Assignment 3 Chi-Square

Martin Ottosson

2023-11-24

Analysis of count data and frequencies using Chi-Square (X^2) tests

Part 1 - Personality and animals

Introduction

This study is about finding out if there is an association between the personality type of students (introvert or extrovert) and which animal they prefer (tiger, bat, rabbit or whale). The research question is if there is a significant association (or relationship) between the personality of students and preferred animal.

Method

The method in the study is a Chi-Square (X^2) test for associations. It is a suitable non-parametric test for categorical data. The Chi-Square test for associations is preferable when expected counts are not known beforehand and we want to test if two or more groups are different from each other when it comes to counts (or frequencies). It is a test where counts of observations are compared to expected counts for different groups and levels of a question (in this case). It is calculated as X^2 = Sum of (Observed counts - Expected counts)^2 / Expected counts.

The data comes from Studiewebben and contains character data of students, their personality type and which animal they prefer.

Result

First a contingency table is created (also called a matrix).

Table 1 - Contingency table for personality type and preferred animal

```
##
##              Bat Rabbit Tiger Whale
##   Extrovert  16     13    20    10
##   Introvert  11     18     2    10
```

The table sets up the data in a form that makes it easy to see how the two groups of students have answered. The Chi-Square test also needs to compare the observed counts in each cell with expected values and with the help of the table it is also easy to sum the counts in each row and column.

Table 2 - Chi-Square test of associations for personality type and preferred animal

```
##
##  Pearson's Chi-squared test
##
## data:  cont_table
## X-squared = 13.662, df = 3, p-value = 0.003403
```

The results of the test are a X^2-value of 13.662, with 3 degrees of freedom and a p-value of 0.003403, which is clearly significant at alpha = 0.05. We can say that there is an association between personality type and preferred animal.

We did not have any expectations beforehand regarding the associations between personality type and preferred animal, but an expected count is being performed. The expected count of each cell is calculated as column sum * row sum / total sum.

Table 3 - Expected counts

```
##
##                Bat Rabbit Tiger Whale
##   Extrovert 15.93  18.29 12.98  11.8
##   Introvert 11.07  12.71  9.02   8.2
```

We can already here see a big difference between the actual frequency and the expected frequency for rabbit and tiger.

We can also calculate the Chi-Square components by checking the residuals.

Table 4 - Chi-Square components rounded to two decimals

```
##
##               Bat Rabbit Tiger Whale
##   Extrovert 0.00   1.53  3.80  0.27
##   Introvert 0.00   2.20  5.46  0.40
```

Here it is confirmed that the largest components come from the rabbit and tiger cells. The sum of each component is equal to the X^2-value.

To show the distribution of the contributions in percantage values each value can be divided by the total value of the test statistic.

Table 5 - Percentage distribution of the Chi-Square components

```
##
##                Bat Rabbit Tiger Whale
##   Extrovert  0.00  11.20 27.79  2.01
##   Introvert  0.00  16.12 39.99  2.89
```

These percentage values can then be shown in a correlation plot.

