Solutions exercises Chapter 6

Exercise 1

Note on the representativeness of the sample: if a sample is randomly drawn from a population, each subject from the population is chosen equally likely. If subpopulations can be identified, it is preferable that the subpopulations are represented proportionally in the sample, but this cannot be guaranteed in advance. In this case, if we sample randomly, one would expect that about 27% of the sampled subjects are drawn from subpopulation 1, 18% from 2, etc.

Applying the Chi-squared test we can verify whether the observed numbers for each subpopulation deviate significantly from the expected number: we define the seven numbers $N_1, N_2, \dots, N_7$ as follows: N_1 is the number of draws from the subpopulation 1, N_2 is the number of draws from the subpopulation 2, etc.

Pearson's χ^2-test in eight steps:

1. The numbers $N_1, N_2, \dots, N_7$ have a multinomial distribution with probabilities $p_1, p_2, \dots, p_7$.
2. Test $H_0: p_1 = 0.27, p_2 = 0.18, p_3 = 0.15, p_4 = 0.14, p_5 = 0.10. p_6 = 0.09$ and $p_7 = 0.07$ against H_1: "at least one of the p_i deviates from the values under H_0", with $\alpha = 5\%$
3. Test statistic: $\chi^2 = \sum_{i=1}^{2} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$, with expected numbers $E_0 N_i = np_{i0} = 150 p_{i0}$ under H_0
4. If H_0 is true, χ^2 a Chi-square distribution with number of degrees of freedom $df = 7 - 1 = 6$.
5. Observed value of χ^2:
   $\chi^2 = \frac{(43-40.5)^2}{40.5} + \frac{(27-27.0)^2}{27.0} + \frac{(31-22.5)^2}{22.5} + \frac{(20-21.0)^2}{21.0} + \frac{(11-15.0)^2}{15.0} + \frac{(10-13.5)^2}{13.5} + \frac{(8-10.5)^2}{10.5} = 5.982$
6. Reject H_0 if $\chi^2 \geq c$: using $\alpha = 0.05$ and the χ_6^2-table, so c = 12.59
7. The observed value 5.982 does **not** lie in the Rejection Region, so do **not** reject H_0.
8. At a 5% significance level we did not prove that the sample is not representative.
   (*In practice it means that assuming representativeness is reasonable, that is, it is not rejected: note that we did not "prove" representativeness. We could not prove it is not the case….*)

Exercise 2

The mentioned 1×2 −table is:

Decision by	mother	daughter	total
number	243	157	400

- The observed numbers $N_1 = 243$ and $N_2 = 157$ should be compared to the expected values under H_0: $E_0 N_1 = E_0 N_2 = 400 \cdot \frac{1}{2} = 200$, so
- $\chi^2 = \sum_{i=1}^{2} \frac{(N_i - E_0 N_i)^2}{E_0 N_i} = \frac{(243-200)^2}{200} + \frac{(157-200)^2}{200} = 18.49$
- The test is upper-tailed with a critical value c taken from the Chi-square table with $df = 2 - 1 = 1$: $c = 3.84$
- Since the observed value of χ^2 falls in the Rejection Region ($\chi^2 \geq 3.84$), we reject H_0.

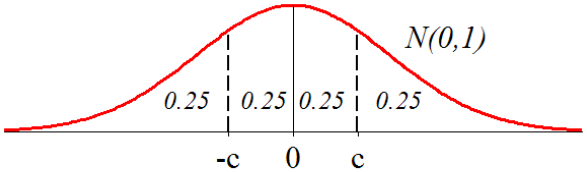
Indeed, $Z^2 = \chi^2$, since $4.3^2 = 18.49$, but $(1.645)^2 \neq c$. We have $1.645^2 \approx 2.71 \neq 3.84$

This is a consequence of the upper-tailed binomial test on p: χ^2 "does not distinguish negative end positive differences of the observed and expected numbers".

Conclusion: for two-sided test the binomial test on p and the χ^2-test are equivalent, for a one-sided test on p the binomial test should be preferred.

Exercise 3

To determine 4 categories (intervals) with probability 0.25 we will use the standard normal table, see the graph.



We are searching a value of c such that $\Phi(c) = 0.75$, so $c = \Phi^{-1}(0.75) = 0.67$.

Interval 1: from $-\infty$ to -0.67

Interval 2: from -0.67 to 0

Interval 3: from 0 to 0.67

Interval 4: from 0.67 to ∞

The observed numbers n_1, n_2, n_3 and n_4 for these 4 intervals are 4, 5, 6 and 5 ($n = 20$ in total), where the expected numbers are $E_0 N_i = n \times 0.25 = 5$ (satisfies the bound of the condition $E_0 N_i \geq 5$).

