Statistical Methods Fall 2021

Assignment 4: Correlation, regression and contingency tables

Deadline: see Canvas

Topics of this assignment

The exercises below concern topics that were covered in Lectures 9 and 10: correlation, regression and contingency tables (see Sections 9.2 (incl. Part 2), 9.3 (incl. Part 3), 9.4, 10.2 and 10.3 (incl. Part 2) of the book and the slides of Lectures 9 and 10). Before making the assignment, study these topics.

How to make the exercises? See Assignment 1.

If you are asked to perform a test, do not only give the conclusion of your test, but report:
- the hypotheses in terms of the population parameter of interest;
- the significance level;
- the test statistic and its distribution under the null hypothesis;
- also check the assumptions required for retrieving the distribution under the null hypothesis;
- the observed value of the test statistic (the observed score);
- the P-value or the critical region;
- whether or not the null hypothesis is rejected and why.

If applicable, also phrase your conclusion in terms of the context of the problem.

Theoretical exercises

For the two theoretical exercises below use Tables 3 and 4 from the Appendix in the book to find probabilities and/or critical values. Do not use R. If you need to use a t-distribution with the number of degrees of freedom not included in Table 3, report the number of degrees of freedom, and use the critical value based on a t-distribution with the next lower number of degrees of freedom found in the table.

Exercise 4.1
A political scientist wishes to analyze the relationship between the 7-day-incidence rates (SARS-CoV-2; data from Dec. 6, 2021) and the recent governmental election results of a certain (extreme) political party. He used 16 data points that correspond to the 16 federal states of that country. He computed a sample correlation of 0.926. Check with the help of a hypothesis test whether there is sufficient evidence to support a claim of a linear correlation of the incidence numbers and the election results. Take $\alpha = 1\%$.
Also argue whether a strong vote for that party seems to go along with a rather high or low incidence rate.
Follow the detailed instructions about testing presented above.

Exercise 4.2
A sport scientist claims that more baseball players have birthdays in the months immediately following July 31, because that was the cutoff date for nonschool baseball leagues. Here is a sample of frequency counts of months of birthdates of American-born major league baseball players (starting with January):

$$64, 55, 61, 58, 56, 52, 52, 84, 70, 72, 66, 62.$$

Check with the help of a hypothesis test whether there is sufficient evidence to warrant rejection of the claim that American-born major league baseball players are born in different months with the same frequency. Take $\alpha = 10\%$.
Follow the detailed instructions about testing presented above.

R-exercises

Do not use tables from the Appendix in the book. Use R to find probabilities and/or critical values.

Hints concerning R:

- The *R*-function `cor()` computes the sample linear correlation coefficient. The *R*-function `cor.test()` can be used to compute a confidence interval for the population correlation coefficient, and to perform a test concerning the population correlation coefficient. At the same time it also gives the sample correlation coefficient; here it is called 'sample estimate' (for the population linear correlation coefficient).
- For analysis of the linear regression model the R-function `lm()` can be used. Let the measurements of the explanatory variable be in the vector `x` and the measurements of the outcome variable be in `y`. Then `lmsim=lm(y~x)` fits a simple linear regression model and stores the output in `lmsim`. The output is a list, which can be studied using `summary(lmsim)`. To obtain the estimated coefficients for the intercept and slope the command `lmsim$coef` can be used. Similarly, `lmsim$res` provides the residuals. The standard errors of the estimated coefficients can be obtained (apart from inspection of `summary(lmsim)`) with the command `summary(lmsim)$coef[,2]`. To visualise the regression equation, the command `abline(lmsim$coef)` can be used. See also the slides of Lecture 9.
- For the analysis of contingency tables the function `chisq.test()` can be used.

  The command `chisq.test(table)$exp` provides the expected frequency count of the data in the fictitious contingency table `table` under the null hypothesis. See also the slides of Lecture 10.
- Recall: for computing probabilities and quantiles of normally, t-, chisquare, etc. distributed random variables the *R*-functions `pnorm, pt, pchisq, ...`, and `qnorm, qt, qchisq, ...` can be used. For the t- and chisquare distributions the number of degrees of freedom needs to be specified.
- You can load an `.RData` file into the workspace using the command `load(...)`.

