Chi-square test of independence in R

How to test independence between two qualitative variables

Antoine Soetewey Jan 26, 2020 · 5 min read ★

Photo by Giorgio Tomassetti

Introduction

This article explains how to perform the Chi-square test of independence in R and how interpret its results. To learn more about how the test works and how to do it by hand, I invite you to read the article "Chi-square test of independence by hand".

To briefly recap what have been said in that article, the Chi-square test of independence tests whether there is a relationship between two categorical variables. The null and alternative hypotheses are:

- H0: the variables are independent, there is **no** relationship between the two categorical variables. Knowing the value of one variable does not help to predict the value of the other variable
- H1: the variables are dependent, there is a relationship between the two categorical variables. Knowing the value of one variable helps to predict the value of the other variable

The Chi-square test of independence works by comparing the observed frequencies (so the frequencies observed in your sample) to the expected frequencies if there was no relationship between the two categorical variables (so the expected frequencies if the null hypothesis was true).

Data

For our example, let's reuse the dataset introduced in the article "Descriptive statistics in R". This dataset is the well-known `iris` dataset slightly enhanced. Since there is only one categorical variable and the Chi-square test requires two categorical variables, we add the variable `size` which corresponds to `small` if the length of the petal is smaller than the median of all flowers, `big` otherwise:

```
dat <- iris
dat$size <- ifelse(dat$Sepal.Length < median(dat$Sepal.Length),
  "small", "big"
)
```

We now create a contingency table of the two variables `Species` and `size` with the `table()` function:

```
table(dat$Species, dat$size)
```

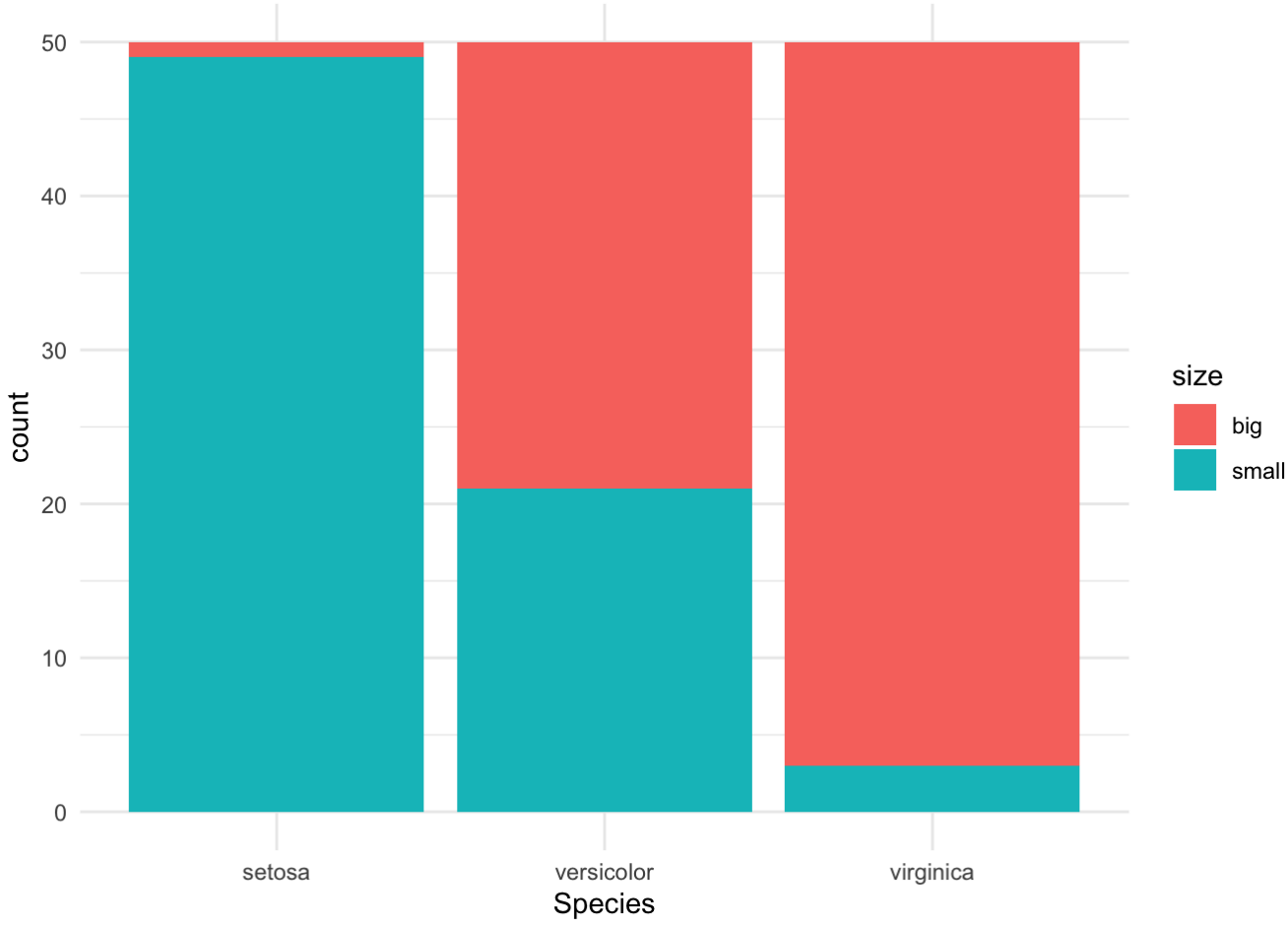
```
##
##              big small
##   setosa       1    49
##   versicolor  29    21
##   virginica   47     3
```

