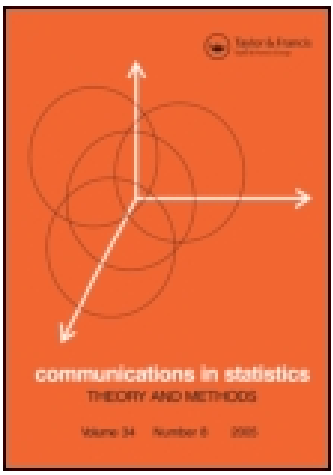

Communications in Statistics - Theory and Methods
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lsta20

A test of goodness-of -fit based on extreme multinomial cell frequencies

Martin T. Wells [a] , S. Rao Jammalamadaka [b] & Ram C. Tiwari [c]

[a] Dept. of Economic and Soc. Statistics , Cornell University , Ithaca, NY, 14851-0952

[b] Statistics and Applied Probability Program , Univ. of California , Santa Barbara, CA, 93106

[c] Dept. of Math , Univ. of North Carolina , Charlotte, NC, 28223

Published online: 27 Jun 2007.

To cite this article: Martin T. Wells , S. Rao Jammalamadaka & Ram C. Tiwari (1989) A test of goodness-of -fit based on extreme multinomial cell frequencies, Communications in Statistics - Theory and Methods, 18:4, 1527-1547, DOI: 10.1080/03610928908829984

To link to this article: http://dx.doi.org/10.1080/03610928908829984

COMMUN. STATIST.-THEORY METH., 18(4), 1527-1547 (1989)

A TEST OF GOODNESS-OF-FIT BASED ON EXTREME MULTINOMIAL CELL FREQUENCIES

Martin T. Wells
Dept. of Economic and Soc. Statistics
Cornell University
Ithaca, NY 14851-0952

S. Rao Jammalamadaka
Statistics and Applied Probability Program
Univ. of California
Santa Barbara, CA 93106

Ram C. Tiwari
Dept. of Math.
Univ. of North Carolina
Charlotte, NC 28223

Keywords and Phrases: *sparse cells*; *crowded cells*; *spacings*; *asymptotic normality*; *efficiency*.

ABSTRACT

For the problem of testing goodness-of-fit of a specified distribution, a new test based on the number of extreme cell frequencies is proposed. A cell is called "sparse" ("crowded") if the corresponding cell frequency is less than (greater than) or equal to a value $u \geq 0 (v \geq 0)$. Then, the proposed test statistic, $SC_N^n(u,v)$, is the number of sparse plus croweded, cells, where n denotes the sample and N is the number of mutually exclusive and collectively exhaustive cells. The exact distribution of $SC_N^n(u,v)$ is derived under the null hypothesis. The asymptotic distribution of $SC_N^n(u,v)$ under a sequence of local alternatives is also derived. The efficiency of this test statistic with respect to several other test statistics is obtained. A discussion of the merits and shortcomings of the proposed test procedure is also given.

1. INTRODUCTION

Let $X_1,\ldots,X_n$ be a sample of size n from a continuous distribution function F. The goodness-of-fit problem of testing whether a specified distribution generated the observations can be reduced to testing if the observations have a uniform distribution, through the probability integral transform. Thus we may (and shall) assume, without any loss of generality, that the support of F is [0,1] and that the null hypothesis of interest is

$$H_0: \quad f(x) = 1, \; x\in[0,1] \tag{1.1}$$

where f denotes the probability density function.

Let P_N be a partition of the interval [0,1] into N mutually exclusive and collectively exhaustive cells with the probability of any sample observation falling into the jth cell being equal to p_{jN}. The hypothesis of uniformity in (1.1) is equivalent to testing that the p_{jN}'s are equal to 1/N (j=1, ..., N). Let O_{jN} be the observed frequency of the jth cell, j=1, ..., N. Then note that $O_{jN} \geq 0 \; \forall_j$ and $\Sigma_{j=1}^{N} O_{jN} = n$. Choose and fix numbers u and v such that $0 \leq u < v \leq N$, and define

$$S_N^n(u) = \Sigma_{j=1}^{N} I(O_{jN} \leq u), \tag{1.2}$$

$$C_N^n(v) = \Sigma_{j=1}^{N} I(O_{jN} \geq v), \tag{1.3}$$

and

$$\begin{aligned} SC_N^n(u,v) &= \Sigma_{j=1}^{N} I(O_{jN} \leq u \text{ or } O_{jN} \geq v) \\ &= S_N^n(u) + C_N^n(v). \end{aligned} \tag{1.4}$$

Then, the statistics $S_N^n(u)$, $C_N^n(v)$, and $SC_N^n(u,v)$ represent the number of sparse cells, crowded cells, and the number of sparse plus crowded cells, respectively. The statistic $S_N^n(u)$

is a generalization of Renyi's(1962) statistic $S_N^n(0)$. The test criterion is to reject H_0 if any of these statistics is too large. Under H_0, the exact distributions of the statistics $S_N^n(u)$ and $C_N^n(v)$ are derived in Section 2 while their asymptotic distributions as $n/N \to m$, $0 < m < \infty$, are derived in Section 3 under the sequence of alternatives

