Statistics 101

Proudly built with R, Quarto and GitHub Copilot

Daniel Sánchez Pazmiño

Table of contents

Non parametric statistics

So far, in order to use the t-test and z-tests, we've made plenty of assumptions for the data. However, there are cases in which we either do not believe the data is distributed in a specific way, or we do not have enough data to make such assumptions. In these cases, we can use non-parametric tests, which do not require us to make any assumptions about the data.

Chi-squared test

The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories. It is used to test the null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution. It is a non-parametric test since it does not require the data to be distributed in a specific way.

When do we use the chi-squared test? On our context, we use it when we have categorical data in the form of frequencies. For example, we could have a sample of 100 people, and we could ask them whether they prefer apples, oranges or bananas. There are several types of chi-squared tests.

Chi-squared test for goodness of fit

One of the most common chi-squared tests is the chi-squared test for goodness of fit. This test is used to determine whether the observed frequencies for a categorical variable match the expected frequencies. For example, we could have a sample of 100 people, and we could ask them whether they prefer apples, oranges or bananas. We could then compare the observed frequencies with the expected frequencies. The null hypothesis is that the observed frequencies match the expected frequencies. The alternative hypothesis is that the observed frequencies do not match the expected frequencies. This test is often also called the *one-sample goodness of fit* test. This is because we only have one sample of observations, and we are testing whether the observed frequencies match the expected frequencies (which are just numbers that we can come up with).

Let us consider an example. Suppose we take a sample of 100 students, and we calculate the proportions of students who pass, fail or withdraw from one specific course. The results from this sampling procedure are:

$$p_{pass} = 0.6$$
$$p_{fail} = 0.3$$
$$p_{withdraw} = 0.1$$

Note that we use probabilities, as we did before when we first learned about them. The probability of a student passing this course will be 0.6, meaning that in 60% of cases a student is predicted to pass the course. We call these the *observed* frequencies.

Now, the expectation is something that comes from outside the sample. Maybe the university has an policy that states that 50% of students should pass the course, 40% should fail and 10% should withdraw. These are the *expected* frequencies.

If you are the professor, you might to argue that your grading guidelines followed the university's rule, and all deviations from it are due to chance. In other words, you want to test the null hypothesis that the observed frequencies match the expected frequencies. The alternative hypothesis is that the observed frequencies do not match the expected frequencies. Present the hypotheses below:

$$H_0 : p_{pass} = 0.5, p_{fail} = 0.4, p_{withdraw} = 0.1$$

$$H_1 : p_{pass} \neq 0.5, p_{fail} \neq 0.4, p_{withdraw} \neq 0.1$$

As every other test, we need to set a significance level. Let us set it to 0.05. We can now calculate the test statistic. The theory tells us that in these cases, the test statistic is called

the chi-square statistic. "Chi" comes from the Greek letter χ, which is pronounced "kai". The chi-square statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed frequency for category i, E_i is the expected frequency for category i, and k is the number of categories. In our case, $k = 3$, since we have three categories: pass, fail and withdraw. Let us calculate the chi-square statistic for our example. We have:

$$\chi^2 = \frac{(60-50)^2}{50} + \frac{(30-40)^2}{40} + \frac{(10-10)^2}{10} = 4$$

The theory says that the chi-square statistic follows a chi-square distribution with $k-1$ degrees of freedom. The chi-square distribution is a family of distributions, just like the t distributions. The chi-square distribution with $k-1$ degrees of freedom is denoted as χ^2_{k-1}. In our case, we have $k = 3$, so we have $k - 1 = 2$ degrees of freedom.

As with our previous tests, after we've calculated the test statistic, we need to calculate the critical value or p-value. The critical value is calculated using the degrees of freedom and the significance level α for the test. We look at the table of chi-square values and find the value that corresponds to the degrees of freedom and the significance level. In our case, we have $\alpha = 0.05$ and $df = 2$. The critical value is 5.99.

