J. R. Statist. Soc. B (1984)
46, *No.* 3, *pp.* 440–464

Multinomial Goodness-of-fit Tests

By NOEL CRESSIE† and TIMOTHY R.C. READ‡

The Flinders University of South Australia

CSIRO, Sydney, Australia and Colorado State University, USA

[Received April 1983]

SUMMARY

This article investigates the family $\{I^\lambda; \lambda \in \mathbb{R}\}$ of power divergence statistics for testing the fit of observed frequencies $\{X_i; i = 1, \ldots, k\}$ to expected frequencies $\{E_i; i = 1, \ldots, k\}$. From the definition

$$2nI^\lambda = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{E_i} \right)^\lambda - 1 \right\} ; \; \lambda \in \mathbb{R},$$

it can easily be seen that Pearson's X^2 ($\lambda = 1$), the log likelihood ratio statistic ($\lambda = 0$), the Freeman–Tukey statistic ($\lambda = -\frac{1}{2}$) the modified log likelihood ratio statistic ($\lambda = -1$) and the Neyman modified X^2 ($\lambda = -2$), are all special cases. Most of the work presented is devoted to an analytic study of the asymptotic difference between different I^λ, however finite sample results have been presented as a check and a supplement to our conclusions. A new goodness-of-fit statistic, where $\lambda = \frac{2}{3}$, emerges as an excellent and compromising alternative to the old warriors, I^0 and I^1.

Keywords: BEST ASYMPTOTICALLY NORMAL ESTIMATOR; BAHADUR EFFICIENCY; GOODNESS-OF-FIT; LIKELIHOOD RATIO; MAXIMUM LIKELIHOOD; MINIMUM DISCREPANCY ESTIMATORS; PITMAN EFFICIENCY; POWER DIVERGENCES; SECOND ORDER EFFICIENCY

1. INTRODUCTION

Many tests of goodness-of-fit can be reduced to testing a hypothesis about the parameters $\Pi = (\pi_1, \ldots, \pi_k)$ from a multinomial distribution:

$$\Pr(\mathbf{X} = \mathbf{x}) = \frac{n!}{\pi_1! \ldots \pi_k!} \pi_1^{x_1} \ldots \pi_k^{x_k}, \tag{1.1}$$

where Π forms a probability distribution and the x's are non negative integers which sum to n. Here the multinomial random variable $\mathbf{X}$ is often derivative in that X_i is often the number out of n of independent and identically distributed (i.i.d.) $Y_1, \ldots, Y_n$ from $F(y; \Theta)$ belonging to the class C_i, where $\{C_i; i = 1, \ldots, k)$ is a set of mutually exclusive classes which exhaust the probability content of F.

To test in (1.1) the null hypothesis

$$H_0 : \Pi = \Pi_0, \tag{1.2}$$

where $\Pi_0 = (\pi_{01}, \ldots, \pi_{0k})$ is a prespecified probability vector, possibly the most commonly used

† *Present address*: Department of Statistics, Iowa State University, Ames, Iowa, 50011, USA.

‡ *Present address*: Hewlett Packard Co., Stanford Park Division, 1501 Page Mill Road, Palo Alto, CA 94304, USA.

0035–9246/84/46440 $2.00

statistic is Pearson's X^2 (Pearson, 1900):

$$X^2 = \sum_{i=1}^{k} (X_i - n\pi_{0i})^2 / n\pi_{0i}.$$

This has asymptotically a chi-squared distribution on $k-1$ degrees of freedom (write χ^2_{k-1}) under H_0, so that rejection occurs when the observed value of X^2 is greater than or equal to a prespecified percentage point found from the χ^2_{k-1} tables. Another popular test procedure is to replace X^2 with the log likelihood ratio statistic

$$G^2 = 2 \sum_{i=1}^{k} X_i \ln (X_i / n\pi_{0i}).$$

Which of these two is "best" has long provided interest, speculation and controversy in the literature. For reasons of space, we can only briefly mention the issues we believe to be important.

Cochran (1952) gives a nice review of the early development of X^2 and concludes that there is little to distinguish it from G^2. Since then various authors have looked at the two tests according to

i. finite sample comparisons under the null hypothesis (e.g. Chapman, 1976; Larntz, 1978),
ii. asymptotic and finite sample power comparisons for various alternative hypotheses, including the effect of varying the class intervals (e.g. Hoeffding, 1965; West and Kempthorne, 1972; Goldstein, Wolf and Dillon, 1976),
iii. asymptotic distribution theory under both null and certain contiguous alternatives, for k increasing with n (e.g. Holst, 1972; Morris, 1975; Koehler and Larntz, 1980),
iv. modifying the statistics to allow for estimation of unknown parameters (e.g. Moore and Spruill, 1975).

As well as X^2 and G^2, the Freeman–Tukey statistic,

$$T^2 = 4 \sum_{i=1}^{k} \{\sqrt{X_i} - \sqrt{(n\pi_{0i})}\}^2,$$

the Neyman modified X^2 statistic

$$NM^2 = \sum_{i=1}^{k} (X_i - n\pi_{0i})^2 / X_i,$$

and the modified log likelihood ratio statistic

$$GM^2 = 2 \sum_{i=1}^{k} n\pi_{0i} \ln (n\pi_{0i} / X_i),$$

are all asymptotically distributed as χ^2_{k-1} under H_0. Various properties and comparisons of these so-called chi-squared tests can be found in one or more of Watson (1959), Lancaster (1969), Moore (1976), Horn (1977), and Fienberg (1979).

There is no uniformly preferable test of H_0, however we believe in this paper that we have cleared away much of the undergrowth surrounding X^2 and G^2 by showing that there is a path between them and beyond. The essential feature which enables us to do this, is the general family of power divergence statistics defined in Section 2. Of course the null hypothesis H_0 is not always simple; in Section 2 we investigate inference procedures for $H_0 : \Pi = \Pi_0(\Theta)$, where $\Theta = (\theta_1, \ldots, \theta_s)$, $s < k-1$, is a row vector of unknown nuisance parameters. Note that here and henceforth Θ, Π are row vectors and Θ', Π' are column vectors.

In Section 2 we define the tests based on the power divergence statistics I^λ, $\lambda \in \mathbb{R}$, and investigate their properties under H_0; the results of Bishop, Fienberg and Holland (1975) (write BFH) are extended to cover this general family of tests. Efficiency and power comparisons are made in Section 3. Section 4 collects together a number of further ways to compare the I^λ statistics, while Section 5 shows that the same ideas have really existed in other goodness-of-fit problems for some time. The final Section gives the practitioner working rules to decide which multinomial goodness-of-fit test should be chosen.

2. TESTS BASED ON POWER DIVERGENCES

2.1. *The Power Divergence Statistics*

Each of the goodness-of-fit statistics defined in Section 1, namely X^2, G^2, T^2, NM^2 and GM^2, tries to indicate in different ways how observed multinomial variables $\{X_i\}$ differ from their expected values $\{n\pi_{0i}\}$, where it is assumed $\pi_{0i} > 0$, each i. In fact all of them are embedded in a family of power divergence statistics indexed by a real parameter λ. Let $E_i = n\pi_{0i}$; $i = 1, \ldots, k$, and define

$$2nI^\lambda(\mathbf{X}/n : \Pi_0) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{E_i} \right)^\lambda - 1 \right\} \; ; \; \lambda \in \mathbb{R} \; , \tag{2.1}$$

where for $\lambda = 0, -1$, $2nI^\lambda$ is defined by continuity; e.g.

$$2nI^0(\mathbf{X}/n : \Pi_0) = \lim_{\lambda \to 0} 2nI^\lambda(\mathbf{X}/n : \Pi_0) = 2 \sum_{i=1}^{k} X_i \ln\left(\frac{X_i}{E_i} \right) \tag{2.2}$$

Clearly $X^2 = 2nI^1$, $G^2 = 2nI^0$, $T^2 = 2nI^{-1/2}$, $GM^2 = 2nI^{-1}$, and $NM^2 = 2nI^{-2}$. The study and comparison of X^2, G^2, etc. and tests based on them by linking them through the index λ, gives a new perspective and understanding to an old problem. It is tempting to generalize even further to a test based on the statistic $\Sigma(X_i/n)\,\beta(X_i/E_i) - \beta(1)$, summing over i from 1 to k, β convex, however it adds little to the subsequent development. General results for all members of the family $\{I^\lambda; \lambda \in \mathbb{R}\}$ shall be proved and particularized to well known and special cases.

2.2. *Limiting Chi-squared Distributions*

In this subsection the number of cells k, is assumed to be fixed. Let us suppose for the moment that under $H_0 : \Pi = \Pi_0$, the multinomial probabilities $\{\pi_{0i}\}$ are completely specified. Then for $\lambda \neq 0, -1$

$$2nI^\lambda(\mathbf{X}/n : \Pi_0) = \frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0i} \left\{ (1 + \frac{X_i - n\pi_{0i}}{n\pi_{0i}})^{\lambda+1} - 1 \right\} .$$

Writing $V_i = (X_i - n\pi_{0i})/n\pi_{0i}$ and expanding in a Taylor series, we see that the above equals

$$\frac{2n}{\lambda(\lambda+1)} \sum_{i=1}^{k} \pi_{0i} \left\{ (\lambda+1)\,V_i + \frac{\lambda(\lambda+1)}{2} V_i^2 + 0_p(V_i^3) \right\}$$

$$= 2n \left\{ \sum_{i=1}^{k} \pi_{0i}\, V_i^2/2 + o_p(1/n) \right\} ,$$

under the model H_0. An identical result holds for $\lambda = 0, -1$, also by a Taylor series expansion. Hence

$$2nI^\lambda(\mathbf{X}/n: \Pi_0) = 2nI^1(\mathbf{X}/n: \Pi_0) + o_p(1);\ \lambda \in \mathbb{R}. \tag{2.3}$$

Thus each power divergence statistic has asymptotically the same distribution as Pearson's X^2 which is well known to be asymptotically distributed as a χ^2_{k-1} random variable, under the model H_0.

We will now develop this result for the more general case of unspecified parameters in the null hypothesis, following BFH's (1975, Chapter 14) approach closely. More specifically, define a parameter (row) vector $\Theta = (\theta_1, \ldots, \theta_s) \in \mathbb{R}^s$, $s < k-1$, and a mapping

$$\mathbf{f}:\ \mathbb{R}^s \to \Delta_k = \{\mathbf{p} = (p_1, \ldots, p_k): p_i \geqslant 0, i = 1, \ldots, k; \sum_{i=1}^{k} p_i = 1\}$$

such that to each parameter vector Θ there corresponds a probability (row) vector $\Pi = (\pi_1, \ldots, \pi_k)$. Hence the following two null hypotheses are equivalent

$$H_0: \Theta \in Q_0 \text{ and } H_0: \Pi \in M_0, \tag{2.4}$$

where $M_0 = \mathbf{f}(Q_0)$. Later we will put regularity conditions on $\mathbf{f}$ and Q_0 to avoid degenerate cases such as when the null model space M_0 is a one point set $\{\Pi_0\}$. There, no estimation of Π under H_0 is necessary in order to calculate the divergence between Π_0 and the unconstrained maximum likelihood estimate of Π, namely $\mathbf{x}/n$. We can immediately base a test of H_0 on the statistic (2.1), where large values define the critical region. If however H_0 is a composite hypothesis then it is necessary to choose an estimate $\hat{\Pi} \in M_0$ to represent the class M_0 in the calculation of (2.1). Associated with $\hat{\Pi}$ will be the estimated parameter $\hat{\Theta} \in Q_0$ with $\mathbf{f}(\hat{\Theta}) = \hat{\Pi}$.

