Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40

**Assignment 4**

**Exercise 4.1**

We are testing to see if an r = 0.876 based on n = 40 pairs are significantly different from 0, so, if the r value indicates that the Weight Watcher program is effective in reducing weight.

**Hypothesis:**

*Null hypothesis:*H0: ρ = 0

*Alternative hypothesis:*

Ha: ρ ≠ 0

**Significance level:**

α = 0.01

**Data:**

Data yields r = 0.876, n = 40.

**Test statistic:**

Tρ has a t-distribution with n – 2 = 40 – 2 = 38 degrees of freedom.

Observed value tρ = r = 11.196

**Critical values:**

We have a two-tailed test with α = 0.01 and n = 40, so we get the critical values

-t38,0.01 and t38,0.01, which gives: -2.712 and 2.712.

Since tρ = 11.196 > 2.712, we reject H0.

**Conclusion:**

There is enough evidence to reject the claim that there is no linear correlation between the before weight and the after weight. So, one could argue that the value of r indicates that the Weight Watcher program is effective in reducing weight.

**Exercise 4.2**

**Hypothesis:**

*Claimed Distribution:*

For each category month, the expected value of baseball players is the same.

Therefore, the Pi of every category is also equal to all other Pi

*Null hypothesis:*H0: The above claimed distribution is agreed by the frequency counts.

*Alternative hypothesis:*

Ha: The above claimed distribution is not agreed by the frequency counts.

**Significance level:**

α = 0.05

**Data:**

Data yields n = 387+329+366+344+336+313+313+503+421+434+398+371 = 4515.

Pi= 4515/12 = 376.25

**Test statistic:**

X2 has a chi-squared distribution with n – 1 = 4515 – 1 = 4514 degrees of freedom.

X2 93.072

**Critical values:**

We have a right-tailed test with α = 0.05 and n = 4515, so we get the critical values

X24514,0.05, The closest amount of degree of freedom found in table 4 is 100, which gives: 140.169

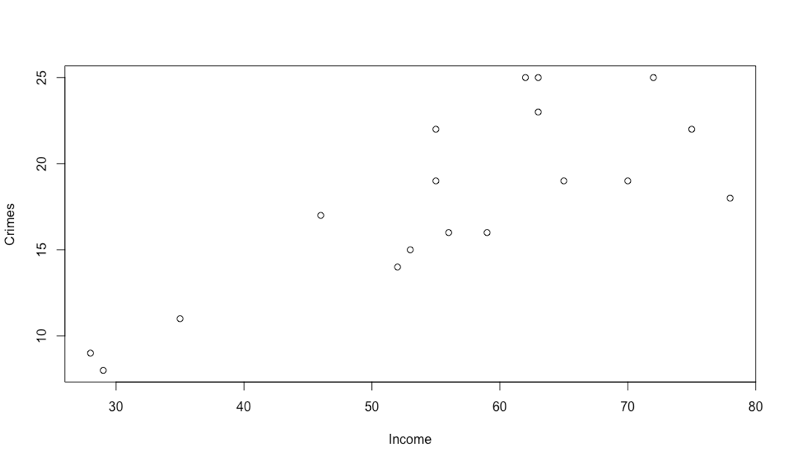
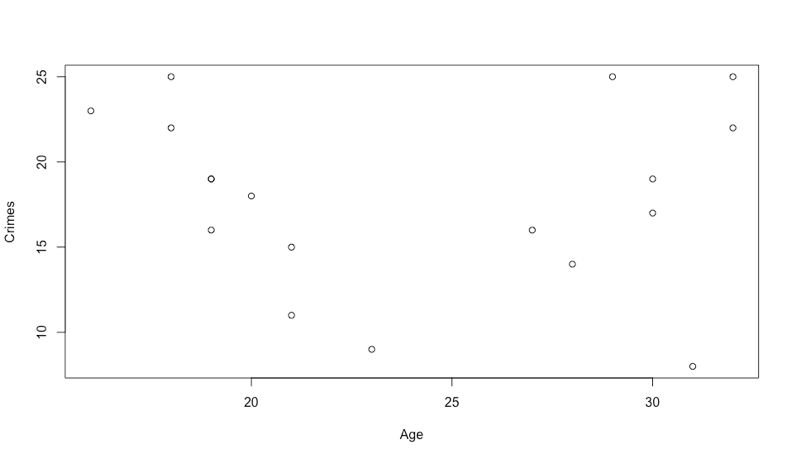
Since X2= 93.072 < 140.169, we fail to reject H0.