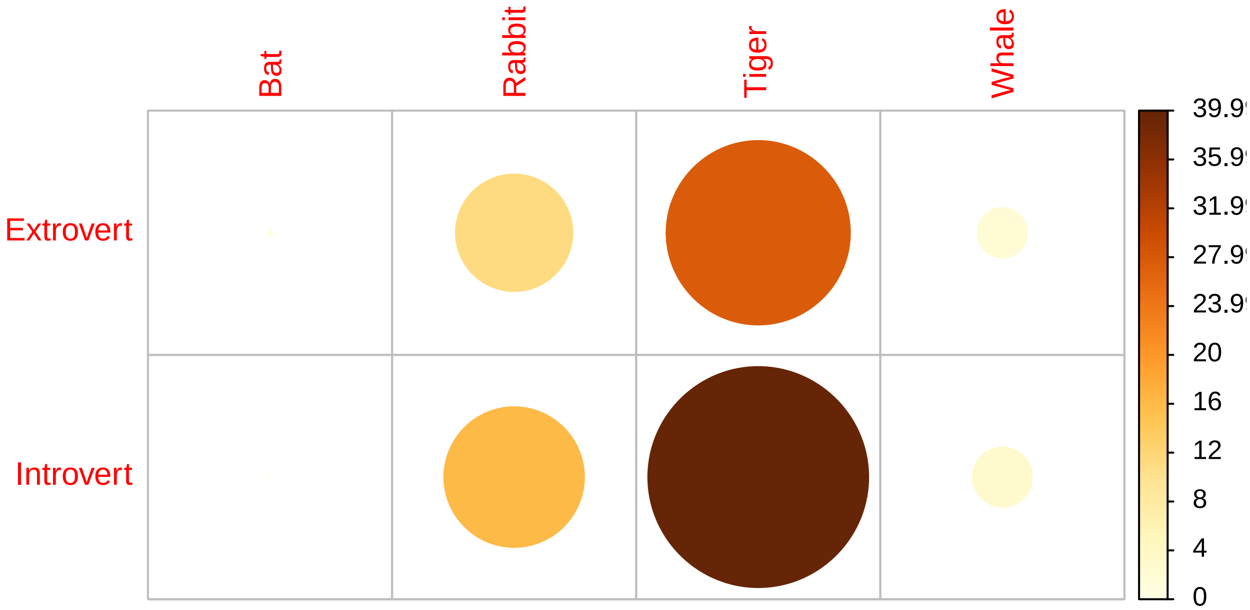

Figure 1 - Correlation plot of Chi-Square contribution shares

Figure 1 above shows a larger contribution the larger and more dark red the circle is. Here it is clear that those considereed introvert and at the same time say tiger stand out most. In second place comes extrovert-tiger and then introvert-rabbit and extrovert-rabbit.

It can be concluded that there is a significant association between the personality of students and preferred animal. Only two introvert students of totally one hundred would prefer tiger and at the same time 20 of those considered to be extrovert would prefer tiger, which is a clear difference. Rabbit on the other hand had 18 of those who are introvert and 13 of those who are extrovert. The expected counts were approximately reversed in the case for rabbits.

Discussion

Since non of the expected counts was below 5 there was no need to do a Fisher's exact test.

The results gave a clearly significant Chi-Square value and it was clear that personality type and preferred animal are associated. A closer look suggested that firstly tiger and secondly rabbit were behind those results. Tiger, a very powerful animal, with no fear, was preferred by those who are considered extrovert. Rabbit, which is a calmer and perhaps more easily scared animal, was preferred by those considered to be introvert. The test does not explain why there is an association, but in this case we could actually assume that the students with the personality type of being extrovert identify themselves more with an animal that is stronger (in muscle power), more direct and courageous like a tiger and the introvert group could very well identify themselves with a smaller, cautious and defensive animal such as the rabbit.

References

Appendix

In this appendix the R code from the study is presented.

Imported libraries

library(corrplot) library(graphics)

importing data and check of dataset

Personality_animal <- read.table(“C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/Statistisk analys och visualisering i R - I/Part 5/Assignment 3/Personality_animal.txt”, header = TRUE, sep = “”)

head(Personality_animal)

Table 1

cont_table <- table(Personality_animal$Personality, Personality_animal$Animal) cont_table

Table 2

chisquare_test1 <- chisq.test(cont_table) chisquare_test1

Table 3

exp_freq1 <- chisquare_test1$expected exp_freq1

Table 4

round(chisquare_test1$residuals^2, 2)

Table 5

contrib1 <- 100*chisquare_test1$residuals^2/chisquare_test1$statistic round(contrib1, 2)

Figure 1

corrplot(contrib1, is.cor = FALSE, cl.align.text = “l”)

Part 2 - Chi-square analysis of parent status and willingness for higher education

Introduction

This study looks at high school students in Portugual and the aim of this study is to find out if there is an association between parent's living status and if the student wants to take higher education. The research hypothesis is that there is an association, perhaps because these students have a more stable situation at home without need for moving between houses, maybe they get more support with home-works or they feel less stress. This could in turn affect how they feel about school and learning in general. The aim, however, is not to find out why, only if there is an association.

Method

The method used is a Chi-Square test for associations which is suitable for categorical data, when expected counts are not known beforehand and we want to test if two or more groups are different from each other when it comes to counts (or frequencies). It is a test where counts of observations are compared to expected counts for different groups and levels of a question (in this case). It is calculated as X^2 = Sum of (Observed counts - Expected counts)^2 / Expected counts.

The dataset comes from Kaggle (Cortez and Silva, 2008). It is a dataset with student achievements for two high schools in Portugal. The dataset contained data of performance in the Portuguese language, but this variable was not used in the Chi-Square test. The dataset was collected from school reports and questionnaires but there is no information about which the schools are. The dataset was already cleaned without much need of pre-processing. It just needed to be downloaded, opened in a text editor and saved as a text file.

The variables used in the Chi-Square test are defined as:

parent_status - parent's cohabitation status (binary: "Living together" or "Apart")

and

higher_ed - wants to take higher education (binary: yes or no)

There are 649 observations in the dataset and it contains students in the ages 15 to 22 (where the majority are between 15-18 years). See figure 1 for a plot over the age distribution.