The fine details of the following development can be found in Read (1982, Section 2.1.1); here we present the main features.

Definition 2.1.

The minimum I^λ-discrepancy estimator of $\Theta \in Q_0$ is any $\hat{\Theta}_\lambda \in \bar{Q}_0$ (the closure of Q_0) for which

$$I^\lambda(\mathbf{X}/n: \mathbf{f}(\hat{\Theta}_\lambda)) = \inf_{\Theta \in Q_0} I^\lambda(\mathbf{X}/n; \mathbf{f}(\Theta)).$$

Note that the term discrepancy rather than distance is used here, to indicate that I^λ is not generally a true distance function. Possible non-uniqueness or unboundedness of the estimator occurs with probability zero as $n \to \infty$, under the following *regularity conditions* (Birch, 1964). We assume the null hypothesis (2.4) is correct so that there exists a $\Theta_* \in Q_0$ with $\Pi_* = \mathbf{f}(\Theta_*)$ where Π_* is the true value of Π. Assume

i. there is an s dimensional neighbourhood of Θ_* completely contained in Q_0,

ii. $f_i(\Theta_*) > 0$ for $i = 1, \ldots, k$,

iii. $\mathbf{f}$ is totally differentiable at Θ_*, so that the partial derivatives of f_i with respect to each θ_j exist at Θ_*,

iv. the Jacobian $\partial\mathbf{f}(\Theta_*)/\partial\Theta$ is of full rank s,

v. the inverse mapping $\mathbf{f}^{-1}$ is continuous at $\mathbf{f}(\Theta_*)$,

vi. the mapping $\mathbf{f}$ is continuous at every point $\Theta \in Q_0$ (see BFH, 1975, p. 510).

Definition 2.2

Any estimator $\hat{\Theta} \in Q_0$ satisfying the expansion

$$\hat{\Theta} = \Theta_* + (\mathbf{X}/n - \Pi_*) D_{\pi_*}^{-1/2} A(A'A)^{-1} + o_p(n^{-1/2}) \tag{2.5}$$

is called a best asymptotically normal (BAN) estimator of Θ, where D_{π^*} is the diagonal matrix with $\pi_{*1}, \ldots, \pi_{*k}$ along the diagonal, and $A = (\pi_{*i}^{-1/2}\ \partial f_i(\Theta_*)/\partial\theta_{*j})$.

By a straightforward generalization of the argument found in Birch (1964) we can prove (see Read, 1982, Appendix A):

Theorem 2.1

Under the above regularity conditions, any minimum I^λ-discrepancy estimator is BAN.

Birch (1964) showed that the maximum likelihood estimator ($\lambda = 0$) is BAN. Another special case, the minimum chi-squared estimator ($\lambda = 1$), was proved to be BAN by Holland (1967).

The following very useful theorem applies for any BAN estimator $\hat{\Theta}$ (BFH, 1975, pp. 517, 518).

Theorem 2.2

If the above regularity conditions hold for $\mathbf{f}$ in $\Pi = \mathbf{f}(\Theta)$, and if $\hat{\Theta}$ is any BAN estimator of Θ, then (under H_0 given by (2.4)) $\sqrt{n}\,[(\mathbf{X}/n, \hat{\Pi}) - (\Pi_*, \Pi_*)]$ converges in distribution to a multivariate normal with zero mean vector and variance matrix

$$\Sigma = \left\{ \begin{array}{cc} D_{\pi_*} - \Pi'_* \Pi_* & (D_{\pi_*} - \Pi'_* \Pi_*) L \\ L'(D_{\pi_*} - \Pi'_* \Pi_*) & L'(D_{\pi_*} - \Pi'_* \Pi_*) L \end{array} \right\}.$$

where $L = D_{\pi_*}^{-1/2} A (A'A)^{-1} A' D_{\pi_*}^{1/2}$.

Now it is simple to show in a similar manner to the way (2.3) was derived, that

$$2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi}) = 2nI^1(\mathbf{X}/n\colon \hat{\Pi}) + o_p(1), \tag{2.6}$$

where $\hat{\Pi} = \mathbf{f}(\hat{\Theta})$ and $\hat{\Theta}$ is a BAN estimator of Θ. But it is a direct consequence of Theorem 2.2, that $2nI^1$ converges in distribution to a χ^2_{k-s-1} random variable under H_0, and hence so does *every* $2nI^\lambda$. Consequently, the following important result can be stated.

Theorem 2.3

If the regularity conditions (i) to (vi) hold and if $\hat{\Theta}$ is any BAN estimator of $\Theta = (\theta_1, \ldots, \theta_s)$ and $\hat{\Pi} = \mathbf{f}(\hat{\Theta})$, then under H_0 given by (2.4), $2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi})$ converges in distribution to a χ^2_{k-s-1} random variable, as $n \to \infty$.

Corollary 2.1

Suppose the regularity conditions (i) to (vi) hold. For any minimum I^μ-discrepancy estimator $\hat{\Theta}_\mu$ (let $\hat{\Pi}_\mu = \mathbf{f}(\hat{\Theta}_\mu)$) consider the statistic $2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi}_\mu)$. Then under H_0, this statistic will converge in distribution to a χ^2_{k-s-1} random variable.

It would be sensible to use the same λ-scale for both testing and estimation, as is done in, say, the generalized likelihood ratio test procedure, or least squares estimation in the general linear model. However, in practice the calculations are frequently complicated (see e.g. BFH, pp. 348–349), whereas the m.l. procedures are well documented in the cases of interest and are generally available; e.g. BMDP4F (BMDP Statistical Software, Los Angeles, CA 90025) for m.l. estimation under the log-linear model.

2.3. *Limiting Normal Distributions*

The asymptotic results of the previous subsection were derived for the number of cells k *fixed*, and $n \to \infty$. If now k is allowed to grow (with n) what limiting distribution for (2.1) results? Intuitively it should be the normal because as df increases, χ^2_{df} treated as a sum, is within the ambit of the central limit theorem. However, a rigorous treatment must take into account k and n simultaneously increasing in (2.1). This type of passage to the limit is sensible for certain types of testing situations (see e.g. Fienberg, 1980, p. 174; Ivchenko and Medvedev, 1978) and for comparisons to related goodness-of-fit statistics (see Section 5.3).

We shall rely heavily on the following general theorem (Holst, 1972).

Theorem 2.4.

Suppose $n \to \infty$ and $k \to \infty$ in such a way that $n/k \to a$ ($0 < a < \infty$), and suppose $k\pi_i \leqslant d < \infty$; $i = 1, \ldots, k$, for all k. Define

$$S_k = \sum_{i=1}^{k} f_i(X_i). \tag{2.7}$$

The functions $f_i(\cdot)$ are assumed real valued functions such that

$$\sigma_n^2 = \sum_{i=1}^{k} \text{var}\,(f_i(Y_i)) - \frac{1}{n}\left\{\sum_{i=1}^{k} \text{cov}\,(Y_i, f_i(Y_i))\right\}^2$$

satisfies

$$0 < \liminf_{n \to \infty} \sigma_n^2/n \leqslant \limsup_{n \to \infty} \sigma_n^2/n < \infty,$$

where $Y_1, \ldots, Y_k$ are independent Poisson random variables with means $n\pi_1, \ldots, n\pi_k$ respectively. Moreover assume the f_i satisfy $|f_i(x)| \leqslant ce^{bx}, i = 1, \ldots, k$, where $c > 0$, $b > 0$. Suppose

$$\mu_n = \sum_{i=1}^{k} E(f_i(Y_i)).$$

Then

$$(S_k - \mu_n)/\sigma_n,$$

converges in distribution to the standard normal random variable.

Corollary 2.2

Suppose n and $k \to \infty$ such that $n/k \to a$ $(0 < a < \infty)$ and $\pi_{0i} = 1/k$ for $i = 1, \ldots, k$. Define

$$\mu_n = \begin{cases} \dfrac{2n}{\lambda(\lambda+1)} E\{(Y/a)^{\lambda+1} - 1\} & \lambda > -1, \neq 0 \\ 2n \quad E\{(Y/a)\ln(Y/a)\} & \lambda = 0, \end{cases}$$

$$\sigma_n^2 = \begin{cases} \left(\dfrac{2a}{\lambda(\lambda+1)}\right)^2 k[\text{var}\,\{(Y/a)^{\lambda+1}\} - a\,\text{cov}^2\,\{Y/a, (Y/a)^{\lambda+1}\}] & \lambda > -1, \neq 0 \\ (2a)^2\,k\,[\text{var}\,\{(Y/a)\ln(Y/a)\} - a\,\text{cov}^2\,\{Y/a, (Y/a)\ln(Y/a)\}] & \lambda = 0, \end{cases}$$

where Y is a Poisson random variable with mean a. Let $\mathbf{1}$ represent a row vector of 1's. Then the power divergence statistic suitably normalized, namely

$$\frac{2nI^{\lambda}(\mathbf{X}/n : \mathbf{1}/k) - \mu_n}{\sigma_n},$$

converges in distribution to the standard normal random variable, when $\lambda > -1$.

Proof

The proof is an application of Theorem 2.4 with

$$f_i(x) = \begin{cases} \dfrac{2}{\lambda(\lambda+1)}\,\dfrac{n}{k}\left\{\left(\dfrac{kx}{n}\right)^{\lambda+1} - 1\right\} & \lambda > -1, \neq 0 \\ 2x\ln(kx/n) & \lambda = 0, \end{cases}$$

Corollary 2.2 is stated for the symmetric null hypothesis, $\pi_{0i} = 1/k$ $(i = 1, \ldots, k)$. Notice that for the extended asymptotics in this subsection, the limiting result is *not* identical for all λ. For example Pearson's X^2 statistic $2nI^1(\mathbf{X}/n\colon 1/k)$ has $\mu_n \sim k$ and $\sigma_n^2 \sim 2k$, whereas the log likelihood ratio statistic $2nI^0(\mathbf{X}/n\colon 1/k)$ has

$$\mu_n \sim 2k \left\{ \sum_{j=1}^{\infty} \ln(j) \, . \, \frac{e^{-a}a^j}{(j-1)!} - \ln(a) \right\},$$

and σ_n^2 is likewise different. It is this sensitivity to different λ which will be exploited in the next section. There we will look at the power of tests based on the I^λ statistics under a sequence of contiguous alternatives.

The assumption of symmetry is important for the development of these efficiencies, and it is this scheme which has been studied most effectively in the literature (see e.g. Holst, 1972; Ivchenko and Medvedev, 1978). The case of non-symmetric probabilities requires more complicated conditions to derive the limiting distribution of the statistics. Furthermore under such non-symmetric schemes, Ivchenko and Medvedev (1978) in considering Pearson's X^2 and the log-likelihood ratio statistics, give notice that each scheme will need individual treatment.