(Pearson's) Chi-square test:

1. The numbers N_1, N_2, N_3 and N_4 are multinomially distributed with total n=10 and unknown success probabilities p_1, p_2, p_3 and p_4, resp.
2. We test $H_0: p_1 = p_2 = p_3 = p_4 = 0.25$ against $H_1: p_i \neq 0.25$ for at least one i with $\alpha = 0.05$.
3. Test statistic: $\chi^2 = \sum_{i=1}^{4} \frac{(N_i - E_0 N_i)^2}{E_0 N_i}$ met $E_i = 5$
4. Under H_0 χ^2 has a Chi-square distribution with $df = k - 1 = 3$.
5. Observed value: $\chi^2 = \frac{(4-5)^2}{5} + \frac{(5-5)^2}{5} + \frac{(6-5)^2}{5} + \frac{(5-5)^2}{5} = \frac{1}{5} + \frac{1}{5} = 0.4$
6. We will reject H_0 if $\chi^2 \geq c$. $\alpha = 0.05$, so from the χ_3^2-table it follows that $c = 7.81$
7. The observed value 0.4 does not lie in the RR, so do not reject H_0.
8. At a 5% significance level we did not prove that the distribution from which the data are drawn is not the standard normal distribution. (*Note that we did not prove that the distribution is standard normal)*

Exercise 4

We denote the exam result a 1 if it is "low", 2 for "medium" and 3 for "high".
The text provides the following information, presented in the cross table:

	Exam result			
	1	2	3	Row total
Level of education 1	11 (10.2)	20 (17.5)	4 (7.3)	35
Level of education 2	15 (13.1)	18 (22.5)	12 (9.4)	45
Level of education 3	9 (11.7)	22 (20)	9 (8.3)	40
Column total	35	60	25	$120 = n$

The numbers between brackets are the estimates of the expected numbers that we will use later. They are determined with the formula $\hat{E}_0 N_{ij} = \frac{row\ total \times column\ total}{n}$, e.g. for cell (1,1): $\hat{E}_0 N_{11} = \frac{35 \times 35}{120} \approx 10.2$

Define: N_{ij} = "number of employees with educational level i and exam result j" $(i, j = 1, 2, 3)$,
p_{ij} = "probability that an arbitrary person in the population has educational level i and exam result j"

We will apply the **test on independence**, since we have only one sample where two variables (educational level and exam result) are scored.

1. Model: The number $N_{11}, N_{12}, \ldots, N_{33}$ have a multinomial distribution with total $n = 120$ and probabilities p_{ij}.
2. Test $H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ against $H_1: p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$ for at least one pair (i, j), with α = 1%.
   (*remark: $p_{i\cdot}$ and $p_{\cdot j}$ are the row total and column total, related to the cell (i, j), for example, $p_{1\cdot} = p_{11} + p_{12} + p_{13}$*)
3. Test statistic: $\chi^2 = \sum_{j=1}^{3} \sum_{i=1}^{3} \frac{(N_{ij} - \hat{E}_0 N_{ij})^2}{\hat{E}_0 N_{ij}}$, with $\hat{E}_0 N_{ij} = \frac{\text{rowtotal} \times \text{columntotal}}{n}$
4. Under H_0 χ^2 has a chi-square distribution with degrees of freedom $df = (c-1)(r-1) = 4$.
5. For $\hat{E}_0 N_{ij}$ see the table above, so $\chi^2 = \frac{(11 - 10.2)^2}{10.2} + \frac{(20 - 17.5)^2}{17.5} + \cdots + \frac{(9 - 8.3)^2}{8.3} = 4.69$
6. We reject H_0 if $\chi^2 \geq c$. $\alpha = 0.05$, so from the χ_4^2-table it follows that $c = 9.49$.
7. The outcome 4.69 is not in the Rejection Region, so do not reject H_0.
8. At a 5% significance level we could not prove that the exam results depend on the educational level

Exercise 5

We start with producing the table with totals and the $E_{ij} = \hat{E}_0 N_{ij}$ between brackets in each cell:

Type of company	No reply	Reply	Row total
Small	102 (117.3)	98 (82.7)	200
Medium	121 (117.3)	79 (82.7)	200
Large	129 (117.3)	71 (82.7)	200
Column total	352	248	600

We have the results of three samples in this case. We are investigating whether the three populations of companies differ in replying to the questionnaire: a test on the homogeneity of the "Reply"-variable. (This only affects the first two steps of the testing procedure)

