Chi-Squared Test for Independence

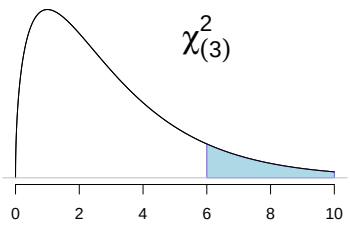

Figure 1: A chi-squared distribution with 3 degrees of freedom.

What it is

The χ^2 test for independence (also called the χ^2 test for homogeneity of proportions) investigates the relationship between two categorical variables (nominal or ordinal). Equivalently, we could say the test compares the distribution of one categorical variable across groups defined by a second categorical variable. The example below compares diabetes across metro status.

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, we could say

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(mydat$metro, mydat$diabetes)
```

which yields[1]

	Yes	No
Urban	55	228
Suburban	27	171
Rural	67	245

[1] To view the row and column 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 or `X-squared`), 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 $P(A) = P(A \mid B)$, or the probability of A is the same for every category defined by B.

Using `addmargins()`, we get

	Yes	No	Sum
Urban	55	228	283
Suburban	27	171	198
Rural	67	245	312
Sum	149	644	793

Take, for example, urban residents with diabetes (Row 1, Column 1). By independence (assumed in H_0), the expected number of urban residents with diabetes is simply the number of urban residents times the probability of anyone in the sample (regardless of metro status) having diabetes[4]:

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

$$E_{urban\ with\ diabetes} = 283 \times \frac{149}{793} = 53$$

We can do this for every cell in the table in the same way. For example, the expected number of rural residents with diabetes would be $312 \times \frac{149}{793} = 58$. For the χ^2 tests to work well, each E should be at least 5.[5]

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

The test statistic then is calculated as

$$X^2 = \sum_{all\ cells} \frac{(O-E)^2}{E}$$

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 urban and suburban diabetes counts are known, the rest of the values are fixed (urban no diabetes = 283 - urban diabetes; suburban no diabetes = 198 - suburban diabetes; rural diabetes = 149 - urban diabetes - suburban diabetes; etc.). Thus, $df = 2$.