A CHI-SQUARE STATISTIC FOR GOODNESS-OF-FIT TESTS

BU-386-M

by

August, 1971

K. C. Rao and D. S. Robson

Cornell University, Ithaca, N. Y.

Abstract

Chernoff and Lehman (Ann. Math. Stat., 1954) have shown that the chi-square goodness-of-fit statistic is asymptotically distributed as $\chi^2_{r-s-1} + \lambda_1 Z^2_{r-s} + \ldots + \lambda_s Z^2_{r-1}$ when the r classes are predetermined and the s unknown parameters are estimated by maximum likelihood from the ungrouped data. The Z_i are NIID(0,1) independent of χ^2_{r-s-1} and the λ_j, $0 < \lambda_j < 1$, depend on the s unknown parameters in $F(x;\theta)$. Subsequent papers have shown that this same result applies in the more realistic and useful case where only the number of classes r and their probability content with respect to $F(x;\hat{\theta})$ are predetermined. In either case the joint conditional distribution of the class frequencies $\nu = (\nu_1, \ldots, \nu_{r-1})$, conditioned on $\hat{\theta}$, is asymptotically nonsingular multinormal and the quadratic form $Q_{r-1}(\nu;\hat{\theta},\theta)$ of this conditional distribution is therefore asymptotically distributed as χ^2_{r-1}.

If $F(x;\theta)$ belongs to the Koopman-Pitman family then this quadratic form does not depend on θ; in other cases the substitution of $\hat{\theta}$ for θ still gives $Q_{r-1}(\nu;\hat{\theta},\hat{\theta}) \xrightarrow{\mathcal{L}} \chi^2_{r-1}$. Structurally, this statistic takes the form $Q_{r-1}(\nu;\hat{\theta},\hat{\theta}) = X^2 + Q^*_s(\nu;\hat{\theta},\hat{\theta})$ where Q^*_s is a function of the estimated information matrix and

$$Q^*_s(\nu;\hat{\theta},\hat{\theta}) \xrightarrow{\mathcal{L}} (1-\lambda_1)Z^2_{r-s} + \ldots + (1-\lambda_s)Z^2_{r-1}.$$

Corresponding results are obtained when maximum likelihood estimators are replaced by other asymptotically normal consistent estimators.

A CHI SQUARE STATISTIC FOR GOODNESS-OF-FIT TESTS

by

BU-386-M

K. C. Rao and D. S. Robson
Cornell University, Ithaca, New York

August, 1971

Suppose we wish to test the hypothesis:

H_0: $x_1,\cdots,x_n$ are i.i.d random variables obtained from $f(x;\theta)$.

We assume that the functional form of f is known but $\theta = (\theta_1,\cdots,\theta_s)$ is unknown. Suppose θ were estimated from the data using maximum likelihood method by solving

$$A_j(\hat{\theta}) = \frac{\partial}{\partial\theta_j} \sum_{\alpha} \log f(x_\alpha;\theta)\Big|_{\theta=\hat{\theta}} = 0 , \quad j=1,\cdots,s .$$

Let the range of X be divided into r disjoint and contiguous class intervals $(I_1,\cdots,I_r)$. Let

$$g_i(x_\alpha) = 1 \text{ if } x_\alpha \in I_i$$

$$= 0 \text{ if } x_\alpha \notin I_i , \quad i=1,\cdots,r$$

Let $\nu_i = \sum_\alpha g_i(x_\alpha)$ = number of $x_\alpha \in I_i$. Let $p_i(\theta) = P_{H_0}(x \in I_i)$. Then $\sum_i p_i(\theta) = 1$. Let $\hat{p}_i = p_i(\hat{\theta})$.

Chernoff and Lehmann (AMS 1954) showed that

$$X^2 = \sum_i (\nu_i - n\hat{p}_i)^2/n\hat{p}_i \xrightarrow{\mathcal{L}} z_1^2+\cdots+z_{r-s-1}^2 + \lambda_1 z_{r-s}^2+\cdots+\lambda_s z_{r-1}^2$$

where $z_i \sim \text{NID}(0,1)$ and $\lambda(0 < \lambda < 1)$ may depend on θ.

Later A. R. Roy (1956) and G. S. Watson (J.R.S.S., Ser. B, 1958) showed that in the case the class intervals are selected as functions of $\hat{\theta}$ a similar result holds.

