
χ^2 tests for goodness of fit

Presented by Xi Zhao

July 19th, 2007

Herman Chernoff and E.L. Lehmann

David S. Moore and M. C. Spruill

David W. Hosmer and Stanley Lemeshow

Outline

1. Introduction
2. The Pearson χ^2 test for goodness of fit and the use of MLE in χ^2 test for fixed cells.
3. The χ^2 test for goodness of fit for random cells
4. Hosmer-Lemeshow goodness of fit test

Introduction

- The original Pearson χ^2 test for goodness of fit to a fixed distribution is based on observed cell frequencies in a set of fixed cells.
- In practice, we are more interested in testing the composite null hypothesis that the observations come from a parametric family $F(x|\theta)$ of distributions.
- If the estimator used is the MLE of θ based on the cell frequencies, the resulting test is the Pearson-Fisher χ^2 .
- If instead, the MLE based on the original data is used, the resulting Chernoff χ^2 does not have a limiting χ^2 null distribution.
- What is worse, the limiting null distribution usually depends on the true value of θ . This unpleasant situation leads to the use of cells which are themselves functions of the data, which we will call *random cells*.

-
- In section 3, we will show that the difference between the random-cell statistic and a fixed-cell statistic of similar form approaches zero in probability as the sample size increases.
 - Finally, we will use this result to prove the Hosmer Lemeshow goodness of fit test.

The Pearson χ^2 test for goodness of fit and the use of MLE in χ^2 test for fixed cells.

p_i — the probability of an observation falling into the i th of the k cells.

m_i — number of observations falling into the i th of the k cells.

\tilde{p}_i — the maximum likelihood estimate of p_i based on the cell frequencies

\hat{p}_i — the maximum likelihood estimate of p_i based on the original data

Let

$$R = \sum (m_i - np_i)^2 / np_i$$

which is the Pearson χ^2 test.

Define

$$\tilde{R} = \sum (m_i - n\tilde{p}_i)^2 / n\tilde{p}_i$$

which is called the Pearson-Fisher χ^2 test.

$$\hat{R} = \sum (m_i - n\hat{p}_i)^2 / n\hat{p}_i$$

which is called the Chernoff χ^2 test.

Regularity conditions:

(i) Every $p_i(\theta)$ has continuous derivatives $\frac{\partial p_i}{\partial \theta_j}$ and $\frac{\partial^2 p_i}{\partial \theta_j \partial \theta_k}$

(ii) The matrix $D = \left\{ \frac{\partial p_i}{\partial \theta_j} \right\}$, where $i = 1, \dots, k$ and $j = 1, \dots, s$, is of rank s

Let

$$m_i - np_i = \sqrt{np_i} \epsilon_i$$

$$n(\tilde{p}_i - p_i) = \sqrt{np_i} \tilde{\nu}_i$$

$$n(\hat{p}_i - p_i) = \sqrt{np_i} \hat{\nu}_i$$

Then

$$R = \sum \epsilon_i^2 = \epsilon' \epsilon$$

$$\tilde{R} = \sum (\epsilon_i - \tilde{\nu}_i)^2 [1 + o_p(1)]$$

$$\hat{R} = \sum (\epsilon_i - \hat{\nu}_i)^2 [1 + o_p(1)]$$

Then we will show that:

CASE 1: $\tilde{R} = (\tilde{F}\epsilon)'(\tilde{F}\epsilon) + o_p(1)$

CASE 2: $\hat{R} = (\hat{F}\epsilon + \hat{G}\eta)'(\hat{F}\epsilon + \hat{G}\eta) + o_p(1)$

First compute $\tilde{\nu}_i$ to show that \tilde{R} is asymptotically a sum of squares of normally distributed random variables.

