Chi-square test of association

Use case

We want to analyze the relationship between two categorical variables, exploring if there is an association between the two. These categorical variables can be—but needn't necessarily be—binary categorical variables.

We represent the possible values of our categorical variables as a two-way table. If one variable can take on
r values, and the other can take on
c values, we can represent our data as an
$r \times c$ two-way table, where each individual entry counts the number of cases with that particular configuration of values.

Preconditions

The approximation to the chi-square distribution is generally valid for a
$r \times c$ table if the following preconditions are met:

1. $\frac{n}{rc} \geq 5$
2. Every expected count is at least 1.

For a
2×2 table, we only need that every expected count is at least 5.

Hypotheses

The hypotheses for a chi-square test of independence look slightly different than our other methods of inference.

Null hypothesis,
H_0**:** there is no association between the two variables (i.e., the conditional distributions are identical).

Alternative hypothesis,
H_a**:** there is an association between the two variables (i.e., the conditional distributions vary).

Test statistic

To compute the test statistic, we first have to compute the expected counts for each cell under
H_0.

$$\text{expected count} = \frac{\text{row total} \times \text{column total}}{n}$$

Next, we compute the test statistic by comparing the expected counts to their actual counts.

$$X^2 = \sum \frac{(\text{observed count} - \text{expected count})^2}{\text{expected count}}$$

p-value

If
H_0 is true, the
X^2 test statistic approximately follows the
$\chi^2(k)$ distribution with degrees of freedom
$k = (r-1)(c-1)$.

We compute our p value as follows:

$$p = P(\chi^2(k) > X^2) \text{ with } k = (r-1)(c-1)$$

$p = 1 - P(\chi^2(df) \geq x^2)$ with $df = (r-1)(c-1)$

p-value (Python)

```
p = stats.chi2.cdf(x2, (r-1) * (c-1))
```

Related tests

If you want to test for association between two binary variables, a two-sample z-test for proportions might be simpler.

Considerations

Performing a chi-square test of association on a
2×2 two-way table will always yield the same result as a two-sided two-sample z-test for proportions. They willl produce the same p-value.