Handout for joint statistical meetings of American Statistical Association, Institute of Mathematical Statistics and Biometric Society, Fort Collins, Colorado, August 23-26, 1971.

We consider these two cases predetermined and variable interval cases, and develop a goodness-of-fit statistic which is asymptotically χ^2 distributed.

In what follows, Σ's, A's, p's are functions of θ.

Predetermined class intervals.

Theorem 1. If $\Sigma = \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{12} \\ \hline \Sigma_{21} & \Sigma_{22} \end{array}\right]$ is the covariance matrix of $\sqrt{n}\left(\frac{\nu_1}{n},\cdots,\frac{\nu_{r-1}}{n}\right.$ $\left. A_1,\cdots,A_s\right)$ and $V = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ is the conditional covariance matrix of $\frac{\nu}{\sqrt{n}} = \frac{1}{\sqrt{n}}(\nu_1,\cdots,\nu_{r-1}$ given $\sqrt{n}\,A = \sqrt{n}\,(A_1,\cdots,A_s)$. Then

$$Q_{r-1}(\nu;\theta) = \frac{1}{n}\,(\nu - np - \Sigma_{12}\Sigma_{22}^{-1}A)'\,V^{-1}\,(\nu - np - \Sigma_{12}\Sigma_{22}^{-1}A) \xrightarrow{\mathcal{L}} \chi^2_{r-1}$$

Theorem 2.

$$Q_{r-1}(\nu;\hat{\theta}) = \sum_i \frac{(\nu_i - n\hat{p}_i)^2}{n\hat{p}_i} + \frac{1}{n}\sum_{j,k}\left(\sum_i \frac{(\nu_i - n\hat{p}_i)}{\hat{p}_i}\,\frac{\partial p_i}{\partial\theta_j}\bigg|_{\theta_j=\hat{\theta}_j}\right)$$

$$\left(\sum_i \frac{(\nu_i - n\hat{p}_i)^2}{\hat{p}_i}\,\frac{\partial p_i}{\partial\theta_k}\bigg|_{\theta_k=\hat{\theta}_k}\right)\hat{a}^{jk}$$

$$\xrightarrow{\mathcal{L}} z_1^2+\cdots+z_{r-s-1}^2 + \lambda_1 z_{r-s}^2+\cdots+\lambda_s z_{r-1}^2 + (1-\lambda_1)z_{r-s}^2+\cdots+(1-\lambda_s)z_{r-1}^2$$

where $(a^{jk}) = -(\tilde{J}-\hat{J})^{-1}$

$$\tilde{J} = \left(\sum_i \frac{1}{p_i}\,\frac{\partial p_i}{\partial\theta_j}\right), \quad \hat{J} = \left(E\left(\frac{\partial \log f(x;\theta)}{\partial\theta_j}\,\frac{\partial \log f(x;\theta)}{\partial\theta_k}\right)\right)$$

and $(\hat{a}^{jk}) = (a^{jk}(\hat{\theta}))$.

Variable intervals.

We suppose that the r classes $(\hat{I}_1,\cdots,\hat{I}_r)$ are formed as functions of $\hat{\theta}$, the estimated parameters. Using multidimensional central limit theorem we observe that the conditional distribution of $\frac{1}{\sqrt{n}}(\nu_1,\cdots,\nu_{r-1})$ given $\hat{\theta}$ is asymptotically multinormal with $E(\nu_i|\hat{\theta}) = nP_i$, $Var(\nu_i|\hat{\theta}) = n(P_i-P_{ii}) + n^2(P_{ii}-P_i^2)$, $Cov(\nu_i,\nu_h|\hat{\theta})$ $= -nP_{ij} + n^2(P_{ih}-P_iP_h)$ where $P_i = \int_{\hat{I}_i} f_n(x_1|\hat{\theta})dx_1$, $P_{ih} = \int_{\hat{I}_i}\int_{\hat{I}_h} f_n(x_1,x_2|\hat{\theta})dx_1dx_2$, $(i,h=1,\cdots,r)$, and $f_n(x|\hat{\theta})$ is the conditional density of x given $\hat{\theta}$.