$$A_n: \quad f_n(x) = 1 + n^{-1/4}\ell(x), \quad x\epsilon[0,1] \qquad (1.5)$$

converging to the hypothesis (1.1) at the rate $n^{-1/4}$, where $\ell(\cdot)$ is a continuously differentiable function on [0,1] that is, $\ell(\cdot)$ belongs to $C^{(1)}[0,1]$. Alternatives of this type have been considered by many authors (see, e.g., Del Pino(1979), Gebert and Kale(1969), Sethuraman and Rao(1970), Rao and Sethuraman(1975), Jammalamadaka and Tiwari(1985,1987), and Jammalamadaka and Wells(1986)). The comparison of asymptotic efficiencies of these tests with respect to the usual χ^2-test statistic given by

$$T_N^n = (N/n)\,\Sigma_{j=1}^N (O_{jN} - n/N)^2$$
$$= (N/n)\,\Sigma_{j=1}^N O_{jN}^2 - n \qquad (1.6)$$

are made in Section 4. Note that the statistic T_N^n depends on all cell frequencies where as the statistic $SC_N^n(u,v)$ depends only on the cell frequencies which are either less than or equal to u or greater than or equal to v. Thus, there will be efficiency loss in using $SC_N^n(u,v)$ against T_N^n, and this loss could be substantial if there are relatively large number of cells with frequencies between u and v exclusively. In Section 5, we treat the problem of testing goodness-of-fit in the presence of nuisance parameters. It is shown that the results of Section 3 remain valid even for the composite hypothesis.

2. EXACT DISTRIBUTIONS OF $S_N^N(u)$, $C_N^N(v)$ AND $SC_N^n(u,v)$

In this section, we derive the exact distribution of the statistics $S_N^n(u)$, $C_N^n(v)$ and $SC_N^n(u,v)$ under H_0 using the following Dirichlet integral of type-I given in Sobel and Uppuluri (1974): For $0 \le p \le 1/b$ and $n \ge br$, define

$$I_p^b(r,n) = \frac{\Gamma(n+1)}{\Gamma^b(r)\,\Gamma(n-br+1)} \int_0^p \cdots \int_0^p (1 - \Sigma_{i=1}^r x_i)^{n-br} \prod_{i=1}^b x_i^{r-1} dx_i . \tag{2.1}$$

The expression (2.1) can be seen to be exactly equal to the probability that b cells each, of a multinomial distribution with all cell probabilities equal to p, Mult(b; p,...,p), has frequency at least r, when n independent observations are drawn from this distribution.

The following results are slight modifications of Sobel and Uppuluri(1974).

Theorem 2.1.

$$P(S_N^n(u) = s) = \binom{N}{s} \Sigma_{j=0}^s (-1)^j \binom{s}{j} I_{1/N}^{N-s+j}(u+1,n) ; \tag{2.2}$$

$$E(S_N^n(u)^{[m]}) = N^{[m]} \Sigma_{j=0}^m (-1)^j \binom{m}{j} I_{1/N}^j(u+1, n) ; \tag{2.3}$$

$$P(C_N^n(v)=c) = \binom{N}{c} \Sigma_{j=0}^{N-c} (-1)^j \binom{N-c}{j} I_{1/N}^{c+j}(v,n) ; \tag{2.4}$$

$$E(C_N^n(v)^{[m]}) = N^{[m]}\, I_{1/N}^m(v,n) , \tag{2.5}$$

where $E(X^{[m]})$ is the factorial moment and $N^{[m]} = N(N-1) \ldots (N-m+1)$. Let

$$Q(s,i) = \Sigma_{j=0}^s (-1)^j \binom{s}{j} I_{1/N}^j(u+1,i)$$

$$F(s,c) = \Sigma_{j=0}^n b_j(n,s/N)\, Q(s,j)\, I_{1/N}^c(1 - \tfrac{s}{N})(v,n-j)$$

where $b_i(n,s/N) = \binom{n}{i} (\frac{s}{N})^i (1 - \frac{s}{N})^{n-1}$. Then

$$P(S_N^n(u) = s, C_N^n(v) = c) =$$
$$\Sigma_{j=0}^{N-s-c}(-1)^j \Sigma_{i=0}^{j} \binom{s+i}{i}\binom{c+j-i}{j-1}\binom{n}{s+i,c+j-i}F(s+i,c+j-1) \tag{2.6}$$

where $\binom{a}{b,c}$ denote the usual multinomial coefficients. Also,

$$E\{(S_N^n(u))^{[\ell]} \cdot (C_N^n(v))^{[m]}\} = N^{[\ell+m]}F(\ell,m). \tag{2.7}$$

Finally, the probability distribution of $SC_N^n(u,v)$ is given by the following theorem.

Theorem 2.2. $P(SC_N^n(u,v) = t) =$

$$\Sigma_{k=0}^{t} \Sigma_{j=0}^{N-t-c-k}(-1)^j \Sigma_{i=0}^{j} \binom{t-k+i}{i}\binom{k+j-i}{j-i}\binom{n}{t-k+i,k+j-i} \times F(t+i-k,k+j-1).$$

Proof. The result follows from (2.1) and Theorem 2.1 by observing that $P(SC_N^n(u,v)=t)=\Sigma_{k=0}^{t}P(S_N^n(u)=t-k,\ C_N^n(v) = k)$. ∎

One could use the tables in Sobel, Uppuluri, and Frankowski(1977) to tabulate the exact values.

3. THE ASYMPTOTIC DISTRIBUTIONS OF $SC_N^n(u,v)$

In this section we establish the asymptotic normality of $SC_N^n(u,v)$ under the sequence of local alternatives A_n given by (1.5). These results will be useful in Section 4 where we compute the Pitman asymptotic relative efficiencies(ARE's). We will assume that the sample size n and the number of cells N tend to infinity in such a way that $n/N \to m$, $0 < m < \infty$.