3. EFFICIENCIES OF THE TESTS

3.1. *Pitman and Bahadur Relative Efficiency when the Number of Cells is fixed*

Any decision as to which of the power divergence statistics should be used to carry out a goodness-of-fit test depends on their performances according to various criteria. We discuss here the relative efficiencies of the tests in detecting certain alternative models.

For a fixed test size α, the Pitman asymptotic relative efficiency (a.r.e) for two tests is calculated below under the sequence of contiguous alternatives

$$H_{1,n}\colon \Pi = \mathbf{f}(\Theta_*) + \mathbf{c}/\sqrt{n}, \text{ some } \Theta_* \in Q_0 \tag{3.1}$$

Here the row vector $\mathbf{c} = (c_1, \ldots, c_k)$ satisfies $\Sigma_{i=1}^k c_i = 0$. In order to evaluate the Pitman a.r.e. between any two divergence statistics we require the asymptotic distribution of $2nI^\lambda$under (3.1).

Assume the regularity conditions (i) to (vi) of Section 2.2 to hold. Let $\hat{\Theta}$ be any BAN estimator of $\Theta \in Q_0$. Then under the contiguous alternative (3.1) it is easily shown that

$$\mathbf{X}/n = \mathbf{f}(\Theta_*) + 0_p(n^{-1/2}) \text{ and } \hat{\Pi} = \mathbf{f}(\Theta_*) + 0_p(n^{-1/2}).$$

But this is exactly the type of result that was needed to derive (2.6) from Theorem 2.2. Hence the members of the family $2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi})$, where $\hat{\Pi} = \mathbf{f}(\hat{\Theta})$, are asymptotically stochastically equivalent; i.e. for $\lambda \in \mathbb{R}$

$$2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi}) = 2nI^1(\mathbf{X}/n\colon \hat{\Pi}) + o_p(1). \tag{3.2}$$

Theorem 3.1

Assume that the regularity conditions of Section 2.2 hold. Let $\hat{\Theta}$ be any BAN estimator of Θ and let $\hat{\Pi} = \mathbf{f}(\hat{\Theta})$. Then under the contiguous alternatives (3.1), $2nI^\lambda(\mathbf{X}/n\colon \hat{\Pi})$ converges in distribution to a $\chi^2_{k-s-1}(\delta)$ random variable for each $\lambda \in \mathbb{R}$, where $\chi^2_{k-s-1}(\delta)$ is a non-central χ^2 random variable on $k-s-1$ degrees of freedom and non-centrality parameter $\delta = \mathbf{c}\, D^{-1}_{f(\Theta*)}\, \mathbf{c}'$.

Proof

Mitra (1958) has proved the result for $\lambda = 1$ and the general result then follows immediately from equation (3.2).

This result indicates that the power divergence family is not only equivalent under the null model in Section 2.1, but also under the contiguous alternative (3.1). The Pitman a.r.e. between any two family members I^{λ_1} and I^{λ_2} is given by the ratio of their non-centrality parameters (see e.g. Kendall and Stuart (1973, pp. 285, 286)) and is therefore equal to one for each λ_1 and λ_2.

Another concept of efficiency, involving different asymptotics from that of Pitman's, was introduced by Bahadur (1960). In this case we assume the null hypothesis (2.4) is simple. If $u = (y_1, y_2, \ldots)$ is a sequence of independent observations on Y taking values in the set $\{a_1, \ldots, a_k\}$ and $\mathbf{x} = (x_1, \ldots, x_k)$, then

$$L_\lambda(u) = \Pr\,[I^\lambda(\mathbf{X}/n : \Pi_0) \geqslant I^\lambda(\mathbf{x}/n : \Pi_0)]$$

is the level obtained by $2nI^\lambda$ where $x_i = \#(y_j = a_i)$, $i = 1, \ldots, k$. Bahadur (1967) points out that typically the attained level converges in distribution to a uniform random variable on $[0, 1]$ as $n \to \infty$ under $H_0 : \Pi = \Pi_0$, and converges to 0 with probability one as $n \to \infty$ when $\Pi \neq \Pi_0$. Furthermore in many cases (and this is one) the rate at which the level converges to 0 is exponential. Define

$$\Delta_k^+ = \{\mathbf{p} = (p_1, \ldots, p_k) : p_i > 0, i = 1, \ldots, k;\ \sum_{i=1}^{k} p_i = 1\},$$

which we use in the following important definition.

Definition 3.1

Let $c_\lambda(\Pi)$ be defined over $\Delta_k^+ - \{\Pi_0\}$ with $0 < c < \infty$ such that the random variable $L_\lambda(U)$ satisfies

$$n^{-1} \log L_\lambda \to -\tfrac{1}{2}\, c_\lambda(\Pi)$$

with probability one as $n \to \infty$. Then c_λ is called the exact Bahadur slope of the sequence $\{2nI^\lambda(\mathbf{X}/n : \Pi_0)\}_{n=1}^{\infty}$. Furthermore the ratio $c_{\lambda_1}(\Pi)/c_{\lambda_2}(\Pi)$ is defined to be the exact Bahadur relative efficiency of the test $2nI^{\lambda_1}$ to $2nI^{\lambda_2}$.

The existence of such a c_λ will be proved below. However, under the assumption that it exists, a justification for using the ratio $c_{\lambda_1}/c_{\lambda_2}$ as a measure of efficiency follows from Bahadur (1967). That is, if $n^{(\lambda_1)}, n^{(\lambda_2)}$ are the sample sizes required to make $2nI^{\lambda_1}$ and $2nI^{\lambda_2}$ significant when $\Pi \neq \Pi_0$, then $n^{(\lambda_2)}/n^{(\lambda_1)} \to c_{\lambda_1}(\Pi)/c_{\lambda_2}(\Pi)$ as the test size tends to 0.

The calculations involved in determining $c_\lambda(\Pi)$ are simplified by the following theorem which is a special case of the results of Bahadur (1971).

Lemma 3.1

If T_n is a test statistic for the simple null hypothesis $H_0 : \Pi = \Pi_0$ based on $(Y_1, \ldots, Y_n)$ with

(a) $n^{-1/2}\, T_n \to b(\Pi)$ with probability one as $n \to \infty$ for each $\Pi \neq \Pi_0$ and $-\infty < b < \infty$,

(b) $n^{-1} \log\{\Pr(T_n \geqslant n^{1/2} t \mid \Pi = \Pi_0)\} \to -f(t)$ as $n \to \infty$ for each t in a specified open interval in which f is continuous and which contains $\{b(\Pi) : \Pi \neq \Pi_0\}$,

then the exact Bahadur slope $c(\Pi) = 2f(b(\Pi))$.

Theorem 3.2

Define $T_n = \{2nI^\lambda(\mathbf{X}/n : \Pi_0)\}^{1/2}$ for λ fixed. Then

1. $n^{-1/2}\, T_n \to \{2I^\lambda(\Pi : \Pi_0)\}^{1/2}$ with probability one as $n \to \infty$,
2. $n^{-1} \log\{\Pr(T_n \geqslant n^{1/2}\, t \mid \Pi = \Pi_0)\} \to -\inf\limits_{\mathbf{v} \in A_{\lambda, t}} I^0(\mathbf{v} : \Pi_0)$ as $n \to \infty$ for each t in an open interval, where $A_{\lambda, t} = \{\mathbf{v} : \mathbf{v} \in \Delta_k^+$ and $[2nI^\lambda(\mathbf{v} : \Pi_0)]^{1/2} \geqslant t\}$.
3. The exact Bahadur slope of the test based on $2nI^\lambda(\mathbf{X}/n : \Pi_0)$ is thus given by

$$c_\lambda(\Pi) = \inf_{\mathbf{v} \in B_\lambda} 2I^0(\mathbf{v} : \Pi_0) \tag{3.3}$$

where $B_\lambda = \{\mathbf{v} : \mathbf{v} \in \Delta_k^+$ and $I^\lambda(\mathbf{v} : \Pi_0) \geqslant I^\lambda(\Pi : \Pi_0)\}$.

Proof

This is a straightforward generalization of Bahadur (1971, p. 31), who proved the result for $\lambda = 0, 1$. For more details see Read (1982, Section 2.2.3).

Example

In the special case $\lambda = 0$ (the log likelihood ratio statistic) we obtain immediately from (3.3) that $c_0(\Pi) = 2I^0(\Pi: \Pi_0)$, $\Pi \neq \Pi_0$. Furthermore since $\Pi \in B_\lambda$,

$$c_\lambda(\Pi) = \inf_{\mathbf{v} \in B_\lambda} 2I^0(\mathbf{v}: \Pi_0) \leqslant 2I^0(\Pi: \Pi_0) = c_0(\Pi),$$

which gives

$$c_\lambda(\Pi)/c_0(\Pi) \leqslant 1 \text{ for all } \Pi \neq \Pi_0.$$

Thus the likelihood ratio test ($\lambda = 0$) obtains maximal Bahadur efficiency amongst all tests based on the power divergence family (2.1). However other family members can be *equally* efficient if there does not exist a probability vector $\mathbf{v}$ satisfying both $I^0(\mathbf{v}: \Pi_0) < I^0(\Pi: \Pi_0)$ and $I^\lambda(\mathbf{v}: \Pi_0) \geqslant I^\lambda(\Pi: \Pi_0)$.

3.2. *Pitman Asymptotic Relative Efficiency when the Number of Cells is Large*

When the number of cells k increases with n, it is apparent from Section 2.3 that the power divergence statistics are no longer asymptotically equivalent under the symmetric null hypothesis. This result can be extended to the general contiguous alternative

$$H_{1,n}: \pi_i = 1/k + \int_{(i-1)/k}^{i/k} c(x)/n^{1/m} dx, \tag{3.4}$$

where c is a known continuous function on $[0, 1]$ and $\int_0^1 c(x)\, dx = 0$, and m is a constant to be determined.

Theorem 3.3

Let $\mu_{n,0}, \mu_{n,1}$ and $\sigma^2_{n,0}, \sigma^2_{n,1}$ denote the means and variances of the power divergence statistic with fixed λ, under the symmetric null and alternative (3.4) respectively. If $n, k \to \infty$ so that $n/k \to a$ $(0 < a < \infty)$, then

$$\Pr\{2nI^\lambda(\mathbf{X}/n: \Pi_0) > K_\alpha\, \sigma_{n,0} + \mu_{n,0} \mid H_0\} = \alpha + o(1)$$

$$\Pr\{2nI^\lambda(\mathbf{X}/n: \Pi_0) > K_\alpha\, \sigma_{n,0} + \mu_{n,0} \mid H_{1,n}\} = \Phi\{-K_\alpha\, \sigma_{n,0}/\sigma_{n,1} + (\mu_{n,1} - \mu_{n,0})/\sigma_{n,1}\} + o(1).$$

where $K_\alpha = \Phi^{-1}(1 - \alpha)$, $0 < \alpha < 1$, and Φ denotes the standard normal distribution function.

Proof. The result can be derived in a way similar to that of Corollary 2.2, using the general theorem of Holst (1972) given by Theorem 2.4.