Let N_{11} and N_{12} be the numbers of the small companies that gave "No reply" and "Reply", resp. We could state that N_{11} and N_{12} are multionially distributed, but since we have only two categories we will prefer to state that N_{11} binomially distributed with $n = 200$ and success probability p_{11} (implying that $N_{12} \sim B(200, 1 - p_{11})$, $p_{12} = 1 - p_{11}$). *Instead of p_{i1} and p_{i2} we might use the symbols p_i and $1 - p_i$.* Similarly N_{21} and N_{22} for the medium companies and N_{31} and N_{32} for the large companies are introduced. The 8 steps of the **test on the homogeneity (**3 samples, all $n = 200$) are in this case:

1. Model: the numbers N_{11}, N_{21} en N_{31} are and binomially distributed with success probabilities p_{i1}.
2. Test $H_0: p_{11} = p_{21} = p_{31}$ against H_1: "$p_{11} \neq p_{21}$ or $p_{21} \neq p_{31}$" with $\alpha = 1\%$.
3. Test statistic: $\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{3} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$ with estimates $\hat{E}_0 N_{ij} = \frac{row\ total \times column\ total}{n}$.
4. Under H_0 χ^2 has a Chi-square distribution with $df = (r-1)(c-1) = 2$
5. Observed value: $\chi^2 = \frac{(102-117.3)^2}{117.3} + \cdots + \frac{(71-82.7)^2}{82.7} \approx 7.93$ ($for\ E_{ij}\ and\ N_{ij}\ see\ the\ tab$le above)
6. We reject H_0 if $\chi^2 \geq c$. $\alpha = 0.01$, so from the χ_2^2-table it follows that c = 9.21.
7. The observed value 7.93 does not fall in the RR, so do not reject H_0.
8. We cannot state, at a 1% level of significance, that the three types of companies responded differently.

Exercise 6

a. We consider the observations to be drawn as one random sample from a population of PhD-students, where for each PhD student the variables "Gender" and "Promotion in 6 years" are scored: a test on independence of these variables can be applied.

b. The observed numbers N_{ij} and the expected numbers $\hat{E}_0 N_{ij} = \frac{row\ total \times column\ total}{n}$ are the following:

		Gender		
		Female	male	Total
Promotion in 6 years	Yes	N_{11}= 98, $\hat{E}_0 N_{11}$= 116.5	N_{12}= 423, $\hat{E}_0 N_{12}$ = 404.5	521
	No	N_{21}=131, $\hat{E}_0 N_{21}$= 112.5	N_{22} = 372, $\hat{E}_0 N_{22}$ = 390.5	503
	Total	229	795	1024 = *n*

1. Model: the numbers N_{11}, N_{12}, N_{21} and N_{22} are multinomially distributed with total $n = 1024$ and (unknown) probabilities p_{11}, p_{12}, p_{21}, p_{22}.
2. Test $H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ versus $H_1: p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$, for at least one pair (i, j), with $\alpha = 1\%$
   (*remark: $p_{i\cdot}$ and $p_{\cdot j}$ are the row and column total, related to the cell (i, j), e.g.. $p_{1\cdot} = p_{11} + p_{12}$*)
3. Test statistic: $\chi^2 = \sum_{j=1}^{2} \sum_{i=1}^{2} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with estimates $\hat{E}_0 N_{ij} = \frac{row\ total \times column\ total}{n}$
4. Under H_0 χ^2 has a Chi-square distribution with number of degrees of freedom $df = (r-1)(c-1) = 1$
5. Observed value: $\chi^2 = \frac{(98-116.5)^2}{116.5} + \cdots + \frac{(372-390.5)^2}{390.5} \approx 7.703$ ($\hat{E}_0 N_{ij}$ is in the table above)
6. We reject H_0 if $\chi^2 \geq c$. $\alpha = 0.05$, so from the χ_1^2-table it follows that c = 3.84.

7. The observed value 7.703 lies in the Rejection Region, so reject H_0.
8. At a 5% level of significance we showed that the probability of a promotion within 6 years depends on the gender.

c. The z-score in exercise 3 of chapter 5: $Z = -2.77$, so $Z^2 \approx 7.67$, which equals $\chi^2 \approx 7.70$ (small difference because of rounding)
The test was: reject H_0, if $|Z| \geq 1.96 \Leftrightarrow \ Z^2 \geq 1.96^2$, which matches the Chi-square test $\chi^2 \geq 3.84$.

Exercise 7

a. We apply the **Chi-square test on independence** toe (we have one sample and 2 variables in this case): We will add the row and column totals and the expected numbers (assuming independence) between brackets.

