Math 408 - Mathematical Statistics

Lecture 32-33. Pearson's χ^2 Test For Multinomial Data

April 19-22, 2013

Agenda

- Multinomial Distribution and its Properties
- Construction the GLRT for Multinomial Data
- The MLE for Parameters of the Multinomial Distribution
- The GLRT with Significance Level α
- Pearson's χ^2 Test
- Asymptotic Equivalence of the GLRT and the Pearson's Test
- Example: Mendel's Peas
- Summary

Multinomial Distribution

The multinomial distribution is a generalization of the binomial distribution.

Consider drawing a ball from a box which has balls with k different colors labeled color 1, color 2,..., color k. Let $p = (p_1, \ldots, p_k)$, where p_i is the probability of drawing a ball of color i,

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{k} p_i = 1$$

Draw n times (independent draws with replacement) and let $X = (X_1, \ldots, X_k)$, where X_i is the number of times that color i appeared.

$$\sum_{i=1}^{k} X_i = n$$

We say that X has a $\text{Multinomial}(n, p)$ distribution.

Application: Multinomial distributions are useful when a "success-failure" description is insufficient to understand a system. Multinomial distributions are relevant to situations where there are more than two possible outcomes. For example, temperature = high, med, low.

Properties of the Multinomial Distribution

$$X \sim \text{Multinomial}(n, p)$$

- n is the number of trials
- k is the number of possible outcomes
- $p = (p_1, \ldots, p_k)$, where p_i is the probability of observing outcome i
- $X = (X_1, \ldots, X_k)$, where X_i is the number of occurrences of outcome i

Theorem

- *The probability mass function of X is*

$$\pi_X(x|n, p) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k}$$

- *The marginal distribution of X_i is* $\text{Binomial}(n, p_i)$
- *The mean and covariance matrix of X are*

$$\mathbb{E}[X] = \begin{pmatrix} np_1 \\ \vdots \\ np_k \end{pmatrix} \qquad \mathbb{V}[X] = \begin{pmatrix} np_1(1-p_1) & -np_1p_2 & \ldots & -np_1p_k \\ -np_1p_2 & np_2(1-p_2) & \ldots & -np_2p_k \\ \vdots & \vdots & \ddots & \vdots \\ -np_1p_2 & -np_2p_k & \ldots & np_k(1-p_k) \end{pmatrix}$$

Constructing the GLRT

Suppose that $X \sim \text{Multinomial}(n, p)$, where p is unknown, and we want to test

$$H_0 : (p_1, \ldots, p_k) = (\tilde{p}_1, \ldots, \tilde{p}_k) \equiv \tilde{p} \qquad \text{v.s.} \qquad H_1 : (p_1, \ldots, p_k) \neq (\tilde{p}_1, \ldots, \tilde{p}_k)$$

To construct the generalized likelihood ratio test, first, we need to determine the likelihood function $\mathcal{L}(p)$. In this case:

$$\mathcal{L}(p_1, \ldots, p_k) = \pi_X(X|n, p) = \frac{n!}{X_1! \ldots X_k!} p_1^{X_1} \ldots p_k^{X_k}$$

The likelihood ratio statistic is

$$\Lambda = \frac{\max_{p \in \Theta_0} \mathcal{L}(p)}{\max_{p \in \Theta} \mathcal{L}(p)} = \frac{\mathcal{L}(\tilde{p})}{\mathcal{L}(\hat{p}_{MLE})}$$

- $\Theta_0 = \{p : p = \tilde{p}\}$, $\dim \Theta_0 = 0$
- $\Theta = \{p : \sum_{i=1}^{k} p_i = 1\}$, $\dim \Theta = k - 1$

Thus, to proceed, we need to find the MLE of p.