Proof of case 1:

In our case of \tilde{R} , the information matrix is given by

$$\tilde{J} = \left\| \sum_{r=1}^k \frac{1}{p_r} \frac{\partial p_r}{\partial \theta_i} \frac{\partial p_r}{\partial \theta_j} \right\| = D' D$$

where

$$D = \left\| \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} \right\|$$

Therefore

$$\begin{aligned}\tilde{\nu}_i &= \frac{\sqrt{n}(\tilde{p}_i - p_i)}{\sqrt{p_i}} = \sum_{j=1}^s \sqrt{n}(\tilde{\theta}_i - \theta_i) \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} + o_p(\mathbf{1}), \\ \tilde{\nu} &= D\sqrt{n}(\tilde{\theta} - \theta) + o_p(\mathbf{1}) = D\tilde{J}^{-1}D'\epsilon + o_p(\mathbf{1})\end{aligned}$$

Finally

$$\tilde{R} = (\tilde{F}\epsilon)'(\tilde{F}\epsilon) + o_p(\mathbf{1})$$

where

$$\tilde{F} = I - D\tilde{J}^{-1}D'$$

Proof of case 2:

Let $z = (z_1, \dots, z_k)$, where $z_i = 1$ if the observation falls in the i th cell and 0 otherwise. Let $f(z, \theta) = \prod p_i^{z_i}$, and let us assume that the value w of our random variable x determines z , and that the density function of x is given by

$$f^*(w, \theta) = \prod p_i^{z_i} g(w|z, \theta)$$

where g is the conditional density of x given z .

Now

$$\frac{\partial \log f^*(w, \theta)}{\partial \theta_j} = \sum_{i=1}^k \frac{z_i \frac{\partial p_i}{\partial \theta_j}}{p_i} + \frac{\partial \log g(w|z, \theta)}{\partial \theta_j}$$

Since the conditional expectation, given z , of

$$\left[\sum_{i=1}^k \frac{z_i \frac{\partial p_i}{\partial \theta_j}}{p_i} \right] \bullet \frac{\partial \log g(w|z, \theta)}{\partial \theta_j}$$

is zero, we have

$$\begin{aligned} \hat{J} &= \tilde{J} + J^* \\ \hat{A} &= \tilde{A} + A^* \end{aligned}$$

where

$$\begin{aligned} J^* &= \left\| E \left[\frac{\partial \log g(w|z, \theta)}{\partial \theta_i} \bullet \frac{\partial \log g(w|z, \theta)}{\partial \theta_j} \right] \right\| \\ A_i^* &= \frac{1}{n} \sum_{\alpha=1}^k \frac{\partial \log g(x|z^{(\alpha)}, \theta)}{\partial \theta_i} \\ \tilde{A}_i &= \frac{1}{n} \sum_{\alpha=1}^k \frac{z_i \frac{\partial p_i}{\partial \theta_j}}{p_i} \end{aligned}$$

and $z^{(\alpha)}$ is the α th observation on z . Now

$$\begin{aligned}\widehat{\nu}_i &= \frac{\sqrt{n}(\widehat{p}_i - p_i)}{\sqrt{p_i}} = \sum \sqrt{n}(\widehat{\theta}_j - \theta_j) \frac{1}{\sqrt{p_i}} \frac{\partial p_i}{\partial \theta_j} + o_p(1), \\ \widehat{\nu} &= D\sqrt{n}(\widehat{\theta} - \theta) + o_p(1) \\ &= D(\widetilde{J} + J^*)^{-1}(D'\epsilon + \sqrt{n}A^*) + o_p(1)\end{aligned}$$

hence

$$\widehat{R} = (\widehat{F}\epsilon + \widehat{G}\eta)'(\widehat{F}\epsilon + \widehat{G}\eta) + o_p(1)$$

where $\eta = \sqrt{n}A^*$, while $\widehat{F} = I - D(\widetilde{J} + J^*)^{-1}D'$ and $\widehat{G} = D(\widetilde{J} + J^*)^{-1}$. The asymptotic distribution of \widehat{R} is $(\widehat{F}\epsilon + \widehat{G}\eta)'(\widehat{F}\epsilon + \widehat{G}\eta)$.

To specify the distribution, we must know the asymptotic distribution of (ϵ, η) .