	Very high	high	perhaps	Not considerably	Not at all	Row total
> 3 times per week	73 (67.5)	140 (141.4)	223 (227.6)	185 (192.4)	132 (124.2)	**753**
1 – 3 times per week	19 (17.7)	39 (37.0)	56 (59.5)	54 (50.3)	29 (32.5)	**197**
3 times a month or less	2 (8.9)	18 (18.6)	38 (29.9)	9 (25.3)	12 (16.3)	**99**
Column total	**94**	**197**	**317**	**268**	**173**	**1049**

1. Model: the numbers $N_{11}, N_{12}, \ldots, N_{35}$ have a multinomial distribution with total $n = 1049$ and probabilities $p_{11}, p_{12}, \ldots, p_{35}$.
2. Test $H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ against $H_1: p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$, for at least one pair (i, j), with α = 5%.
   (*remark: $p_{i\cdot}$ and $p_{\cdot j}$ are the row and column total, e.g.. $p_{1\cdot} = p_{11} + \cdots + p_{15}$*)
3. Test statistic: $\chi^2 = \sum_{j=1}^{5} \sum_{i=1}^{3} \frac{\left(N_{ij} - \hat{E}_0 N_{ij}\right)^2}{\hat{E}_0 N_{ij}}$, with estimates $\hat{E}_0 N_{ij} = \frac{row\ total \times column\ total}{n}$
4. Under H_0 χ^2 has a Chi-square distribution with $df = (r-1)(c-1) = 8$
5. Observed value (using the values in the table above) $\chi^2 = \frac{(73-67.5)^2}{67.5} + \cdots + \frac{(12-16.3)^2}{16.3} \approx 11.63.$
6. We reject H_0 if $\chi^2 \geq$. $\alpha = 0.05$, so from the χ_8^2-table it follows: c = 15.51.
7. The outcome 11.63 is ***not*** in the rejection region (11.63 < 15.51), so do not reject H_0.
8. At a 5% significance level we could not prove that the need for automatic adjustment of the speed depends on the frequency of the car use.

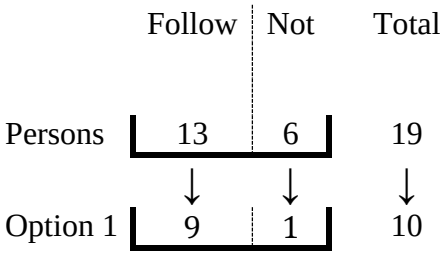
b. 2 is de **observed number** n_{31}, but the estimated **expected value** $\hat{E}_0 N_{31} = 8.9 \geq 5$ is large enough.

Exercise 8

The reasoning is as follows: if the options are equally attractive, the number of 9+4 persons, who follow, will be arbitrarily divided into the two groups of 10 and 9 persons in the two tests. We will reject "arbitrary division" in favour of "option 1 is more attractive" if the probability of 9 or even 10 followers among the option 1 test persons is small.

If $X =$ "# followers among the 10 persons in the option 1 test", then:
the p-value $= P(X \geq 9) = P(X = 9) + P(X = 10)$ (see the diagram)

$$= \frac{\binom{13}{9}\binom{6}{1}}{\binom{19}{10}} + \frac{\binom{13}{10}\binom{6}{0}}{\binom{19}{10}} \approx 4.64\% + 0.31\% = 4.95\%$$



Fisher's exact test for a 2×2 cross table with small numbers:

1. Model: the numbers N_{ij} have a multinomial distribution with total $n = 19$ and success rates p_{ij}
2. Test $H_0: p_{ij} = p_{i\cdot}\, p_{\cdot j}$ (independence) against
   $H_1: p_{ij} \neq p_{i\cdot}\, p_{\cdot j}$, for at least one pair (i, j), with α = 5%.
3. Test statistic: $X =$ "the number of followers among the 10 persons in the option 1 test"

4. Under H_0 X has a an hypergeometric distribution (see diagram for the parameters)
5. Observed: $X = 9$
6. We will reject H_0 if the p-value $\leq \alpha = 0.05$.

   p-value = $P(X \geq 9|H_0) = P(X = 9) + P(X = 10) = \frac{\binom{13}{9}\binom{6}{1}}{\binom{19}{10}} + \frac{\binom{13}{10}\binom{6}{0}}{\binom{19}{10}} \approx 4.64\% + 0.31\% = 4.95\%$
7. p-value < $\alpha = 5\%$, so the null hypothesis is rejected.
8. At a 5% level of significance the hypothesis "equal effect in options 1 and 2" is rejected in favour of the alternative "option 1 is more attractive than option 2".