Report using Chi-square as a statistical test

Maria Kunevich

2022-11-24

Chi-square test

The chi-square test of independence is used to analyse the frequency table (i.e. contingency table) formed by two categorical variables. The chi-square test evaluates whether there is a significant association between the categories of the two variables.

In the dataset *Life in Estonia* I can explore several variables for associations, as there is a number of categorical variables. I want to know if there is a relationship between what people do as their main activity in life and the level of satisfaction with life (on a 10.point scale, including 0).

There are several categories for the variable *Main Activity*:
1. work/have a paid job
2. study
3. jobless, but looking for a job
4. jobless, but not looking for a job
5. disabled
6. retired
7. are in military service 8. stay at home 9. other

Research Question:

For the analysis, the following research question is formulated:
Is there a statistically significant relationship between what people in Estonia do as their main activity during the day and their level of satisfaction with life?

To answer this question we need to create a contingency table for these two variables:

```
##
##          0    1    2    3    4    5    6    7    8    9   10  Sum
##   1     31   21   58  116  117  348  264  501  809  512  159 2936
##   2      1    3    4    8   19   30   41   82  168  123   53  532
##   3      9   13   21   25   30   49   26   36   28   14   11  262
##   4      3    6   12    7   13   24    9   15   21    6    1  117
##   5      8    5    7   12   14   25   10   13   21    8    5  128
##   6     27   22   46   78   96  216  155  206  300  214  124 1484
##   7      0    0    0    0    0    1    2    3    2    0    1    9
##   8      6    4   11   16   13   21   22   42   53   35   17  240
##   9      3    0    3    6    6    9    4   14   16   11    9   81
##   Sum   88   74  162  268  308  723  533  912 1418  923  380 5789
```

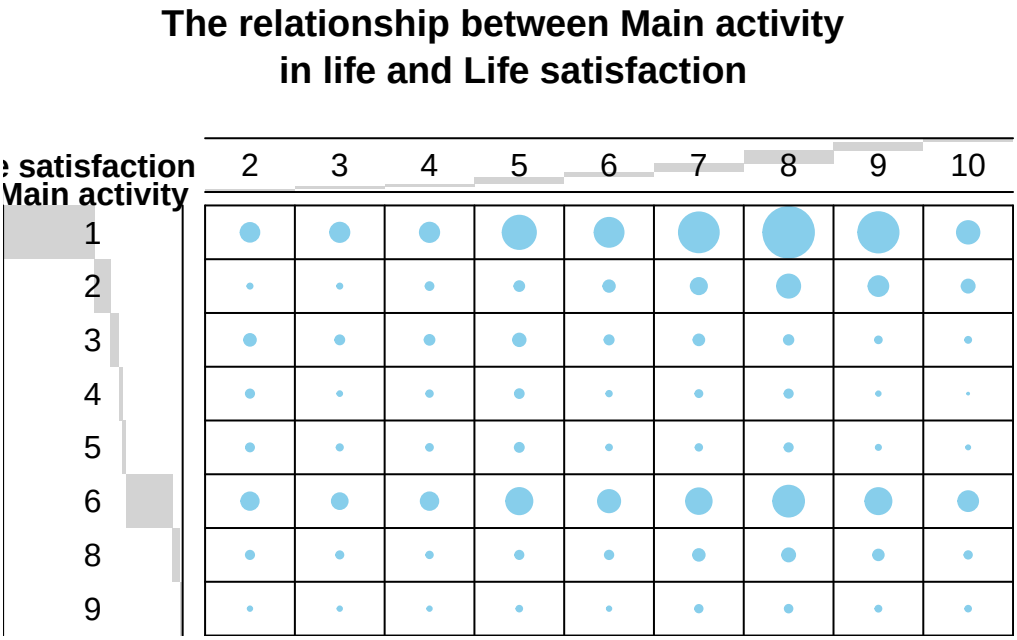
The contigency table provides us with information that for some categories there are only few answers (in particular, less than 5). This may result in violation of the assumption for chi-square test, so I need to check what are the expected values for each cell.
The Chi square test used for the contingency table requires at least 80% of the cells to have an expected count

greater than 5 or else the sum of the cell Chi squares will not have a Chi square distribution and the test p-value will not be valid.

```
chisq <- chisq.test(observed.table)
chisq
chisq$expected
```

By examining the expected values, I can conclude that for categories 7 and 9 there is not enough data, so I combine the category `military service` with the category `other`. I also combine the first three levels of satisfaction (0 to 2) into one level `2`.