Exercise 4.3 There is considerable variation among individuals in their perception of crime, and in particular of which specific acts constitute a crime. A study was made to investigate which variables, like age, level of education, parental income, etc., may influence this perception. The file `crimemale.txt` contains part of the results of this study: data are given for 18 male college students who were asked how many of the following 25 acts they perceive as being a crime: *aggravated assault, armed robbery, arson, atheism, auto theft, burglary, civil disobedience, communism, drug addiction, embezzlement, forcible rape, gambling, homosexuality, land fraud, nazism, payola, price fixing, prostitution, sexual abuse of child, sex discrimination, shoplifting, striking, strip mining, treason, vandalism.* The column `crimes` contains the number of acts the students perceive as crimes, `age` the ages of the students, and `income` the incomes of the parents (in $1000).

a) Make for the data of the male students a scatterplot in which the x variable is `age` and the y variable is `crimes`. Compute also the sample linear correlation coefficient. Based on the plot and the sample linear correlation coefficient (without conducting a hypothesis test), do you think there is linear correlation between the two variables?

b) Repeat part a with `income` as the x variable.

c) Perform a linear regression analysis – i.e. formulate the regression model and compute the estimates of the unknown parameter values – with the variable `income` as the explanatory variable and the variable `crimes` as the response variable. Report the estimated values of the intercept and slope that determine the 'best' line and draw this 'best' line in the corresponding scatterplot.

d) Using the results of the regression analysis of part c, test the claim that there is no linear relationship between the two variables `income` and `crimes`. Take significance level 1%.
Follow the detailed instructions about testing presented in the first page of this assignment.

e) In order to perform the test of part d, certain requirements have to be met. What are these requirements? Provide a suitable plot (or plots) and report and argue whether the requirements are indeed met.

Exercise 4.4 Andy uses a mobile app to play games of trivia with his friends. On different evenings, he made appointments with either of Bob, Cecilia, David, Emma, or Freddy and played one-on-one with the chosen friend for the whole evening. They have previously agreed on playing, respectively, 283, 149, 83, 69, and 160 rounds of the game. Both players have to answer a number of randomly selected questions, and the player who correctly answers more questions wins. The table below contains results of 744 games Andy played with his friends (e.g., Andy won 179 games against Bob).

	Won	Lost	Draw	Total
Bob	179	47	57	283
Cecilia	96	17	36	149
David	52	13	18	83
Emma	39	15	15	69
Freddy	84	37	39	160
Total	450	129	165	744

a) In order to investigate whether Andy's friends are equally strong opponents, should you use a test of independence or a test of homogeneity? Motivate your answer and formulate the null and alternative hypothesis.

b) Create a matrix `results` containing the data and use it to perform the test of part (a). Take significance level $\alpha = 5\%$. (See the first page of the assignment for detailed instructions about testing).

c) How many games against Freddy would Andy be expected to win, if the null hypothesis were true and he played 160 games against Freddy?

d) Use a suitable test to test whether the probability to win against Freddy are smaller than the probability to win against Bob. Use $\alpha = 10\%$.
Follow the detailed instructions about testing presented in the first page of this assignment.

Assignment 4
Statistical Methods 2021

Project Group 3:
Caspar M. S. Grevelhörster (2707848)
Zülfiye Tülin Manisa (2664758)
Sidharth R. Singh (2708756)

Vrije Universiteit Amsterdam,
De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

Exercise 1

Are the incidence numbers in the states of the given country related to the election results in the states of the given country?
Given: A sample of 16 pairs and sample correlation coefficient $r = 0.926$.

Step 0: Population Parameter:
The population parameter is the linear correlation coefficient ρ.

Step 1: Hypotheses and significance level:
The null-hypothesis is that there is no linear correlation of the incidence numbers and the election results: $H_0: \ \rho = 0$ vs. The alternative hypothesis is that there is a linear correlation between the incidence numbers and the election results: $H_1: \ \rho \neq 0$ with a signifcance level of $\alpha = 0.01$.

