

## Lecture 20 $\chi^2$ Goodness of Fit Tests

Various  $\chi^2$  tests will be presented as special cases of the asymptotic LR tests given in Lecture 19.

**Example 20.1** (K. Pearson's  $\chi^2$  test for a multinomial distribution with categorical data) Let  $X_1, X_2, \dots$  be iid random variables with unknown parameters  $p_j = P(X_1 = j)$ ,  $j = 1, \dots, r$ . Consider a simple null hypothesis  $H_0 : p_j = p_{j0}$ ,  $j = 1, \dots, r$ . Note that  $\dim\Theta = r - 1$  and  $\dim\Theta_0 = 0$ .

Define  $K_n = \sum_{j=1}^r \frac{(S_n^{(j)} - np_{j0})^2}{np_{j0}}$ , call a  $\chi^2$  statistic, which is a measure of the departure of the data from  $H_0$ .  $K_n$  represents "goodness of fit", more precisely, the approximate  $p$ -value  $P(\chi_{r-1}^2 > c)$  is referred to as the goodness of fit when  $K_n = c$ . In what follows, we will show that  $K_n \approx 2 \log \Lambda(X^n)$  for large  $n$  under  $H_0$ .

For the parameter  $\theta = (p_1, \dots, p_{r-1})$ , the MLE  $\hat{p}_j = S_n^{(j)}/n$ ,  $j = 1, \dots, r$ . Since  $L(\theta|X^n) = \log \prod_{j=1}^r p_j^{S_n^{(j)}}$ , we have  $2 \log \Lambda(X^n) = 2 \sum_{j=1}^r S_n^{(j)} [\log S_n^{(j)} - \log(np_{j0})]$ . A Taylor expansion yields

$$\log S_n^{(j)} - \log(np_{j0}) = \log \left( 1 + \frac{S_n^{(j)} - np_{j0}}{np_{j0}} \right) = \frac{S_n^{(j)} - np_{j0}}{np_{j0}} - \frac{a_n}{2} \left( \frac{S_n^{(j)} - np_{j0}}{np_{j0}} \right)^2,$$

where  $a_n = (1+b_n)^{-2}$  and  $b_n$  is between 0 and  $\frac{S_n^{(j)} - np_{j0}}{np_{j0}}$ . Note that  $\sup_{1 \leq j \leq r} \left| \frac{S_n^{(j)} - np_{j0}}{np_{j0}} \right| \xrightarrow{P} 0$  under  $H_0$  as  $n \rightarrow \infty$ . Hence

$$\begin{aligned} 2 \log \Lambda(X^n) &= 2 \sum_{j=1}^r (S_n^{(j)} - np_{j0}) \left[ \frac{S_n^{(j)} - np_{j0}}{np_{j0}} - \frac{a_n}{2} \left( \frac{S_n^{(j)} - np_{j0}}{np_{j0}} \right)^2 \right] \\ &\quad + 2 \sum_{j=1}^r (S_n^{(j)} - np_{j0}) - a_n \sum_{j=1}^r \frac{(S_n^{(j)} - np_{j0})^2}{np_{j0}} \\ &= 2K_n - a_n \sum_{j=1}^r \frac{(S_n^{(j)} - np_{j0})^3}{(np_{j0})^2} - a_n K_n \\ &= (1 + o(1)) K_n, \end{aligned}$$

where  $o(1)$  represents a factor that converges to zero in probability.

Therefore, Wilks' Theorem and Slutsky's Theorem imply that  $K_n \xrightarrow{\mathcal{D}} \chi_{r-1}^2$  under  $H_0$  as  $n \rightarrow \infty$ .

**Example 20.2** (an extension of Example 20.1, due to R.A. Fisher) Consider testing  $H'_0 : p_j = \pi_j(\eta)$ ,  $j = 1, \dots, r$ , where  $\eta = (\eta_1, \dots, \eta_m)$  is a  $m$ -dimensional unknown parameter ( $m < r - 1$ ), and  $\pi_1, \dots, \pi_r$  are known functions.