**Conclusion:**

There is not enough evidence to reject the claim that the frequency counts agree the claimed distribution (American-born major league baseball players are born in different months with the same frequency). Therefore we can reject the claim that “there is sufficient evidence to warrant the rejection of the claim that American-born major league baseball players are born in different months with the same frequency”.

**Exercise 4.3**

a)

Scatterplot of variables *age* and *crimes.*  
The sample linear coefficient is -0.071 (rounded).   
  
If we take a look at the scatterplot, it does not seem like there is any relationship between the two variables. The sample linear coefficient is -0.071 - very close to 0, which also suggest that there is no linear relationship between the two.  
  
b)   
Scatterplot of variables *income* and *crimes.* The sample linear coefficient is 0.792 (rounded).

If we take a look at the scatterplot, we do see signs of a positive linear relationship. Since the sample coefficient is 0.792, which is very close to 1. This suggests that there is indeed a positive linear relationship.

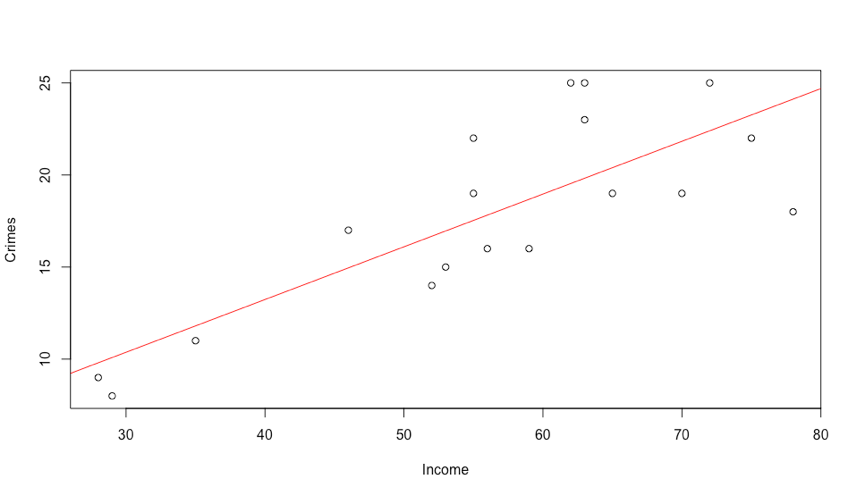
c) The fitted regression equation is given by

The R function *lm(crime$crimes ~ crime$income, data = crime)*yields

Intercept: b0 = 1.781

Slope: b1 = 0.286

So, the fitted regression equation becomes =



Scatterplot of variables *income* and *crimes*, with also plotted.

d) Claim: There is no linear relationship between the variables *income* and *crimes*.

**Hypothesis:**

*Null hypothesis:*

*Alternative hypothesis:*

**Significance level:**

α = 0.05

**Data:**

Data yields b1 = 0.286 and = 0.055

**Test statistic:**

The test statistic = b1 / has a t-distribution with n-2 = 18 – 2 = 16 degrees of freedom under H0. The observed value: = 0.286 / 0.055 = 5.2

**Critical values:**

We have a two-tailed test with α = 0.05 and n = 18, so we get the critical values –t16,0.05 and t16,0.05 which gives: -2.120 and 2.120.

Since = 5.2 > 2.120, we reject H0.

