Categorical Data Analysis

California State University - Northridge

Ovande Furtado, Jr., Ph.D.

April 03, 2022

CSUN® | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

```
xaringanExtra::use_webcam()
```

The χ^2 (chi-square) goodness-of-fit test

- It tests whether an observed frequency distribution of a nominal variable matches an expected frequency distribution.
- For example: a group of casual runners have been training for the LA Marathon and have their motivation assessed throughout to see whether motivation has improved, stayed the same or worsened.

A goodness-of-fit test could be used to determine whether the numbers in each category - improved, remained the same, or worsened.

The χ^2 (chi-square) goodness-of-fit test[1]

What it is?

It is a statistical hypothesis test used to determine whether a variable is likely to come from a specified distribution or not. It is often used to evaluate whether a sample data is representative of the population the sample was taken from.

When can it be used?

When testing a single dependent categorical variable and values are expressed in counts.

Variables

Dependent: 1 categorical with J independent groups
Independent: none

[1] Click here to compare with Chi-square test for relationship between 2 categorical variables.

The χ^2 (chi-square) goodness-of-fit test

Hypotheses

Null

H_0: the population proportions in each of the *J* conditions are $p_1, p_2, p_3, \ldots p_j$
or stated differently,
H_0: $p_1 = p_2 = p_3 \ldots p_j$

Alternative

H_a: at least one p_j not equal

Note: When not testing for equal proportions, use the following:
H_0: $p_1 = .3$; $p_2 = .35$; $p_3 = .35$
H_a: at least one p_j not equal to expected value

[1] JMP - Statistical Discovery

The χ² (chi-square) goodness-of-fit test

Assumptions

- Sample size is large enough. All *J* expected cell counts must be 5 or more
- Sample is a simple random sample - observations are independent of one another

Test Statistics

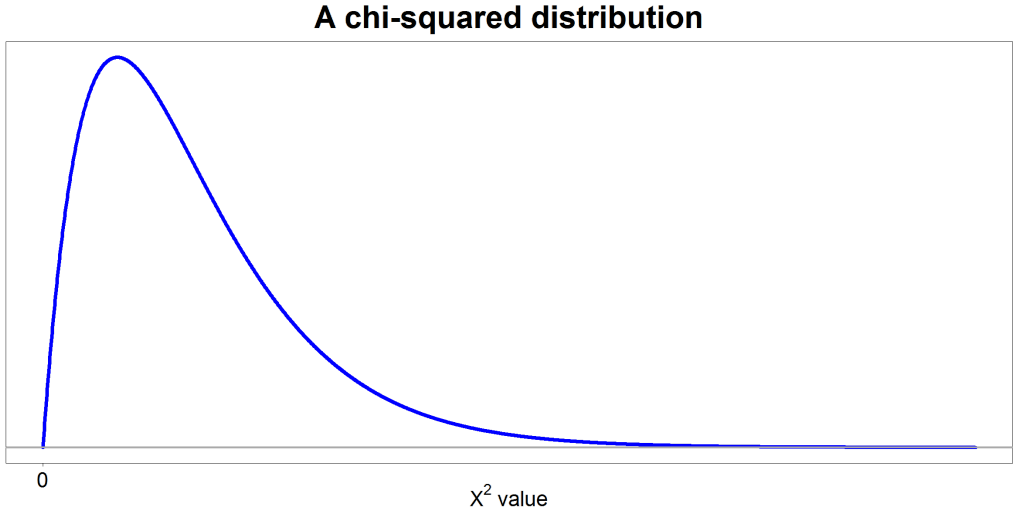
$$\chi^2 = \sum \left[\frac{(f_o - f_e)^2}{f_e} \right]$$

with $k - 1$ degrees of freedom, where:

- k is the number of categories.
- f_o is an observed frequency in a particular category.
- f_e is an expected frequency in a particular category.

The χ^2 (chi-square) goodness-of-fit test

Sampling Distribution



Sampling distribution of X^2 if H_0 were true[1]

- Small scores of X^2 tend to occur most to the time
- Large scores of X^2 tend to occur less often
- So, if we find a X^2 value that is large in our sample and the H_0 is TRUE
- Then, we found evidence against the H_0

[1] www.statkat.com

Practice

- First, read the excerpt below
- Then, find the critical X^2 value[1] associated with the DF, N, and α = 0.05
- Then, confirm if the hypothesis test statement is correct by comparing the observed X^2 value with the X^2 critical value
- Then, using the **Interactive Graphs** for the Chi-square distribution found in StatsKat[2], find:
  - the p-value associated with a X^2 = 3.97

> The observed quarterly birth rate distribution of the participants is displayed in Table 1. There was no statistically significant difference between the observed and expected distribution values, X^2 = (3, *N* = 150) = 3.97, p = [omitted]. While the results do not illustrate a statistically significant variation from the expected random distribution of births, it is interesting to note that the final three quarters were evenly distributed while the first quarter contained the highest number of participants[3]

[1] https://bit.ly/3Kc5NNO

[2] https://statkat.com/interactive-graphs.php

[3] Beals, T. C., Furtado, O., Jr., & Fontana, F. E. (2017). Relative Age Effect and Academic Timing in American Junior College Baseball. Perceptual and Motor Skills. https://doi.org/10.1177/0031512517724260

Doing it in jamovi

- Data set up
- Data file

Phrasing results

Of the 200 participants in the experiment, 64 selected hearts for their first choice, 51 selected diamonds, 50 selected spades, and 35 selected clubs. A χ^2-goodness-of-fit test was conducted to test whether the choice probabilities were identical for all four suits. The results were significant ($\chi^2(3) = 8.44$, $p < 0.05$), suggesting that people did not select suits purely at random.

The χ^2 (chi-square) Test of Independence (Association)

What it is?

The Chi-square test of independence is a statistical hypothesis test used to determine whether two categorical or nominal variables are likely to be related or not.

When can it be used?

You can use the test when you have counts of values for two categorical variables.

Variables

Dependent: One categorical with J independent groups ($J \geqslant 2$)
Independent: One categorical with I independent groups ($I \geqslant 2$)

[1] Click here to compare with Chi-square test for relationship between 2 categorical variables.

The χ^2 (chi-square) Test of Independence (Association)

- Tutorial from StatsKat
- Comparing Goodness of fit with the Test of Association

Doing it

- Data set up
- File can be found here

Phrasing Results

Pearson's χ^2 revealed a significant association between species and choice ($\chi^2(2) = 10.7$, $p < 0.01$). Robots appeared to be more likely to say that they prefer flowers, but the humans were more likely to say they prefer data.

The χ^2 (chi-square) - more

Effect Size

Best option is Cramer's V

The Fisher exact test

Do it if cell counts are too small - less than 5

McNemar's test

More information here

- Doing it in jamovi

Thanks!

CSUN | CALIFORNIA STATE UNIVERSITY NORTHRIDGE

ovandef@csun.edu

@ofurtado

@drfurtado