171

**f5) the P-value or the critical region;**

- p-value = 0.005644

**f6) whether or not the null hypothesis is rejected and why.**

- The expected value of this random variable is $\frac{mk}{N} = \frac{35 \cdot 39}{63} = 15.476 < 27$
- $\alpha = 0.01 > p - value = 0.005644$

therefore,

- we reject the null hypothesis $H_0$;
- there is sufficient evidence to support the claim that the German men are *more likely* to lose against/tie with Italy than the German women.

# *Appendix*

```
# E4.3
crimeData <- read.table('crimemale.txt', header=TRUE)

# a)
plot(crimeData$age, crimeData$crimes, xlab = "Student age",
ylab = "Selected #crimes", col = "red")
age_cor = cor(crimeData$age, crimeData$crimes)

# b)
plot(crimeData$income, crimeData$crimes, xlab = "Incomes of student's parents",
ylab = "Selected #crimes", col = "red")
income_cor = cor(crimeData$income, crimeData$crimes)

# c)
lmCrime = lm(crimeData$crimes ~ crimeData$income)
summary(lmCrime)
coefs = lmCrime$coef

plot(crimeData$income, crimeData$crimes, xlab = "Incomes of student's parents",
ylab = "Selected #crimes", col = "blue")
abline(coefs, col = "red")

# d)
income_test = cor.test(crimeData$income, crimeData$crimes)
income_test

# e)
chi = chisq.test(crimeData$income, crimeData$crimes)
chi$p.value

lmCrime = lm(crimeData$crimes ~ crimeData$income);
par(mfrow=c(1,2))
qqnorm(lmCrime$res,main="Normal Q-Q plot of residuals", col = "red")
plot(crimeData$income,lmCrime$res,ylab="Residuals", xlab="Incomes of student's parents",
main="Residual plot", col="blue")

# E4.4
```

```r
# d)
o11=8; o12=12; o13=15; o21=16; o22=8; o23=4;
result = c(o11, o12, o13, o21, o22, o23) # the
soccer=matrix(result,nrow=2,byrow=T);

chisq.test(soccer)$exp # check the requirement of chi-test
chisq.test(soccer)

# f)
soccer2=matrix(c(27,8,12,16),nrow=2,ncol=2,byrow=T)
fisher.test(soccer2,alt="greater")
```