**Conclusion:**

The is sufficient evidence to warrant the rejection of the claim that there is no linear relationship between the variables *income* and *crimes*.

e) The requirements that have to be met are:

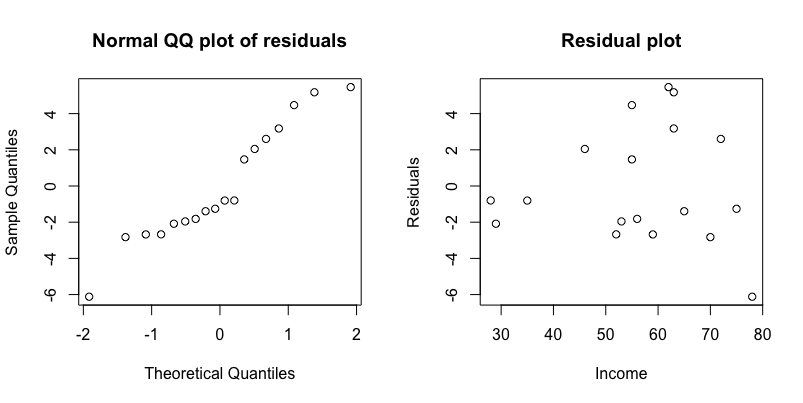
Errors are should be:

- Independent, something we can assume to be true

- Normally distributed

- Have a fixed standard deviation

With R, using qqnorm and plot for the residuals, we get the following plots:



From the QQ plot, we see that the plot follows approximately a straight line, so this probably comes from a normal distribution.

For the residual plot, there is no obvious pattern in the residuals. This has to be the case, otherwise there is something wrong: because our residuals look randomly placed, we can say that we have a fixed standard deviation. So we can positively say that the requirements for testing linearity are met.

**Exercise 4.4**

a) n = 8+12+15=35

Pwin  = 0.4

Ewin = 0.4\*35 = 14

b)

*Claimed Distribution:*

*Pwin = 0.4*

*Pdraw = 0.3*

*Pdefeat= 0.3*

*Null hypothesis:*H0: The above claimed distribution is agreed by the frequency counts.

*Alternative hypothesis:*

Ha: The above claimed distribution is not agreed by the frequency counts.

**Significance level:**

α = 0.1

**Data:**

Data yields n = 35

Ewin = 14

Edraw = 10.5

Edefeat =10.5

**Test statistic:**

X2 has a chi-squared distribution with n – 1 = 35 – 1 = 34 degrees of freedom.

X2 4.714

**Critical values:**

We have a right-tailed test with α = 0.1 and n = 35, so we get the critical values

X234,0.1, The closest amount of degree of freedom found in table 4 is 30, which gives 43.773 of area to the right.

Since X2= 4.714 < 43.773, we fail to reject H0.

**Conclusion:**

There is not enough evidence to reject the claim that the frequency counts agree the claimed distribution of Dennis.

c)

We should use a test of homogeneity. Dennis claims that one team should have the same performance of the other team, which is a different population. We are not interested in the independence of 2 variables (eg. winning and the womens soccer players. We are interested if both teams have the same proportions of winning, playing draw, and losing.

H0:The mens and womens German soccer team have both the same chance of winning (and losing).

Ha:The mens and womens German soccer team do NOT have the same chance of winning (and losing).

d)

*Null hypothesis:*H0:The mens and womens German soccer team have both the same chance of winning (and losing).

*Alternative hypothesis:*

Ha:The mens and womens German soccer team do NOT have the same chance of winning (and losing).

**Significance level:**

α = 0.05

**Data:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Won | Draw | Lost | **Total** |
| Men | 8 | 12 | 15 | **35** |
| Women | 15 | 8 | 4 | **27** |
| **Total** | **23** | **20** | **19** | **62** |

**Expected values:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Won | Draw | Lost |
| Men | 35\*(23/62) = 12.984 | 35\*(20/62)=11.290 | 35\*(19/62)=10.726 |
| Women | 27\*(23/62)=10.016 | 27\*(20/62)=8.709 | 27\*(19/62)=8.274 |

**Test statistic:**

Requirements: All Eij are larger than 5, requirements met.

X2 has a chi-squared distribution with (3-1)(2-1) = 2 degrees of freedom.

X2 8.406

**Critical values:**

We have a right-tailed test with α = 0.1 , r = 2 and c = 3, so we get the critical values

X22,0.1, which gives 5.991 of area to the right.

Since X2= 8.406 > 5.991, we succeed to reject H0.

**Conclusion:**

There is enough evidence to reject the claim that both populations (men and womens German soccer team) have equal proportions of winning (and losing).

e) Emen,won = 12.984

f)

*Null hypothesis:  
H0: “Men have the same chances to win a soccer math against Italy as women have.”*

*Alternative hypothesis:*

Ha : “Men have worse chances to win a soccer match against Italy than women.”

