Chi-square test: hypotheses, assumptions, formula, and calculation

Renesh Bedre 6 minute read Follow @renbedre (https://twitter.com/renbedre?ref_src=twsrc%5Etfw)

Chi-square (χ2) test for independence (Pearson Chi-square test)

- Chi-square test is a non-parametric (distribution-free) method used to compare the relationship between the two categorical (nominal) variables (https://reneshbedre.com/blog/others.html#variable-types) in a contingency table.
- For example, we have different treatments (treated and nontreated) and treatment outcomes (cured and noncured), here we could use the chi-square test for independence to check whether treatments are related to treatment outcomes.
- Note: Chi-square test for independence is different than the chi-square goodness of fit test (https://www.reneshbedre.com/blog/chi-square-test.html#-chi-square-%CF%872-goodness-of-fit-test-)

Formula

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

The χ^2 with Yates' correction for continuity

$$\chi^2{}_{corrected} = \sum_{i}^{n} \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Where, O_i = Observed value in contingency table,
E_i = Expected value for each cell in contingency table,
i = 1,2,..., n

The χ^2 is always positive as it based on squared differences of observed and expected counts

Expected value calculation,

$$E = \frac{Row_T * Column_T}{N}$$

Where, N = Total sample size,
Row_T = Total sample size for row,
$Column_T$ = Total sample size for column,

Hypotheses

- *Null hypothesis*: The two categorical variables are independent (no association between the two variables) (H_0: $O_i = E_i$)
- *Alternative hypothesis*: The two categorical variables are dependent (there is an association between the two variables) (H_a: $O_i \neq E_i$)
- Note: There are no one or two-tailed (https://www.reneshbedre.com/blog/hypothesis-testing.html#one--and-two-tailed-sided-alternate-hypothesis) *p* value. Rejection region of the chi-square test is always on the right side of the distribution.

Learn more about hypothesis testing and interpretation (https://www.reneshbedre.com/blog/hypothesis-testing.html)

Assumptions

- The two variables are categorical (nominal) (https://reneshbedre.com/blog/others.html#variable-types) and data is randomly sampled
- The levels of variables are mutually exclusive
- The expected frequency count for at least 80% of the cell in a contingency table is at least 5
- The expected frequency count should not be less than 1
- Observations should be independent of each other
- Observation data should be frequency counts and not percentages, proportions or transformed data

Calculate a chi-square test for independence in Python

- We will use `bioinfokit` v0.9.5 or later and scipy python packages
- Check bioinfokit documentation (https://github.com/reneshbedre/bioinfokit) for installation and documentation
- Download a hypothetical dataset for chi-square test for independence

Note: If you have your own dataset, you should import it as pandas dataframe. Learn how to import data using pandas (https://www.reneshbedre.com/blog/import-data-pandas.html)

```
# I am using interactive python interpreter (Python 3.7.4)
from bioinfokit.analys import stat, get_data
# load example dataset
df = get_data('drugdata').data
df.head()
# output
   treatments  cured  noncured
0     treated     60        10
1  nontreated     30        25
# set treatments column as index
df = df.set_index('treatments')
# output
df.head()
            cured  noncured
treatments
treated        60        10
nontreated     30        25

# run chi-square test for independence
res = stat()
res.chisq(df=df)

# output
print(res.summary)
# corrected for the Yates' continuity
Chi-squared test for independence

Test              Df    Chi-square      P-value
--------------  ----  ------------  -----------
Pearson            1       13.3365  0.000260291
Log-likelihood     1       13.4687  0.000242574

print(res.expected_df)
```

```
Expected frequency counts

      cured    noncured
--  -------  ----------
 0     50.4        19.6
 1     39.6        15.4

# using chi2_contingency function from scipy package
import numpy as np
from scipy.stats import chi2_contingency
# using Pearson's chi-squared statistic
# corrected for the Yates' continuity
observed = np.array([[60, 10], [30, 25]])
chi_val, p_val, dof, expected =  chi2_contingency(observed)
chi_val, p_val, dof, expected
# output
(13.3364898989899, 0.0002602911116400899, 1, array([[50.4, 19.6],
       [39.6, 15.4]]))

# without Yates' correction for continuity
chi_val, p_val, dof, expected =  chi2_contingency(observed, correction=False)
chi_val, p_val, dof, expected
# output
(14.842300556586274, 0.00011688424010613195, 1, array([[50.4, 19.6],
       [39.6, 15.4]]))

# for log-likelihood method run command as below
chi_val, p_val, dof, expected =  chi2_contingency(observed, lambda_="log-likelihood")
```