For deriving the asymptotic distributions of the statistics $S_N^n(u)$, $C_N^n(v)$, and $SC_N^n(u,v)$ under the alternatives A_n in (1.5) we make use of Theorem 2.1 of Holst and Rao(1980, p. 25). This theorem gives a general technique for finding the distribution of functions of multinomial frequencies. We state this result for completeness.

Let $\{n_\nu\}$ and $\{N_\nu\}$ be sequences of nondecreasing positive numbers. Assume that as $\nu \to \infty$,

$$N_\nu \longrightarrow \infty,\ n_\nu \longrightarrow \infty \text{ and } m_\nu = n_\nu/N_\nu \longrightarrow m,\ 0 < m < \infty. \quad (3.1)$$

Let $(O_{1\nu},\ldots,O_{N_\nu\nu})$ be Mult $(n_\nu; p_{1\nu},\ldots,p_{N_\nu\nu})$, where $\Sigma_{j=1}^{N_\nu} O_{j\nu} = n_\nu$. Under the null hypothesis, we are interested in the asymptotic distribution of the random variable

$$W_\nu = \Sigma_{k=1}^{N_\nu} h_{k\nu}(O_{k\nu}) \text{ as } \nu \longrightarrow \infty, \quad (3.2)$$

where $\{h_{k\nu};\ k=1,\ldots,N_\nu,\ \nu \geq 1\}$ are real-valued Borel measurable functions satisfying certain regularity conditions (see condition (A) on P. 23 of Holst and Rao(1980)). Let $\{\xi_{1\nu},\ldots,\xi_{N_\nu\nu}\}$, $\nu \geq 1$ be a sequence of independent random variables, where $\xi_{k\nu}$ has a Poisson distribution with parameter $n_\nu p_{k\nu}$, $k=1,\ldots,N_\nu$, $\nu \geq 1$. Define

$$\lambda_\nu = \Sigma_{k=1}^{N_\nu} h_{k\nu}(\xi_{k\nu}), \quad (3.3)$$

$$\mu_\nu = E(\lambda_\nu). \quad (3.4)$$

For $0 < q < 1$, let $M = [Nq]$, the integer part of Nq, and define

$$\lambda_{\nu q} = \Sigma_{k=1}^{M} h_{k\nu}(\xi_{k\nu}). \quad (3.5)$$

Then we have the following.

Theorem 3.1 (Holst and Rao(1980)). Let λ_ν, μ_ν and $\lambda_{\nu q}$ be as defined by (3.3), (3.4) and (3.5). Assume that there exists a $q_0 < 1$ such that for $q \geq q_0$, $\Sigma_{k=1}^{M} p_k \longrightarrow p_q$, $0 < p_q < 1$, and

$$\mathcal{L}\begin{bmatrix} (\lambda_{\nu q} - E\lambda_{\nu q})/N^{1/2} \\ \sum_{k=1}^{M} (\xi_{k\nu} - np_{k\nu})/n^{1/2} \end{bmatrix} \longrightarrow N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} A_q & B_q \\ B_q & P_q \end{bmatrix}\right],$$

where $A_q \to A_1$, $B_q \to B_1$ and $P_q \to 1$ as $q \to 1-0$. Then as $\nu \to \infty$, $\mathcal{L}((W_\nu - \mu_\nu)/n^{1/2}) \longrightarrow N(0, A_1 - B_1^2)$.

Consider the partition of [0,1] into N mutually exclusive and collectively exhaustive cells with kth cell being $[(k-1)/N, k/N)$. Then

$$p_{kN} = \text{Probability of the kth cell}$$
$$= \int_{(k-1)/N}^{k/N} f_n(x)dx \simeq [1+n^{-1/4}\ell(k/N)]/N \qquad (3.6)$$

under the alternatives (1.5). Define

$$\lambda_N(S) = \Sigma_{k=1}^N I(\xi_{kN} \leq u) ; \qquad (3.7)$$

$$\lambda_N(C) = \Sigma_{k=1}^N I(\xi_{kN} \geq v) ; \qquad (3.8)$$

$$\lambda_N(SC) = \lambda_N(S) + \lambda_N(C) , \qquad (3.9)$$

where $(\xi_{1N}, \ldots, \xi_{NN})$ are independent random variables with ξ_{kN} having a Poisson distribution with parameter Np_{kN}, $k=1,\ldots,N$. Then, from (3.6) we have

$$\mu_N(S) = E\lambda_N(S) = \Sigma_{k=1}^N \Sigma_{j=0}^u e^{-np_{kN}}(np_{kN})^j/j!$$
$$\simeq N \Sigma_{j=0}^u m^j e^{-m} \int_0^1 \{\ell(t)/n^{1/4}\}(1+\ell(t)/n^{1/4})^j \, dt$$
$$\simeq N \Sigma_{j=0}^u \frac{m^j}{j!} e^{-m}\{1+[\binom{j}{2} - jm + m^2/2]\int_0^1 (\ell^2(t)/n^{1/2})dt, \qquad (3.10)$$

where $\binom{k}{j} = 0$ if $k<j$. The third relation in (3.10) follows by expanding the integrand and ignoring the terms which are smaller than $n^{-1/2}$. Similarly, we have

$$\mu_N(C) = E\lambda_N(C) \simeq N \Sigma_{j=v}^\infty \frac{m^j}{j!} e^{-m}\{1+[\binom{j}{2} - j + m^2/2]\int_0^1 (\ell^2(t)/n^{1/2})dt \qquad (3.11)$$

