Chi-squared test for independence

Date: 2017-10-10

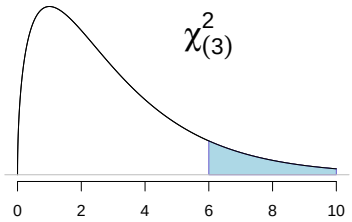

Figure 1: A chi-squared distribution with 3 degrees of freedom.

What it is

The χ^2 test for independence (mathematically equivalent to the χ^2 test for homogeneity of proportions) investigates the relationship between two categorical variables (nominal or ordinal). The example below compares BMI category between genders. This could be equivalently stated as comparing the distribution of a categorical variable across multiple populations (i.e., how BMI is distributed across genders).

When to use it

The χ^2 test requires observations to be sampled independently as well as a large sample. The standard large sample rule of thumb is an expected count of at least 5 in each table cell (see *Why it Works* for details).

How to use it

As with all tests, begin by stating the hypotheses:

H_0: X and Y are independent.

H_A: X and Y are not independent.

Equivalently,

H_0: Y has the same distribution for all levels of X.

H_A: Y does not have the same distribution for all levels of X.

In `R`, create a two-way table using the `table()` function:

```
tab = table(d$sex, d$bmicat)
```

which yields[1]

	underweight	normal	overweight	obese
female	6	102	81	78
male	6	65	99	63

[1] To view the margin totals, use `addmargins(tab)`.

Then pass the table to the `chisq.test()` function:[2]

[2] You want to use the initial table, not the table with the margins.

```
chisq.test(tab)
```

The output gives you, in order, the test statistic (X^2), degrees of freedom, and p-value. More information is kept hidden and can be accessed by saving the test as a new object.

Why it works

The χ^2 test for independence is based on the concept of independence from probability theory[3].

[3] If A and B are independent, then $P(A, B) = P(A) \times P(B)$

Using `addmargins()`, we get

	underweight	normal	overweight	obese	Sum
female	6	102	81	78	267
male	6	65	99	63	233
Sum	12	167	180	141	500

Take for example, the cell in Row 1, Column 1. By independence (assumed in H_0),

$$P(female, underweight) = P(female) \times P(underweight)$$

Using the magins, each probability can be estimated by data:

$$P(female) \approx \frac{267}{500} \quad \text{and} \quad P(underweight) \approx \frac{12}{500}$$

So,

$$P(female, underweight) \approx \frac{267}{500} \times \frac{12}{500} = \frac{(267)(12)}{500^2}$$

The expected count for that cell, E, is [4]

[4] In general, $E = \frac{(row\ total) \times (column\ total)}{n}$

$$E = 500 \cdot \frac{(267)(12)}{500^2} = 500 \cdot \frac{0.013}{500^2} = \frac{0.013}{500}$$

For the χ^2 tests to work well, each E should be at least 5.[5]

[5] To view the expected counts, use `chisq.test(tab)$expected`.

The test statistic then is calculated as

$$X^2 = \sum_{cell\ i} \frac{(O_i - E_i)^2}{E_i}$$

This is always positive. Additionally, the farther the observed O is from the expected E in each cell, the bigger X^2. Therefore, large values of X^2 provide evidence against the null hypothesis.

The degrees of freedom affect the shape of the χ^2 distribution and come from the minimum number of cells needed in order to fix the entire table with the given margin totals. For example, in the table above if female underweight, female normal, and female overweight are known, the rest of the values are fixed (male underweight = 12 - female underweight; male normal = 167 - female normal; male overweight = 180 - female overweight; etc.). Thus, $df = 3$.