Step 2: Data and required assumptions:
We have $r = 0.926$ and $n = 16$. We must assume that the data pairs come from a bivariate normal distribution to conduct this test. If a scatterplot was given for the data-pairs we could check whether the points are approxiamtely on a straight line to proceed with the test.

Step 3: Test statistic and critical region:
We use the test statistic:

$$T_\rho = \frac{R}{\sqrt{\frac{1-R^2}{n-2}}} \ \sim \ t_{n\,-\,2\,=\,t_{14}\text{ under }H_0}$$

The observed value of the test statistic is:

$$t_\rho = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} \ = \ \frac{0.926}{\sqrt{\frac{1-0.926^2}{14}}} \ \approx \ 9.178$$

The test is two-tailed, in **Table 3** of the book we see that the critical values for a $t-$ distribution with 14 degrees of freedom and significance level of $\alpha = 0.01$ are: $-t_{14,0.005} \ = \ -2.977$ and $t_{14,0.005} = 2.977$
$t_\rho = 9.178 \ > 2.977$ is in the critical region.

Step 4: Conclusion:
The null-hypothesis H_0 that claims that there is no correlation between the incidence numbers of a state and the election results of a certain political party, is rejected. We have enough evidence to confirm that there is a linear correlation between the incidence number and the election result in a state of the given country. Since the correlation is positive, a strong vote for that party seems to go along with a high incidence rate.

Exercise 2

Are the birth months of American-born baseball players in the major league baseball equally distributed over the months of a year?
Given: Sample of frequency counts of birth dates of American-born baseball players in the major league baseball.
Step 0: Population Parameter:
The population parameter is the proportions of birth dates per month: p_i for all $i \in \{1,2,3,4,5,6,7,8,9,10,11,12\}$.

Step 1: Hypotheses and significance level α:
The claim is that American-born major league baseball players are born in different months with the same frequency. It follows that: $H_0 : p_i = \frac{1}{12} \approx 0.083$ for all $i \in \{1,2,3,4,5,6,7,8,9,10,11,12\}$ vs. The alternative hypothesis which claims that there is at least one month where the frequency differs:
$H_1 : p_i \neq e_i$ for at least one $i \in \{1,2,3,4,5,6,7,8,9,10,11,12\}$, with significance level $\alpha = 0.1$.

Step 2: Data and requirements:
The researcher observed the birth dates of 752 American-born baseball players so $n = 752$. Since the claim is that there is no difference in the frequency of the birth dates per month we expect that all months should have an equal number of births: $E_i = 752 * 0.083 \approx 62.67$ for all $i \in \{1,2,3,4,5,6,7,8,9,10,11,12\}$.

Table 1: Observed frequency vs. Expected frequency

Month	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
Observed frequency	64	55	61	58	56	52	52	84	70	72	66	62
Expected frequency	62.67	62.67	62.67	62.67	62.67	62.67	62.67	62.67	62.67	62.67	62.67	62.67

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

Step 3: Test statistic and critical region:
We use the test statistic:

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

The observed value of the statistic is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - E_i)^2}{E_i} = \frac{(64-62.67)^2}{62.67} + \frac{(55-62.67)^2}{62.67} + \frac{(61-62.67)^2}{62.67} + \frac{(58-62.67)^2}{62.67} + \frac{(56-62.67)^2}{62.67}$$
$$+ \frac{(52-62.67)^2}{62.67} + \frac{(84-62.67)^2}{62.67} + \frac{(70-62.67)^2}{62.67} + \frac{(72-62.67)^2}{62.67} + \frac{(66-62.67)^2}{62.67} + \frac{(62-62.67)^2}{62.67} \approx 15.393$$