The contingency table can be visualised using the function *balloonplot()* in *gplots package*. This function draws a graphical matrix where each cell contains a dot whose size reflects the relative magnitude of the corresponding component.



From the plot we can see that people who have a paid job and retired people tend to be more satisfied with life.

However, I'm interested in statistically significant associations between rows and columns of the contingency table. So I need to perform a *Chi-square test* to examine this.

Null hypothesis: the row and the column variables of the contingency table are independent.

Alternative hypothesis: row and column variables are dependent.

```
## Warning in chisq.test(observed.table.1): Chi-squared approximation may be
## incorrect
```

```
##
##  Pearson's Chi-squared test
##
## data:  observed.table.1
## X-squared = 457.87, df = 56, p-value < 2.2e-16
```

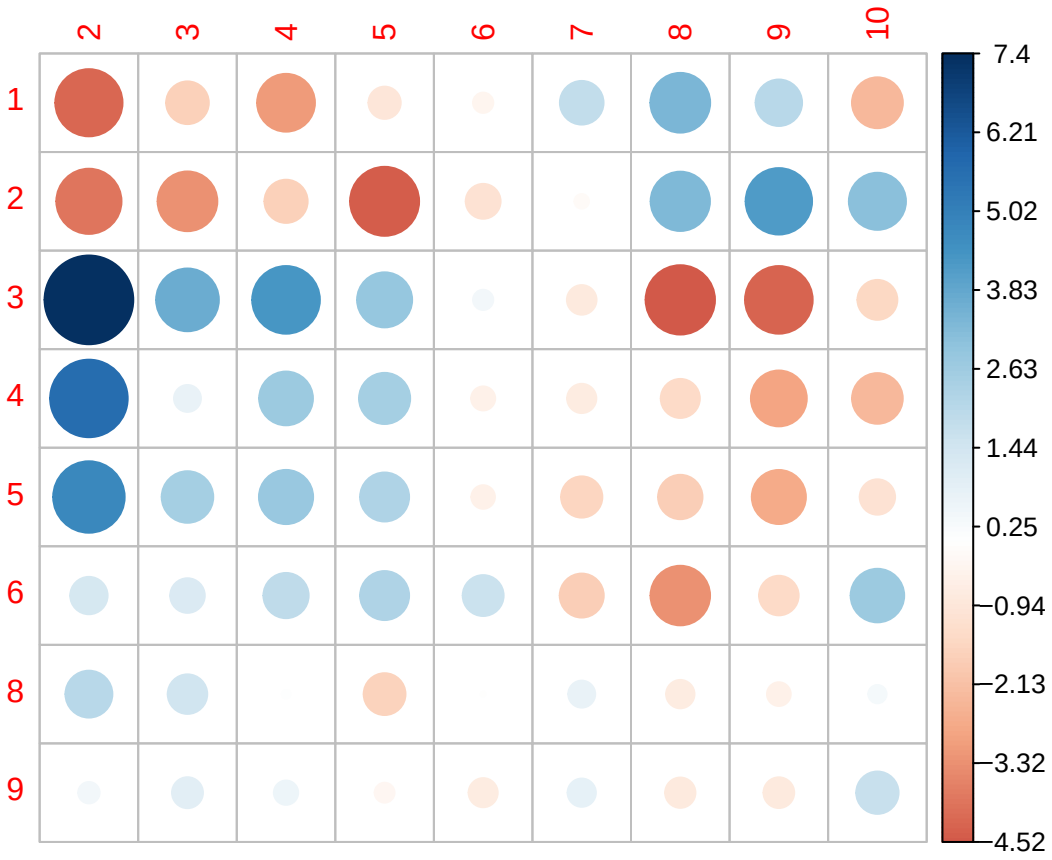
```
##
##              2          3          4         5          6         7         8
##   1 164.322681 135.921230 156.207981 366.68302 270.320954 462.53792 719.16531
##   2  29.775091  24.628779  28.304716  66.44256  48.981862  83.81137 130.31197
```

```
##    3  14.663672  12.129211  13.939541  32.72171  24.122646  41.27552  64.17620
##    4   6.548281   5.416480   6.224909  14.61237  10.772327  18.43220  28.65884
##    5   7.163932   5.925721   6.810157  15.98618  11.785110  20.16514  31.35326
##    6  83.056832  68.701330  78.955260 185.33978 136.633615 233.78960 363.50181
##    8  13.432372  11.110727  12.769045  29.97409  22.097081  37.80964  58.78736
##    9   5.037139   4.166523   4.788392  11.24028   8.286405  14.17861  22.04526
##
##             9         10
##    1 468.11677 192.724132
##    2  84.82225  34.921403
##    3  41.77336  17.198134
##    4  18.65452   7.680083
##    5  20.40836   8.402142
##    6 236.60943  97.412334
##    8  38.26568  15.754016
##    9  14.34963   5.907756
```

