The Annals of Statistics
1976, Vol. 4, No. 2, 409-412

A COMPARISON OF CHI-SQUARE GOODNESS-OF-FIT TESTS BASED ON APPROXIMATE BAHADUR SLOPE[1]

By M. C. Spruill

Georgia Institute of Technology

The Pearson-Fisher χ^2 statistic is asymptotically chi-square under the null hypothesis with $M - m - 1$ degrees of freedom where $M =$ number of cells and $m =$ dimension of parameter. The Chernoff-Lehmann statistic is a weighted sum of chi-squares and the Kambhampati statistic is χ^2 with $M - 1$ degrees of freedom. The approximate Bahadur slopes of the tests based on these statistics are computed. It is shown that the Kambhampati test always dominates the Chernoff-Lehmann and that no such dominance exists between the Pearson-Fisher test and Kambhampati test, or the Pearson-Fisher and Chernoff-Lehmann.

1. Introduction. The original Pearson χ^2 test of fit to a fixed distribution is based on observed cell frequencies in a set of fixed cells. This test may be modified to test the hypothesis that the observations come from a member of the parametric family $F(y \mid \theta)$ by estimating the parameter θ from the data and allowing cells whose boundaries are functions of the data. It is the purpose of this note to investigate the performance against fixed noncontiguous alternatives of three basic modifications of the Pearson test. These are the Pearson-Fisher (P-F) χ^2 with parameters estimated from grouped data, the Chernoff-Lehmann (C-L) χ^2 with parameters estimated from the original data, and Kambhampati's (K) nonstandard χ^2 statistic. These statistics are discussed in detail in Moore and Spruill (1975). The measure of performance is approximate Bahadur slope.

It should be noted that the results which follow could differ from those based on exact Bahadur slope since Bahadur (1967) has pointed out that there may be large differences between exact and approximate slopes especially at alternatives far from the null hypothesis.

2. Notation and assumptions. The notation used is that of Moore and Spruill (1975). Specifically, let $\{F(\cdot \mid \theta)\}$ be a family of df's on R^k indexed by the m-dimensional real parameter θ. The cells for the χ^2 tests are rectangles in R^k with sides parallel to the coordinate axes. They are functions of the variable γ. The resulting cells are denoted by $\{I_\sigma(\gamma)\}_{\sigma=1}^M$. The number of $y_1, y_2, \cdots, y_n$ falling in $I_\sigma(\gamma)$ is $N_{n\sigma}(\gamma)$ and the probability of $I_\sigma(\gamma)$ under $F(\cdot \mid \theta)$ is $p_\sigma(\theta, \gamma)$. The unknown parameter θ is estimated by $\theta_n = \theta_n(y_1, \cdots, y_n)$ while the value of γ

Received August 1974; revised July 1975.

[1] This research was in part sponsored by the Air Force Office of Scientific Research under Grant No. 72-2350 at Purdue University.

AMS 1970 *subject classifications.* Primary 62G20; Secondary 62F10.

Key words and phrases. Chi-square test, Bahadur slope.

is determined by $\gamma_n = \gamma_n(y_1, \cdots, y_n)$. Letting $V_n(\theta, \gamma)$ be the M-vector whose σth component is

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

the three χ^2 tests are based on the three statistics

$$T_{1n} = \|V_n(\bar{\theta}_n, \gamma_n)\|^2 \quad \text{(P–F)}$$
$$T_{2n} = \|V_n(\hat{\theta}_n, \gamma_n)\|^2 \quad \text{(C–L)}$$
$$T_{3n} = V_n'(\hat{\theta}_n, \gamma_n)Q(\hat{\theta}_n, \gamma_n)V_n(\hat{\theta}_n, \gamma_n) \quad \text{(K)}.$$

Here $\bar{\theta}_n$ maximizes $\sum_{\sigma=1}^{M} N_{n\sigma}(\gamma_n) \log p_\sigma(\theta, \gamma_n)$, $\hat{\theta}_n$ maximizes $\sum_{j=1}^{n} \log f(y_j \mid \theta)$, $f(y \mid \theta)$ is the density of $F(y \mid \theta)$ with respect to a fixed measure μ,

$$Q(\theta, \gamma) = (I_m - B(\theta, \gamma)J^{-1}(\theta)B'(\theta, \gamma))^{-1},$$

$B(\theta, \gamma)$ is the $M \times m$ matrix with (i, j)th entry $p_i^{-\frac{1}{2}}(\theta, \gamma)\partial p_i/\partial\theta_j(\theta, \gamma)$, and finally $J(\theta)$ is the $m \times m$ matrix with (i, j)th entry $E((\partial \log f(y \mid \theta)/\partial\theta_i)(\partial \log f(y \mid \theta)/\partial\theta_j))$.

We make the following assumptions.