Applying the central limit theorem to

$$\left[z_1, z_2, \dots, z_k, \frac{\partial \log g(w|z, \theta)}{\partial \theta_1}, \dots, \frac{\partial \log g(w|z, \theta)}{\partial \theta_s} \right]$$

we see that

$$d\infty(\epsilon, \eta) = N \left[0, \begin{pmatrix} I - qq' & 0 \\ 0 & J^* \end{pmatrix} \right]$$

where q is the vector whose i th component is $\sqrt{p_i}$

Now, we know that

$$\begin{aligned} d(\epsilon) &= N(0, \Sigma), \\ d(\tilde{F}\epsilon) &= N(0, \tilde{\Sigma}) \\ d(\hat{F}\epsilon + \hat{G}\eta) &= N(0, \hat{\Sigma}) \end{aligned}$$

where

$$\begin{aligned} \Sigma &= I - qq' \\ \tilde{\Sigma} &= I - qq' - D\tilde{J}^{-1}D' \\ \hat{\Sigma} &= I - qq' - D(\tilde{J} + J^*)^{-1}D' \end{aligned}$$

If for symmetric matrices we write $K \geq L$ whenever $K - L$ is nonnegative definite, then

$$\Sigma \geq \hat{\Sigma} \geq \tilde{\Sigma}$$

We now need the following important lemma:

Lemma 1 *If $d(y) = N(0, U)$ where the characteristic roots of U are $\lambda_1, \lambda_2, \dots, \lambda_k$, then*

$$d(y'y) = d\left(\sum \lambda_i z_i^2\right)$$

where $d(z) = N(0, I)$.

As a consequence of this Lemma, we have

- Σ has for characteristic roots $k - 1$ ones and 1 zero while $\tilde{\Sigma}$ has for characteristic roots $k - s - 1$ ones and $s + 1$ zeros.
- **NOTE:** A direct proof of the properties of the characteristic roots of Σ and $\tilde{\Sigma}$ may be given by showing that qq' and $D\tilde{J}^{-1}D'$ are projection operators on orthogonal manifolds of dimension 1 and s .

$$\begin{aligned} (qq')(qq') &= q\left(\sum p_i\right)q' = qq' \\ (D\tilde{J}^{-1}D')(D\tilde{J}^{-1}D') &= (D\tilde{J}^{-1}\tilde{J}\tilde{J}^{-1}D') = D\tilde{J}^{-1}D' \\ (D\tilde{J}^{-1}D')(qq') &= 0 \end{aligned}$$

-
- Since $\Sigma \geq \widehat{\Sigma} \geq \widetilde{\Sigma}$, it follows that $\widehat{\Sigma}$ has for characteristic roots : $k - s - 1$ ones, 1 zero and s roots $\lambda_1, \lambda_2, \dots, \lambda_s$ between zero and one.
 - The roots $\lambda_1, \lambda_2, \dots, \lambda_s$ which determine the distribution of the test criterion \widehat{R} can be obtained from the following lemma.
 - If $\mu_i = 1 - \lambda_i$, then the μ_i are the characteristic roots of the determinant equation $|\widetilde{J} - \mu\widehat{J}| = 0$
 - The conclusions are as follows:

CONCLUSION 1: The asymptotic distribution of R is that of χ^2 with $k - 1$ degrees of freedom,

CONCLUSION 2: The asymptotic distribution of \widetilde{R} is that of χ^2 with $k - s - 1$ degrees of freedom, where s is the number of population parameters being estimated.

CONCLUSION 3: The limiting distribution of \widehat{R} lies between those of \widetilde{R} and of R . More generally, we shall show that under suitable regularity conditions, the asymptotic distribution of \widehat{R} is that of

$$\chi^2(k - s - 1) + \sum_{i=k-s}^{k-1} \lambda_i \chi_{1i}^2$$

where χ_{1i}^2 are independent χ^2 variables with 1 degree of freedom and the $0 < \lambda_i < 1$

The χ^2 test for goodness of fit for random cells

Basic notation, definitions

- This section deals with the case of random cells and they can be employed in the Pearson-Fisher statistic and Chernoff statistic as well.
- The essential technique is to show that the difference between the random-cell statistic and a fixed-cell statistic of similar form approaches zero in probability as the sample size increases.
- Let Y_1, Y_2, \dots, Y_n be observed independent R^k -valued random variables with df $F(x|\theta)$. The parameter θ ranges over an open set Ω_1 in R^m
- The random cells for the χ^2 tests are rectangles in R^k with edges parallel to the coordinate axes, which are denoted by $I_\sigma(\varphi)$, $\sigma = 1, 2, \dots, M$ and φ is a variable defined on an open set Ω_2

-
- In forming the χ^2 statistic, the unknown parameter θ is estimated by $\theta_n = \theta_n(Y_1, Y_2, \dots, Y_n)$ and the cells are chosen by $\varphi_n = \varphi_n(Y_1, Y_2, \dots, Y_n)$ with φ_0 fixed.

