Lec 12. Goodness-of-fit Tests

Zhang W.P.

Contents

12.1 Goodness-of-fit test 1
12.2 Pearson's χ^2 Test 2
12.2.1 Test of independence 7
12.2.2 Test of homogeneity 9
12.3 Kolmogorov-Smirnov Tests 10

Learning objectives

After this section, you should be able to ...

- understand Pearson χ^2 test
- do the tests for independence and homogeneity
- understand the goodness-of-fit test, especially the normality tests

12.1 Goodness-of-fit test

- The "goodness-of-fit test" is a way of determining whether a set of categorical data came from a claimed distribution or not. The null hypothesis is that they did and the alternate hypothesis is that they didn't.

$$H_0 : X \sim F$$

  - F is discrete distribution: Pearson χ^2 test
  - F is continuous distribution: Kolmogorov test
  - Two sample case: Smirnov test

$$H_0 : F_1(x) = F_2(x)$$

12.2 Pearson's χ^2 Test

Goodness-of-Fit in a Multinomial Model.

Let $X_1, \ldots, X_n$ be a simple sample from population

$$X \sim \begin{pmatrix} a_1 & a_2 & \cdots & a_r \\ p_1 & p_2 & \cdots & p_r \end{pmatrix}$$

where $p_i \geq 0, p_1 + \cdots + p_r = 1$. Because $p_r = 1 - \sum_{j=1}^{r-1} p_j$, we consider the parameter $\theta = (p_1, \ldots, p_{r-1})^T$ and we test the following hypothesis

$$H_0 : p_j = p_{0j}, j = 1, \ldots, r, \quad i.e., \quad H_0 : \theta = \theta_0$$

for specified $\theta_0 = (p_{01}, \ldots, p_{0r-1})^T$.

The likelihood function is

$$L_n(\theta) = L_n(p_1, \ldots, p_{r-1}) = p_1^{n_1} \cdots p_{r-1}^{n_{r-1}} (1 - \cdots - p_{r-1})^{n_r}$$

and the MLE $\hat{p}_j = n_j/n$, where $n_j = \sum_{i=1}^{n} 1\{X_i = a_j\}$.

- **Wald test:** Let $\hat{\theta} = (\hat{p}_1, \ldots, \hat{p}_{r-1})^T$, then the Wald test that rejects $H_0 : \theta = \theta_0$ when

$$W_n(\theta_0) = n\left(\widehat{\theta}_n - \theta_0\right)^T I(\theta_0)\left(\widehat{\theta}_n - \theta_0\right) > \chi^2_{r-1}(\alpha)$$

has asymptotic level α under H_0.

The Fisher information $I(\theta) = (I_{ij})$ with

$$I_{ij} = -E\left[\frac{\partial^2}{\partial p_i \partial p_j}\sum_{k=1}^{r} I(X = a_k) log p_k\right]$$

$$= \begin{cases} \frac{1}{p_i} + \frac{1}{p_r}, & i = j \\ \frac{1}{p_r}, & i \neq j \end{cases}$$

Thus,

$$W_n(\theta_0) = n\sum_{j=1}^{r-1}\left[\widehat{p}_j - p_{0j}\right]^2 / p_{0j} + n\sum_{j=1}^{r-1}\sum_{i=1}^{r-1}\left(\widehat{p}_i - p_{0i}\right)\left(\widehat{p}_j - p_{0j}\right) / p_{0r}$$

$$= n\sum_{j=1}^{r}\left[\widehat{p}_j - p_{0j}\right]^2 / p_{0j} = n\sum_{j=1}^{r}\left[n_j - np_{0j}\right]^2 / np_{0j}$$

The term on the right is called Pearson's chi-square (χ^2) statistic and is the statistic that is typically used for this multinomial testing problem. It is easily remembered as

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

- **LRT**

$$2\log\Lambda_n(\mathbf{X}) = 2\sum_{j=1}^{r} n_j \log\left(n_j/np_{0j}\right) > \chi^2_{r-1}(1-\alpha)$$

- **Score test**

$$LM_n\left(\theta_0\right) = U_n^T\left(\theta_0\right) I^{-1}\left(\theta_0\right) U_n\left(\theta_0\right) > \chi^2_{r-1}(\alpha).$$

where $U_n(\theta) = \frac{1}{\sqrt{n}}\nabla\ell_n(\theta)$. It follows that the Rao statistic equals Pearson's χ^2.