(A1) Under any member $F(\cdot \mid \theta)$ of the family $\{F(\cdot \mid \theta)\}$

T_{1n} has limiting distribution χ^2_{M-m-1},
T_{2n} has limiting distribution $\chi^2_{M-m-1} + \sum_{j=M-m}^{M-1} \lambda_j \chi^2_{1j}$,
T_{3n} has limiting distribution χ^2_{M-1}.

(A2) $p_\sigma(\theta, \gamma)$ and $Q(\theta, \gamma)$ are continuous in (θ, γ) and $\int_{I_\sigma(\gamma)} g\, d\mu$ is continuous in γ.

(A3) Under G, $\gamma_n \to \gamma_0$ w.p. 1, $\hat{\theta}_n \to \hat{\theta}_0$ w.p. 1, and $\bar{\theta}_n \to \bar{\theta}_0$ w.p. 1.

Here χ^2_{M-m-1} and χ^2_{1j} represent independent central χ^2 rv's with $M - m - 1$ and 1 degrees of freedom and the λ_j are fixed constants. Moore and Spruill (1975) give conditions under which the T_{in} have these limiting null distributions. The remaining assumptions involve a specified alternative df $G \notin \{F(\cdot \mid \theta)\}$ and having density g w.r.t. μ. Perlman (1972) gives conditions under which (A3) holds.

3. Comparison of the tests. The first lemma is an easy consequence of Lemma 1 in Spruill (1975).

LEMMA. *Under the assumptions above and subject to the* rhs *being positive and finite, the approximate Bahadur slope of the test based on*

$$T_{1n} \quad \textit{is} \quad \psi_1(G) = \|V_0(\bar{\theta}_0, \gamma_0)\|^2$$
$$T_{2n} \quad \textit{is} \quad \psi_2(G) = \|V_0(\hat{\theta}_0, \gamma_0)\|^2$$
$$T_{3n} \quad \textit{is} \quad \psi_3(G) = V_0'(\hat{\theta}_0, \gamma_0)Q(\hat{\theta}_0, \gamma_0)V_0(\hat{\theta}_0, \gamma_0),$$

where $V_0(\theta, \gamma)$ is the M-vector with σth *component*

$$V_{0\sigma}(\theta, \gamma) = [\textstyle\int_{I_\sigma(\gamma)} g\, d\mu - p_\sigma(\theta, \gamma)]/[p_\sigma(\theta, \gamma)]^{\frac{1}{2}}.$$

It is known (see Moore and Spruill (1975)) that when the matrix Q exists it is symmetric with eigenvalues $\beta_i \geqq 1$. This proves the following theorem.

THEOREM. (K *dominates* C–L.) *Whenever* ψ_2 *and* ψ_3 *exist* $\psi_3 \geqq \psi_2$.

By means of the following two examples it is shown that no other dominance relationship holds without conditions on the alternative G. That is, it is shown that both of $\psi_1 < \psi_2$ and $\psi_3 < \psi_1$ are possible.

EXAMPLE 1. ($\psi_1 < \psi_2$). Consider testing fit to $F(y \mid \theta) = \Phi(y - \theta)$. Let $\varepsilon > 0$ and t_1, t_2 be such that $-\infty < t_1 + \varepsilon < 0 < t_2 - \varepsilon < +\infty$ and $|t_1| \neq |t_2|$. Let G be a df satisfying

$$G(t_2 - \varepsilon) - G(t_1 + \varepsilon) = 1 \qquad \text{and}$$

$$\int_{t_1+\varepsilon}^{t_2-\varepsilon} x \, dG(x) = 0 \,.$$

Using the random cells $(-\infty, t_1 + \bar{x}]$, $(t_1 + \bar{x}, t_2 + \bar{x}]$, $(t_2 + \bar{x}, +\infty)$ T_{2n} has slope $\psi_2 = [\Phi(t_2) - \Phi(t_1)]^{-1} - 1$ and T_{1n} has slope $\psi_1 = [\Phi(t_2 - \bar{\theta}_0) - \Phi(t_1 - \bar{\theta}_0)]^{-1} - 1$. Since $\bar{\theta}_0$ is the unique (see Kulldorf (1961) Theorem 12.1) point which maximizes $\log [\Phi(t_2 - \theta) - \Phi(t_1 - \theta)]$ and $\bar{\theta}_0 \neq 0$ here, it follows that $\psi_2 > \psi_1$.