The contingency table gives the observed number of cases in each subgroup. For instance, there is only one big setosa flower, while there are 49 small setosa flowers in the dataset.

It is also a good practice to draw a barplot to visually represent the data:

```
library(ggplot2)

ggplot(dat) +
  aes(x = Species, fill = size) +
  geom_bar() +
  scale_fill_hue() +
  theme_minimal()
```



See the article "Graphics in R with ggplot2" to learn how to create this kind of barplot in `{ggplot2}`.

Chi-square test of independence in R

For this example, we are going to test in R if there is a relationship between the variables `Species` and `size`. For this, the `chisq.test()` function is used:

```
test <- chisq.test(table(dat$Species, dat$size))
test

##
##  Pearson's Chi-squared test
##
## data:  table(dat$Species, dat$size)
## X-squared = 86.035, df = 2, p-value < 2.2e-16
```

Everything you need appears in this output: the title of the test, what variables have been used, the test statistic, the degrees of freedom and the p-value of the test. You can also retrieve the χ2 test statistic and the p-value with:

```
test$statistic # test statistic

## X-squared
##  86.03451

test$p.value # p-value

## [1] 2.078944e-19
```

If you need to find the expected frequencies, use `test$expected`.

If a warning such as "Chi-squared approximation may be incorrect" appears, it means that the smallest expected frequencies are lower than 5. To avoid this issue, you can either:

- gather some levels (especially those with a small number of observations) to increase the number of observations in the subgroups, or
- use the Fisher's exact test

Fisher's exact test does not require the assumption of a minimum of 5 expected counts. It can be applied in R thanks to the function `fisher.test()`. This test is similar to the Chi-square test in terms of hypothesis and interpretation of the results. Learn more about this test in this article dedicated to this type of test.

For your information, there are two other methods to perform the Chi-square test of independence. One with the `summary()` function and one with the `assocstats()` function from the `{vcd}` package:

```
# second method:
summary(table(dat$Species, dat$size))

## Number of cases in table: 150
## Number of factors: 2
## Test for independence of all factors:
##  Chisq = 86.03, df = 2, p-value = 2.079e-19

# third method:
library(vcd)
assocstats(table(dat$Species, dat$size))

##                     X^2 df P(> X^2)
## Likelihood Ratio 107.308  2        0
## Pearson           86.035  2        0
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.604
## Cramer's V        : 0.757
```

All three methods give the same results.