Define  $K'_n = \sum_{j=1}^r \frac{(S_n^{(j)} - n\pi_j(\hat{\eta}))^2}{n\pi_j(\hat{\eta})}$  where  $\hat{\eta}$  is a MLE of  $\eta$  that satisfies the likelihood equation

$$\sum_{j=1}^r \frac{S_n^{(j)}}{\pi_j(\eta)} \cdot \frac{\partial \pi_j(\eta)}{\partial \eta} = 0, \quad l = 1, \dots, m.$$

Denote the set of parameters for  $H'_0$  by  $\Theta'_0$ . Then  $K'_n \xrightarrow{\mathcal{D}} \chi_{r-1-m}^2$  under  $P_\eta$  for every  $\eta \in \Theta'_0$ .

**Note:**  $H_0$  in Example 20.1 is a simple hypothesis, but  $H'_0$  in Example 20.2 is composite. Note that there are  $r$  categories, and  $m$  free parameters to be estimated. Usually,  $m$  is considerably smaller than  $r$  which makes Example 20.2 more useful than Example 20.1.

**Example 20.3** ( $\chi^2$  tests for independence in two-way tables) Consider two categorical random variables  $X$  and  $Y$ . Let  $(X_1, Y_1), \dots, (X_n, Y_n)$  be iid samples that follow the distribution  $p_{ij} = P(X_1 = i, Y_1 = j)$ ,  $i = 1, \dots, I$  and  $j = 1, \dots, J$  with the following setting:

*Parameters*

$$\begin{aligned} p_{ij} &= P(X_1 = i, Y_1 = j) \\ p_{i\cdot} &= P(X_1 = i) \\ p_{\cdot j} &= P(Y_1 = j) \end{aligned}$$

*Data*

$$\begin{aligned} n_{ij} &= \sum_{l=1}^n I_{\{X_l=i, Y_l=j\}} \\ n_{i\cdot} &= \sum_{l=1}^n I_{\{X_l=i\}} \\ n_{\cdot j} &= \sum_{l=1}^n I_{\{Y_l=j\}} \end{aligned}$$

*MLEs:*  $\widehat{p}_{ij} = n_{ij}/n$ ,  $\widehat{p}_{i\cdot} = n_{i\cdot}/n$ ,  $\widehat{p}_{\cdot j} = n_{\cdot j}/n$ .

For testing  $H_0$ :  $X$  and  $Y$  are independent, i.e.  $p_{ij} = p_{i\cdot}p_{\cdot j} \quad \forall i = 1, \dots, I, j = 1, \dots, J$ , use the statistic

$$K_n = \sum_{i=1}^I \sum_{j=1}^J \frac{(n_{ij} - n \frac{n_{i\cdot} n_{\cdot j}}{n})^2}{n \frac{n_{i\cdot} n_{\cdot j}}{n}} = n \left( \sum_{i=1}^I \sum_{j=1}^J \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right).$$

To apply Example 20.2, set  $r = IJ$  and  $m = (I - 1) + (J - 1)$ . Then  $r - 1 - m = (I - 1)(J - 1)$ , and  $K_n \xrightarrow{\mathcal{D}} \chi_{(I-1)(J-1)}^2$  under  $H_0$ , as  $n \rightarrow \infty$ .

**Note:**

(i) The exact LR test statistic in Example 20.3 is

$$2 \log \Lambda(\{n_{ij}\}) = 2 \sum_{i=1}^I \sum_{j=1}^J n_{ij} \log \frac{n_{ij}}{n \frac{n_{i.}}{n} \frac{n_{.j}}{n}} \xrightarrow{\mathcal{D}} \chi_{(I-1)(J-1)}^2$$

under  $H_0$ . In fact,  $2 \log \Lambda(\{n_{ij}\}) = K_n + O(1/n)$  for large  $n$ .

(ii)  $\chi^2$  goodness of fit tests have extensive applications, not only to discrete distributions, but also to continuous distributions by partitioning their supports. In principle, they can be used to compare any two distributions.

(iii) A common flaw shared by various  $\chi^2$  tests is that once  $H_0$  is rejected, those test statistics  $K_n$  and  $K'_n$  contain almost no information about the alternatives, i.e. statisticians receive little guidance as to what models actually generated the data.