Enrico Dal Pos (2708533) & Osama Obaidat(2704856) & Justus Beck(2717047) - Group 64

Assignment 4

Exercise 4.1

The sample linear correlation coefficient r is given by the task description and is defined as $r = 0.926$. For this task we assume that the data pairs are sampled from a bivariate normal distribution, which means that the scatterplot approximates a straight line.

Hypothesis:
H_0: $p = 0$
Alternative Hypothesis:
H_a: $p \neq 0$
Significance level:
$\alpha = 1\%$
Test statistic:
$T_p = \frac{R-p}{\sqrt{\frac{1-R^2}{n-2}}}$ with a t-distribution with $n - 2 = 14$ degrees of freedom.
Observed value:
$t_p = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0.926}{\sqrt{\frac{1-0.926^2}{16-2}}} = 9.18$
Critical values:
Two-tailed test, $\alpha = 1\%$ and $n = 16$ so the critical values are: $-t_{14,0.005} = -2.977$ and $t_{14,0.005} = 2.977$
Conclusion:
Since $t_p = 9.18 > 2.977$ we reject H_0. There is sufficient evidence to reject the claim that there is no linear correlation between the 7-day-incidence rates and recent government election results of a certain (extreme) political party.

Since we know that $r = 0.926$ we know that the data pairs approximate a perfect positive linear relationship, which means that higher values of variable 1 are associated with higher values of variable 2. From this we can infer that if the incidence-rates becomes higher, also the vote for the party becomes stronger.

Exercise 4.2

To check whether 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, we can use the Goodness-of-Fit test:

Suppose: *k* different categories (month player born in); random sample of size *n* (total players).

Step 0: Indicate population parameter:

Proportions of Jan, Feb, Mar, Apr, May, Jun, Jul, Aug, Sep, Oct, Nov, Dec: p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8, p_9, p_{10}, p_{11}, p_{12}.

Step 1: Formulate H_0 and H_1 and choose significance level α:

H_0: $p_1 = 0.0833$ (1/12), $p_2 = 0.0833$, $p_3 = 0.0833$, $p_4 = 0.0833$, $p_5 = 0.0833$, $p_6 = 0.0833$, $p_7 = 0.0833$, $p_8 = 0.0833$, $p_9 = 0.0833$, $p_{10} = 0.0833$, $p_{11} = 0.0833$, $p_{12} = 0.0833$ vs.

H_1: $p_i \neq e_i$ for at least one *i*, with significance level $\alpha = 0.1$.

Step 2: Collect data and check requirements:

Let O_i be the observed frequency count of category i,

and expected frequency $E_i = n * p_i$.

Random sample of n = 752

From n and the expected percentages, we can make a table for the expected frequencies.

We expect p_1 is 752 * 0.833 = 62.64. We do the same for the rest of the different categories (months).

Month	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
Observed frequency	64	55	61	58	56	52	52	84	70	72	66	62
Expected frequency	62.64	62.64	62.64	62.64	62.64	62.64	62.64	62.64	62.64	62.64	62.64	62.64

All $E_i \geq 5$, so requirements met.

Step 3: Test statistic and critical region:

$$X^2 = \sum_{i=1}^{k} \frac{(Oi - Ei)^2}{Ei} \sim \chi^2_{k-1} = \chi^2_{11} \text{ under } H_0$$

The observed value of the test statistic is

$$\chi^2 = \sum_{i=1}^{12} \frac{(Oi - Ei)^2}{Ei} = \frac{(64-62.64)^2}{62.64} + \frac{(55-62.64)^2}{62.64} + \frac{(61-62.64)^2}{62.64} + \frac{(58-62.64)^2}{62.64} + \frac{(56-62.64)^2}{62.64} + \frac{(52-62.64)^2}{62.64}$$

$$+ \frac{(52-62.64)^2}{62.64} + \frac{(84-62.64)^2}{62.64} + \frac{(70-62.64)^2}{62.64} + \frac{(72-62.64)^2}{62.64} + \frac{(66-62.64)^2}{62.64} + \frac{(62-62.64)^2}{62.64} \approx 15.4$$

