

Unit 7 *Chi-Squared*

Introduction

Learning objectives

Here we see how we can test particular hypotheses using the chi-squared test, sometimes referred to as 'goodness of fit' testing. After studying this unit you should

- understand how the chi-squared test can be used to test a uniform distribution, using critical values
- understand how the chi-squared test can be used to test the dependency between different attributes by using a contingency table.

Notes

The first use of the test was by the English mathematician Karl Pearson (1857-1936) who was working on correlation and regression problems. He wanted to test his methods using real data and, for example, used over 4000 trials of roulette and tosses of a coin. Although he published his work in 1900 it not until many decades later that it was recognised as an important advance in statistical analysis.

Key points

- In testing whether data fits a particular distribution, using chi-squared, you need to have observed data, O_i , and theoretical (expected) data, E_i
- Critical values for the chi-squared test are based on the significance level and number of degrees of freedom
- For 2×2 contingency tables, you need to calculate the chi-squared value using Yates' continuity correction
- Groups in contingency tables must be combined if expected values are less than 5.

Facts to remember

- The formula for chi-squared (χ^2) is given by

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$$

where O_i are the observed values and E_i are the expected values.

- The degrees of freedom for testing distribution is given by

$$v = \text{number of classes} - \text{number of constraints}.$$

- For 2×2 contingency tables, Yates' continuity correction is used to calculate χ^2 ; that is

$$\chi^2 = \sum_{i=1}^n \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

Unit 7 *Chi-Squared*

Introduction

- For contingency tables, the expected frequencies are calculated by

$$\text{Expected frequency} = \frac{(\text{row total}) \times (\text{column total})}{(\text{grand total})}$$

- For $h \times k$ contingency tables (h rows and k columns), the number of degrees of freedom, σ , is given by

$$\nu = (h - 1) \times (k - 1)$$

Glossary of terms

Observed and expected frequencies: observed data, O_i , and corresponding theoretical values, E_i , based on the distribution being tested

Chi-squared: distribution based on the chi-squared values $\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i}$ and the degrees of freedom

Degrees of freedom: calculated from

number of groups – number of constraints

Contingency table: gives the numbers in different categories for 2 factors (that may or may not be related), e.g.

	<i>French</i>	<i>Russian</i>	<i>Total</i>
<i>Male</i>	39	16	55
<i>Female</i>	21	14	35
<i>Total</i>	60	30	90