Again,

$$A_q^{(S)} = \lim_{N\to\infty} N^{-1} \operatorname{Var}\Big(\sum_{k=1}^{M} I(\xi_{kN} \le u)\Big)$$

$$= \lim_{N\to\infty} N^{-1} \sum_{k=1}^{M} P(\xi_{kN} \le u) P(\xi_{kN} \ge u+1)$$

$$\simeq \lim_{N\to\infty} (qN/N) \sum_{j=0}^{u} \frac{e^{-n/N}}{j!} (n/N)^j \sum_{i=u+1}^{\infty} \frac{e^{-n/N}}{i!} (n/N)^i$$

$$= q\, e^{-2m} \sum_{j=0}^{u} (m^j/j!) \sum_{i=u+1}^{\infty} m^i/i!$$

and hence

$$A_1(S) = \lim_{q\to 1-0} A_q(S) = e^{-2m} \sum_{j=0}^{u} \sum_{i=u+1}^{\infty} m^{i+j}/i!j!. \quad (3.12)$$

Similarly,

$$B_q(S) = \lim_{N\to\infty} N^{-1} \operatorname{Cov}\Big(\sum_{k=1}^{N} I(\xi_{kN} \le u)\ ,\ \sum_{k=1}^{M} \xi_{kN}\Big)$$

$$= q\, e^{-m} \sum_{j=0}^{u} m^j (j-m)/j!$$

and

$$B_1(S) = \lim_{q\to 1-0} B_q(S) = e^{-m} \sum_{j=0}^{u} m^j (j-m)/j!. \quad (3.13)$$

Also,

$$A_1(C) = \lim_{q\to 1-0} A_q(C) = e^{-2m} \sum_{i=0}^{v-1} \sum_{j=v}^{\infty} m^{i+j}/i!j! \quad (3.14)$$

and

$$B_1(C) = \lim_{q\to 1-0} B_q(C) = e^{-m} \sum_{j=v}^{\infty} m^j (j-m)/j! \ . \quad (3.15)$$

The joint asymptotic normality required in Theorem 3.1 is established if we verify that, for any real a, the triangular sequence

$$\{Y_{kN} = aI(\xi_{kN} \le u) + \xi_{kN};\ k=1,\ldots,N,\ N \ge 1\} \quad (3.16)$$

satisfies the Liapounov condition (see Chung(1968), p. 200). We need to show that

$$\sum_{k=1}^{M} E|Y_{kN}|^3 \ / \ [\operatorname{Var}(\sum_{k=1}^{M} Y_{kN})]^{3/2} \quad (3.17)$$

goes to zero as $M \to \infty$. Since $N^{-1} \operatorname{Var}(\sum_{k=1}^{M} Y_{kN})$ has a finite nonzero limit, it follows that $\operatorname{Var}(\sum_{k=1}^{M} Y_{kN})$ is

$O(M)$. It is easily checked that the numerator in (3.17) is $O(M)$ so that the ratio in (3.17) goes to zero as $M \longrightarrow \infty$.

Similarly, the triangular sequence $\{aI(\xi_{kN} \geq v) + \xi_{kN};$ $k=1,\ldots,N,\ N \geq 1\}$, for any real a, also satisfies the Liapounov condition. Hence we have proved the following theorem.

Theorem 3.2. Under the alternatives (1.5), the asymptotic distributions of $S_N^n(u)$ and $C_N^n(v)$ are given by

$$\mathcal{L}(n^{-1/2}(S_N^n(u)-\mu_N(S)) \longrightarrow N(0,\ A_1(S) - B_1^2(S))$$

and

$$\mathcal{L}(n^{-1/2}(C_N^n(v) - \mu_N(C)) \longrightarrow N(0,A_1(C) - B_1^2(C)),$$

respectively, as $n,N \longrightarrow \infty$, $n/N \longrightarrow m$, $0 < m < \infty$.

Now, note that $SC_N^n(u,v)$ is the convolution of $S_N^n(u)$ and $C_N^n(v)$. For this convolution to be asymptotically normal we must verify that, the random vector $(n^{-1/2}(S_N^n(u) - \mu_N(S)),\ n^{-1/2}(C_N^n(v) - \mu_N(C))' = (S_N^*, C_N^*)'$, say, is asymptotically normal. The joint normality of $(S_N^*, C_N^*)'$ is established if we verify that the sequence $\{aS_N^* + C_N^*\}$ satisfies the Liapounov condition for any real number a. The verification of this is similar to that for the sequence in (3.16) and hence is omitted.

Using the independence of the sequences $\{\xi_{1N},\ldots,\xi_{NN}\}$, $N \geq 1$ of Poisson random variables, we have

$$\begin{aligned}\mu_N(SC) &= E(\lambda_N(SC)) \\ &= E(\lambda_N(S) + \lambda_N(C)) \\ &= \mu_N(S) + \mu_N(C)\ , \end{aligned} \tag{3.18}$$

$$\begin{aligned}A_1(SC) &= \lim_{q \to 1-0} A_q(SC) \\ &= \lim_{N \to \infty} N^{-1}\ \mathrm{Var}(\lambda_N(SC))\end{aligned}$$

$$= A_1(S) + A_1(C) - 2e^{-2m} \Sigma_{j=0}^{u} m^j/j! \; \Sigma_{i=v}^{\infty} m^i/i! \qquad (3.19)$$

and

$$B_1(SC) = \lim_{q \to 1-0} B_q(SC)$$
$$= \lim_{N \to \infty} N^{-1} \operatorname{Cov}(\lambda_N(SC), \Sigma_{k=1}^{N} \xi_{kN})$$
$$= B_1(S) + B_1(C) \; . \qquad (3.20)$$

Thus, we have the following main result of this section.