- The number of Y_1, Y_2, \dots, Y_n falling into the cell $I_\sigma(\varphi)$ will be denoted by $N_{n\sigma}(\varphi)$. The cell probability for this cell under (θ, φ) is $p_\sigma(\theta, \varphi) = \int_{I_\sigma(\varphi)} dF(x|\theta)$.

- The standardized cell frequencies are

$$v_{n\sigma}(\theta, \varphi) = \frac{N_{n\sigma}(\varphi) - np_\sigma(\theta, \varphi)}{[np_\sigma(\theta, \varphi)]^{\frac{1}{2}}}$$

which is the σ th component of an M -vector $V_n(\theta, \varphi)$.

- If $K(\theta, \varphi)$ is a symmetric $M \times M$ matrix for each (θ, φ) in $\Omega_1 \times \Omega_2$, a general χ^2 statistic has the form

$$T_n = V_n'(\theta_n, \varphi_n)K(\theta_n, \varphi_n)V_n(\theta_n, \varphi_n)$$

- Standard χ^2 statistics have this form with $K(\theta_0, \varphi_n) \equiv I_M$, so the statistic is $\|V_n(\theta_0, \varphi_n)\|^2$.

-
- The Pearson-Fisher statistic uses $\theta_n = \widetilde{\theta}_n$, where $\widetilde{\theta}_n$ is the MLE based on the cell frequencies.
 - The Chernoff statistic uses $\theta_n = \widehat{\theta}_n$, where $\widehat{\theta}_n$ is the MLE based on the original data.

General assumptions:

A1. Under $\theta = \theta_0$, $\theta_n - \theta_0 = O_p(n^{-\frac{1}{2}})$ and $\varphi_n - \varphi_0 = o_p(1)$.

A2. For each σ , $p_\sigma(\theta, \varphi)$ is continuous in (θ, φ) and continuously differentiable in θ in a neighborhood of (θ_0, φ_0)

A3 $F(x) = F(x|\theta_0)$ is continuous at every vertex $x(\varphi_0)$ of every cell $I\sigma(\varphi_0)$.

A4 $K(\theta, \varphi) = S(\theta, \varphi)S(\theta, \varphi)'$ for an $M \times M$ matrix $S(\theta, \varphi)$ with entries continuous in (θ, φ) at (θ_0, φ_0)

A5 Under $\theta = \theta_0$,

$$n^{\frac{1}{2}}(\theta_n - \theta_0) = n^{-\frac{1}{2}} \sum_{i=1}^n h(Y_i) + A\gamma + o_p(1)$$

for some m -vector A and measurable function $h(x)$.

Preliminary results:

- The basic tool used to relate random-cell chi-squared tests to fixed-cell tests is the weak convergence of empirical df process on the unit cube
- Suppose $G(x|\theta)$ is a family of continuous df's on E^k having all univariate marginal df's uniform on $[0, 1]$ and such that $G(x|\theta_0)$ is continuous. Let G_n be the empirical df after n observations from G and define the process

$$y_n(x) = n^{\frac{1}{2}}\{G_n(x) - G(x|\theta_0)\}$$

- Under the assumptions stated, y_n converges weakly to a Gaussian process y_0 such that $P(y_0 \in C_k) = 1$, where C_k is the set of continuous functions on E^k