Goodness-of-Fit to Composite Multinomial Models. Contingency Tables

- We test the hypothesis

$$H_0 : \theta = \theta(\eta) = (p_1(\eta), \ldots, p_{r-1}(\eta))^T$$

where $\eta = (\eta_1, \ldots, \eta_s)^T$ with $s < r - 1$ is the unknown parameter.

- **LRT** If a maximizing value, $\widehat{\eta} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_s)$ exists, the log likelihood ratio is given by

$$\log \Lambda\left(n_1, \ldots, n_k\right) = \sum_{i=1}^{r} n_i \left[\log\left(n_i/n\right) - \log \theta_i(\widehat{\eta})\right]$$

then reject H_0 when $2 \log \Lambda > \chi^2_{r-1-s}(\alpha)$.

- **Score test**

$$LM_n(\theta(\widehat{\eta})) = \sum_{j=1}^{r} \frac{\left[n_i - n p_j(\widehat{\eta})\right]^2}{n p_j(\widehat{\eta})} > \chi^2_{r-1-s}(\alpha).$$

Notice that the score test statistic is (see Lecture 11)

$$LM_n(\theta(\widehat{\eta})) = U_n^{\top}(\widehat{\theta}_n^R) I^{-1}(\widehat{\theta}_n^R) U_n(\widehat{\theta}_n^R)$$

and

(1) $U_n(\theta) = \frac{1}{\sqrt{n}}[\frac{n_1}{p_1} - \frac{n_r}{p_r}, \ldots, \frac{n_{r-1}}{p_{r-1}} - \frac{n_r}{p_r}]^T$

(2) $I^{-1}(\theta) = [A_{r-1} + \frac{1}{p_r} 11^T]^{-1} = diag(\theta) - \theta\theta^T$, where $A_{r-1} = diag(p_1, \ldots, p_{r-1})$. (using fact $\left(\mathbf{A} + \mathbf{bc}^T\right)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{bc}^T\mathbf{A}^{-1}}{1+\mathbf{c}^T\mathbf{A}^{-1}\mathbf{b}}$)

- **Wald test** the Wald statistic is also equal to Pearson's χ^2 test statistic. WLOG, let $\theta_1 = p_1 - p_1(\eta), \ldots, \theta_{r-1-s} = p_{r-1-s} - p_{r-1-s}(\eta); p_{r-s} = p_{r-s}(\eta), \ldots, p_{r-1} = p_{r-1}(\eta)$, then

$$(p_1, \ldots, p_{r-1}) \mapsto (\theta_1, \ldots, \theta_{r-1-s}, \eta^T)$$

is one-to-one mapping. And

$$H_0 : \theta_j = 0, j = 1, \ldots, r-1-s$$

12.2.1 Test of independence

- Suppose that Variable A has r levels, and Variable B has c levels. We consider the following hypothesis:

  H_0 : A and B are independent.

  H_a : A and B are not independent.

- Equivalently,

$$H_0 : P(A = i, B = j) = p_{ij} = P(A = i)P(B = j) = u_i v_j$$

  where $\{p_{ij}\}, \{u_i\}, \{v_j\}$ are pmfs. That is,

$$H_0 : p_{ij} = p_{ij}(\eta) = u_i v_j$$

  where $\eta = (u_1, \ldots, u_{r-1}, v_1, \ldots, v_{c-1})$ by noting $\sum u_i = 1, \sum v_j = 1$.

- Thus the Pearson Chi-square test is applicable with $df = rc - 1 - (r - 1) - (c - 1) = (r - 1)(c - 1)$.

H_0 : A and B are independent

$Data:$

$A \backslash B$	1	$\cdots$	c	
1	n_{11}	$\cdots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r\cdot}$
	$n_{\cdot 1}$	$\cdots$	$n_{\cdot c}$	n

When to Use Chi-Square Test for Independence:

- The sampling method is simple random sampling.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

12.2.2 Test of homogeneity

- In the test of homogeneity, we consider two or more populations (P) and a single categorical variable (B). e.g.,

$P\backslash A$	1	$\cdots$	c
1	p_{11}	$\cdots$	p_{1c}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	p_{r1}	$\cdots$	p_{rc}

Data:

$P\backslash B$	1	$\cdots$	c	
1	n_{11}	$\cdots$	n_{1c}	$n_{1\cdot}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	n_{r1}	$\cdots$	n_{rc}	$n_{r\cdot}$

and consider the hypothesis:

H_0 : The distributions of the r populations are the same., i.e.,

$$H_0 : p_{1j} = \cdots = p_{rj}, j = 1, \ldots, c$$

- If we treat P as a categorical variable, then the above problem is testing for independence.