Theorem 3.3. Under the sequence of alternatives in (1.5),

$$\mathcal{L}(n^{-1/2}(SC_N^n(u,v) - \mu_N(SC))) \longrightarrow N(0, A_1(SC) - B_1^2(SC)),$$

as $n,N \longrightarrow \infty$, and $n/N \to m$, $0 \leq m < \infty$.

Under the null hypothesis the asymptotic distributions of $S_N^n(u)$, $C_N^n(v)$ and $SC_N^n(u,v)$ are given by the following.

Corollary 3.4. Under the null hypothesis, the random variables $n^{-1/2}(S_N^n(u) - \mu_N^*(S))$, $n^{-1/2}(C_N^n(v) - \mu_N^*(S))$ and $n^{-1/2}(SC_N^n(u,v) - \mu_N^*(SC))$ are normally distributed as $n,N \to \infty$, and $n/N \to m$, $0 \leq m < \infty$ with means zero and variances $A_1(S)-B_1^2(S)$, $A_1(C)-B_1^2(C)$ and $A_1(SC) - B_1^2(SC)$, respectively, where from (3.10) and (3.11), $\mu_N^*(S) = Ne^{-m} \Sigma_{j=0}^{u} m^j/j!$, $\mu_N^*(C) = Ne^{-m}\Sigma_{i=v}^{\infty} m^i/i!$, and $\mu_N^*(SC) = \mu_N^*(S) + \mu_N^*(C)$.

As an example, we shall analyze a data set from Hald(1967, p. 329). He gave data that represent the range in terms of percentage concentrations of calcium carbonate for 52 sets of 5 samples each, taken from a mixing plant of raw metal. These data were formed into a 50 cell histogram as reported in Simonoff(1985). Letting m=1, u=0, and v=4 (see Section 4 for the optimal choices of u and v) computation gives $SC_{50}^{52}(0,4) = 23$. This gives a p-value of $p = .06$ for testing uniformity. If one uses the χ^2-test one can compute that $T_{50}^{52} = 65.3$ which has a p-value of $p = .08$. Therefore,

for this data set the test based on extreme cell frequencies gives results similar to the χ^2-test, that is, the hypothesis of uniformity can not be rejected.

Simonoff(1985) also analyzed this data set with his proposed test statistic M^2 and found that M^2 rejects the hypothesis of uniformity with a p-value which is less than .001. Clearly, M^2 is a much more powerful test statistic than either $SC_N^n(u,v)$ or T_N^n. However, the test procedure for M^2 is quite complicated to carry out and the critical values may be difficult to find in practice.

4. PITMAN ASYMPTOTIC RELATIVE EFFICIENCY OF $S_N^n(u)$, $C_N^n(v)$ AND $SC_N^n(u,v)$

The Pitman asymptotic relative efficiency (ARE) of a test relative to another test is defined to be the limit of the inverse ratio of sample sizes required to obtain the same limiting power at a sequence of alternatives converging to the null hypothesis. The limiting power should be a value between the limiting test size, α, and the maximum power, 1. If the limiting power of the test at a sequence of alternatives is α, then its ARE with respect to any other test with the same test size and with limiting power greater than α, is zero. On the other hand, if the limiting power of a test at a sequence of alternatives converges to a number in the open interval $(\alpha,1)$, then a measure of the rate of convergence, called "efficacy", can be computed. Under certain standard regularity assumptions (see for example Serfling(1980)) which include a condition about the nature of the alternative, asymptotic normality of the test statistic under the sequence of alternatives, etc., the efficacy is given by

$$\text{eff} = \mu_\Delta^4/\sigma^4 \ .$$

Here, μ_Δ and σ^2 are the mean and variance of the limiting normal distribution under the sequence of alternatives when the test statistic has been normalized to have a limiting standard normal distribution under the null hypothesis. In such a situation, the ARE of one test with respect to another is simply the ratio of their efficacies.

From Theorems 3.2, 3.3 and Corollary 3.4, the efficacies of the test statistics $S_N^n(u)$, $C_N^n(v)$ and $SC_N^n(u,v)$ are

$$\text{eff } (S_N^n(u)) = \frac{\{\Sigma_{j=0}^{u}(m^j/j!)[(\frac{j}{2})-jm-(m^2/2)]\int_0^1 \ell^2(t)dt\}^4}{[\Sigma_{i=0}^{u}(m^i/i!)\Sigma_{j=u+1}^{\infty}(m^j/j!)-(\Sigma_{j=0}^{u} m^j(j-m)/j^2]^2}, \tag{4.1}$$

$$\text{eff } (C_N^n(v)) = \frac{\{\Sigma_{j=v}^{\infty}(m^j/j!)[(\frac{j}{2})-jm+(m^2/2)]\int_0^1 \ell^2(t)dt\}^4}{[\Sigma_{i=v}^{\infty}\Sigma_{j=0}^{v-1}(m^{i+j}/i!j!)-(\Sigma_{j=v}^{\infty} m^j(j-m)/j!)^2]^2}, \tag{4.2}$$

