Significance Tests

Part 4: Chi-square Test (Chi-2 test)

By: Noureddin Sadawi, PhD

University of London

Chi-square test (Chi-2 test)

- ‘The chi-square test is used with **count data** to test how well it fits some expected distribution.
- The most common use of the chi-square statistic in statistical practice is with **r × c contingency tables**, to assess whether the null hypothesis of independence among variables is reasonable.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Contingency table

- ‘**Contingency table:** A tally of counts between two or more categorical variables.
- A useful way to summarize two categorical variables.
- Contingency tables can look only at counts, or they can also include column and total percentages.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

- Sometimes called the *confusion matrix* (more on it later in the course).
- Covid-19: Imagine you carry out a study of 1000 vaccinated people at Heathrow Airport to count how many of them had which vaccine?
  A contingency table is a useful way to summarise the results.

	British	Non British
Oxford-AstraZeneca	291	189
Pfizer	244	276

Key terms for chi-square test

- ‘**Chi-square statistic:** A measure of the extent to which some observed data departs from expectation.
- **Expectation or expected:** How we would expect the data to turn out under some assumption, typically the null hypothesis.’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Chi-square test: a resampling approach

- ‘Suppose you are testing three different headlines—A, B, and C—and you run them each on 1,000 visitors.
- A resampling procedure can test whether the click rates differ to an extent greater than chance might cause.
- For this test, we need to have the “expected” distribution of clicks, and in this case, that would be under the null hypothesis assumption that all three headlines share the same click rate, for an overall click rate of 34/3,000.

	Headline A	Headline B	Headline C
Click	14	8	12
No-click	986	992	988

Chi-square test: a resampling approach

- Under this assumption, our contingency table would look like [the one on the right].’

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

	Headline A	Headline B	Headline C
Click	11.33	11.33	111.33
No-click	988.67	988.67	988.67

Chi-square test: a resampling approach

- 'The *Pearson residual* is defined as:

$$\mathrm{R} = \frac{Observed - Expected}{\sqrt{Expected}}$$

R measures the extent to which the actual counts differ from these expected counts.

- The chi-square statistic is defined as the sum of the squared Pearson residuals:

$$X = \sum_{i}^{r} \sum_{j}^{c} R^2$$

where *r* and *c* are the number of rows and columns, respectively.'

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

	Headline A	Headline B	Headline C
Click	14	8	12
No-click	986	992	988

	Headline A	Headline B	Headline C
Click	11.33	11.33	111.33
No-click	988.67	988.67	988.67

	Headline A	Headline B	Headline C
Click	0.792	-0.990	0.198
No-click	-0.085	0.106	-0.021

Chi-square test: a resampling approach

	Headline A	Headline B	Headline C
Click	0.792	-0.990	0.198
No-click	-0.085	0.106	-0.021

- 'The chi-square statistic for this example is 1.666.
- Is that more than could reasonably occur in a chance model?'

- 'We can test with this resampling algorithm:
    1. Constitute a box with 34 ones (clicks) and 2,966 zeros (no clicks).
    2. Shuffle, take three separate samples of 1,000, and count the clicks in each.
    3. Find the squared differences between the shuffled counts and the expected counts and sum them.
    4. Repeat steps 2 and 3, say, 1,000 times.
    5. How often does the resampled sum of squared deviations exceed the observed?

  That's the p-value.'

  (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

Chi-square test: statistical theory

- ‘Asymptotic statistical theory shows that the distribution of the chi-square statistic can be approximated by a chi-square distribution.
- The appropriate standard chi-square distribution is determined by the degrees of freedom.
- For a contingency table, the degrees of freedom are related to the number of rows (r) and columns (c) as follows:

$$\textit{degrees of freedom} = (r - 1) \times (c - 1)'$$

(Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).

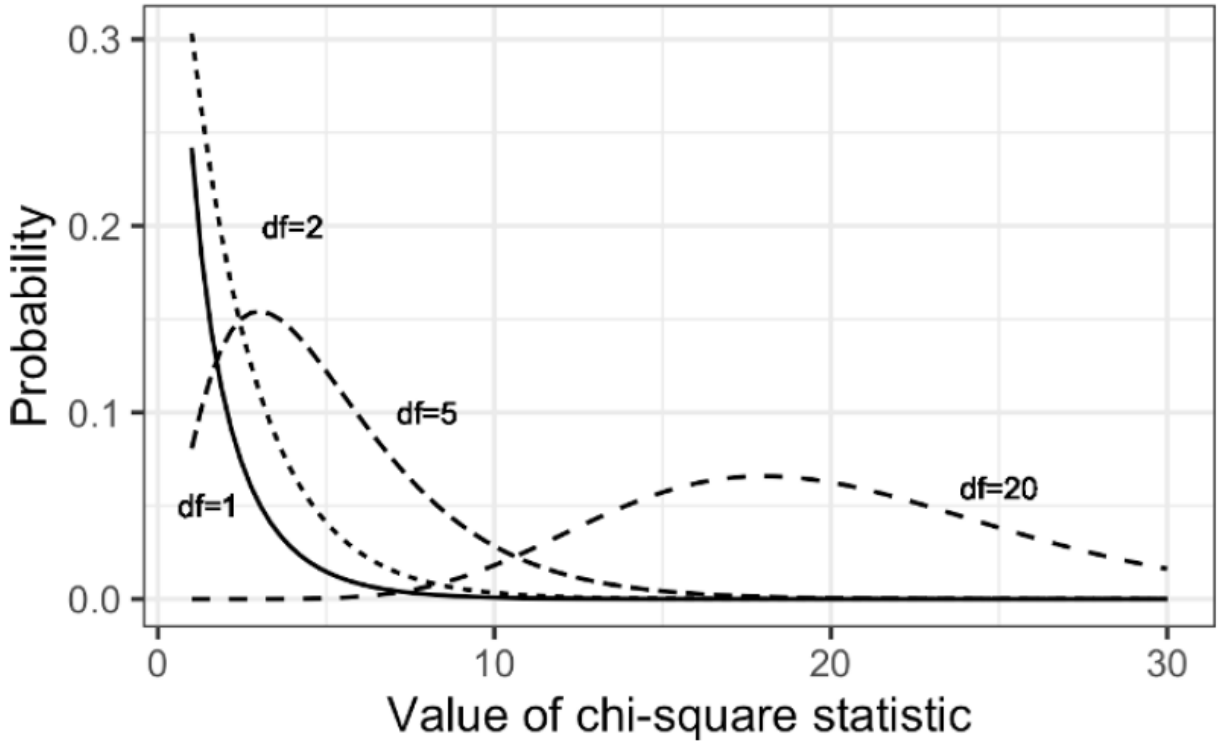
Chi-square test: statistical theory

- ‘The chi-square distribution is typically skewed, with a long tail to the right.
- The further out on the chi-square distribution the observed statistic is, the lower the p-value.’

(Bruce and Bruce *Practical statistics for data scientists, second edition*, 2020).

- There are existing tools and functions that compute the p-value based on the chi-square distribution.

Chi-square distribution with various degrees of freedom



Source: (Bruce and Bruce *Practical statistics for data scientists*, second edition, 2020).