12.3 Kolmogorov-Smirnov Tests

Kolmogorov test

- Let $X_1, \ldots, X_n$ be a simple sample from population $X \sim F$, we test the following hypothesis

$$H_0 : F(x) = F_0(x)$$

where $F_0(x)$ is some known distribution.

- Let F_n be the empirical distribution function based on $X_1, \ldots, X_n$. One form of the Kolmogorov test statistic for H_0 is

$$D_n := \sup_x |F_n(x) - F_0(x)| .$$

- The p-value is

$$p\text{-value} = P(D_n \geq D_{obs} | H_0)$$

Theorem 1. *If $X_1, X_2, \ldots,$ are i.i.d. with the continuous distribution function F_0 then*

(1). The null distribution of D_n does not depend on F_0; it depends only on n.

(2). If $n \to \infty$ the distribution of $\sqrt{n}D_n$ is asymptotically Kolmogorov's distribution with the c.d.f.

$$K(x) = 1 - 2\sum_{k=1}^{\infty}(-1)^{k-1}e^{-2k^2x^2}, x > 0 \tag{12.1}$$

that is

$$\lim_{n\to\infty} P\left(\sqrt{n}D_n \leq x\right) = K(x)$$

Smirnov test

- Let $X_{11}, \ldots, X_{1n_1}$ be $i.i.d$ with a c.d.f. F_1, and $X_{21}, \ldots, X_{2n_2}$ $i.i.d$ with a c.d.f. F_2. We are interested in testing the null hypothesis of the form

$$H_0 : F_1(x) = F_2(x) \quad \text{for all} \quad x$$

- Let $F_{kn_k}(x) = \frac{1}{n_k} \sum_{i=1}^{n_k} I(X_{ki} \leq x), k = 1, 2$, be the empirical distribution functions, then consider the following statistics

$$D_{n_1,n_2} = \sup_{-\infty<x<\infty} |F_{1n_1}(x) - F_{2n_2}(x)| = \min\{D^+_{n_1,n_2}, D^-_{n_1,n_2}\}$$

$$D^+_{n_1,n_2} = \sup_{-\infty<x<\infty} (F_{1n_1}(x) - F_{2n_2}(x))$$

$$D^-_{n_1,n_2} = \sup_{-\infty<x<\infty} (F_{2n_2}(x) - F_{1n_1}(x))$$

Theorem 2 (Smirnov). *Under the null hypothesis H_0, we have*

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D^+_{n_1,n_2} \le x\right) = 1 - e^{-2x^2}, x > 0$$

$$\lim_{n_1,n_2\to\infty} P\left(\sqrt{\frac{n_1n_2}{n_1+n_2}}D_{n_1,n_2} \le x\right) = K(x)$$

where $K(x)$ is the Kolmogorov's distribution given by (12.1).

- Test for $H_0 : F_1 = F_2 \leftrightarrow H_0 : F_1 \neq F_2$:

$$\text{Reject } H_0 \text{ if } D_{n_1,n_2} > \frac{\lambda}{\sqrt{n_1n_2/(n_1+n_2)}}$$

- Test for $H_0 : F_1 \le F_2 \leftrightarrow H_0 : F_1 > F_2$:

$$\text{Reject } H_0 \text{ if } D^+_{n_1,n_2} > \frac{\lambda_1}{\sqrt{n_1n_2/(n_1+n_2)}}, \lambda_1 = \sqrt{-ln\alpha/2}$$

- Test for $H_0 : F_1 \ge F_2 \leftrightarrow H_0 : F_1 < F_2$:

$$\text{Reject } H_0 \text{ if } D^-_{n_1,n_2} > \frac{\lambda_1}{\sqrt{n_1n_2/(n_1+n_2)}}, \lambda_1 = \sqrt{-ln\alpha/2}$$