and

$$\text{eff}(SC_N^n(u,v)) = A/B, \tag{4.3}$$

where

$$A=\{[\Sigma_{i=0}^{u}(m^i/i!)[(\tfrac{i}{2})-jm+m^2/2]$$

$$+\Sigma_{j=v}^{\infty}(m^j/j!)[(\tfrac{j}{2})-jm+m^2/2]]\int_0^1 \ell^2(t)dt\}^4$$

$$B=[\Sigma_{i=0}^{u}(m^i/i!)\Sigma_{j=u+1}^{\infty}(m^j/j!)+\Sigma_{i=v}^{\infty}(m^i/i!)\Sigma_{j=0}^{v-1}(m^j/j!)$$
$$- 2\,\Sigma_{j=0}^{u}(m^j/j!)\,\Sigma_{i=v}^{\infty}(m^i/i!)$$
$$- (\Sigma_{i=0}^{u} m^i(i-m)/i! + \Sigma_{j=v}^{m} m^j(j-m)/j!)^2]^2 .$$

A well known special case of $S_N^n(u)$ is $S_N^n(0)$, Renyi's(1962) empty cell test-statistic. From (4.1) we see that

$$\text{eff}(S_N^n(0)) = m^8\{\int_0^1 \ell^2(t)dt\}^4/(e^m - 1 - m^2)^2/16. \quad (4.4)$$

The efficacy of the χ^2-type test statistic T_N^n as defined in (1.6) has been obtained by Jammalamadaka and Tiwari(1987). This is given by

$$\text{eff}(T_N^n) = m^2(\int_0^1 \ell^2(t)dt)^4/4 . \quad (4.5)$$

The test statistic T_N^n is known to be most efficient among the class of tests based on symmetric functions.

Using (4.1)-(4.4) one can make tables of the efficacies of the test statistics $S_N^n(u)$, $C_N^n(v)$ and $SC_N^n(u,v)$ for any choice of m, u and v. Also, for fixed m, the optimal choice of (u,v) can be obtained on a computer. For example, if m = 1, one can find that u=0 and v=4 is optimal. Note that $\text{eff}(S_N^n(0)) = .121142$, and $\text{eff}(SC_N^n(0,4)) = .176386$, thus using the crowded cells adds efficiency. Using (4.4) one can show that among the empty cell tests, the efficacy is maximized at m = 2.8. Using (4.3) and (4.5) asymptotic relative efficiency of $SC_N^n(0,4)$ with respect to T_N^n for m=1 is given by $\text{ARE}(SC_N^n(0,4), T_N^n) = 70.55\%$. The efficacies of the tests is increasing in m. Though the test statistic $SC_N^n(0,4)$ is computationally simple and appealing, note that there is substantial loss in the efficiency with respect to the most efficient symmetric test T_N^n.

The dual of the tests based on frequencies are the tests based on disjoint spacings. The duality is that in the tests based on frequencies the width of the cell is fixed and the number of observations contained in a cell is random and has expected value m. In the tests based on m-step disjoint spacings, the length of the cell, that is the spacing, is random, but the number of observations covered by the

spacings is fixed at m. For more on this duality see Jammalamadaka and Tiwari(1985, 1987). Jammalamadaka and Wells(1986) considered one-step spacings which are the dual of the tests discussed in this paper with m = 1. They have shown that the most efficient spacings test of this type has efficacy equal to .79415. These results agree with the conclusions of Jammalamadaka and Tiwari(1985,1987) that tests based on spacings seem to be more efficient than tests based on cell frequencies.

All of the test procedures discussed here are unable to detect alternatives converging to the uniform at a rate faster than $n^{-1/4}$. There are quite a few goodness-of-fit tests, such as the Kolmogorov-Smirnov or Cramer-von Mises test that can distinguish alternatives converging at a rate no faster than $n^{-1/2}$. Therefore the proposed tests have an asymptotic relative efficiency of zero as compared to these more powerful tests. Other tests with this lack of efficiency include "symmetric" functions of spacings and multinomial frequencies; see Holst(1972), Cressie(1979), Hall(1986) and the references contained therein for a full discussion on this problem. Also, the proposed tests are not useful for testing any null hypothesis other than the uniform if the data is provided in tabled form <u>apriori</u> since the uniform distribution for this discretized data cannot be assumed without loss of generality.

There may be situations where the goodness-of-fit tests which are more efficient may not be applicable, for example, when the number of cells is forced on the user(see the example at the end of Section 3). In such a situation one may use the tests based on the cell frequencies. The critical values of most of the goodness-of-fit tests depend on large sample theory; however, for the test based on extreme cell frequencies one may tabulate the exact critical values even

for small sample sizes using the tables of Sobel, Uppulari and Frankowski(1977).

As will be seen in the next section these test procedures adapt easily to the problem of testing goodness-of-fit in presence of nuisance parameters.

5. COMPOSITE HYPOTHESIS

Let $Y_1, \ldots, Y_n$ be a sample from a continuous distribution with d.f. F^*. In this section we will study the goodness-of-fit testing problem for the composite null hypothesis H_0^c: $F^*(y) = G(y; \theta, \beta_0)$ versus the sequence of alternatives A_n^c: $F_n^*(y) = G(y; \theta, \beta_n)$ where $\theta \in \Theta \subseteq \mathbb{R}^p$ is a vector of unknown parameters which must be estimated from the data and $\beta_0, \beta_n \in \Gamma \subseteq \mathbb{R}^q$. It will be shown that H_0^c and A_n^c may be transformed into the form of (1.1) and (1.5). We shall propose a modified version of $SC_N^n(u,v)$ for the composite hypothesis and discuss its asymptotic behavior under the sequence of alternatives $\{A_n^c\}$.