If you do not have the same *p*-values with your data across the different methods, make sure to add the `correct = FALSE` argument in the `chisq.test()` function to prevent from applying the Yate's continuity correction, which is applied by default in this method.1

Conclusion and interpretation

From the output and from `test$p.value` we see that the p-value is less than the significance level of 5%. Like any other statistical test, if the p-value is less than the significance level, we can reject the null hypothesis.

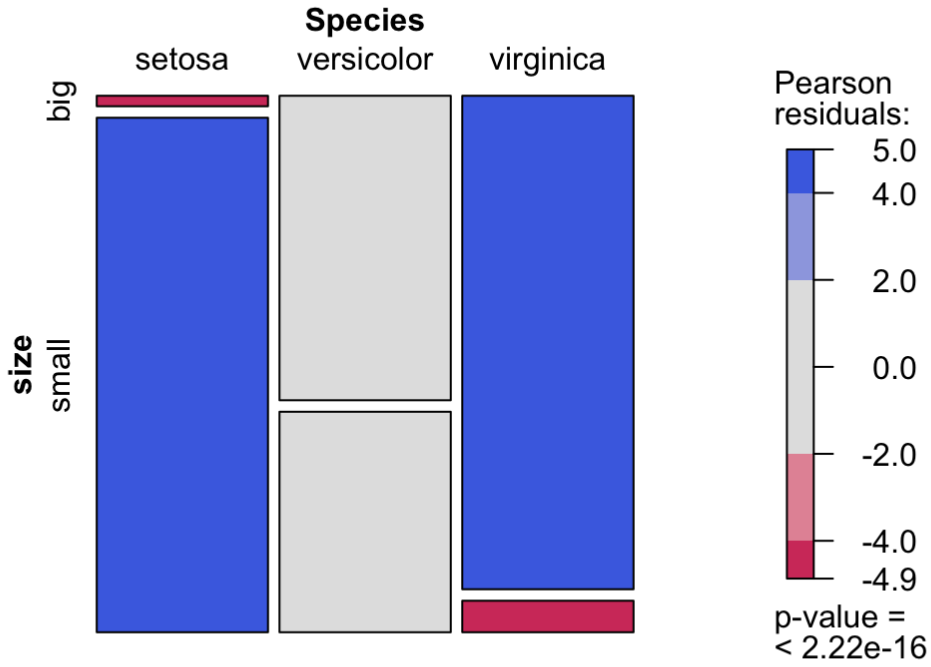
⇒ In our context, rejecting the null hypothesis for the Chi-square test of independence means that there is a significant relationship between the species and the size.

Therefore, knowing the value of one variable helps to predict the value of the other variable.

Combination of plot and statistical test

I recently discovered the `mosaic()` function from the `{vcd}` package. This function has the advantage that it combines a mosaic plot (to visualize a contingency table) and the result of the Chi-square test of independence:

```
library(vcd)
mosaic(~ Species + size,
       direction = c("v", "h"),
       data = dat,
       shade = TRUE)
```



As you can see, the mosaic plot is similar to the barplot presented above, but the *p*-value of the Chi-square test is also displayed at the bottom right.

Moreover, this mosaic plot with colored cases shows where the observed frequencies deviates from the expected frequencies if the variables were independent. The red cases

means that the observed frequencies are *smaller* than the expected frequencies, whereas the blue cases means that the observed frequencies are *larger* than the expected frequencies.

Thanks for reading. I hope the article helped you to perform the Chi-square test of independence in R and interpret its results. If you would like to learn how to do this test by hand and how it works, read the article "Chi-square test of independence by hand".

As always, if you have a question or a suggestion related to the topic covered in this article, please add it as a comment so other readers can benefit from the discussion.

1. Thanks Herivelto for pointing it out.↩️

Related articles:

- An efficient way to install and load R packages
- Do my data follow a normal distribution? A note on the most widely used distribution and how to test for normality in R
- Fisher's exact test in R: independence test for a small sample
- How to create a timeline of your CV in R

Originally published at https://statsandr.com on January 27, 2020.