To compute P_i and P_{ih} we need an approximation for $f_n(x|\hat{\theta})$ in terms of $f(x;\hat{\theta})$ which is obtained by substituting $\hat{\theta}$ for θ in $f(x;\theta)$. For brevity, we denote $f = f(x;\hat{\theta})$, $\partial' f = \left(\frac{\partial f}{\partial\hat{\theta}_1},\cdots,\frac{\partial f}{\partial\hat{\theta}_s}\right)$ and $f_i = f(x_i;\hat{\theta})$.

Lemma: If $var(\hat{\theta}) \to 0$ as $n \to \infty$ then

$$f_n(x|\hat{\theta}) = f - \frac{1}{2n}\partial' f\,\hat{\hat{J}}^{-1}\,\partial f + o_p(\tfrac{1}{n})$$

where

$$\hat{\hat{J}} = \hat{J}(\hat{\theta}) .$$

Using this approximation we obtain

$$P_i = \hat{p}_i - \frac{1}{2n}\int_{\hat{I}_i}\partial'\,\hat{\hat{J}}^{-1}\,\partial f_1 dx_1$$

$$P_{ih} = \hat{p}_i\hat{p}_h - \frac{1}{2n}\left(\hat{p}_i\int_{\hat{I}_h}\partial'\hat{\hat{J}}^{-1}\partial f_2 dx_2 + 2\int_{\hat{I}_h}\int_{\hat{I}_i}\partial' f_1\hat{\hat{J}}^{-1}\partial f_2 dx_1dx_2\right.$$

$$\left. + \hat{p}_h\int_{\hat{I}_i}\partial'\hat{\hat{J}}^{-1}\partial f_1 dx_1\right) .$$

[Here $\hat{p}_i$'s may depend on $\hat{\theta}$.]

Using these expressions for P_i and P_{ih} we obtain, correct up to terms of $0_p(\frac{1}{n})$,

$$E\left(\frac{\nu_i}{\sqrt{n}} \,|\hat{\theta}\right) = \sqrt{n}\,\hat{p}_i$$

$$\mathrm{Var}\left(\frac{1}{\sqrt{n}}\,\nu_i|\hat{\theta}\right) = \hat{p}_i(1-\hat{p}_i) - \int_{\hat{I}_i}\int_{\hat{I}_i} \partial' f_1 \hat{\hat{J}}^{-1} \partial f_2 dx_1 dx_2$$

$$\mathrm{Cov}\left(\frac{1}{\sqrt{n}}\,\nu_i, \frac{1}{\sqrt{n}}\,\nu_h|\hat{\theta}\right) = -\hat{p}_i\hat{p}_h - \int_{\hat{I}_i}\int_{\hat{I}_h} \partial' f_1 \hat{\hat{J}}^{-1} \partial f_2 dx_2 \quad .$$

Writing $u_{ij} = \int_{\hat{I}_i} \frac{\partial f}{\partial \hat{\theta}_j} dx$, we obtain the conditional mean and covariance matrix of $\frac{\nu}{\sqrt{n}}$ given $\hat{\theta}$ as $\sqrt{n}\,\hat{p} = (\hat{p}_1,\cdots,\hat{p}_{r-1})$ and $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$ where $\Sigma_{11} = (\hat{p}_i(\delta_{ih}-\hat{p}_h))$ (where $\delta_{ih} = 1$ if $i=h$, 0 otherwise), $\Sigma_{12} = (u_{ij})$ and $\Sigma_{22} = \hat{\hat{J}}$. Now applying theorem 1 and a little algebra we obtain

$$Q_{r-1}(\nu;\hat{\theta}) = \sum \frac{(\nu_i - n\hat{p}_i)^2}{n\hat{p}_i} + \frac{1}{n}\sum_{j,k}\left(\sum_i \frac{(\nu_i - n\hat{p}_i)}{\hat{p}_i}\mu_{ij}\right)\left(\sum \frac{(\nu_i - n\hat{p}_i)}{\hat{p}_i} u_{ik}\right)\hat{a}^{jk}$$

where $(\hat{a}^{jk}) = -(\tilde{\tilde{J}}-\hat{\hat{J}})^{-1}$, $\tilde{\tilde{J}} = (\Sigma_i \frac{1}{p_i} u_{ij}u_{ik})$, $\hat{\hat{J}} = \hat{J}(\hat{\theta})$.