For some differentiable function $h(x)$, $x = (x_1,\ldots,x_d) \in \mathbb{R}^d$, let $\nabla_{x_0} h(x)$ denote the vector of partial derivatives $(\frac{\partial}{\partial x_1} h(x), \ldots, \frac{\partial}{\partial x_d} h(x))$ evaluated at $x_0 \in \mathbb{R}^d$.

Let N_0 be the closure of a neighborhood of θ_0, the true unknown value of θ, and β_0 , the value of β specified under the null hypothesis H_0^c. Assume $\nabla_{\theta_0} G(y; \theta,\beta)$ and $\nabla_{\beta_0} G(y; \theta,\beta)$ exist and are finite all for $(\theta, \beta) \in N_0$. Since $\theta \in \Theta$ is unknown it must be estimated from the data. Assume that $\hat{\theta}_n$,

the estimate of θ_0, is $\sqrt{n}$-consistent that is, $n^{1/2}(\hat{\theta}_n-\theta_0) = O_p(n^{-1/2})$. It will be clear in what follows that this assumption is crucial.

Concerning the nature of the alternatives we shall assume that $\beta_n = \beta_0 + \gamma n^{-1/4}$, for some $\gamma \in \mathbb{R}^q$. This type of alternative was considered by Durbin(1973) for the composite goodness-of-fit test in the context of tests based on the empirical distribution function. This is a somewhat restrictive situation. An example of this setup is testing whether a distribution is Gaussian with some mean and a specified variance β_0 against the alternative that the variance is shifted by a factor of $n^{-1/4}$.

In the case of a simple hypothesis we were able to transform the goodness-of-fit problem to the problem of testing uniformity on [0,1] by applying a probability integral transformation (p.i.t.). In the case of the composite hypothesis no such transformation is exactly possible because θ is unknown. However, it will be shown that such a p.i.t. is asymptotically possible. If $\hat{\theta}_n$ is a $\sqrt{n}$-consistent estimate of θ_0, than one may define "estimated" uniform [0,1] random variables as $\hat{X}_i = G(Y_i; \hat{\theta}_n, \beta_0)$. Let $F^C(x) = G(G^{-1}(x; \hat{\theta}_n,\beta_0); \theta_0,\beta_0)$. Clearly F^C is defined on [0,1]. A Taylor series of F^C about θ_0 under H_0^C yields

$$F^C_{(x)} = x+(\hat{\theta}_n-\theta_0)\nabla_{\theta_0} G(G^{-1}(x;\theta,\beta_0);\theta_0,\beta_0)+O_p(n^{-1/2}) .$$

Similarly, under A_n^C this estimated p.i.t. yields $F_n^C(x)=G(G^{-1}(x; \hat{\theta}_n,\beta_0); \theta_0, \beta_n)$ and a Taylor series about θ_0 and β_0 yields $F_n^C(x) = x + L(x,\theta_0)n^{-1/4} + (\hat{\theta}_n-\theta_0)\nabla_{\theta_0} G(G^{-1}(x;$

$\theta,\beta_0)$; $\theta_0,\beta_0)+O_p(n^{-1/2})$, where $L(x;\ \theta_0) = \gamma\cdot\nabla_{\beta_0} G(G^{-1}(x;$ $\theta_0,\beta_0);\ \theta_0,\beta)$. Since $\hat{\theta}_n$ is an $\sqrt{n}$-consistent estimate $F^c(x)$ $= x + O_p(n^{-1/2})$ and $F_n^c(x) = x + L(x;\ \theta_0)n^{-1/4} + O_p(n^{-1/2})$. Therefore $F_n^c(k)$ is a $O_p(n^{-1/2})$ perturbation of $F_n(x)$ in (1.5). If $\frac{d}{dx} L(x;\ \theta_0) = \ell(x;\ \theta_0)$, $\frac{d}{dx} F^c(x) = f^c(x)$, and $\frac{d}{dx} F_n^c(x) = f_n^c(x)$, then $f^c(x)$ is as in (1.1) aside from the $O_p(n^{-1/2})$ term and $f_n^c(x)$ is an in (1.5) aside from $O_p(n^{-1/2})$ term. Since the alternatives of concern are only a distance proportional to $n^{-1/4}$ away from the null hypothesis, the $O_p(n^{-1/2})$ term vanishes as $n \to \infty$. Note that $\lim_{n\to\infty} n^{1/4}(f(x)-$ $f^c(x)) = 0$ and $\lim_{n\to\infty} n^{1/4}(f_n(x)-f_n^c(x)) = 0$ for all $x \in [0,1]$.

To carry out the test of H_0^c versus A_n^c one needs to partition the interval [0,1] into N disjoint cells. The analysis in the preceding paragraph indicates that these cells will be asymptotically equally probable under H_0^c. Therefore, the counting statistic, $SC_N^c(u, v; \hat{\theta}_n)$ (defined in similar fashion as $SC_N^n(u,v)$ in (1.4)), for the composite hypothesis will be asymptotically equal to $SC_N^n(u,v)$. To compute the centering mean, $\mu_N(SC;\ \hat{\theta}_n)$, one would follow steps similar to those in (3.6) - (3.11). However, one can see that the extra $O_p(n^{-1/2})$ term in the composite hypothesis problem would be one of the terms ignored as in (3.10). Hence $|\mu_N(SC;\ \hat{\theta}_n) - \mu_N(SC)\ | \xrightarrow{P} 0$ as $n \to \infty$. Therefore, the asymptotic distribution of the test statistic for the composite hypothesis is the same as the test statistic for the simple hypothesis.