EXAMPLE 2. ($\psi_3 < \psi_1$). Consider testing fit to $F(x \mid \theta) = \Phi(x/\theta)$, $\theta > 0$, using the random cells $(-\infty, a_1 s_n]$, $(a_1 s_n, a_2 s_n]$, $\cdots$, $(a_{M-1} s_n, +\infty)$,

$$s_n = (n-1)^{-1} \sum_{j=1}^{n} (y_j - \bar{y})^2 \,.$$

For reasons to be seen later take $a_1^2 > 2a_{M-1}^2$ and $-\infty < a_1 < a_2 < \cdots < a_{M-1} < 0$. For each $\tau > 1$ let $G(y \mid \tau)$ be the df with a jump of τ^{-1} at $x = -(\tau - 1)^{\frac{1}{2}}$ and a jump of $1 - \tau^{-1}$ at $x = (\tau - 1)^{-\frac{1}{2}}$. We shall compute $\psi_1(G(\cdot \mid \tau))$ and $\psi_2(G(\cdot \mid \tau))$ and show that

(1) $\limsup_{\tau \to \infty} \psi_3(G(\cdot \mid \tau)) < \infty$ and

(2) $\psi_1(G(\cdot \mid \tau)) \to \infty$ as $\tau \to \infty$.

First (1) is easily seen, for since $\int y \, dG(y \mid \tau) \equiv 0$ and $\int y^2 \, dG(y \mid \tau) \equiv 1$ neither the matrix Q nor the denominators of the entries in V_0 depend upon τ. Turning to (2), we have that for sufficiently large τ, $\bar{\theta}_0(\tau)$ maximizes

$$\tau^{-1} \log [\Phi(a_1/\theta)] + (1 - \tau^{-1}) \log [1 - \Phi(a_{M-1}/\theta)] \,.$$

It follows from Kulldorf (1961) Theorem 13.1 that this expression is uniquely maximized for $\tau > 1 + a_1/a_{M-1}$ by $\bar{\theta}_0(\tau)$ satisfying

$$\frac{-a_1 \tau^{-1}}{\Phi(a_1/\theta)} \exp\left(\frac{-a_1^2}{2\theta^2}\right) + \frac{a_{M-1}(1 - \tau^{-1})}{1 - \Phi(a_{M-1}/\theta)} \exp\left(\frac{-a_{M-1}^2}{2\theta^2}\right) = 0 \,. \tag{3}$$

Manipulation of (3) yields

$$\frac{1}{\tau^2 \Phi(a_1/\theta)} \equiv \frac{(1 - \tau^{-1})^2 a_{M-1}^2 \Phi(a_1/\theta)}{(1 - \Phi(a_{M-1}/\theta))^2 a_1^2} \exp\left(\frac{2(a_1^2 - a_{M-1}^2)}{2\theta^2}\right) . \tag{4}$$

Using $1 - \Phi(x) \sim (2\pi)^{-\frac{1}{2}} x^{-1} e^{-x^2/2}$ for x large, we have

$$\exp \frac{2(a_1^2 - a_{M-1}^2)}{2\theta^2(\tau)} \Phi(a_1/\theta(\tau)) \sim \frac{\theta(\tau)}{|a_1|} \exp\left(\frac{(a_1^2 - 2a_{M-1}^2)}{2\theta^2(\tau)}\right) \tag{5}$$

since (3) implies $\theta(\tau) \to 0$ as $\tau \to \infty$. In view of our assumption that $a_1^2 > 2a_{M-1}^2$, (5) gives

$$\lim_{\tau\to\infty} \psi_1(G(\cdot \mid \tau)) \geqq \lim_{\tau\to\infty} \frac{1}{\tau^2\Phi(a_1/\theta(\tau))} = +\infty .$$

4. Remark. The points $\hat{\theta}_0$ and $\bar{\theta}_0$ in (A3) typically uniquely maximize

$$\int \log f(y \mid \theta) g(y)\, d\mu(y)$$

and

$$\sum_{\sigma=1}^{M} [\int_{I_\sigma(\gamma_0)} g(y)\, d\mu(y)] \log p_\sigma(\theta, \gamma_0)$$

respectively, so that the slopes in the lemma may be computed for nontrivial cases. The reader is referred to Perlman (1972) for relevant material.

Acknowledgment. The research reported here contains part of the author's doctoral thesis written at Purdue University. I wish to thank Professor D. S. Moore for his guidance and helpful comments.

REFERENCES

[1] KULLDORF, G. (1961). *Estimation from Grouped and Partially Grouped Samples.* Wiley, New York.

[2] MOORE, D. S. and SPRUILL, M. C. (1975). Unified large-sample theory of general chi-squared statistics for tests of fit. *Ann. Statist.* **3** 599–616.

[3] PERLMAN, M. D. (1972). On the strong consistency of approximate maximum likelihood estimators. *Proc. Sixth Berkeley Symp. Math. Statist. Prob.* **1** 263–281. Univ. of California Press.

[4] SPRUILL, M. C. (1976). Cell selection in the Chernoff–Lehmann chi-square statistic. *Ann. Statist.* **4** 375–383.

SCHOOL OF MATHEMATICS
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332