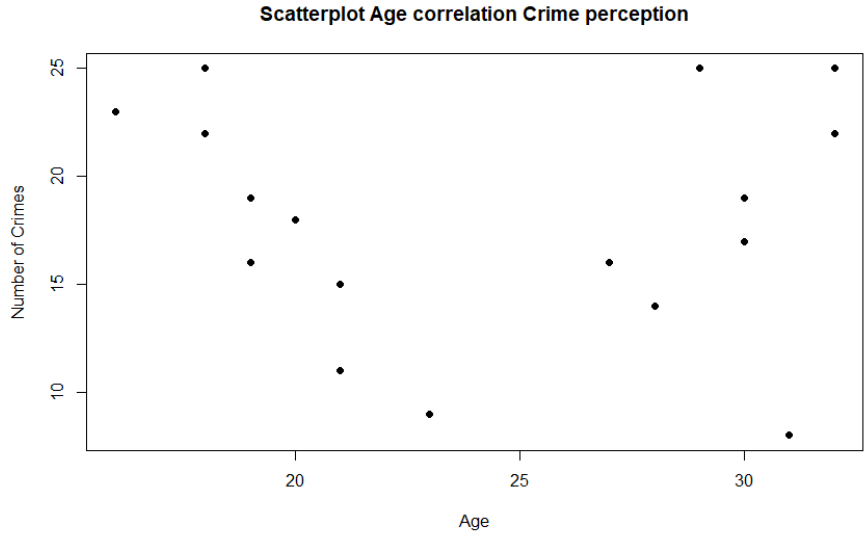
The test is right-tailed, in **Table 4** of the book we see that the critical value for a Chi-Square (χ^2) distribution with 11 degrees of freedom and significance level of $\alpha = 0.1$ is: $\chi^2_{k-1,\,\alpha} = \chi^2_{11,0.1} = 17.275$.
$\chi^2 = 15.393 < 17.275$ is not in the critical region.

Step 4: Conclusion:
H_0 is not rejected. There is not sufficient evidence to reject the claim that American-born major league baseball players are born in different months with the same frequency.

Exercise 3

a) Figure 1 shows a scatter plot in which the x variable is age and the y variable is crimes. It was computed using R, the code can be found in the appendix.

Fig. 1: Graphical representation of the Scatter plot showing correlation between age and crime perception.



To get the correlation coefficient, R is used. The program uses the following formula to compute this value:

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x, s_y}$$

The code in R:

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
cor(t$age, t$crimes)

## [1] -0.07095301
```

The sample linear correlation coefficient r shows the strength of a linear relationship. If $r = 0$, there is no relationship and if $r = 1$ or $r = -1$ there is a perfect positive or negative correlation. The closer the observed test statistic is to 0, the "weaker" the correlation, the closer to 1, the stronger it is. In this instance, $t = -0.07095301$. This means that there is a very weak linear relationship between the two variables. However, looking at the scatter plot, it becomes obvious that the linear relationship between the variables is negligible because the data points are randomly scattered across the plot.

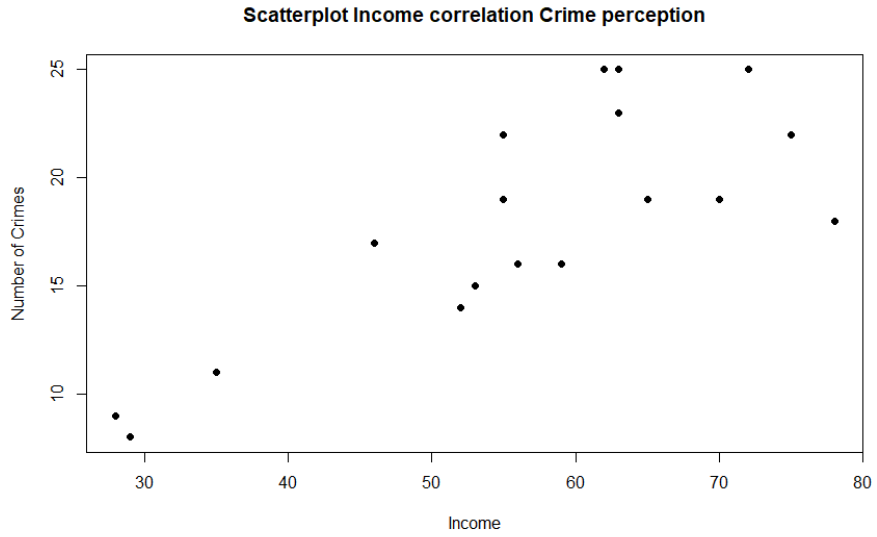
b) Figure 2 shows a scatter plot in which the explanatory x variable is income and the response y variable is crimes. It was computed using R, the code can be found in the appendix.
To get the correlation coefficient, R is used:

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
cor(t$income, t$crimes)

## [1] 0.7915573
```

Based on the explanation in part **a)**, the observed test statistic suggests an existing linear correlation between the two tested variables income and crime perception, since it is relatively close to 1. Furthermore, it can be seen that in fig. 2, the data points are scattered at the left bottom and right top with little outliers which suggests a stronger positive linear relationship between the two variables.