From this result it follows that the asymptotic power of the test is monotonic in

$$e^{(m)}_{\lambda,a} = \lim_{\substack{n\to\infty \\ k\to\infty}} \frac{\mu_{n,1} - \mu_{n,0}}{\sigma_{n,1}}, \tag{3.5}$$

if it exists, for m defined in (3.4). Putting $m = 2$ results in the contiguous alternatives similar to those in (3.1). However, in this case, from the results of Holst (1972) it can be shown that $e^{(2)}_{\lambda,a} = 0$, and hence the alternatives (3.4) are too close to H_0 for any power discrimination. Putting $m = 4$ we can derive from the results of Ivchenko and Medvedev (1978) that

$$e^{(4)}_{\lambda,a} = \rho_\lambda \sqrt{\frac{a}{2}} \int_0^1 c(x)^2\, dx,$$

where $\rho_\lambda = \text{sgn}\,(\lambda).\text{corr}\,\{Y^{\lambda+1} - b_\lambda Y, Y^2 - (2a+1)Y\}$ and $b_\lambda = a^{-1}\,\text{cov}\,(Y^{\lambda+1}, Y)$ for Y a Poisson random variable with mean a. Since $-1 \leqslant \rho_\lambda \leqslant 1$, the power of the test will be maximal when $\rho_\lambda = 1$. In particular it is easily shown that $\rho_1 = 1$ and hence the Pearson X^2 test (for which $\lambda = 1$) obtains maximum Pitman a.r.e. amongst tests based on the family of power divergence statistics for alternatives (3.4) with $m = 4$.

In order to assess the magnitude of the loss in efficiency resulting from using other values of $\lambda \neq 1$ ($\lambda > -1$), the values of $e^{(4)}_{\lambda, a}/\int_0^1 c(x)^2 dx$ have been calculated for various λ and a in Table 1.

TABLE 1
Values of $e^{(4)}_{\lambda, a}/\int_0^1 [c(x)]^2 dx$ *for various* λ, a

						a				
λ	0	0.1	0.5	1.0	1.5	2.0	3.0	10	20	50
−2/3	0	.22	.47	.62	.72	.78	.89	2.07	3.08	4.95
−1/2	0	.22	.47	.63	.74	.83	.98	2.12	3.10	4.96
−1/3	0	.22	.48	.65	.77	.87	1.05	2.15	3.11	4.97
0	0	.22	.49	.67	.81	.93	1.15	2.19	3.13	4.98
1/3	0	.22	.49	.69	.84	.97	1.20	2.22	3.15	4.99
1/2	0	.22	.50	.70	.85	.99	1.21	2.23	3.16	5.00
2/3	0	.22	.50	.70	.86	.99	1.22	2.23	3.16	5.00
1	0	.22	.50	.71	.87	1.00	1.22	2.24	3.16	5.00
1½	0	.22	.50	.70	.86	.99	1.21	2.23	3.16	5.00
2	0	.22	.48	.68	.83	.96	1.19	2.21	3.14	4.98
2½	0	.22	.46	.65	.80	.93	1.15	2.17	3.11	4.96
3	0	.21	.43	.61	.76	.89	1.10	2.13	3.07	4.94
4	0	.18	.37	.53	.67	.79	1.00	2.01	2.97	4.86
5	0	.15	.30	.44	.57	.67	.87	1.87	2.85	4.76

For fixed a, $e^{(4)}_{\lambda, a}$ is maximal for $\lambda = 1$ as indicated above. Notice that $e^{(4)}_{\lambda, a}$ starts to decrease rapidly for $\lambda > 3$, and the largest difference between $e^{(4)}_{\lambda, a}$ and $e^{(4)}_{1, a}$ occurs for $a = \lim n/k$ "moderate". For large a (i.e. $n \gg k$) we approach the classical theory for fixed number of classes k (Section 3.1). In this case the table shows that ρ_λ approaches 1 for all λ and the values of $e^{(4)}_{\lambda, a}$ become indistinguishable. This finding is in accord with the efficiency results of Section 3.1 for contiguous alternatives. For a small (i.e. $n \ll k$) the powers of all the tests tend to zero.

We conclude that the maximum discrimination between family members occurs for moderate a. In this case $\lambda = 1$ (Pearson's X^2) results in optimal efficiency and large values of λ perform poorly in comparison. There is however clearly a wide band of λ values for which $e^{(4)}_{\lambda,a}$ stays near the optimal value $e^{(4)}_{1,a}$.

3.3. *Finite Sample Power Results*

The most important criterion for comparing tests, namely the power function, is for finite samples often mathematically intractible. However here it is accessible on the computer for the family of tests based on I^λ (given specific choices of sample size, class size, null hypotheses, and alternative hypothesis). In Table 2 the exact power is illustrated for tests of the symmetric null hypothesis, at the 0.05 level, against the alternative

$$H_1: \pi_i = \begin{cases} \{1-\delta(k-1)\}/k & i=1,\ldots,k-1 \\ (1+\delta)/k & i=k \end{cases} \tag{3.6}$$

where $-1 \leqslant \delta \leqslant k-1$ is fixed.

TABLE 2
Exact power functions for the randomized size 0.05 test of the symmetric hypothesis. Alternatives defined by model (3.6); $n = 20$, $k = 4$.

λ	δ=1.5	δ=.5	δ=-.9
-5.0	0.6316	0.1228	0.7434
-2.0	0.6500	0.1231	0.7434
-1.0	0.7960	0.1384	0.7342
-0.5	0.8009	0.1412	0.7263
-0.3	0.8525	0.1538	0.7108
0.0	0.8640	0.1567	0.7045
0.3	0.8640	0.1567	0.7045
0.5	0.8640	0.1567	0.7045
0.7	0.8640	0.1567	0.7045
1.0	0.8745	0.1629	0.5150
1.5	0.8855	0.1682	0.3844
2.0	0.8962	0.1725	0.3291
2.5	0.8982	0.1733	0.2780
5.0	0.9025	0.1743	0.2422

This alternative results from the kth probability being perturbed by δ/k, while the rest are adjusted so that they still sum to one. For these calculations a randomized test has been used to give the *exact* 0.05 level (since the attainable levels of the non-randomized test are discrete).

The results in Table 2 together with those of Read (1984a) indicate that for $\delta > 0$ (i.e. a "bump" alternative) the exact power increases with λ. Conversely for $\delta < 0$ (i.e. a "dip" alternative) the exact power decreases with λ. In the special cases of Pearson's X^2 and the log-likelihood ratio statistics (i.e. $\lambda = 1, 0$) these results coincide with those of Koehler and Larntz (1980) under Monte Carlo comparisons. Read (1984a) concludes further that when testing against such perturbation alternatives, it is always possible to improve upon the power of these well known tests by choosing other members of the family I^λ. However, as $|\lambda|$ increases there is a "plateau" effect evident in Table 2 from which it is clear that for $|\lambda|$ large there is little change in the power curve as λ varies.

4. COMPARISON OF FAMILY MEMBERS VIA MOMENTS AND DISTRIBUTION FUNCTIONS

4.1 *Moments under the Simple Null Hypothesis and Corrections for the Critical Region*

Under the simple null hypothesis

$$H_0 : \Pi = \Pi_0, \tag{4.1}$$

the results of Section 2.2 indicate that the power divergence statistics are asymptotically equivalent and follow a χ^2 distribution on $k-1$ degrees of freedom. One method of assessing the speed of convergence to this asymptotic result is to calculate the second order asymptotic expansions of the exact moments for the family $2nI^\lambda(\mathbf{X}/n : \Pi_0)$. The sizes of the correction terms give us some information about the approximation error incurred in using the asymptotic χ^2 distribution in place of the exact probability distribution.

Defining $W_i = (X_i - n\pi_{0i})/\sqrt{n}$, (2.1) can be expanded as a Taylor series to give

$$2nI^\lambda(\mathbf{X}/n : \Pi_0) = \sum_{i=1}^{k} \frac{W_i^2}{\pi_{0i}} + \frac{(\lambda-1)}{3\sqrt{n}} \sum_{i=1}^{k} \frac{W_i^3}{\pi_{0i}^2} + \frac{(\lambda-2)(\lambda-1)}{12n} \sum_{i=1}^{k} \frac{W_i^4}{\pi_{0i}^3} + 0_p(n^{-3/2}). \tag{4.2}$$

Subsequently we can obtain the first three moment expansions as

$$\begin{aligned} E\{2nI^\lambda(\mathbf{X}/n : \Pi_0)\} &= k-1 \\ &\quad + \frac{1}{n}\left\{\frac{\lambda-1}{3}(2-3k+S) + \frac{(\lambda-1)(\lambda-2)}{4}(1-2k+S)\right\}. \\ &\quad + 0(n^{-3/2}) \end{aligned}$$

$$\begin{aligned} E\{2nI^\lambda(\mathbf{X}/n : \Pi_0)\}^2 &= k^2-1 \\ &\quad + \frac{1}{n}[2-2k-k^2+S+\frac{2(\lambda-1)}{3}\{10-13k-6k^2+(k+8)S\} \\ &\quad + \frac{(\lambda-1)^2}{3}(4-6k-3k^2+5S) + \frac{(\lambda-1)(\lambda-2)}{2}\{3-5k-2k^2 \\ &\qquad + (k+3)S\}] \\ &\quad + 0(n^{-3/2}) \end{aligned}$$

$$\begin{aligned} E\{2nI^\lambda(\mathbf{X}/n : \Pi_0)\}^3 &= k^3+3k^2-k-3 \\ &\quad + \frac{1}{n}[26-24k-21k^2-3k^3+(19+3k)S \\ &\qquad + (\lambda-1)\{70-81k-64k^2-9k^3+(65+18k+k^2)S\} \\ &\qquad + (\lambda-1)^2\{20-26k-21k^2-3k^3+(25+5k)S\} \\ &\qquad + \frac{3(\lambda-1)(\lambda-2)}{4}\{15-22k-15k^2-2k^3+(15+8k+k^2)S\}] \\ &\quad + 0(n^{-3/2}), \end{aligned}$$

where $S = \Sigma_{i=1}^{k} (1/\pi_{0i})$. The expressions for the first two moments agree with those given by Johnson and Kotz (1969, p. 286) for Pearson's X^2 (substituting $\lambda = 1$) and Smith, Rae, Manderscheid and Silbergeld (1981) for the log likelihood ratio statistic (substituting $\lambda = 0$). The first order terms can be seen to be the first three moments of a χ^2 random variable on $k-1$ degrees of freedom. For given k and S, we have solved for the values of λ for which the second order correction factors are zero, and these are presented in Table 3 (note that always $S \geqslant k^2$). These values of λ (when they exist) are small and positive for all cases considered and furthermore converge to the same limits $\lambda = 1$, $2/3$ for k large. The value $\lambda = 1$ is not surprising from the Taylor expansion (4.2), however the value $\lambda = 2/3$ is not so intuitive and gives us a new competitor to the well known statistics. Note that for the *first* moment, the correction factor is *always* zero if $\lambda = 1$ (i.e. the Pearson X^2 statistic). The case $S = k^2$ is interesting since it arises when (4.1) is symmetric (i.e. $\pi_{0i} = 1/k, i = 1, \ldots, k$). In this case the solutions for λ are within ±0.1 of the limiting solutions $\lambda = 2/3$, 1 for $k > 20$.