General χ^2 statistics

- Both cell frequencies and cell probabilities for $I\sigma(\varphi)$ can be expressed in terms of the difference operator $\Delta_\sigma(\varphi)$ defined by

$$p_\sigma(\varphi) = \int_{I\sigma(\varphi)} dF(x) = \Delta_\sigma(\varphi)F$$

- $\Delta_\sigma(\varphi)$ can be expressed explicitly as a linear combination of $F(x(\varphi))$ for vertices $x(\varphi)$ of $I\sigma(\varphi)$
- The empirical process is defined by $W_n(x) = n^{\frac{1}{2}}\{F_n(x) - F(x|\theta_0)\}$.
- Suppose A1, A2 and A3 hold, then using the preliminary results, we have

$$\Delta_\sigma(\varphi_n)W_n - \Delta_\sigma(\varphi_0)W_n = o_p(1)$$

- If A1, A2 and A3 hold, then by the above result, we can show that

$$V_n(\theta_n, \varphi_n) = V_n(\theta_0, \varphi_0) - Bn^{\frac{1}{2}}(\theta_n - \theta_0) + B_{12}\gamma + o_p(1)$$

- This permits some immediate conclusions:

φ_n affects the large sample theory only through its limit φ_0 . Random-cell versions of all statistics of form T_n therefore differ by $o_p(1)$ from the corresponding statistics with fixed cells .

- If A1 through A5 hold, the statistic T_n has its limiting distribution of

$$\sum_{j=1}^M \lambda_j \chi_{1j}^2$$

where λ_j are the characteristic roots of Σ and the χ_{1j}^2 are independent χ^2 variables with 1 degree of freedom.

$$\begin{aligned} \Sigma &= I_M - qq' + BL'B - BE[h(Y)W(Y)'] \\ &\quad - E[W(Y)h(Y)']B' \\ q' &= (p_1^{\frac{1}{2}}, \dots, p_M^{\frac{1}{2}}), L = E[h(Y)h(Y)']; \end{aligned}$$

$W(y)$ is the M -vector with σ th component $[\mathbf{1}_\sigma(y) - p_\sigma]/p_\sigma^{\frac{1}{2}}$,

B is defined in the above;

One sample χ^2 tests

- We can facilitate a unified derivation of the limiting distribution of the one-sample statistics

$$T_{0n} = \|V_n(\theta_0, \varphi_n)\|^2 \quad (\text{Original Pearson})$$

$$T_{1n} = \|V_n(\bar{\theta}_n, \varphi_n)\|^2 \quad (\text{Pearson-Fisher})$$

$$T_{2n} = \|V_n(\hat{\theta}_n, \varphi_n)\|^2 \quad (\text{Chernoff-Lehmann})$$

- Suppose $m \leq M$ and the matrix with entries $\partial p_i / \partial \theta_j$ has rank m , then we have the following conclusion.

CONCLUSION 1: T_{0n} has a limiting distribution χ_{M-1}^2

CONCLUSION 2: T_{1n} has a limiting distribution χ_{M-m-1}^2

CONCLUSION 3: T_{2n} has a limiting distribution $\chi_{M-m-1}^2 + \sum_{j=M-m}^{M-1} \lambda_j \chi_{1j}^2$

Hosmer-Lemeshow goodness of fit test

Introduction:

Within the logistic regression setting, the commonly used goodness of fit tests are the Pearson chi-square test, the deviance statistic and the Hosmer-Lemeshow test. All these tests are based on comparing the observed vs. expected responses for the various combinations of the independent variables. However, the Pearson and deviance tests have the limitation that they are not appropriate when there are continuous independent variables. The Hosmer and Lemeshow approach first orders observed responses according to their fitted probabilities and then group them into g deciles. The statistic is given by

$$\chi_{HL}^2 = \sum_{k=0}^1 \sum_{j=1}^g \frac{(n_{kj} - np_{kj})^2}{np_{kj}}$$

Notations and assumptions

- Let $Y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ be the outcome variable and $X^T = \{X_1, X_2, \dots, X_n\}$ be the vector of independent variables. Also let

$$\pi(x) = P(Y = 1|X = x) = \frac{1}{1 + \exp(\beta_0 + \beta \cdot x)}$$

where $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$

- Assume that β_0 and β are known. Define a random variable W where $W_i = j$ if $c_{j-1} \leq \pi < c_j$, $j = 1, \dots, g$, $i = 1, \dots, n$, where c'_j 's are known constants such that $0 = c_0 < c_1 < \dots < c_{g-1} < c_g = 1$.
- Let n_{kj} be the frequency of occurrences of the pair $(y_i = k, W_i = j)$ in the sample, $k = 0, 1$ and $j = 1, 2, \dots, g$.