Fig. 2: Graphical representation of the Scatter plot showing correlation between income and crime perception.



c) A linear regression analysis is based on the approximate regression line given by $y_1 = \beta_0 + \beta_1 x_i + \text{error}_i$. The unknown population parameters β_0 and β_1 can be approximated with the "best" sample statistics b_0 and b_1. The latter values constitute to the linear regression line given by $\hat{y} = b_0 + b_1 x$. In order to construct this linear regression line, the sample linear correlation coefficient r needs to be computed using R. The computation happens based on the following formula:

$$r = \frac{1}{n-1} \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{s_x, s_y}$$

The coefficient was already obtained in the previous task, $r = 0.7915573$. With this value on hand, the linear model can be obtained through computing b_0 and b_1:

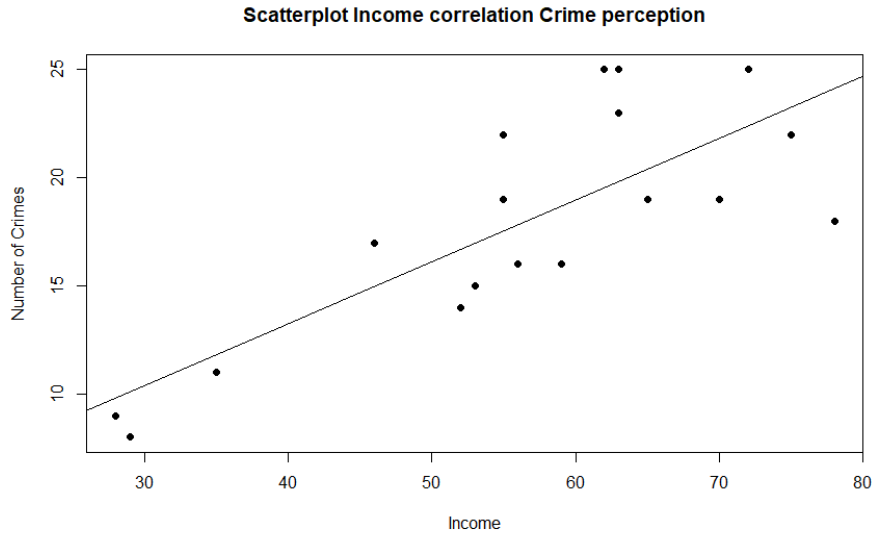
$$b_1 = r\frac{s_y}{s_x} = 0.7915573\frac{s_{\text{crimes}}}{s_{\text{income}}} = 0.7915573\frac{5.263327}{14.54899} \approx 0.29 \text{ and } b_0 = \bar{y} - b_1\bar{x} = 17.94444 - 0.29 * 56.44444 \approx 1.78$$

These values come together in the following model function:

$$\hat{y} = b_0 + b_1 x = 1.78 + 0.29x$$

plotting this line in the scatter plot results in the graph depicted in fig. 3.

Fig. 3: Scatter plot showing income and crime perception with a fitted linear model.



To summarize, the estimates of the unknown population parameters are $b_0 = 1.78$ and $b_1 = 0.29$. The standard error of b_1, so s_{b_1}, can be computed using R:

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
summary(lm(t$crime ~ t$income))$coefficients["t$income", "Std. Error"]