Where the first two moments are not close to $k-1$ and $k^2 - 1$ respectively, a corrected distribution function can be defined as follows. Suppose a_λ and b_λ are given by

$$E\{2nI^\lambda(\mathbf{X}/n: \Pi_0)\} = k - 1 + a_\lambda/n + o(n^{-1})$$

$$\operatorname{var}\{2nI^\lambda(\mathbf{X}/n: \Pi_0)\} = 2(k-1) + b_\lambda/n + o(n^{-1}).$$

Then setting $\gamma_\lambda = (k-1)(1-\sqrt{\delta_\lambda}) + a_\lambda/n$ and $\delta_\lambda = 1 + b_\lambda/(n2(k-1))$, it follows that the corrected statistic

$$\{2nI^\lambda(\mathbf{X}/n: \Pi_0) - \gamma_\lambda\}/\sqrt{\delta_\lambda}$$

TABLE 3

Entries give the two roots λ_1, λ_2 where the second order correction factors of the first three moments $M_i = E[2nI^\lambda(\mathbf{X}/n: \Pi_0)]^i$, $i = 1, 2, 3$, are zero.

						$S = k^2$							
	k	2	3	4	5	10	20	30	40	50	100	200	...∞
M_1	λ_1	1.00	1.00	1.00	1.00	0.81	0.74	0.71	0.70	0.69	0.68	0.67	2/3
	λ_2	2.00	1.33	1.11	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1
M_2	λ_1	0.48	0.50	0.52	0.54	0.59	0.62	0.64	0.64	0.65	0.66	0.66	2/3
	λ_2	2.52	1.60	1.35	1.24	1.08	1.02	1.01	1.01	1.00	1.00	1.00	1
M_3	λ_1	0.32	0.35	0.38	0.40	0.48	0.55	0.58	0.60	0.61	0.64	0.65	2/3
	λ_2	2.68	1.56	1.31	1.20	1.06	1.02	1.01	1.00	1.00	1.00	1.00	1

TABLE 3 (continued) $s = k^5$

	k	2	3	4	5	10	20	30	40	50	100	200	... ∞
M_1	λ_1	0.71	0.68	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	0.67	2/3
	λ_2	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1
M_2	λ_1	0.42	0.50	–	–	–	–	0.69	0.68	0.68	0.67	0.67	2/3
	λ_2	0.63	0.58	–	–	–	–	0.82	0.86	0.89	0.94	0.97	1
M_3	λ_1	0.31	–	–	–	–	–	–	–	–	0.68	0.67	2/3
	λ_2	0.45	–	–	–	–	–	–	–	–	0.88	0.94	1

(Note: A dash indicates that no roots exist.)

will have a mean and variance which coincide with $k-1$ and $2(k-1)$ to $o(n^{-1})$. Therefore setting

$$F_A(c) = F_\chi((c-\gamma_\lambda)/\sqrt{\delta_\lambda}), \tag{4.3}$$

where $F_\chi(\cdot)$ is the distribution function of a χ^2 random variable on $k-1$ degrees of freedom, it follows that F_A should be a closer approximation to the distribution function of the power divergence family. A numerical assessment of this approximation is discussed in Section 4.4.

4.2. *Moments under the Simple Contiguous Alternative Hypothesis*

Under the sequence of contiguous alternatives

$$H_{1,n}: \Pi = \Pi_0 + \mathbf{c}/\sqrt{n}, \tag{4.4}$$

where $\Sigma_{i=1}^k c_i = 0$, the results of Section 3.1 indicate that the power divergence statistics are asymptotically equivalent but follow a non-central χ^2 distribution on $k-1$ degrees of freedom with non-centrality parameter $\delta = \mathbf{c}\, D_{\pi_0}^{-1}\, \mathbf{c}'$. Therefore by paralleling Section 4.1 we can assess the speed of convergence to this asymptotic limit through the second order correction factors to the moments of the family. In this case we obtain

$$E\{2nI^\lambda(\mathbf{X}/n: \Pi_0)\} = k-1+\delta + \frac{1}{\sqrt{n}}\left\{\sum_{i=1}^k \frac{c_i}{\pi_{0i}} + \frac{\lambda-1}{3}\left(\sum_{i=1}^k \frac{c_i^3}{\pi_{0i}^2} + 3\sum_{i=1}^k \frac{c_i}{\pi_{0i}}\right)\right\} + 0(n^{-1})$$

$$E\{2nI^\lambda(\mathbf{X}/n: \Pi_0)\}^2 = k^2-1+2(k+1)\delta+\delta^2 + \frac{1}{\sqrt{n}}[2(k+3+\delta)\sum_{i=1}^k \frac{c_i}{\pi_{0i}} + 4\sum_{i=1}^k \frac{c_i^2}{\pi_{0i}^2} + (\lambda-1)\{2(k+3+\delta)\sum_{i=1}^k \frac{c_i}{\pi_{0i}} + \frac{2}{3}(k+5+\delta)\sum_{i=1}^k \frac{c_i^3}{\pi_{0i}^2}\}] + 0(n^{-1}).$$

The sheer enormity of algebra has prevented us from calculating the third moment. In the case of the symmetric null hypothesis where $\Pi_0 = 1/k$, the $0(n^{-1/2})$ correction factors in the first and second moments are zero for $\lambda = 1$ and

$$\lambda = 1 - 6/\left(k \sum_{i=1}^{k} c_i^2 + k + 5\right)$$

respectively. Furthermore since $k \geqslant 1$ the second solution tends to one as k increases. This result hints at Pearson's X^2 statistic ($\lambda = 1$) having closest distribution to the approximate non-central χ^2 under the contiguous alternatives (4.4).

4.3. *Second Order Approximate Distribution Functions*

Under the simple null hypothesis, Section 4.1 gives us an indication as to how the choice of λ will affect the convergence of the exact moments to the asymptotic χ^2 moments. Read (1984b) proves a theorem which extracts the λ dependent second order component from the $o(1)$ term in the null distribution

$$\Pr(2nI^\lambda(\mathbf{X}/n: \Pi_0) < c) = \Pr(\chi^2_{k-1} < c) + o(1).$$

He uses this to motivate the following approximation. Under the simple null hypothesis (4.1),

$$\Pr(2nI^\lambda(\mathbf{X}/n: \Pi_0) < c) \cong J_1^\lambda + J_2^\lambda,$$

where

$$\begin{aligned} J_1^\lambda &= \Pr(\chi^2_{k-1} < c) \\ &+ \frac{1}{24n}\,[\Pr(\chi^2_{k-1} < c)\{2(1-s)\} \\ &\quad + \Pr(\chi^2_{k+1} < c)\{3(3s - k^2 - 2k) + (\lambda - 1)\,6(S - k^2) \\ &\quad + (\lambda - 1)^2(5S - 3k^2 - 6k + 4) - (\lambda - 2)(\lambda - 1)\,3(S - 2k + 1)\} \\ &\quad + \Pr(\chi^2_{k+3} < c)\{-6(2S - k^2 - 2k + 1) - (\lambda - 1)\,4(4S - 3k^2 - 3k + 2) \\ &\quad - (\lambda - 1)^2\,2(5S - 3k^2 - 6k + 4) + (\lambda - 1)(\lambda - 2)\,3(S - 2k + 1)\} \\ &\quad + \Pr(\chi^2_{k+5} < c)\{\lambda^2\,(5S - 3k^2 - 6k + 4)\}], \end{aligned}$$

where $S = \Sigma_{i=1}^k (1/\pi_{0i})$. Define the lattice

$$L = \{\mathbf{w} = (w_1, \ldots, w_{k-1}) \colon w_i = (m_i - n\,\pi_{0i})/\sqrt{n}, \text{ where } m_i \geqslant 0 \text{ are integers, } i = 1, \ldots, k-1 \text{ with } \sum_{j=1}^{k-1} m_j \leqslant n\}$$

and the set

$$B_\lambda(c) = \{\mathbf{w} = (w_1, \ldots, w_{k-1}) \colon w_i = (x_i - n\pi_{0i})/\sqrt{n} \text{ for } \mathbf{x}/n = (x_1/n, \ldots, x_k/n) \in \Delta_k \text{ satisfying } 2nI^\lambda(\mathbf{x}/n: \Pi_0) < c\}.$$

Then

$$J_2^\lambda = \{N_\lambda(c) - n^{(k-1)/2}\,V_\lambda(c)\}\,e^{-c/2}/\{(2\pi n)^{(k-1)} \prod_{i=1}^{k} \pi_{0i}\}^{1/2},$$

where

$$N_\lambda(c) = \{\#\mathrm{w} \in L \text{ such that } \mathrm{w} \in B_\lambda(c)\}$$

$$= \text{the number of lattice points in } B_\lambda(c),$$

$$V_\lambda(c) = \text{the volume of } B_\lambda(c)$$

$$= \frac{(\pi c)^{(k-1)/2}}{\Gamma((k+1)/2)} \left(\prod_{i=1}^{k} \pi_{0i} \right)^{1/2} [1 + \frac{c}{24(k+1)\,n} \{(\lambda-1)^2(5S-3k^2-6k+4)$$

$$-3(\lambda-1)(\lambda-2)(S-2k+1)\}] + 0(n^{-3/2}).$$

The J_1^λ term would be the Edgeworth expansion term if $2nI^\lambda$ had a continuous distribution function. However since it has a lattice distribution the Cramér condition C, which ensures the validity of the Edgeworth approximation, is not satisfied. Yarnold (1972) evaluated these extra terms for the second order approximation in the case of Pearson's X^2 statistic, and his results can be obtained as a special case of the above by setting $\lambda = 1$.

The usefulness of this closer approximation to the exact distribution of members of the power divergence family is examined briefly in Section 4.4 and in more detail by Read (1984a). There it is shown to be very close in small samples and provides a substantial improvement over the (first order) χ^2 approximation.

4.4. *Finite Sample Comparison of the Exact Test Size with Four Approximations*

In order to use the power divergence statistics (2.1) to test the null hypothesis (4.1), it is necessary to calculate the appropriate critical region for a size α test. Sections 2.2, 2.3, 4.1 and 4.3 give us four different approximations for this calculation. The closeness of these for finite sample size has been discussed by Read (1984a). Due to the fact that most small sample studies have assumed the symmetric hypothesis, and also that the normal approximation in Section 2.3 has only been proved in this case, Read (1984a) has specifically enumerated the exact distribution of the statistics $2nI^\lambda(\mathbf{X}/n\colon 1/k)$ and compared the results with those obtained from the four approximations. Fig. 1 gives us an example of the errors incurred by the various approxi-

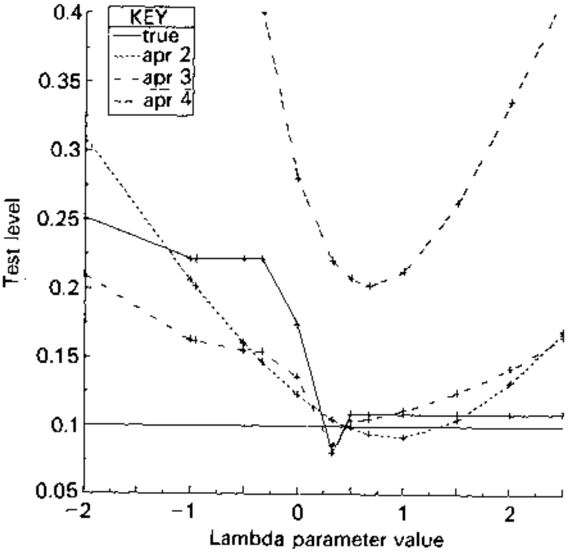

Fig. 1(a). True and approximate significance levels – symmetric hypothesis at nominal chi square level 0.100; $n = 10$, $k = 4$.