What rules should we use for comparing to the critical value? This is a two tailed test, and it is all we will see for the chi-square test of goodness of fit. If the test statistic is greater than the critical value, we reject the null hypothesis. If the test statistic is less than the critical value, we fail to reject the null hypothesis. Since our test statistic is 4, which is less than 5.99, we fail to reject the null hypothesis. We conclude that the observed frequencies match the expected frequencies.

Chi-squared test for independence

The chi-squared test for independence is used to determine whether there is a significant relationship between two categorical variables. Because we are comparing two different categorical variables (read, random variables), this test is also often called the two sample chi-squared test.

For example, we have a sample of 100 students for which we've calculated a pass rate for the course and we've also calculated the proportion/percentage of students who report suffering from mental health issues. We want to know whether there is a relationship between the pass rate and the proportion of students who report suffering from mental health issues. The null hypothesis is that there is no relationship between the pass rate and the proportion of students

who report suffering from mental health issues. The alternative hypothesis is that there is a relationship between the pass rate and the proportion of students who report suffering from mental health issues. Present the hypotheses below:

H_0 : There is no relationship between the pass rate and the proportion of students who report suffering from n

H_1 : There is a relationship between the pass rate and the proportion of students who report suffering from me

For this test, it is important that we tabulate, or summarise, our data in a very specific way. For the example below, we'd do that as follows:

Pass rate	Mental health issues	Total
Pass	Yes	30
Pass	No	30
Fail	Yes	20
Fail	No	20

We're assuming that the withdrawal rate is 0. We need to calculate something called the *expected frequency*. The expected frequency is the frequency that we would expect to see if there was no relationship between the two categorical variables. The expected frequency is calculated as follows:

$$E_{ij} = \frac{R_i \times C_j}{N}$$

where E_{ij} is the expected frequency for row i and column j, R_i is the total number of observations in row i, C_j is the total number of observations in column j, and N is the total number of observations. Let us calculate the expected frequencies for our example. We have:

$$E_{11} = \frac{60 \times 30}{100} = 18$$
$$E_{12} = \frac{60 \times 70}{100} = 42$$
$$E_{21} = \frac{40 \times 30}{100} = 12$$
$$E_{22} = \frac{40 \times 70}{100} = 28$$

We can put this data ina table as follows:

Pass rate	Mental health issues	Total
Pass	Yes	18
Pass	No	42
Fail	Yes	12
Fail	No	28

We can now calculate the test statistic. The test statistic is calculated as follows:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

where O_{ij} is the observed frequency for row i and column j, E_{ij} is the expected frequency for row i and column j, k is the number of rows, and m is the number of columns. In our case, we have $k = 2$ and $m = 2$. Let us calculate the test statistic for our example. We have:

$$\chi^2 = \frac{(30-18)^2}{18} + \frac{(30-42)^2}{42} + \frac{(20-12)^2}{12} + \frac{(20-28)^2}{28} = 8.33$$

As every other test, we need to set a significance level. Let us set it to 0.05.The theory says that the chi-square statistic follows a chi-square distribution with $(k-1)(m-1)$ degrees of freedom. In our case, we have $k = 2$ and $m = 2$, so we have $(k-1)(m-1) = 1$ degree of freedom.

As with our previous tests, after we've calculated the test statistic, we need to calculate the critical value or p-value. The critical value is calculated using the degrees of freedom and the significance level α for the test. We look at the table of chi-square values and find the value that corresponds to the degrees of freedom and the significance level. In our case, we have $\alpha = 0.05$ and $df = 1$. The critical value is 3.84.

What is the comparison rule? If the test statistic is greater than the critical value, we reject the null hypothesis. If the test statistic is less than the critical value, we fail to reject the null hypothesis. Since our test statistic is 8.33, which is greater than 3.84, we reject the null hypothesis. We conclude that there is a relationship between the pass rate and the proportion of students who report suffering from mental health issues.

Summary

- The chi-square test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies.
- The chi-square test for goodness of fit is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies for a single categorical variable.
- The chi-square test for independence is used to determine whether there is a significant relationship between two categorical variables.

Up next

- SPSS I and II