In our dataset, the row and the column variables are statistically significantly associated (p-value = 0). The total `Chi-square statistic` is *457.87*.

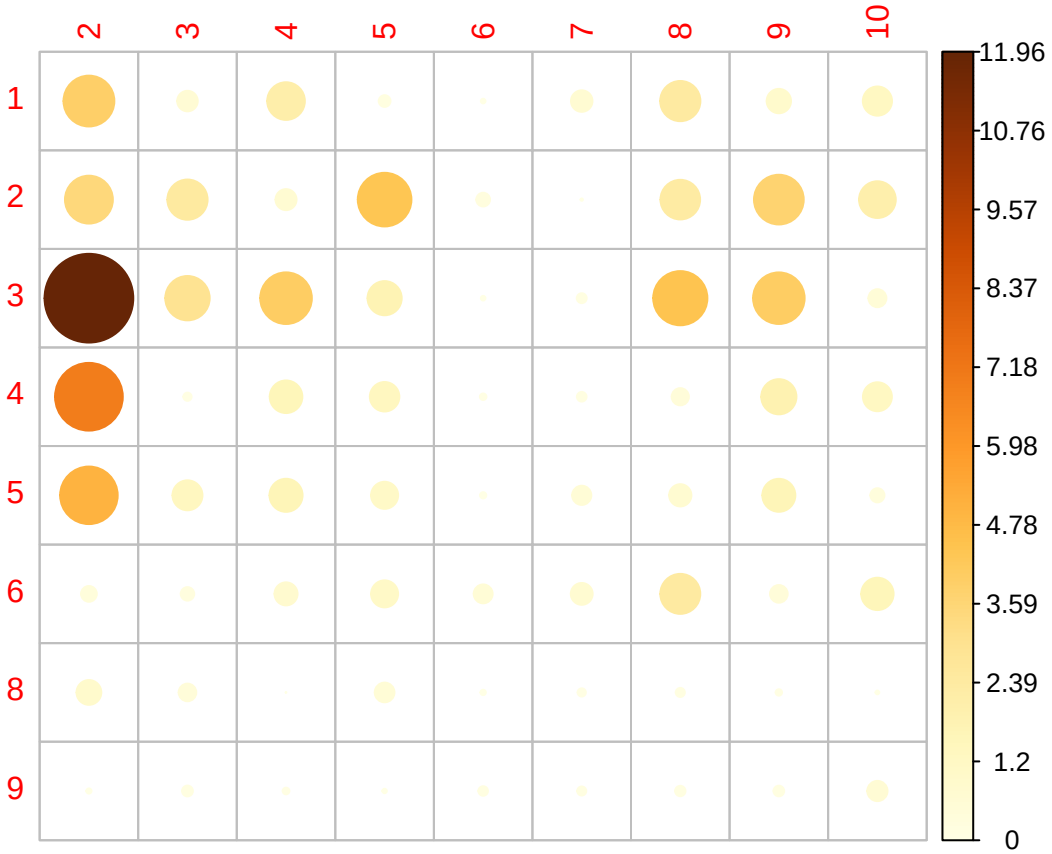
There is still a warning message that Chi-square approximation may be incorrect, so I check the expected values: there are only two cells where expected values are lower than 5, it's less than 20%, so we may disregard the warning.