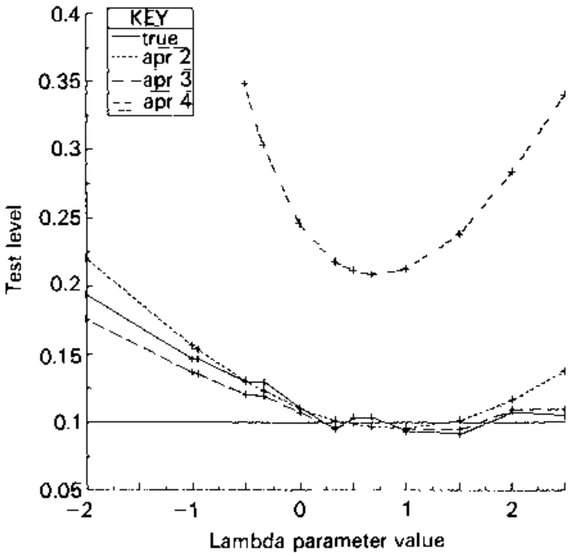

Fig. 1(b). True and approximate significance levels – symmetric hypothesis at nominal chi square level 0.100; $n = 20$, $k = 4$.

mations for $n = 10, 20$, $k = 4$ when the nominal level using the χ^2 approximation on $k-1$ degrees of freedom, is set at 0.1 for all λ. (The qualitative pictures at the nominal 0.05 and 0.01 levels look very much the same.) Here the χ^2 approximation is the straight line through $y = 0.1$, apr 2 is the moment corrected χ^2 (see Section 4.1), apr 3 is the second order expanded χ^2 (see Section 4.3), apr 4 is the normal approximation (see Section 2.3) and "true" is the exact (enumerated) significance level. Read (1984a) concludes that the usual χ^2 approximation is reasonable at levels between 0.1 and 0.01, for $\lambda \in (\frac{1}{3}, 1\frac{1}{2})$ provide $k \leqslant 6$, $n \geqslant 10$. However, as $|\lambda|$ increases, it becomes increasingly conservative, and this conservatism is magnified as k increases for fixed n. For other values of $\lambda \in [-5, 5]$ both the corrected χ^2 approximation of Section 4.1 and the second order approximation of Section 4.3 perform well–however the second order approximation is much more complicated to calculate. The normal approximation of Section 2.3 does not perform well for $n = 10, 20$ when compared to the corrected χ^2 approximation. However when n and k become large simultaneously it should be used in preference to any of the χ^2 approximations for $\lambda > -1$; see Read (1984a).

5. RELATED AREAS OF INTEREST

5.1. *Generalized Directed Divergences*

If we think of $\mathbf{x}/n$ and Π_0 in (2.1) as two discrete probability distributions in Δ_k then the following definition is a natural one.

Definition 5.1

For $\mathbf{p}, \mathbf{q} \in \Delta_k$, $\lambda \in \mathbb{R}$, define

$$I_k^\lambda (\mathbf{p} : \mathbf{q}) = \frac{1}{\lambda(\lambda+1)} \sum_{i=1}^{k} p_i \left\{ \left(\frac{p_i}{q_i}\right)^\lambda - 1 \right\}, \tag{5.1}$$

to be the directed divergence of order λ, where the values at $\lambda = 0, -1$ are defined by continuity.

Notice that $I_k^0(\mathbf{p}:\mathbf{q}) = \Sigma_{i=1}^k p_i \ln(p_i/q_i)$, is a constant multiple of Kullback's directed divergence (Kullback, 1959).

Rényi (1961) defined the directed divergence

$$\hat{I}_k^\alpha(\mathbf{p}:\mathbf{q}) = \begin{cases} (\alpha-1)^{-1} \log_2 \left(\sum_{i=1}^k p_i^\alpha q_i^{1-\alpha} \right), \ \alpha \neq 1 \\ \sum_{i=1}^k p_i \log_2 (p_i/q_i), \ \alpha = 1, \end{cases} \tag{5.2}$$

which has come to be known as Rényi's directed divergence (or information) of order α. It is additive in the sense that $\hat{I}_{kl}^\alpha(\mathbf{p}\times\mathbf{r}:\mathbf{q}\times\mathbf{s}) = \hat{I}_k^\alpha(\mathbf{p}:\mathbf{q}) + \hat{I}_l^\alpha(\mathbf{r}:\mathbf{s})$, where $\mathbf{p},\mathbf{q} \in \Delta_k$, $\mathbf{r},\mathbf{s} \in \Delta_l$ and $\mathbf{p}\times\mathbf{r} = (p_1 r_1, p_1 r_2, \ldots, p_1 r_l, \ldots, p_k r_l) \in \Delta_{kl}$ (see Mathai and Rathie, 1975, p. 48).

There are a number of properties one might *a priori* demand of a general directed divergence (or information measure) G_k. Surprisingly just a few are needed to uniquely characterize I_k^λ given by (5.1). For example, suppose (Rathie and Kannappan, 1972)

$$G_k(\mathbf{p}:\mathbf{q}) = G_{k-1}(p_1+p_2, p_3, \ldots, p_k : q_1+q_2, q_3, \ldots, q_k)$$
$$+ (p_1+p_2)^{\lambda+1}(q_1+q_2)^{-\lambda} G_2\left(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2} : \frac{q_1}{q_1+q_2}, \frac{q_2}{q_1+q_2}\right);$$

suppose

$$G_3(p_1, p_2, p_3 : q_1, q_2, q_3) = G_3(p_{\sigma(1)}, p_{\sigma(2)}, p_{\sigma(3)} : q_{\sigma(1)}, q_{\sigma(2)}, q_{\sigma(3)});$$

and suppose

$$G_2(1, 0 : \tfrac{1}{2}, \tfrac{1}{2}) = \frac{2^\lambda - 1}{\lambda(\lambda+1)}; \ \lambda \neq 0, -1.$$

Then $G_k(\mathbf{p}:\mathbf{q})$ *must* equal $I_k^\lambda(\mathbf{p}:\mathbf{q})$. The first postulate expresses (in a very specific way) a grouping property which says that information can only decrease if any classes are grouped; the second is a rather mild symmetry condition that says information is unchanged if classes are taken in different order; finally a normalizing equation fixes the multiplying constant. This is not as trivial a condition as it seems, since if chosen badly one might characterize a divergence which can be negative (see Read, 1982, Section 4.3.2). Other characterizations (of I_k^λ and of $\hat{I}_k^\alpha$) are possible; see Sharma and Taneja (1975), Mathai and Rathie (1975, p. 48), Rathie (1973). Also, entropy versions of (5.1) have been discussed by Rényi (1961) and Mathai and Rathie (1975, p. 12).

Once characterized, we can show the directed divergence I_k^λ will also satisfy properties of non-negativity ($I_k^\lambda(\mathbf{p}:\mathbf{q}) \geqslant 0$ with equality iff $p_i = q_i$, $i = 1, \ldots, k$), symmetry (for an arbitrary number of classes), continuity (in each of its variables $(p_1, \ldots, p_k)$, $(q_1, \ldots, q_k)$), zero indifference (the addition of a cell with zero probability does not change the value of the directed divergence), and log additivity

$$(\ln\{(1+\lambda(\lambda+1) I_{kl}^\lambda(\mathbf{p}\times\mathbf{r}:\mathbf{q}\times\mathbf{s})\} = \ln\{(1+\lambda(\lambda+1) I_k^\lambda(\mathbf{p}:\mathbf{q})\} + \ln\{(1+\lambda(\lambda+1) I_l^\lambda(\mathbf{r}:\mathbf{s})\}).$$

Although I_k^λ satisfies non-negativity, it only satisfies the triangle inequality when $-1 < \lambda < 0$, and then only becomes a true metric on Δ_k when $\lambda = -\frac{1}{2}$ (Hellinger distance or Matusita distance).

5.2. *Minimum Distance Estimators*

Minimum "distance" estimation provides a simple alternative method to that of maximum likelihood for estimating the unknown parameters in a model. In the case of a multinomial probability model $(\pi_1(\Theta), \ldots, \pi_k(\Theta))$ where the unknown vector of parameters

$$\Theta = (\theta_1, \ldots, \theta_s) \subset Q \subseteq \mathbb{R}^s$$

is to be estimated, the family of minimum I^λ-discrepancy estimators $\{\hat{\Theta}_\lambda\}$ is defined in Definition 2.1. The maximum likelihood estimator occurs when $\lambda = 0$, the minimum chi-squared estimator (Neyman, 1949) when $\lambda = 1$, the minimum modified chi-squared estimator (Neyman, 1949) when $\lambda = -2$, etc. Theorem 2.1 gives the result that *any* minimum I^λ-discrepancy estimator is BAN, which says that all such estimators are first order efficient.

To compare first order efficient estimators for $\theta \in \mathbb{R}$, Rao (1963) introduced the concept of second order efficiency. In the case of the multinomial distribution, this becomes equivalent to comparing the variance of the estimators, after first correcting for bias to $0(n^{-1})$. Read (1984c) has obtained the bias term $b_\lambda(\theta)/n = E(\hat{\theta}_\lambda - \theta)$ and thence var $(\tilde{\theta}_\lambda)$, where $\bar{\theta}_\lambda = \hat{\theta}_\lambda - b_\lambda(\ddot{\theta}_\lambda)/n$. This led him to the calculation of Rao's second order efficiency, $E_\lambda = T_1 + \lambda^2 T_2$, where T_1 and T_2 are positive valued functions only of θ. Clearly then, the m.l.e. $(\lambda = 0)$ is optimal amongst all (bias corrected) minimum I^λ-discrepancy estimators.

5.3. *Alternative Types of Goodness-of-fit Statistics*

In order to test goodness-of-fit using the power divergence statistics (2.1), the data must be discrete or if continuous, grouped in some way. More specifically, tests of $H_0 : F = F_0(.\,, \Theta)$ based on (2.1) are formed by defining boundaries $-\infty \leqslant b_1 < b_2 < \ldots < b_{k+1} \leqslant \infty$, such that $F_0(b_1, \Theta_*) = 0$ and $F_0(b_{k+1}; \Theta_*) = 1$ where Θ_* is the true value of Θ. Put

$$X_i = \#Y's \in (b_i, b_{i+1}); i = 1, \ldots, k,$$

and $\pi_{0i}(\Theta) = F_0(b_{i+1}; \Theta) - F_0(b_i; \Theta); i = 1, \ldots, k$. Then the statistic

$$2nI^\lambda(\mathbf{X}/n : \Pi_0(\hat{\Theta})) = \frac{2}{\lambda(\lambda+1)} \sum_{i=1}^{k} X_i \left\{ \left(\frac{X_i}{n\pi_{0i}(\hat{\Theta})} \right)^\lambda - 1 \right\};\tag{5.3}$$

(where $\hat{\Theta}$ is some estimate of Θ) is well defined. This subsection will discuss alternative approaches for a continuous sample $Y_1, \ldots, Y_n$ from distribution function F, in relation to the statistics I^λ. We have identified three approaches.