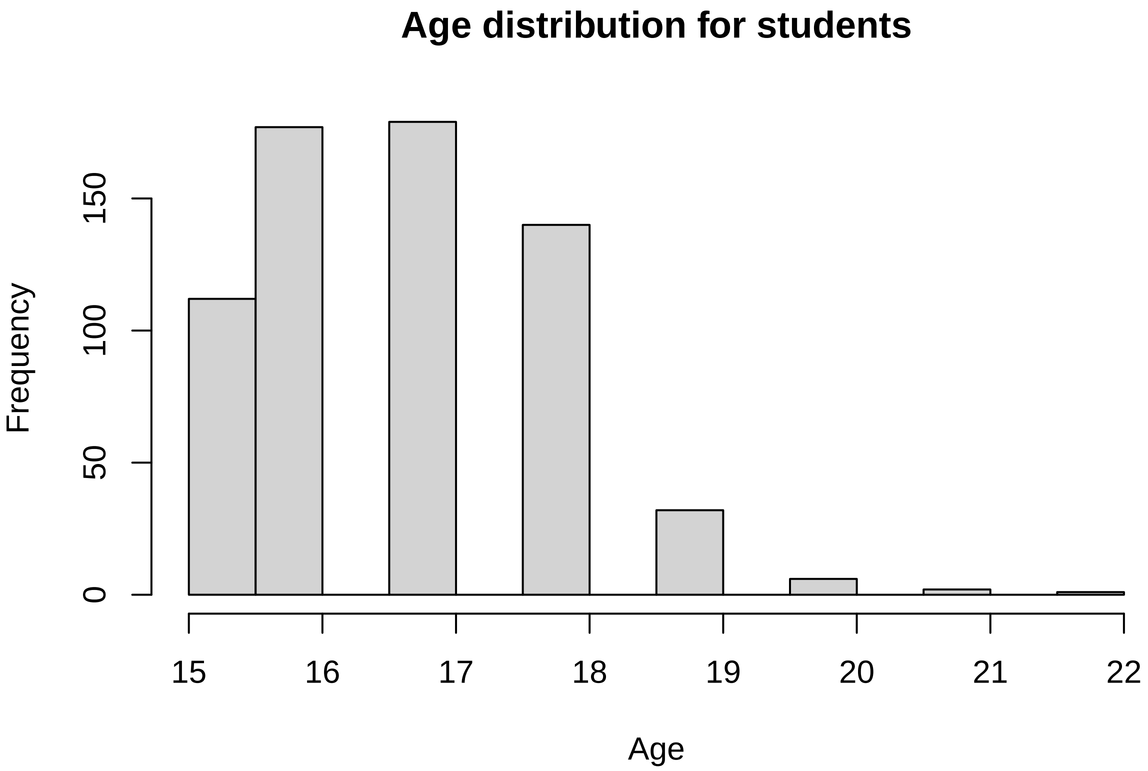

Figure 1 - Age distribution of students

Result

In order to perform the Chi-Square test a contingency table is needed.

Table 1 - Contingency table for parent's living status and willingness for higher education among high school students

```
##
##                    no yes
##   Apart            10  70
##   Living together  59 510
```

The contingency table shows that around 88% of the students with parents living apart had the willingness to study at a higher level. Among the students with parents living together the share was around 90%. The two groups differ in size where the latter group is about 7 times bigger, but this glimpse of the data already indicate that the groups do not differ much when it comes to willingness for higher education. We can suspect that there is no association.