Example. Exponential case:

$$H_0: f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \cdot \quad x \geq 0$$

$$= 0 \qquad x < 0 \ , \quad \theta \text{ unknown} \ .$$

The sample mean $\bar{x}$ is a sufficient statistic for θ. Let the r class intervals be $(\bar{x}z_{i-1}, \bar{x}z_i)$, $i=1,\cdots,r-1$, with $z_0=0$ and $z_r=\infty$. Let z_i's be determined from

$$\int_{\bar{x}z_{i-1}}^{\bar{x}z_i} f(x;\bar{x})dx = \frac{1}{r}$$

which implies $z_i = -\log(1-\frac{i}{r})$. Let $v_i = \int_{\bar{x}z_{i-1}}^{\bar{x}z_i} \frac{\partial f(x;\bar{x})}{\partial \bar{x}}\, dx = \frac{1}{\bar{x}}\left(z_{i-1}e^{-z_{i-1}} - z_i e^{-z_i}\right)$ and $u_i = \bar{x}v_i$, $(i=1,\cdots,r)$. After some simplification, we obtain the test statistic as

$$Q_{r-1}(\nu;\bar{x}) = \frac{r}{n}\sum_i (\nu_i - \frac{n}{r})^2 + \frac{r^2}{n}\,\frac{[\Sigma_i(\nu_i - n/r)u_i]^2}{(1-r\Sigma_i u_i^2)} \quad .$$

Table 1. Simulated sampling distribution of goodness-of-fit statistics for samples (3500) of size 50 from an exponential distribution with unknown mean.

r \ α	.975	.95	.90	.80	.70	.50	.30	.20	.10	.05	.025	.01
4	.978	.94943	.90571	.798	.69971	.51543	.30686	.19686	.10257	.05514	.02714	.00771
	1.000	.96943	.96943	.85229	.85229	.55771	.33971	.26543	.12486	.06857	.032	.01286
6	.97257	.94371	.89829	.79743	.69571	.49514	.29971	.184	.08857	.04029	.02057	.00771
	.98343	.95829	.92743	.81714	.73371	.51886	.32486	.20857	.096	.04486	.02114	.00886
8	.982	.95514	.90429	.79800	.69657	.49743	.288	.18714	.09371	.04714	.02429	.00886
	.97886	.96343	.90486	.83571	.70086	.52571	.31571	.19629	.10743	.04886	.02829	.012
10	.97943	.95343	.90771	.80971	.70686	.49	.278	.18714	.09029	.04629	.02571	.01057
	.98143	.96686	.92429	.82143	.73857	.50143	.30371	.19486	.09686	.04914	.02914	.00886
12	.97914	.95457	.90743	.80657	.69943	.48943	.27771	.18486	.090857	.04514	.02486	.00829
	.98829	.95857	.90971	.8	.70657	.51114	.29829	.18486	.09657	.04943	.02857	.01114

α: nominal size: In each cell first line gives the sampling distribution of Q and the second line that of X^2.

Table 2. Simulated sampling distribution of goodness-of-fit statistics for samples (3500) of size 500 from an exponential distribution with unknown mean.

r	α	.990	.975	.95	.90	.80	.75	.70	.60	.50	.40
3	Q	.99133	.97567	.94967	.89967	.80067	.74267	.701	.595	.49633	.393
	R	.988	.97767	.95233	.89267	.803	.750	.69267	.60067	.50233	.40633
	X^2	1.00	1.00	1.00	.993	.944	.92033	.87667	.79267	.678	.56233
2	Q	.94733	.94733	.94733	.848	.75167	.75167	.66067	.58567	.50067	.42367
	R	.961	.961	.961	.89733	.83167	.76333	.70067	.63433	.50967	.41333

r	α	.30	.25	.20	.10	.05	.025	.02	.01	.005	.001
3	Q	.299	.24333	.192	.09367	.05167	.025	.02133	.01067	.00467	.00067
	R	.29933	.24867	.20633	.10667	.05267	.02833	.02367	.01067	.00433	.0
	X^2	.42767	.35233	.27533	.13333	.06567	.03267	.024	.01267	.00067	.00067
2	Q	.30367	.24333	.19133	.09167	.05433	.02133	.02133	.01167	.00533	.002
	R	.317	.27833	.20633	.10033	.05833	.025	.01833	.01133	.00533	.001

R: Chi square goodness-of-fit statistic when the parameter is assumed known and equal to 1.