Chi-Square Test

Young W. Lim

2019-10-05 Sat

Outline

1. Based on
2. Chi-Square Test

Based on

"Understanding Statistics in the Behavioral Sciences" R. R. Pagano

Chi-Square test (1)

- A chi-squared (χ^2) test, is
  any statistical hypothesis test
  where the sampling distribution of
  the test statistic is a chi-squared distribution
  when the null hypothesis is true.

Chi-Square test (2)

- Without other qualification, 'chi-squared test' often is used as short for Pearson's chi-squared test.
- The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more categories.

Chi-Square test (3)

- In the standard applications of this test,
  the observations are classified into
  mutually exclusive classes,
- and there is some theory, or say null hypothesis,
  which gives the probability that
  any observation falls into the corresponding class.

Chi-Square test (4)

- The purpose of the test is to evaluate
  how likely the observations that are made would be,
  assuming the null hypothesis is true.

Chi-Square test (5)

- Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance.
- A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Chi-Square test (6)

- Test statistics that follow
  a chi-squared distribution
  arise from an assumption of
  independent normally distributed data,
  which is valid in many cases
  due to the central limit theorem.

Pearson's Chi-Square test (1)

- Pearson's chi-squared test χ^2 is
  a statistical test applied to sets of categorical data
  to evaluate how likely it is that any observed difference
  between the sets arose by chance.

Pearson's Chi-Square test (2)

- It is the most widely used of many chi-squared tests
(e.g., Yates, likelihood ratio, portmanteau test in time series, etc.)
- statistical procedures whose results are evaluated
by reference to the chi-squared distribution.
Its properties were first investigated by Karl Pearson in 1900.
- In contexts where it is important to improve a distinction
between the test statistic and its distribution,
names similar to Pearson χ-squared test or statistic are used.

Pearson's Chi-Square test (3)

- It tests a null hypothesis stating
  that the frequency distribution of certain events
  observed in a sample is consistent
  with a particular theoretical distribution.
- The events considered must be mutually exclusive
  and have total probability 1.
-

Pearson's Chi-Square test (4)

- A common case for this is where
  the events each cover an outcome of
  a categorical variable.
- A simple example is the hypothesis
  that an ordinary six sided die is fair
  (i.e, all six outcomes are equally likely to occur.)

Pearson's Chi-Square test (5)

- Pearson's chi-squared test is used to assess three types of comparison:
  - goodness of fit,
  - homogeneity, and
  - independence.

Pearson's Chi-Square test (6)

- A test of goodness of fit establishes whether an observed frequency distribution differs from a theoretical distribution.

Pearson's Chi-Square test (7)

- A test of homogeneity compares the distribution of counts for two or more groups using the same categorical variable (e.g. choice of activity—college, military, employment, travel—of graduates of a high school reported a year after graduation, sorted by graduation year, to see if number of graduates choosing a given activity has changed from class to class, or from decade to decade).[1]

[1] DEFINITION NOT FOUND.

Pearson's Chi-Square test (8)

- A test of independence assesses whether observations consisting of measures on two variables, expressed in a contingency table, are independent of each other (e.g. polling responses from people of different nationalities to see if one's nationality is related to the response).

- Suppose that n observations
  in a random sample from a population
  are classified into k mutually exclusive classes
  with respective observed numbers x_i (for $i = 1, 2, ..., k$),

Pearson's Chi-Square test (b)

- a null hypothesis gives the probability p_i
  that an observation falls into the i-th class.
  So we have the expected numbers $m_i = npi$ for all i, where

$$\sum_{i=1}^{k} p_i = 1$$

$$\sum_{i=1}^{k} m_i = n \sum_{i=1}^{k} p_i = \sum_{i=1}^{k} x_i$$

Pearson's Chi-Square test (c)

$$X^2 = \sum_{i=1}^{k} \frac{(x_i - m_i)^2}{m_i} = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - n$$

$$X^2 - X'^2 = \sum_{i=1}^{k} \frac{x_i^2}{m_i} - \sum_{i=1}^{k} \frac{x_i^2}{m'_i}$$

Example Chi-Squared test for categorical data (1)

- Suppose there is a city of 1,000,000 residents with four neighborhoods: A, B, C, and D.
- a random sample of 650 residents of the city is taken and their occupation is recorded as "white collar", "blue collar", or "no collar".
- the null hypothesis is that each person's neighborhood of residence is independent of the person's occupational classification.

Example Chi-Squared test for categorical data (2)

	A	B	C	D	total
White collar	90	60	104	95	349
Blue collar	30	50	51	20	151
No collar	30	40	45	35	150
Total	150	150	200	150	650

Example Chi-Squared test for categorical data (3)

- Let us take the sample living in neighborhood A, 150, to estimate what proportion of the whole 1,000,000 live in neighborhood A.

- Similarly we take 349/650 to estimate what proportion of the 1,000,000 are white-collar workers.
- By the assumption of independence under the hypothesis we should "expect" the number of white-collar workers in neighborhood A to be

$$150 \times \frac{349}{650} \approx 80.54$$

Example Chi-Squared test for categorical data (4)

- Then in that "cell" of the table, we have

$$\frac{(\text{observed} - \text{expected})^2}{\text{expected}} = \frac{(90 - 80.54)^2}{80.54} \approx 1.11$$

- The sum of these quantities over all of the cells is the test statistic; in this case, ≈ 24.6
- Under the null hypothesis, this sum has approximately a chi-squared distribution whose number of degrees of freedom are

$$(\text{number of rows} - 1)(\text{number of columns} - 1) = (3 - 1)(4 - 1) = 6$$

Example Chi-Squared test for categorical data (5)

- If the test statistic is improbably large
  according to that chi-squared distribution,
  then one rejects the null hypothesis of independence.
- A related issue is a test of homogeneity.
- Suppose that
  instead of giving every resident of each of the four neighborhoods
  an equal chance of inclusion in the sample,
  we decide in advance how many residents of each neighborhood
  to include.

Example Chi-Squared test for categorical data (6)

- Then each resident has the same chance of being chosen as do all residents of the same neighborhood, but residents of different neighborhoods would have different probabilities of being chosen if the four sample sizes are not proportional to the populations of the four neighborhoods.

Example Chi-Squared test for categorical data (7)

- In such a case, we would be testing "homogeneity" rather than "independence".
- The question is whether the proportions of blue-collar, white-collar, and no-collar workers in the four neighborhoods are the same.
- However, the test is done in the same way.