Then the $P(Y = k, W = j) = p_{kj}$ can be computed as

$$p_{1j} = P(Y = 1, W = j) = \int_{c_{j-1}}^{c_j} \pi f(\pi) d\pi = \alpha_j,$$

$$j = 1, 2, \dots, g$$

$$p_{0j} = P(Y = 0, W = j) = \int_{c_{j-1}}^{c_j} (1 - \pi) f(\pi) d\pi$$

$$= \gamma_j - \alpha_j, \quad j = 1, 2, \dots, g$$

where $\gamma_j = P(c_{j-1} \leq \pi(x) < c_j)$

- With p_{kj} defined as above, the test statistic is

$$H_g = \sum_{k=0}^1 \sum_{j=1}^g \frac{(n_{kj} - np_{kj})^2}{np_{kj}}$$

General test statistics using MLE

- Suppose the sample density function \hat{f} assigns probability $\frac{1}{n}$ to each observed π and 0 to others. For the maximum likelihood estimates $\hat{\beta}_0$ and $\hat{\beta}$, this yields

$$\hat{\alpha}_j = \int_{c_{j-1}}^{c_j} \hat{\pi} \hat{f}(\pi) d\pi = \frac{1}{n} \sum_{r \in \hat{I}_j} \hat{\pi}_r$$

where $\hat{I}_j = \{i : c_{j-1} \leq \hat{\pi}_i < c_j\}$, $j = 1, \dots, g$, and

$$\hat{\gamma}_j = \int_{c_{j-1}}^{c_j} \hat{f}(\pi) d\pi = \hat{n}_{\cdot j} / n$$

where $\hat{n}_{\cdot j}$ is the number of indices falling in \hat{I}_j . Defining $\hat{p}_{1j} = \hat{\alpha}_j$ and $\hat{p}_{0j} = \hat{\gamma}_j - \hat{\alpha}_j$, $j = 1, \dots, g$, the test statistic is

$$\hat{H}_g = \sum_{k=0}^1 \sum_{j=1}^g \frac{(\hat{n}_{kj} - n\hat{p}_{kj})^2}{n\hat{p}_{kj}}$$

Large sample distribution of \widehat{H}_g

- The distribution of \widehat{H}_g cannot be obtained from a straightforward application of the usual theory used for goodness of fit tests.
- Parameter estimates are determined using maximum likelihood and the frequencies \widehat{n}_{kj} depend on the estimated parameters, namely the cells are random not fixed.

- The matrix $\Sigma(\widehat{H}_g)$ may be expressed as

$$\Sigma(\widehat{H}_g) = I - \widehat{q}\widehat{q}' - \widehat{B}\widehat{J}^{-1}\widehat{B}'$$

where $\widehat{q}' = (\widehat{\alpha}_1, \widehat{\gamma}_1 - \widehat{\alpha}_1, \dots, \widehat{\alpha}_g, \widehat{\gamma}_g - \widehat{\alpha}_g)$,

- The matrix \widehat{B} is $2g \times (n+1)$ and has the general element $\frac{1}{\sqrt{p_{kj}}} \frac{\partial p_{kj}}{\partial \beta_l}$, $k = 0, 1; j = 1, \dots, g; l = 0, \dots, n$
- The matrix \widehat{J}^{-1} is $(n+1) \times (n+1)$ and is the inverse of the information matrix.

-
- Suppose $\lambda_1, \lambda_2, \dots, \lambda_{n+1}$ are the non-zero or one eigenvalues of the matrix $\Sigma(\widehat{H}_g)$. Then the asymptotic distribution of \widehat{H}_g is

$$\chi^2(g - n - 1) + \sum_{i=1}^{n+1} \lambda_i \chi_{1i}^2$$

- The simulations indicate that the contribution of $\sum_{i=1}^{n+1} \lambda_i \chi_{1i}^2$ is approximately that of $\chi^2(n - 1)$ and a good approximation to the distribution is $\chi^2(g - 2)$