Although it seems unlikely that the two test statistics have the same asymptotic behavior there is a simple intuitive explanation for this phenomenon. Recall that the type of alternatives under consideration are at a distance of $n^{-1/4}$ away from the null hypothesis. Also the estimates used are $\sqrt{n}$-consistent estimates, hence at a distance proportional to $n^{-1/2}$ away from the true value. Therefore, the test statistic can not distinguish between the estimate and the true value of the nuisance parameter. If the test statistics under consideration could discriminate alternatives at a distance of $n^{-1/2}$, which the test statistics discussed here can not, then one would find that the parameter estimation truly matters. For instance, the Kolmogorov-Smirnov and Cramer-von Mises test statistics can discriminate alternatives that are at a distance proportional to $n^{-1/2}$ away. For these two cases it is well known that the asymptotic distribution theory for the composite hypothesis and the simple hypothesis are quite different.

Similar results have been found for statistics that are functions of the sample spacings. If one uses a statistic based on "symmetric" functions of spacings, the test procedure can only distinguish alternatives at a distance of $n^{-1/4}$ away from the null hypothesis. However, if one uses "nonsymmetric" functions of spacings, then the test procedure can distinguish alternatives at a distance of $n^{-1/2}$ away from the null hypothesis. In Wells(1987) it is shown that the asymptotic behavior for the composite and simple hypotheses of "symmetric" functions of spacings are the same. However, the asymptotic behavior for the "nonsymmetric" functions is quite different when testing the composite hypothesis as compared to the simple hypothesis.

In summary, to test H_0^c versus A_n^c one has to estimate the unknown parameter by a $\sqrt{n}$-consistent estimate and use it in

a p.i.t. to transform the data with values in [0,1], and then proceed as if one is testing a simple hypothesis. One may, therefore, tabulate asymptotic critical values using Theorem 3.3. Once again, as in the case of the simple null hypothesis, the proposed tests will not be as powerful as the ones based on the empirical distribution function for testing the composite null hypothesis.

ACKNOWLEDGMENTS

The authors wish to thank Professor John E. Boyer and the two referees for helpful suggestions on an earlier version of this paper.

BIBLIOGRAPHY

Chung, K. L.(1968). A Course in Probability. Academic Press, New York.

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

Del Pino, G. E.(1979). On the asymptotic distribution of k-spacings with applications to goodness of fit-tests. Ann. Statist., 7, 1058-1065.

Durbin, J.(1973). Weak convergence of the sample distribution function when parameters are estimated. Ann. Statist., 1, 279-290.

Gebert, J. B. and Kale, B. K.(1969). Goodness of fit tests based on discriminatory information. Statistische Hefte, 10, 192-200.

Hald, A.(1967). Statistical Theory with Engineering Applications. John Wiley, New York.

Hall, P.(1986). On powerful distributional tests based on sample spacings. J. Multivariate Anal., 19, 201-224.

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

Holst, L. and Rao, J.S.(1980). Asymptotic theory for some families of two-sample nonparametric statistics, Sankhyā, Ser. A, 42, 19-52.

Jammalamadaka, S. R. and Tiwari, R. C.(1985). Asymptotic comparison of three tests for goodness of fit. Jour. Stat. Plan. Inf., 12, 295-304.

Jammalamadaka, S. R. and Tiwari, R. C.(1987). Efficiencies of some disjoint spacings tests relative to a χ^2 test. New Perspectives in Theoretical and Applied Statistics. (eds. Madan Puri, Jose Perez Vilaplana & Wolfgang Wertz). John Wiley, New York.

Jammalamadaka, S. R. and Wells, M. T.(1988). A test of goodness-of-fit based on extreme spacings with some efficiency comparisons. Metrika,35, 223-232.

Rao, J. S. and Sethuraman, J.(1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors, Ann. Statist., 3, 299-313.

Renyi, A.(1962). Three new proofs and a generalization of a theorem of Irving Weiss, Publ. Math. Inst. Hungar. Acad. Sci., 7, 203-214.

Serfling, R. J.(1980). Approximation Theorems of Mathematical Statistics. John Wiley, New York.

Sethuraman, J. and Rao, J. S.(1970). Pitman efficiencies of tests based on spacings. Nonparametric Techniques in Statistical Inference (ed. M. L. Puri), Cambridge.

Simonoff, J.S.(1985). An improved goodness-of-fit statistic for sparse multinomials. J. Amer. Statist. Assoc., 80, 671-677.

Sobel, M. and Uppuluri, V. P. R.(1974). Sparse and crowded cells and Dirichlet distributions, Ann. Statist., 12, 977-987.

Sobel, M., Uppuluri, V. P. R., and Frankowski, K.(1977). Selected Tables in Mathematical Statistics. Amer. Math. Soc., Providence, RI

Wells, M.T.(1987). Contributions to the Theory of Goodness-of-fit Testing. Unpublished Ph.D. dissertation, University of California, Santa Barbara.

Received February 1988; Revised December 1988.

Recommended by John E. Boyer Jr., Kansas State University, Manhattan, KS.

Refereed by Richard A. Lockhart, Simon Fraser University, CANADA and Anonymously.