**Significance level:**

α = 0.01

**Data:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Won | Draw & Defeat | **Total** |
| Men | 8 | 27 | **35** |
| Women | 15 | 12 | **27** |
| **Total** | **23** | **39** | **62** |

**Fisher’s test**

When testing we would like to reject H0 when the frequency count in (1,1) is too low, so Ha becomes true.

We test fisher in R with: “fisher.test(matrixDouble,alt="less")”

The result of Fisher’s test is:

*Fisher's Exact Test for Count Data*

*data: matrixDouble*

*p-value = 0.008582*

*#Other output*

The P-value is 0.008582 which is below the significance level of 0.01, therefore we reject H0.

**Conclusion**

There is enough evidence to reject the claim that both populations (men and womens German soccer team) have equal proportions of winning. We now know Ha is true and the men’s team has worse chance of winning a match against Italy then the womens team has.

**Appendix**

4.3 a)

> crime = read.table("/Users/lucasfaijdherbe/Library/Mobile Documents/com~apple~CloudDocs/Computer Science/Statistical Methods/Assignments/Assignment 4/Excersises/crimemale.txt", header = T)

> plot(crime$age, crime$crimes, xlab = 'Age', ylab = 'Crimes')

> cor(crime)

age income crimes

age 1.00000000 -0.4145025 -0.07095301

income -0.41450249 1.0000000 0.79155727

crimes -0.07095301 0.7915573 1.00000000

b)

> plot(crime$income, crime$crimes, xlab = 'Income', ylab = 'Crimes')

> cor(crime)

age income crimes

age 1.00000000 -0.4145025 -0.07095301

income -0.41450249 1.0000000 0.79155727

crimes -0.07095301 0.7915573 1.00000000

c and d)

> lmsim = lm(crime$crimes ~ crime$income, data = crime)

> summary(lmsim)

Call:

lm(formula = crime$crimes ~ crime$income, data = crime)

Residuals:

Min 1Q Median 3Q Max

-6.117 -2.054 -1.031 2.462 5.465

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.78111 3.21597 0.554 0.587

crime$income 0.28636 0.05527 5.181 9.1e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.315 on 16 degrees of freedom

Multiple R-squared: 0.6266, Adjusted R-squared: 0.6032

F-statistic: 26.85 on 1 and 16 DF, p-value: 9.097e-05

e)

> lmsim = lm(crime$crimes ~ crime$income, data = crime)

> par(mfrow=c(1,2))

> qqnorm(lmsim$residuals, main="Normal QQ plot of residuals")

> plot(crime$income, lmsim$residuals, main = "Residual plot", ylab = "Residuals", xlab = "Income")

4.4b)

> observed = c(8,12,15)

> expected = c(14,10.5,10.5)

>

> x = 0

> for (i in 1:length(observed)){

+ x = x+((observed[[i]]-expected[[i]])^2/expected[[i]])

+ }

> print(paste("result = ",x))

[1] "result = 4.71428571428571"

d)

> Omen = c(8,12,15)

> Owomen = c(15,8,4)

> totalO = rbind(Omen,Owomen)

> results = matrix(totalO,ncol = 3,byrow=F)

>

>

> Emen = c(12.984,11.290,10.726)

> Ewomen = c(10.016,8.709,8.274)

> totalE = rbind(Emen,Ewomen)

> Expected = matrix(totalE,ncol = 3,byrow=F)

>

> x = 0

> for (j in 1:3){

+ for (i in 1:2) {

+ x = x+((results[i,j]-Expected[i,j])^2/Expected[i,j])

+ }

+

+ }

> print(paste("result = ",x))

[1] "result = 8.40640442413929"

f)

> matrixDouble = matrix(c(8,27,15,12),nrow = 2,byrow = F)

> fisher.test(matrixDouble,alt="less")

Fisher's Exact Test for Count Data

data: matrixDouble

p-value = 0.008582

alternative hypothesis: true odds ratio is less than 1

95 percent confidence interval:

0.0000000 0.6804577

sample estimates:

odds ratio

0.2431269