The MLE of p and the GLRT with level α

Theorem

Let $X \sim \text{Multinomial}(n, p)$. The maximum likelihood estimator of p is

$$\hat{p}_{MLE} = \begin{pmatrix} \frac{X_1}{n} \\ \vdots \\ \frac{X_k}{n} \end{pmatrix} = \frac{X}{n}$$

Therefore, the likelihood ratio statistic is

$$\Lambda = \prod_{i=1}^{k} \left(\frac{n\tilde{p}_i}{X_i} \right)^{X_i}$$

and

$$-2 \log \Lambda = 2 \sum_{i=1}^{k} X_i \log \left(\frac{X_i}{n\tilde{p}_i} \right) \ \dot{\sim} \ \chi^2_{k-1}, \quad \text{when } n \to \infty$$

The GLRT with significance level α rejects H_0 if and only if

$$\boxed{2 \sum_{i=1}^{k} X_i \log \left(\frac{X_i}{n\tilde{p}_i} \right) > \chi^2_{k-1}(\alpha)}$$

Pearson's χ^2 Test

In practice, the Pearson's χ^2 test is often used. The test is based on the following statistic:

$$T = \sum_{i=1}^{k} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

- $O_i = X_i$ is the observed data
- $E_i = \mathbb{E}[X_i] = n\tilde{p}_i$ is the expected value of X_i under H_0
- T is called the **Pearson's χ^2 statistic**

The Pearson's χ^2 statistic and $-2\log\Lambda$ are asymptotically equivalent under H_0

Theorem

- *Under H_0, $T \xrightarrow{\mathcal{D}} \chi^2_{k-1}$.*
- *Pearson's test: reject H_0 if $T > \chi^2_{k-1}(\alpha)$ has asymptotic significance level α.*
- *The p-value is $\mathbb{P}(\xi > t)$, where $\xi \sim \chi^2_{k-1}$ and t is the observed value of T.*

Remark: Pearson's test has been more commonly used than the GLRT, because it is easier to calculate (especially without a computer!)

Mendel's Peas

Example

Mendel bred peas with round yellow seeds and wrinkled green seeds.
There are four types of progeny:

- round yellow, wrinkled yellow, round green, wrinkled green.

The number of each type is multinomial with probability (p_1, p_2, p_3, p_4).
According to Mendel's theory:

$$H_0 : (p_1, p_2, p_3, p_4) = \left(\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}\right) \equiv \tilde{p}$$

In $n = 556$ trials he observed $X = (315, 101, 108, 32)$.
Question: Based on these data, should we accept or reject the Mendel's theory?
Solution:

- The observed value of Pearson's χ^2 statistic is $t = \sum_{i=1}^{4} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} = 0.47$
- Let $\alpha = 0.05$. Then $\chi^2_3(\alpha) = F^{-1}_{\chi^2_3}(1-\alpha) \approx 7.8$.
- Since $T < \chi^2_3(\alpha)$, we accept H_0.
- The p-value is $p\text{-value} = \mathbb{P}(\xi > 0.47) = 1 - F_{\chi^2_3}(0.47) \approx 0.92$.
- No evidence against Mendel's theory.

Summary

- Multinomial distribution: $X \sim \text{Multinomial}(n, p)$
  - The probability mass function of X is

$$\pi_X(x|n,p) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k}$$

  - The marginal distribution of X_i is $\text{Binomial}(n, p_i)$
  - The maximum likelihood estimator of p is $\hat{p}_{MLE} = X/n$
- Suppose that $X \sim \text{Multinomial}(n, p)$, p is unknown, and we want to test

$$H_0 : (p_1, \dots, p_k) = (\tilde{p}_1, \dots, \tilde{p}_k) \equiv \tilde{p} \quad \text{v.s.} \quad H_1 : (p_1, \dots, p_k) \neq (\tilde{p}_1, \dots, \tilde{p}_k)$$

  - GLRT with significance level α rejects H_0 if $\boxed{2\sum_{i=1}^{k} X_i \log\left(\frac{X_i}{n\tilde{p}_i}\right) > \chi^2_{k-1}(\alpha)}$
  - Pearson's test: reject H_0 if $\boxed{T = \sum_{i=1}^{k} \frac{(X_i - n\tilde{p}_i)^2}{n\tilde{p}_i} > \chi^2_{k-1}(\alpha)}$
    - ★ Under H_0, the Pearson's χ^2 statistic $T \xrightarrow{\mathcal{D}} \chi^2_{k-1}$.
    - ★ Pearson's test has asymptotic significance level α.
    - ★ The p-value is $\mathbb{P}(\xi > t)$, where $\xi \sim \chi^2_{k-1}$ and t is the observed value of T.