First, consider tests of H_0 based on sample quantiles $\{Y_{(n_i)}\}$ where $n_i = [(n+1)\gamma_i]$; $i = 1, \ldots, k+1$ and $0 = \gamma_1 < \ldots < \gamma_{k+1} = 1$. Consider the test statistic

$$X_q^2 = n \sum_{i=1}^{k} \{F_0(Y_{(n_{i+1})}; \bar{\Theta}) - F_0(Y_{(n_i)}; \vec{\Theta}) - \pi_{0i}\}^2 / \pi_{0i},\tag{5.4}$$

where $\pi_{0i} = \gamma_{i+1} - \gamma_i$, and $\bar{\Theta}$ is an estimate of Θ. This is essentially Pearson's X^2 (defined in Section 1), with $X_i = \#Y's \in (b_i, b_{i+1})$ where b_i solves $\gamma_i = F_0(b_i; \Theta_*)$. Furthermore, under certain standard regularity conditions on Θ and F, (5.4) is asymptotically indistinguishable in distribution from Pearson's X^2 (i.e. (5.3) with $\lambda = 1$) under both the null and alternatives "close" to the null; see Bofinger (1973), Miyamoto (1976) and Durbin (1978). Thus there is no real advantage in taking this approach based on sample quantiles.

Secondly, consider tests of H_0 based on sample spacings. Assume that Θ is completely-specified within F_0, although this restriction does not affect the general thrust of the development below. By applying the probability integral transformation $U_i = F_0(Y_i), i = 1, \ldots, n$, to the data H_0 is equivalent to testing whether $0 = U_{(0)} \leqslant U_{(1)} \leqslant U_{(2)} \ldots \leqslant U_{(n)} \leqslant U_{(n+1)} = 1$ are the order statistics from a uniform distribution. Then $V_i = U_{(i)} - U_{(i-1)}; i = 1, \ldots, n+1$ defines the set of (first order) spacings. Notice that $E(V_i) = 1/(n+1); i = 1, \ldots, n+1$, and consider the statistics

$$2nI^{\lambda}(\mathbf{V}:E(\mathbf{V}))=\frac{2n}{\lambda(\lambda+1)}\sum_{i=1}^{n+1}V_i\{((n+1)V_i)^{\lambda}-1\},\tag{5.5}$$

which (modulo constants) have been considered in various forms by Greenwood (1946) for $\lambda = 1$. Darling (1953) for $\lambda \to -1$, Kale and Godambe (1967) for $\lambda \to 0, \lambda \to -1$, and Kirmani and Alam (1974) for $\lambda > -1$. There is a close analogy between the asymptotic theory of spacings statistics of the form (5.5) and the theory of large multinomials as discussed in Section 2.3. Roughly speaking, tests based on sample quantiles (already seen to be essentially equivalent to tests based on counts) depend on a *subset* of spacings. The statistics (5.5) have a "k" ($= n+1$) which grows with n, and symmetric "π_{0i}" $= 1/(n+1), i = 1, \ldots, n+1$. From the results in Section 2, where the form of the asymptotic distribution of the goodness-of-fit statistic I^{λ} changed from chi-squared to normal as k went from being fixed to increasing with n, it should come as no surprise that although (5.4) is asymptotically chi-squared, (5.5) is asymptotically normal (Darling, 1953). For further discussions on spacings and higher order spacings see Pyke (1965) and Cressie (1979), respectively.

Thirdly, consider a continuous analogue to the discrete statistic (5.3). Suppose now that the density $f_0(y) = dF_0(y)/dy$ exists, and define the empirical density function

$$f_n(y)=\{F_n(y)-F_n(y-h)\}/h; -\infty<y<\infty,$$

which is a simple kernel estimator. Define

$$2nH^{\lambda}(f_n:f_0)=\frac{2nh}{\lambda(\lambda+1)}\int_{-\infty}^{\infty}f_n(y)\left\{\left(\frac{f_n(y)}{f_0(y)}\right)^{\lambda}-1\right\}dy.\tag{5.6}$$

This can be thought of as a continuous version of I^{λ} which for h small, looks very similar to (5.3) with n and k large together (see Read, 1982, Section 4.2.4). In view of this it again should come as no surprise that Bickel and Rosenblatt (1973) have shown in the case $\lambda = 1$ that (5.6) is asymptotically normally distributed, provided certain conditions hold. Beran (1977) considered the $\lambda = -\frac{1}{2}$ version of (5.6) for some suitable empirical density function f_n, and used the name Hellinger distance. Results for other values of λ in (5.6) have yet to be considered, but we conjecture an asymptotic normal limit. Note that (5.6) compares *increments* in distribution functions, in contrast with other well known statistics such as the Kolmogorov-Smirnov and the Cramer-von-Mises, which compare the distribution functions directly.

6. WHICH TEST STATISTIC?

6.1. *An Example*

To illustrate how the value of the power divergence statistic $2nI^{\lambda}(\mathbf{X}/n:\hat{\Pi})$ varies with λ, we consider an example due to Haberman (1978, Section 1.1) on the relationship between time passage and memory recall. As part of a larger study into the relationship between life stresses and illnesses, respondents were asked to note if any stressful events out of a list of 41 had occurred within the last 18 months. If so, then the number of months prior to interview was recorded. The totals for each month are shown in Table 4. It is clear that the respondents' recall decreases as the number of months increases from 1 to 18, and Haberman proposes a log-linear time trend model to explain this phenomenon, i.e.

$$H_T: \log \pi_i = \alpha + \beta i; i = 1, \ldots, 18,$$

where α and β are to be estimated. The maximum likelihood estimate $\hat{\Pi}$ of Π is calculated iteratively and is also given in Table 4. The values of various power divergence statistics are presented in Table 5 for the model H_T. Using the approximate 5 per cent level obtained from the χ^2_{16} tables, both the Pearson X^2 ($\lambda = 1$) and the log likelihood ratio ($\lambda = 0$) statistics indicate satisfactory agreement between this model and the data. In fact if $0 \leqslant \lambda \leqslant 3$ we would accept H_T at

TABLE 4

Observed and expected distribution (categorized by months prior to interview) of stressful events reported by subjects together with values of $g(x_i, E_i)$ defined in (6.1). Subjects were limited to those reporting one stressful event between 1 and 18 months prior to interview. Expected values are based on the log-linear time trend model H_T (Haberman, 1978, pp. 2-15).

Months before interview	Observed number of subjects x_i	Expected number of subjects E_i	$g(x_i,E_i)$
1	15	15.171	-1.01
2	11	13.952	-1.27
3	14	12.831	1.09
4	17	11.800	1.44
5	5	10.852	-2.17
6	11	9.9796	1.10
7	10	9.1777	1.09
8	4	8.4402	-2.11
9	8	7.7620	1.03
10	10	7.1383	1.40
11	7	6.5647	1.07
12	9	6.0371	1.49
13	11	5.5520	1.98
14	3	5.1059	-1.70
15	6	4.6956	1.28
16	1	4.3183	-4.32
17	1	3.9713	-3.97
18	4	3.6522	1.10
Total	147.00	147.00	

the nominal χ^2 5 per cent level. For λ outside this range we would reject H_T, however from the discussion in Section 4.4 and also Read (1984a), for $|\lambda|$ increasing, using the nominal χ^2 rejection region leads to increasingly liberal test levels (for the symmetric null hypothesis). Combining this result with the calculations of moment corrections in Section 4.1 for the general simple hypothesis, it appears reasonable to assume that this will also be the case in this example.

TABLE 5
Values of the power divergence statistics $2nI^\lambda(\mathbf{X}/n:\hat{\Pi})$ calculated from Table 4 for $\lambda \in [-10, 10]$. $\Pr\{\chi^2_{16} \geqslant 26.30\} = 0.05$.

λ	$2nI^\lambda(X/n:\hat{\Pi})$
−10.0	72.2×10^3
−5.0	28.9×10
−3.0	65.6
−2.0	40.6
−1.5	34.0
−1.0	29.5
−0.5	26.5
0.0	24.6
0.5	23.4
0.67	23.1
1.0	22.7
1.5	22.6
2.0	22.9
3.0	24.8
5.0	35.5
10.0	21.4×10

6.2. *Sensitivity to Large Ratios of x_i/E_i or E_i/x_i*

It is clear from the form of the power divergence statistic in (2.1) that a large ratio x_i/E_i will result in an increasingly inflated value of the statistic as λ increases, $\lambda > 0$. Similarly a large ratio E_i/x_i will result in an increasingly inflated value of the statistic as $|\lambda|$ increases, $\lambda < 0$. This point is discussed in more depth by Read (1984a) with reference to contributions to the statistic for a single cell deviation x_j/E_j, fixed (see also Larntz, 1978). Therefore if we wish to guard against the effects on the statistic of single large ratios, we should choose $|\lambda|$ small.

In Table 4 the values of the function

$$g(x_i, E_i) = \begin{cases} x_i/E_i & \text{if } x_i \geqslant E_i \\ -E_i/x_i & \text{if } x_i < E_i \end{cases} \tag{6.1}$$

are tabulated for the example in Section 6.1. From this table we can see that cells 16 and 17 obtain the largest ratios E_i/x_i and will dominate the statistic for $|\lambda|$ large, $\lambda < 0$. On the other hand, cells 12 and 13 obtain the largest ratios x_i/E_i and will dominate the statistic for λ large.

Furthermore since E_{16}/x_{16} and E_{17}/x_{17} are substantially larger than x_{12}/E_{12} and x_{13}/E_{13} it follows that the rate of increase of the statistic $2nI^{\lambda}(\mathbf{x}/n: \hat{\Pi}_0)$ with $|\lambda|$ will be faster for $\lambda < 0$ than for $\lambda > 0$. This can be seen to occur in Table 5.

Any decision as to which member of the family we should use to finally test H_0 must depend on the type of departure we wish to detect. In this example there appears to be no reason to expect a deviation from H_0 to be the result of a large ratio of x_i/E_i as against a large ratio of E_i/x_i. Therefore we would do best to choose a statistic with $|\lambda|$ small which is less sensitive to the direction of such deviations.

6.3. *Recommendations*

Under regularity conditions which are generally regarded as standard (Section 2.2) it was shown that the family of power divergence statistics are asymptotically equivalent under both the null hypothesis and under certain sequences of contiguous alternatives given by (3.1). Furthermore the common asymptotic distribution was shown to be a central and non-central χ^2 distribution on $k - s - 1$ degrees of freedom respectively, where k is the number of classes and s is the number of independent parameters that need to be BAN estimated under H_0 (see Theorems 2.3 and 3.1).