The test is right tailed, so critical value: $\chi^2_{k-1,\,\alpha} = \chi^2_{11,\,0.1}$ (from Table 4 in book) = 17.275 > 15.4

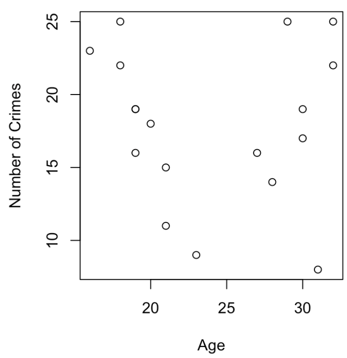
Step 4: conclude:

Our observed value is less than the critical value, so it is not in the critical region therefore H_0 is not rejected. There is no sufficient evidence to reject the claim that American-born major league baseball players are born in different months with the same frequency.

Meaning that the claim from the sport scientist that more baseball players have birthdays in the months immediately following July 31, because that was the cut-off date for non-school baseball leagues is **false**.

Exercise 4.3

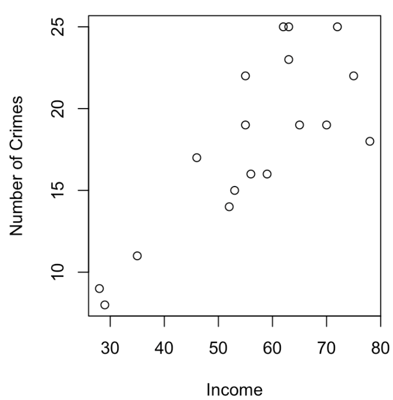
a)



Linear coefficient: r = -0.07095301

Since r ≈ 0 and there's no relationship between the number of crimes and age in the scatterplot we can safely assume that there's no linear correlation between x and y.

b)



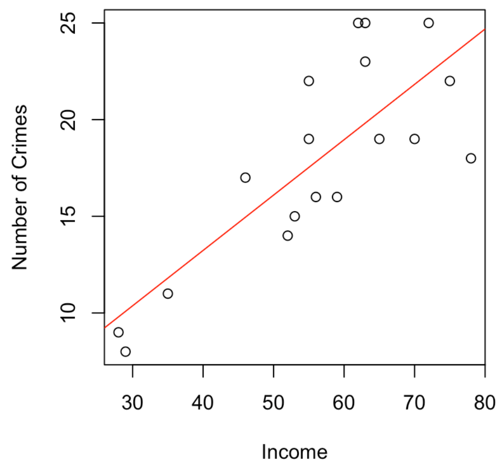
Linear coefficient: r = 0.7915573

From the scatterplot it looks like there's a positive linear relationship since as the income increases the number of crimes also seems to increase and this is confirmed by the linear coefficient since r > 0 and close to 1.

c) To calculate the intercept and slope we did the regression analysis using the R function *lm*

Estimated intercept: 1.78111
Estimated slope: 0.28636

This is the 'best' line plotted in the scatterplot (using the function *abline*):



d) H_0: slope = $\beta_1 = 0$, H_a: slope = $\beta_1 \neq 0$
Significance level $\boldsymbol{\alpha} = 0.01$

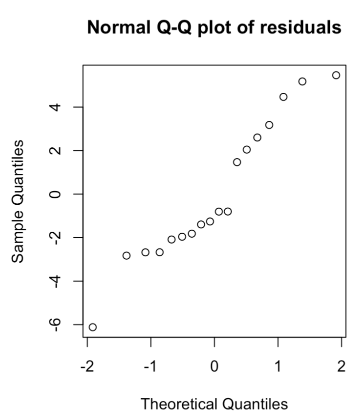
$$T_\beta = \frac{b_1}{s_{b_1}} \sim T_{n-2} \text{ under } H_0$$