## [1] 0.05526831
```

So $s_{b_1} \approx 0.0553$.

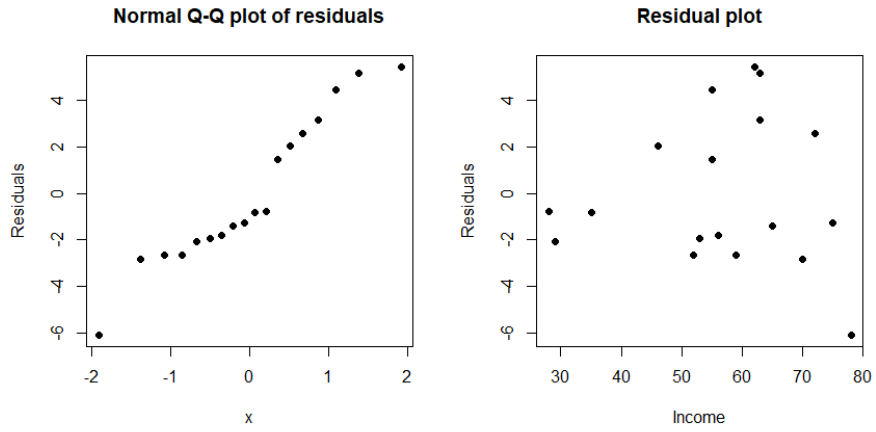
d) To test the linear relationship of income and crimes, it is assumed that $H_0 : \beta_1 = 0$ against $H_a : \beta_1 \neq 0$ (where β_1 is the slope of the regression equation) with a significance level of $\alpha = .01$. The test statistic is given by the t-distributed $T_\beta = \frac{b_1}{s_{b_1}}$ with $df = 18 - 2 = 16$ degrees of freedom. Both needed factors were already computed in the last task: $b_1 = .29$ $s_{b_1} = 0.0553$. With these two values on hand, the observed test score can be computed using the formula of

$$T_\beta = \frac{b_1}{s_{b_1}} = \frac{0.29}{0.0553} \approx 5.2441$$

The confidence interval can be found using **table 3**: $CI = [-t_{df,\alpha/2}, t_{df,\alpha/2}] = [-t_{16,0.005}, t_{16,0.005}] = [-2.921, 2.921]$. Using this interval, it becomes obvious that the observed test statistic of $T_\beta = 5.2441$ lies outside of the critical boundaries (i.e. outside of the interval) since $T_\beta = 5.2441 > t_{16,0.005} = 2.921$. Therefore, there is sufficient evidence to warrant rejection of the claim that there is no linear relationship between income and crime perception; i.e. H_0 is rejected.

e) There are certain requirements that have to be met to perform the test in task **d)**. These requirements are 1. independence, 2. normal distribution, and 3. fixed standard deviation. Independence is difficult to test and thus assumed to be true. Normal distribution can be checked using a normal QQ-plot of the residuals, so the difference between observed and predicted values. The residuals res can be computed using the following formula: $\text{res}_i = y_i - (b_0 + b_1 x)$. For sake of simplicity, they were computed using R. The resulting QQ-plot can be seen in the left plot of fig. 4. Fixed standard deviation is checked by plotting the residual values against the income values of the data set.

Fig. 4: Normal QQ-plot of residuals.



Since the values in the normal QQ-lot in figure 4 follow approximately a straight line, it can be said that the error values are normally distributed. Furthermore, there is no pattern visible in the residual plot in fig. 4, which means that the errors have a fixed standard deviation. The inference from t-test can be used when the errors are independent and from a normal distribution with fixed standard deviation. Since all of those requirements are met (as shown above), the t-test may be used in task **d)**.

Exercise 4

a) Homogeneity test:
The goal is to investigate if Andy's friends are equally strong, which means the dependency is not at question here. The homogeneity test will be applied to find out whether the distribution between the 5 friends are the same or not. Therefore, our hypotheses will be as following:
H_0 : All of the opponents are equally strong
H_1 : All of the opponents are not equally strong

b)

- The null hypothesis is that all of the opponents are equally strong and the alternative hypothesis is that all of the opponents are not equally strong. If the null hypothesis were to be correct the ratio of win-loss-draw should be equal for all opponents.
- It is given that $\alpha = 5\%$, this means that the significance level is 0.05.
- The test statistics and distribution calculated in R:
  - X-squared = 10.931
  - df = 8
  - p-value = 0.2056
    Chisquared distribution with 8 degrees of freedom.
- All $E_{ij} \geq 1$ and 80% of $E_{ij} \geq 5$, so requirements are met.
- The observed score of the test statistic is X-squared = 10.931
- The P-value is 0.2056
- The null hypothesis (H_0) cannot be rejected, since the P-value is larger than α. This means that we cannot reject that all opponents are equally strong.

c) If the null hypothesis is true we can calculate the expected frequency by dividing the total games won by Andy by the total amount of games played, which gives us P(Andy wins) = $\frac{450}{744}$.
Since there are 160 games played against Freddy the answer will be $160 \cdot \frac{450}{744} = 96.7741935484 \approx 97$ games are expected to be won by Andy against Freddy.
A table with the expected values can also be computed in R, this table shows the same outcome as we calculated. The code will be provided in the appendix.

d) To test this we will use the Fisher exact test, because the alternative hypothesis is that the two populations (Bob and Freddy) have different proportions of a characteristic
H_0: The null hypothesis is that the probability to win against Bob or Freddy is the same.
H_1: The alternative hypothesis is that the probability to win against Bob is larger than the probability of winning against Freddy.
It is given that $\alpha = 10\%$, this means that the significance level is 0.1.
The test statistics calculated in R: p-value = 0.07421
The null hypothesis (H_0) can be rejected, since the P-value is smaller than α, this means that the probability to win against Bob or Freddy is not the same.

1 Appendix

1.1 Code of figure 1

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
plot(t$age, t$crimes, main = "Scatterplot Age correlation Crime perception", xlab = "Age",
    ylab = "Number of Crimes", pch = 19)
```