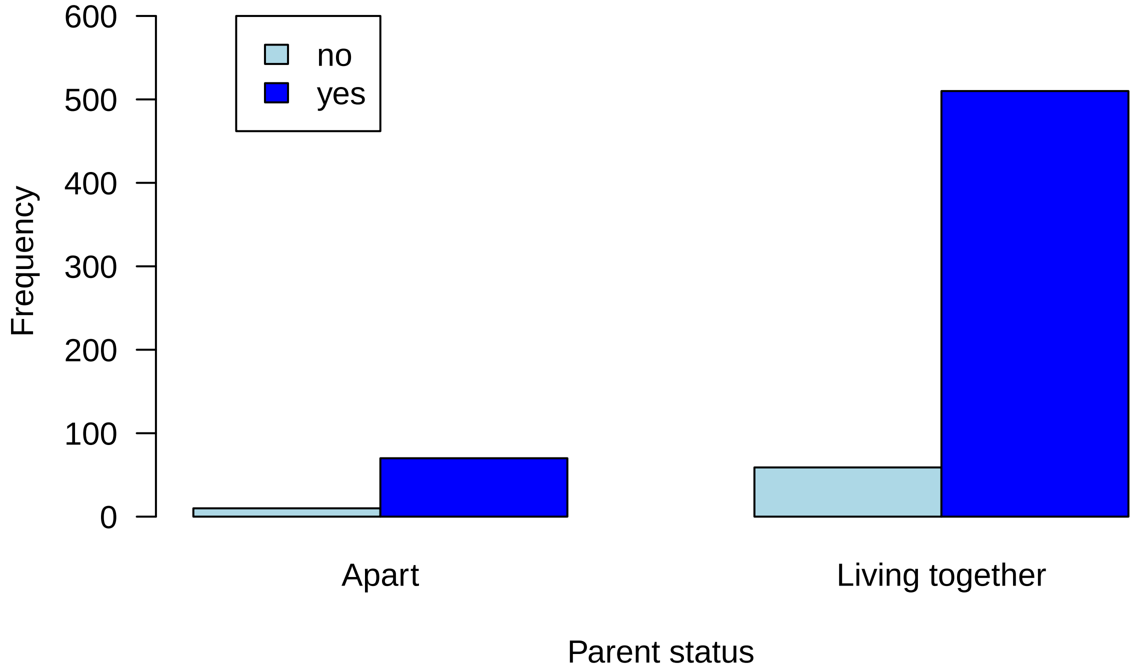

Figure 2 - Contingency table presented as a bar chart

A barplot also visualize the frequencies of the two groups of students. Here the size of the groups are clearly visible and it can also be seen that the willingness to attend higher education is in clear majority for both groups.

The Chi-Square test will tell with almost certainty if the interpretation of the pre-results are correct.

Table 2 - Chi-Square test for association between parent's living status and willlingness for higher education

```
## 
##  Pearson's Chi-squared test
## 
## data:  cont_table2
## X-squared = 0.3352, df = 1, p-value = 0.5626
```

The Xˆ2 value is 0.3352, with 1 df and with p-value = 0.5626. Since the obtained test value does not exceed the critical value of 3.84, at the alpha = 0.05 level we can not reject the null hypothesis. Parental status and higher education are independent, i.e. there is no association whether if parents live apart or together and if the student wants to take higher education.

Table 3 - Expected frequencies

```
## 
##                          no       yes
##   Apart            8.505393  71.49461
##   Living together 60.494607 508.50539
```

Since non of the expected counts in any cell was less than 5, it was not necessary with any correction. The smallest expected count was for parents living apart where the student had no ambitions to go on to higher education. It was counted as row sum * column sum / total sum = 80 * 69 / 649 = 8.505393.

The expected counts are also close to the actual counts in every cell. The actual counts are slightly higher for apart-no and living together-yes and just below for the other two cells. The difference is around 1.5 for each cell.

Table 4 - Pearson residuals from Chi-Square test results

```
##
##                           no        yes
##   Apart            0.5124835 -0.1767626
##   Living together -0.1921625  0.0662795
```

The Pearson residuals show the direction of the difference against the actual counts. As stated earlier the group with students who have parents living apart and do not want to take higher education has a positive value, just as the group with parents living together and who want to continue study further on. The other two are negative.