The rate of convergence to this asymptotic result was assessed through second order approximations and small sample studies. Second order approximations for the first three moments under the simple null hypothesis indicate that the moments of $2nI^{\lambda}(\mathbf{X}/n: \Pi_0)$ converge most rapidly to the asymptotic χ^2 moments for $\lambda \in [0.3, 2.7]$ (see Table 3). Finite sample studies carried out under the symmetric null hypothesis indicate that (for $k \leqslant 6$, $n \geqslant 10$) the χ^2 approximation is only appropriate for choosing the level of the test based on $2nI^{\lambda}(\mathbf{X}/n: \mathbf{1}/k)$, when $\lambda \in [1/3, 1\frac{1}{2}]$. For λ outside this range the χ^2 approximation tends to underestimate the true test size and two alternative approximations are proposed. One is based on correcting the mean and variance of the asymptotic distribution function to second order (Section 4.1) while the second is obtained by approximating the second order term in the asymptotic distribution function directly (Section 4.3). Both approximations perform well for $\lambda \in [-5, 5]$ however the former approximation is recommended since it is much easier to calculate in practice.

Under different asymptotics to those discussed above (namely the number of cells k increases at a rate proportional to the sample size n), the family of power divergence statistics was shown to follow an asymptotic normal distribution under the symmetric null hypothesis and a sequence of contiguous alternatives given by (3.4). In these cases the asymptotic mean and variance of $2nI^{\lambda}(\mathbf{X}/n: \mathbf{1}/k)$ is λ dependent and so the statistics are no longer asymptotically equivalent. Thus the behaviour of k as $n \to \infty$ is an instrument which allows us to assess the local power of the tests based on I^{λ}, $\lambda \in \mathbb{R}$. The focus is sharp when k grows with n, giving $\lambda = 1$ to be the best test, but for fixed k the tests are indistinguishable. In an attempt to sharpen this fuzziness we used Bahadur efficiency for fixed k, which gives $\lambda = 0$ to be the best test. Other criteria such as matching moments gives $\lambda = 2/3$ and $\lambda = 1$, matching actual to nominal significance levels gives $\lambda \in [1/3, 1\frac{1}{2}]$, and power in finite samples gives $\lambda \in [1/3, 2/3]$ as a compromise for alternatives which might be peaked *or* dipped (Read, 1984a).

Clearly there are many λ recommendations, some of which conflict. Based on their respective sensitivities we can however put together a final recommendation. Multinomial goodness-of-fit testing using statistic I^{λ} (see (2.1)) is best performed:

(i) for any $\lambda \in [0, 1\frac{1}{2}]$, when no knowledge of the type of alternative is available;

(ii) for $\lambda = 0$ (i.e. G^2) if the alternative is thought to be dipped; but the approximate percentage point should be determined by matching moments (see Section 4.1);

(iii) for $\lambda = 1$ (i.e. X^2), if the alternative is thought to be peaked, where the approximate percentage point can be found from the chi-squared tables.

Notice that the test based on G^2 ($\lambda = 0$) is on the very edge of the recommended interval, and that T^2 ($\lambda = -\frac{1}{2}$) is not considered at all, whereas X^2 ($\lambda = 1$) is safely surrounded by other possible λ values. In practice one will be in situation (i) most of the time. We recommend that provided $\min\{E_i\} \geqslant 1$ (Larntz, 1978) and $n \geqslant 10$, then the test based on $\lambda = 2/3$:

$$2nI^{2/3} = \frac{9}{5} \sum_{i=1}^{k} X_i \{(X_i/E_i)^{2/3} - 1\},$$

will be an excellent compromise for testing whether the observed multinomial variables $\{X_i\}$ are sufficiently close to their null expected values $\{E_i\}$, where the approximate percentage point can be found from the chi-squared tables. Applying $I^{2/3}$ to the data of Table 4 yields $2nI^{2/3}$ $(\mathbf{x}/n\text{:}\ \hat{\boldsymbol{\pi}}) = 23.08$. Now $\Pr\{\chi^2_{16} \geqslant 26.30\} = 0.05$, resulting in acceptance of the null hypothesis.

ACKNOWLEDGEMENT

The authors are grateful to Dr G. K. Eagleson for helpful comments.

REFERENÇES

Bahadur, R. R. (1960) Stochastic comparison of tests. *Ann. Math. Statist.*, **31**, 276–295.

——— (1967) Rates of convergence of estimates and test statistics. *Ann. Math. Statist.*, **38**, 303–324.

——— (1971) *Some Limit Theorems in Statistics.* Philadelphia: Society for Industrial and Applied Mathematics.

Beran, R. (1977) Minimum Hellinger distance estimates for parametric models. *Ann. Statist.*, **5**, 445–463.

Bickel, P. J. and Rosenblatt, M. (1973) On some global measures of the deviations of density function estimates. *Ann. Statist.*, **1**, 1071–1095.

Birch, M. W. (1964) A new proof of the Pearson–Fisher theorem. *Ann. Math. Statist.*, **35**, 817–824.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* Cambridge, Mass.: MIT Press.

Bofinger, E. (1973) Goodness-of-fit tests using sample quantiles. *J. R. Statist. Soc.* B, **35**, 277–284.

Chapman, J. W. (1976) A comparison of the X^2, $-2 \log R$, and the multinomial probability criteria for significance testing when expected frequencies are small. *J. Amer. Statist. Ass.*, **71**, 854–863.

Cochran, W. G. (1952) The χ^2 test of goodness-of-fit. *Ann. Math. Statist.*, **23**, 315–345.

Cressie, N. (1979) An optimal statistic based on higher order gaps. *Biometrika*, **66**, 619–627.

Darling, D. A. (1953) On a class of problems related to the random division of an interval. *Ann. Math. Statist.*, **24**, 239–253.

Durbin, J. (1978) Goodness-of-fit tests based on the order statistics. In *Conference on Information Theory, Statistical Decision Functions, Random Processes: Transactions, Prague*, 7, 109–118, Dordrecht: Riedel.

Fienberg, S. E. (1979) The use of Chi-squared statistics for categorical data problems. *J. R. Statist. Soc.* B, **41**, 54–64.

——— (1980) *The Analysis of Cross-classified Categorical Data.* Second edition. Cambridge, Mass.: MIT Press.

Goldstein, M., Wolf, E. and Dillon, W. (1976) On a test of independence for contingency tables. *Commun. in Statist., Part* A5, 159–169.

Greenwood, M. (1946) The statistical study of infectious diseases. *J. R. Statist. Soc.* A, **109**, 85–110.

Haberman, S. J. (1978) *Analysis of Qualitative Data 1.* New York: Academic Press.

Hoeffding, W. (1965) Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.*, **36**, 369–408.

Holland, P. W. (1967) A variation on the minimum chi-square test. *J. Math. Psychol.*, **4**, 377–413.

Holst, L. (1972) Asymptotic normality and efficiency for certain goodness-of-fit tests. *Biometrika*, **59**, 137–145.

Horn, S. D. (1977) Goodness-of-fit tests for discrete data: A review and an application to a health impairment scale. *Biometrics*, **33**, 237–248.

Ivchenko, G. I. and Medvedev, Y. I. (1978) Separable statistics and hypothesis testing. The case of small samples. *Theory of Probability and its Applications*, **23**, 764–775.

Johnson, N. L. and Kotz, S. (1969) *Distributions in Statistics: Discrete Distributions.* Boston: Houghton Mifflin.

Kale, B. K. and Godambe, V. P. (1967) A test of goodness-of-fit. *Statistiche Hefte*, **8**, 165–172.

Kendall, M. G. and Stuart, A. (1973) *The Advanced Theory of Statistics*, Vol. 2. London: Griffin.

Kirmani, S. and Alam, S. (1974) On goodness-of-fit tests based on spacings. *Sankhya*, A, **36**, 197–203.

Koehler, K. J. and Larntz, K. (1980) An empirical investigation of goodness-of-fit statistics for sparse multinomials. *J. Amer. Statist. Ass.*, **75**, 336–344.

Kullback, S. (1959) *Information Theory and Statistics.* New York: Wiley.

Lancaster, H. O. (1969) *The Chi-squared Distribution.* New York: Wiley.

Larntz, K. (1978) Small sample comparisons of exact levels of chi-squared goodness-of-fit statistics. *J. Amer. Statist. Ass.*, **73**, 253–263.

Mathai, A. M. and Rathie, P. N. (1975) *Basic Concepts in Information Theory and Statistics.* New Delhi: Wiley.

Mitra, S. K. (1958) On the limiting power functions of the frequency chi-square test. *Ann. Math. Statist.*, **29**, 1221–1233.

Miyamoto, Y. (1976) Optimum spacing for goodness fit test based on sample quantiles. In *Essays in Probability and Statistics* (Volume in honour of Professor J. Ogawa), (S. Ikeda *et al.*, eds) pp. 475–483, Tokyo.

Moore, D. S. (1976) Recent developments in chi-square tests for goodness-of-fit. Mimeograph series 459, Department of Statistics, Purdue University.

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.

Morris, C. (1975) Central limit theorems for multinomial sums. *Ann. Statist.*, **3**, 165–188.

Neyman, J. (1949) Contribution to the theory of the χ^2 test. *Proc. 1st Berkley Symp. Math. Statist. Prob.*, pp. 239–273.

Pearson, K. (1900) On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophy Magazine Series (5)*, **50**, 157–172.

Pyke, R. (1965) Spacings. *J. R. Statist. Soc.* B, **27**, 395–449.

Rao, C. R. (1963) Criteria of estimation in large samples. *Sankhya*, A, **25**, 189–206.

Rathie, P. N. (1973) Some characterization theorems for generalized measures of uncertainty and information. *Metrika*, **20**, 122–130.

Rathie, P. N. and Kannappan, P. (1972) A directed-divergence function of type β. *Information and Control*, **20**, 38–45.

Read, T. R. C. (1982) On choosing a goodness-of-fit test. Unpublished PhD Thesis, Flinders University, South Australia.

——— (1984a) Small sample comparison for the power divergence goodness-of-fit statistics. *J. Amer. Statist. Ass.*, **79**, to appear.

——— (1984b) Closer asymptotic approximations for the distributions of the power divergence goodness-of-fit statistics. *Ann. Inst. Statist. Math.*, **36**, 59–69.

——— (1984c) Minimum distance parameter estimation for the multinomial model. Submitted.

Rényi, A. (1961) On measures of entropy and information. *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, **1**, 547–561.

Sharma, B. D. and Taneja, I. J. (1975) Entropy of type (α, β) and other generalized measures in information theory. *Metrika*, **22**, 205–215.

Smith, P. J., Rae, D. S., Manderscheid, R. W. and Silbergeld, S. (1981) Approximating the moments and distribution of the likelihood ratio statistic for multinomial goodness-of-fit. *J. Amer. Statist. Ass.*, **76**, 737–740.

Watson, G. S. (1959) Some recent results in chi-square goodness-of-fit tests. *Biometrics*, **15**, 440–468.

West, E. N. and Kempthorne, O. (1972) A comparison of the Chi^2 and likelihood ratio tests for composite alternatives. *J. Statist. Computation and Simulation*, **1**, 1–33.

Yarnold, J. K. (1972) Asymptotic approximations for the probability that a sum of lattice random vectors lies in a convex set. *Ann. Math. Statist.*, **43**, 1566–1580.