Yates' correction for continuity

- In the χ2 test, the discrete probabilities (https://www.reneshbedre.com/blog/probability-distributions.html) of observed counts can be approximated by the continuous chi-squared probability distribution (https://www.reneshbedre.com/blog/probability-distributions.html). This can cause errors and needs to be corrected using continuity correction.
- Yates' correction for continuity modifies the 2x2 contingency table and adjust the difference of observed and expected counts by subtracting the value of 0.5 (see formula (https://www.reneshbedre.com/blog/chi-square-test.html#formula)).
- Yates' correction for continuity increases the *p* value by reducing the χ2 value. The corrected *p* value is close to exact tests such as the Fisher exact test. Sometimes, Yates' correction may give an overcorrected *p* value.
- χ2 and Yates' corrected χ2 produce similar results on large samples, but Yates' corrected χ2 can be conservative on smaller samples and gives a higher *p* value.

Interpretation

The *p* value obtained from chi-square test for independence is significant ($p < 0.05$), and therefore, we conclude that there is a significant association between treatments (treated and nontreated) with treatment outcome (cured and noncured)

Chi-square (χ2) Goodness of Fit test

- Chi-square Goodness of Fit Test test is a non-parametric (distribution-free) method used to compare the observed and expected values from one categorical variable (https://reneshbedre.com/blog/others.html#variable-types). The expected values are calculated based on the known theoretical expectation.
- For example, we have resistant (A) and susceptible (B) genotypes for some disease. The crosses between these two genotypes will produce offspring in 3:1 (75% A and 25% B genotype) as per Mendelian ratio assuming resistance to disease is a dominant trait. Here, we could use the chi-square Goodness of Fit Test test to check whether observed counts of A and B genotypes are similar to expected counts of A and B genotypes as per the Mendelian ratio.

Formula

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

Where, O_i = Observed value for category i,
E_i = Expected value for category i,
i = 1,2,..., n

Expected value calculation,

$$E_i = N * p_i$$

Where, N = Total sample size,
p_i = Known theoretical expectation for category i,
i = 1,2,..., n

Hypotheses

- *Null hypothesis*: The observed and expected counts in each group are equal (H_0: $O_i = E_i$)
- *Alternative hypothesis*: The observed and expected counts in each group are different (H_a: $O_i \neq E_i$)

Learn more about hypothesis testing and interpretation (https://www.reneshbedre.com/blog/hypothesis-testing.html)

Assumptions

- The variable should be categorical (nominal) (https://reneshbedre.com/blog/others.html#variable-types) and data is randomly sampled
- The groups of variables are mutually exclusive
- The expected count should be at least 5 for each group
- Observations should be independent of each other
- Observation data should be frequency counts and not percentages or transformed data

Calculate a Goodness of Fit test in Python

- We will use `bioinfokit` v0.9.5 or later
- Check bioinfokit documentation (https://github.com/reneshbedre/bioinfokit) for installation and documentation

```
# I am using interactive python interpreter (Python 3.7)
>>> from bioinfokit.analys import stat
>>> import pandas as pd
# create or import pandas dataframe of observed counts
>>> df = pd.DataFrame({'genotypes':['A', 'B'], 'observed':[155, 45]})
>>> df = df.set_index(['genotypes'])
>>> df.head()
           observed
genotypes
A               155
B                45

# run chi-square test
>>> res = stat()
# p should be known theoretical expectation and must sum to 1
>>> res.chisq(df=df, p=(0.75, 0.25))

# output
>>> print(res.summary)

Chi-squared goodness of fit test

  Chi-Square    Df    P-value    Sample size
------------  ----  ---------  -------------
    0.666667     1   0.414216            200

# get expected counts
>>> print(res.expected_df)
           observed  expected_counts
genotypes
A               155            150.0
B                45             50.0
```

Interpretation

The *p* value obtained from the chi-square Goodness of Fit test is non-significant ($p > 0.05$ and fail to reject the null hypothesis), and therefore, we conclude that the observed genotypes counts after crosses is similar to that of expected counts as per the Mendelian ratio.

References

- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature methods. 2020 Mar;17(3):261-72.
- Bewick V, Cheek L, Ball J. Statistics review 8: Qualitative data–tests of association. Critical care. 2003 Feb 1;8(1):46.
- Serra N, Rea T, Di Carlo P, Sergi C. Continuity correction of Pearson's chi-square test in 2x2 Contingency Tables: A mini-review on recent development. Epidemiology, Biostatistics and Public Health. 2019 Jun 21;16(2).

Tags: Statistics

Updated: January 13, 2021