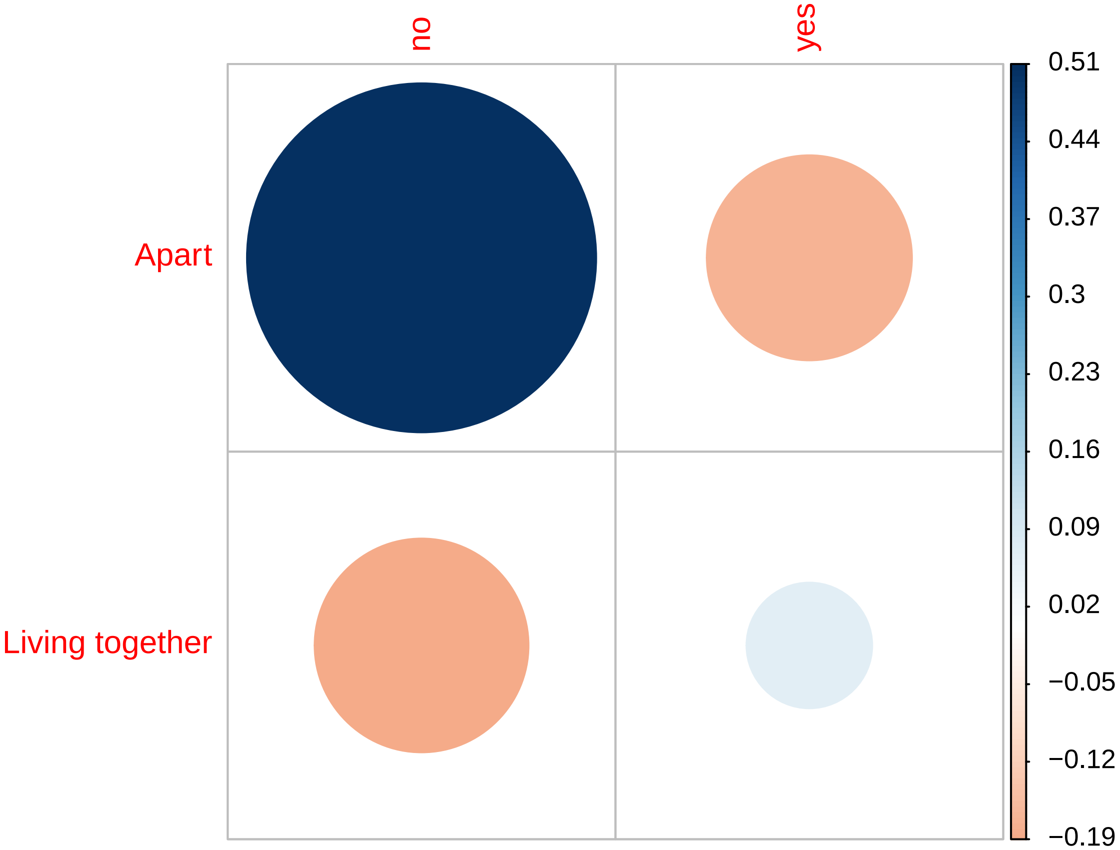

Figure 3 - Pearson residuals visualized as a correlation plot

A correlation plot shows this more easily. The blue circles have positive residuals with actual counts above expected counts. The big and dark one differs from the other circles due to its relative difference, since all deviatiates about one and a half count between the observed and expected value.

Discussion

The research hypothesis for this study was that there is an association between the parent's living status and whether the student wanted to go on to further education. No association was found and there were similar shares in the two groups who wanted to go on to further education. The Chi-Square test was far away from significance and the null hypothesis of no association could not be rejected. We conclude there is no difference.

The data set only consisted of data from two unknown Portuguese schools and it is hard to tell if the sample is representative of Portuguese students at a national level, therefore another test with another dataset could tell a different story. The two groups were also very different in size and the big part of the students had parents living together. Perhaps the difference between the two groups would have been smaller in a less conservative country like for example Sweden. Small group sizes can become a statistical problem, however since the Chi-Square test is non-parametric it is less sensitive to this issue. We do for example not need to assume normally distributed residuals.

The results showed higher observed counts compared to expected counts for the group with parents living apart and said no to higher education and the same for the group with parents living together and said yes to higher education (figure 3). However, these differences were not significant according to the Chi-Square test. The figure actually could give the impression of an association and it is a good example of the need for an actual hypothesis test.

References

P. Cortez and A. Silva. (2008). Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

https://www.kaggle.com/datasets/dillonmyrick/high-school-student-performance-and-demographics/

Appendix

In this appendix the R code from the study is presented.

Imported libraries

library(corrplot) library(graphics)

Importing a checking data

student <- read.table("C:/Users/mart0/OneDrive/Dokument/Martin/Södertörns högskola/Statistisk analys och visualisering i R - I/Part 5/Assignment 3/student_portuguese_clean.txt", header = TRUE, sep = ",")

head(student)

Figure 1

hist(student$age, main = "Age distribution for students", xlab = "Age")

Table 1

cont_table2 <- table(student$parent_status, student$higher_ed) cont_table2

Figure 2

```
barplot(t(cont_table2), beside=T, xlab=“Parent status”, ylab=“Frequency”, legend.text=T, args.legend =
list(x=2, y=600), ylim=c(0,1.2*max(cont_table2)), col=c(“lightblue”,“blue”), las=1)
```

Table 2

```
chisquare_test2 <- chisq.test(cont_table2, p = c(1/4, 1/4, 1/4, 1/4), correct = FALSE) chisquare_test2
```

Table 3

```
exp_freq2 <- chisquare_test2$expected exp_freq2
```

Table 4

```
chisquare_test2$residuals
```

Figure 3

```
corrplot(chisquare_test2$residuals, is.cor = FALSE, cl.align.text = “l”)
```