1.2 Code of figure 2

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
plot(t$income, t$crimes, main = "Scatterplot Income correlation Crime perception",
    xlab = "Income", ylab = "Number of Crimes", pch = 19)
```

1.3 Code of figure 3

```
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
plot(t$income, t$crimes, main = "Scatterplot Income correlation Crime perception",
    xlab = "Income", ylab = "Number of Crimes", pch = 19)
y = t$crimes
x = t$income
b_1 = cor(x, y) * sd(y)/sd(x)
b_0 = mean(y) - mean(x) * b_1
x_l = seq(0, 100, 100/18)
y_l = b_0 + b_1 * x_l
lines(x_l, y_l)
```

1.4 Code of figure 4

```
par(mfrow = c(1, 2))
t = read.table("crimemale.txt", header = T, sep = "\t", dec = ".")
y = t$crimes
x = t$income
b_1 = cor(x, y) * sd(y)/sd(x)
b_0 = mean(y) - mean(x) * b_1
residuals_list = y - (b_0 + b_1 * x)
qqnorm(residuals_list, main = "Normal Q-Q plot of residuals", pch = 19, ylab = "Residuals",
    xlab = "x")
plot(x, residuals_list, ylab = "Residuals", main = "Residual plot", xlab = "Income",
    pch = 19)
```

1.5 Code of Exercise 4b

```
Results = matrix(c(179, 47, 57, 96, 17, 36, 52, 13, 18, 39, 15, 15, 84, 37, 39),
    nrow = 3)
colnames(Results) = c("Bob", "Cecilia", "David", "Emma", "Freddy")
rownames(Results) = c("Won", "Lost", "Draw")
res = chisq.test(Results)
res

##
##  Pearson's Chi-squared test
##
## data:  Results
## X-squared = 10.931, df = 8, p-value = 0.2056
```

1.6 Code of Exercise 4c

```
res$expected

##            Bob  Cecilia    David     Emma   Freddy
## Won  171.16935 90.12097 50.20161 41.73387 96.77419
## Lost  49.06855 25.83468 14.39113 11.96371 27.74194
## Draw  62.76210 33.04435 18.40726 15.30242 35.48387
```

1.7 Code of Exercise 4d

```
Results = matrix(c(179, 47, 57, 84, 37, 39), nrow = 3)
colnames(Results) = c("Bob", "Freddy")
rownames(Results) = c("Won", "Lost", "Draw")
res = chisq.test(Results)
fisher.test(Results)

##
##  Fisher's Exact Test for Count Data
##
## data:  Results
## p-value = 0.07421
## alternative hypothesis: two.sided
```