Introduction to Statistics Note

2024 Spring Semester

21 CST H3Art

Chapter 9: Analysis of Two-Way Tables

9.1 Inference for Two-Way Tables

Two-Way Tables（双向表） can be used to describe the **relationship** between **two categorical variables**.

- When the data are obtained from random sampling, two-way tables of counts can be used to formally **test the hypothesis** that the two categorical variables are **independent** in the population from which the data were obtained.

To test this hypothesis, we compare **actual counts（实际计数）** from the sample data with **expected counts（预期计数）**, the expected count in any cell of a two-way table when H_0 is true is:

$$\textbf{expected count} = \frac{\textbf{row total} \times \textbf{column total}}{n}$$

where $\textbf{row total}$ represents **the sum of a cell row（单元格所在行的计数和）**, $\textbf{column total}$ represents **the sum of a cell column（单元格所在列的计数和）**

The **test statistic** that makes the comparison is the **chi-square statistic（卡方统计量）**, the chi-square statistic is a measure of **how far the observed counts are from the expected counts（检验观察值和期望值的差距）**. The formula for the statistic is:

$$\mathcal{X}^2 = \sum \frac{(\textbf{observed} - \textbf{expected})^2}{\textbf{expected}}$$

where $\textbf{observed}$ represents **an observed cell count（一个单元格的值）**

The **chi-square distributions（卡方分布）** are a family of distributions that take **only positive values** and are **skewed to the right**. A particular $\mathcal{X}^2$ distribution is specified by giving its **degrees of freedom（自由度）**.

The $\mathcal{X}^2$ test for a two-way table with r rows and c columns uses critical values from the $\mathcal{X}^2$ distribution with $(r-1)(c-1)$ **degrees of freedom**.

The **P-value is the area under the density curve** of this $\mathcal{X}^2$ distribution to the **right（右尾）** of the value of the test statistic.

Cell Counts Required（单元格数量要求） for the Chi-Square Test

- The average of the expected counts is 5 **or more**
- All individual expected counts are 1 **or greater**
- In a 2 × 2 table, all four expected cell counts should be **at least 5**.

The **expected count** in any cell of a two-way table when H_0 is true is:

$$\textbf{expected count} = \frac{\textbf{row total} \cdot \textbf{column total}}{\textbf{table total}}$$

Testing for independence（独立性检验）

Suppose we have a single sample from a single population. For each individual in this SRS of size n, we measure two categorical variables. The results are then summarized in a two-way table.

The **null hypothesis** is that the **row and column variables are independent（零假设是行列无关/独立）**. The **alternative hypothesis** is that the **row and column variables are dependent（备选假设是行列相关）**.

9.2 Goodness of Fit

The idea of the chi-square test for goodness of fit is this:
We compare the **observed counts** from our sample with the counts that would be **expected** if H_0 is true.
The **more** the **observed counts** differ from the **expected counts**, the more evidence we have against the null hypothesis.
（比较观测值和预测值的差值，差值越大就越能对抗原假设）

E.g.

A categorical variable has k possible outcomes, with probabilities $p_1, p_2, p_3, \ldots, p_k$. That is, p_i is the probability of the i^{th} outcome. We have n independent observations from this categorical variable.

To test the null hypothesis that the probabilities have specified values:

$$H_0 : p_1, p_2, \ldots, p_k$$

find the **expected count** for each category assuming that H_0 is true. Then calculate the chi-square statistic:

$$\mathcal{X}^2 = \sum \frac{(\mathbf{observed} - \mathbf{expected})^2}{\mathbf{expected}}$$

where the sum is over the k different categories. The P-value is the area to the right of $\mathcal{X}^2$ under the density curve of the chi-square distribution with $k - 1$ degrees of freedom.