I would also like to know the most contributing cells to the total Chi-square score, which is *457.87*, so I calculate the `Chi-square statistic` for each cell with the formula that returns the so-called `Pearson residuals (r)` for each cell (or standardized residuals). Cells with the highest absolute standardized residuals contribute the most to the total Chi-square score.



For each cell, the size of the circle is propotional to the amount of the cell contribution. Blue color on the plot indicates that the observed value is higher than the expected value if the data were random, i.e. these are positive residuals. They specify positive association between the corresponding row and column variables. Red color represents negative residuals, which implies negative association between the corresponding row and column variables. For instance, the column `Life satisfaction: 2-3-4-5` are negatively associated (i.e. "not associated") with the row `Main activity: studying`.
From the residuals plot we can see the main contributors: students for life satisfaction (scale 8, 9, 10) as well as people who have a paid job (scale 8) and retired people (scale 10). But people who are jobless and are looking for a job together with disable people and people who are jobless and are not looking for a job are main contributors to the scale 2 (very dissatisfied with their life) and to some degree to scales 3-4-5.
Visualising the same contribution in percentages, the group people who are jobless and are looking for a job seems to be the most dissatisfied as its contribution to `Chi-square statistic` is the largest.

```
##
##         2     3     4     5     6     7     8     9    10
##   1  3.92  0.64  2.15  0.21  0.03  0.70  2.45  0.90  1.29
##   2  3.48  2.45  0.67  4.37  0.28  0.01  2.38  3.75  2.04
##   3 11.96  2.98  4.04  1.77  0.03  0.15  4.45  4.03  0.49
##   4  6.97  0.10  1.61  1.32  0.06  0.14  0.45  1.87  1.27
##   5  5.02  1.36  1.66  1.11  0.06  0.56  0.75  1.65  0.30
##   6  0.38  0.27  0.80  1.11  0.54  0.72  2.42  0.47  1.58
##   8  0.93  0.47  0.00  0.59  0.00  0.10  0.12  0.06  0.02
##   9  0.04  0.18  0.07  0.03  0.14  0.12  0.16  0.17  0.62
```



The relative contribution of each cell to the total Chi-square score gives some indication of the nature of the dependency between rows and columns of the contingency table. From the image above, it can be seen that the most contributing cells to the Chi-square are *Jobless looking for a job/2* (11.96%),*Jobless but not looking for a job/2* (6.97%), *Disabled/2* (5.02%), *Student/5* (4.37%),*Student/9* (3.75%), *Jobless looking for a job/4*

(4.04%), *Jobless looking for a job/8* (4.45%), *Jobless looking for a job/9* (4.03%).
These cells contribute about 39% to the total Chi-square score and thus account for most of the difference between expected and observed values.

Conclusion:

A chi-square test of independence was performed to examine the relation between `Main activity` and `Satisfaction with life`. The relation between these variables was statistically significant, X^2 (80, N = 5789) = 457.87, p < 0. Students and people with a paid job as well as pensioners were more likely to rate their level of life satisfaction higher. At the same time, jobless people and disabled people were more likely to be dissatisfied with their lives.