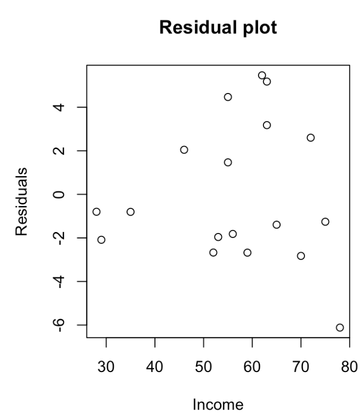
$$T_\beta = \frac{1.78111}{3.21597} \sim T_{n-2} \text{ under } H_0$$

Since (calculated using r) tρ = 0.00009097 < 0.01 we don't reject H_0.

Conclusion:
There is sufficient evidence to reject the claim that there is linear correlation between the income of the criminals and the number of crimes

e) In order to perform the test on d, the errors are assumed to be independent and from a normal distribution with fixed standard deviation





From the qqplot, the errors seem to follow an S-shape, but it also cannot exclude that they follow a normal distribution. In the residual plot, the residuals don't follow any pattern. So, as a result, we can state that a linear regression model probably fits and that the requirements for the test are met

Exercise 4.4

a) We should use a test of homogeneity since we have different samples from different populations. We can 't use an independence test since we don't have only 2 variables.

   H_0 : Andy friends have the same proportion of wins/losses
   H_a : Andy friends do not have the same proportion of wins/losse

b) Significance level α = 0.05

   Pearson's Chi-squared test

   data: results
   X-squared = 6.4865, df = 8, p-value = 0.5929

c) Andy would be expected to win around 29 games out of 160 against Freddy if all players are equally strong.

d) We use Fisher's exact test since we are evaluating a one-sided claim.

   H_0 : probability to win against Freddy is equal to the probability to win against Bob

   H_a : probability to win against Freddy is smaller than the probability to win against Bob

   α = 0.01

   Output if fisher test:

      data: results

      p-value = 0.02993

      alternative hypothesis: true odds ratio is less than 1

      95 percent confidence interval:

      0.0000000 0.9405475

      sample estimates:

      odds ratio

      0.5970088

Since 0.02993 is less than 0.1, H_0 can't be rejected so we can say that the probability of winning against Bob and Freddy is almost the same.

Appendix

Exercise 4.3

a)

```
table = read.table("crimemale.txt", header = TRUE)

x = table$age
y = table$crimes

plot(x,y, xlab = "Age", ylab = "Number of Crimes")
cor(x,y)
```

b)

```
table = read.table("crimemale.txt", header = TRUE)

x = table$income
y = table$crimes

plot(x,y, xlab = "Income", ylab = "Number of Crimes")
cor(x,y)
```

c)

```
table = read.table("crimemale.txt", header = TRUE)

x = table$income
y = table$crimes

model = lm(y~x)

summary(model)
```

```
plot(x,y)

abline(model$coef, col="red")
```

Summary output:

```
Call:
lm(formula = y ~ x)

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
x            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
```

d)

```
par(mfrow=c(1,2));
qqnorm(model$res,main="Normal Q-Q plot of residuals");
plot(x,model$res,ylab="Residuals",main="Residual plot", xlab =
"Income")
```

Exercise 4.4

b)

```
Bob     = c(179,47,57)
Cecilia = c(96,27,36)
David   = c(52,13,18)
Emma    = c(39,15,15)
Freddy  = c(84,37,39)

results = matrix(c(Bob,Cecilia, David, Emma, Freddy), ncol=3,
byrow = T)

chisq.test(results)
```

c)

```
chisq.test(results)$exp
```

d)

```
Bob = c(179,47)
Freddy  = c(84,37)

results = matrix(c(Freddy, Bob), ncol=2, byrow = T)

results

fisher.test(results, alt='less')

?fisher.test
```