The Canadian Journal of Statistics
Vol. 17, No. 4, 1989, Pages 419–426
La Revue Canadienne de Statistique

On the asymptotic distributions of high-order spacings statistics

P. GUTTORP and R.A. LOCKHART

University of Washington and Simon Fraser University

Key words and phrases: Greenwood's statistic, Moran's statistic, goodness of fit.
AMS 1985 subject classifications: Primary 62G10, secondary 62G20.

ABSTRACT

Goodness-of-fit tests for the uniform distribution based on sums of smooth functions of m-spacings are studied. A limiting sum-of-weighted-chi-squareds approximation is shown to be accurate uniformly in m for the special cases of analogues of Greenwood's statistic and Moran's statistic. Asymptotic critical points are provided; theory and Monte Carlo studies show they are accurate for all m provided n is moderately large.

RÉSUMÉ

On étudie des tests d'ajustement pour la loi uniforme basés sur des sommes de fonctions lisses de m-espacements. Dans les cas particuliers de statistiques analogues à celles de Greenwood et de Moran, il est démontré qu'une approximation asymptotique par une somme pondérée de khi-deux est précise uniformément en m. Des valeurs critiques asymptotiques sont fournies. La théorie et des études Monte-Carlo indiquent que ces valeurs sont précises pour chaque m si n est modérément grand.

1. INTRODUCTION

Suppose $U_1 \leq \cdots \leq U_n$ are the order statistics from a sample of size n from a distribution F. The m-spacings $D_0, \ldots, D_n$ are defined by $D_i = U_{i+m} - U_i$, where we set $U_0 = 0$ and $U_{n+k+1} = 1 + U_k$. Tests of the hypothesis that F is the uniform distribution on the unit interval or the circle of unit circumference have been based on statistics of the form

$$C = \sum_0^n \phi\left(\frac{(n+1)D_i}{m}\right).$$

Here and throughout we suppress the dependence of most quantities on n and m if no confusion can result; where necessary we use notation such as $C(m)$.

Examples with $m = 1$ include Greenwood's statistic (Greenwood 1946), where $\phi(x) = x^2$, and Moran's statistic (Moran 1947), where $\phi(x) = \log x$. Cressie (1976, 1979) and Del Pino (1979) suggested using higher values of m. For fixed m Holst (1979) has obtained limiting asymptotic normality for the statistics under very weak conditions. Cressie (1976) obtained asymptotic normality when m grows with n subject to the restriction that m^3/n tends to 0; he speculated that asymptotic normality would fail for m growing too fast. Recently, Hall (1986) has shown that C is asymptotically normally distributed provided that $m = o(n)$ and that m^r/n is bounded away from 0 for an r which depends on the smoothness of ϕ near 1. When $m/n \longrightarrow c \in (0, 1)$, Hall shows the limiting distribution is that of the integral of the square of a Gaussian process.

The statistic C uses "wrapped-around" spacings D_i where $i+m$ exceeds $n+1$. This may be avoided by using

$$L = \sum_{0}^{n+1-m} \phi\left(\frac{(n+1)D_i}{m}\right).$$

The statistic L is discussed by Cressie (1976, 1979) and Hall (1986). We do not support the use of L, since it assigns more weight to central spacings than to those in the tails. Hall (1986) shows that the power of L deteriorates with increasing m when m exceeds $n^{\frac{1}{2}}$. Power is considered more fully in Guttorp and Lockhart (1988) and in a technical report available from the authors.

In this paper we concentrate on two special cases. In the first case $\phi(x) = (x-1)^2$ and the statistic becomes

$$G = (n+1)^2 \sum \frac{D_i^2}{m^2} - n - 1.$$

Cressie (1979) has shown that for fixed m, among statistics symmetric in the m-spacings, G uses the most powerful choice of function ϕ against local alternatives converging to the null hypothesis at a suitable rate. In the second case $\phi(x) = \log x$ and the statistic is

$$M = \sum \log \frac{(n+1)D_i}{m}.$$

When m/n tends to $c > 0$, G and M have asymptotically the distribution of a sum of weighted chi-squareds. We obtain explicit expressions for the weights and show that by centering and standardizing properly the distinction between $c > 0$ and $c = 0$ disappears. Our work thus completes that of Cressie and that of Hall on the null distribution of G and M. The asymptotic distribution theory for L is essentially the same as for C except that we do not have closed-form solutions for the asymptotic weights when m/n tends to $c \in (0,1)$.

In Section 2 we obtain the asymptotic distribution of G uniformly in $m \leq n$ with no other assumptions on the rate of growth of m. We indicate how the results of Hall and Cressie extend our results for G to the more general statistic C. In particular we show how our results apply to M.

In Section 3 we tabulate critical points of the sum-of-weighted-chi-squareds approximation. The theory of Section 2 shows the critical points should be useful for all m provided n is sufficiently large; a Monte Carlo study confirms this conclusion for an n of 40 or more.

Our analysis is based on the following standard construction of uniform order statistics. Suppose $V_i + 1$ for $i = 1, \dots, n$ are independent random variables each with the standard exponential distribution, $F(x) = 1 - e^{-x}$. We may construct uniform order statistics as

$$U_i = \frac{V_1 + \cdots + V_i + i}{V_1 + \cdots + V_{n+1} + n + 1}.$$

Let $\bar{V} = \sum V_i/(n+1)$ and $\bar{V}_i = \sum_{i+1}^{i+m} V_j/m$, where we set $V_{n+1+j} = V_j$. Then

$$(n+1)\frac{D_i}{m} = \frac{1+\bar{V}_i}{1+\bar{V}}.$$

2. LARGE-SAMPLE THEORY

2.1. The Quadratic Statistic, G.

The statistic G is equivalent to Greenwood's statistic $\sum D_i^2$. We may write $G = N/(1+\bar{V})^2$, where $N = \sum(\bar{V}_i - \bar{V})^2$. Since G and $\bar{V}$ are independent, by Basu's theorem we have $\mathcal{E}(G^k) = \mathcal{E}(N^k)/\mathcal{E}\{(1+\bar{V})^{2k}\}$. Define

$$\begin{aligned}
a_n &= \frac{(n+1)^4}{(n+2)^2(n+3)(n+4)}, \\
b_n &= \frac{(n+1)^6}{(n+2)^3(n+3)(n+4)(n+5)(n+6)}, \\
c_n &= \frac{(n+1)^6}{(n+2)(n+3)(n+4)(n+5)(n+6)}, \\
p(m,1) &= 160m^4 - 72m^3 - 168m^2 + 48m + 32, \\
p(m,2) &= 40m^4 + 232m^3 + 8m^2 - 168m - 32, \\
p(m,3) &= 22m^3 + 98m^2 + 72m + 8.
\end{aligned}$$

Lengthy calculations establish that the first three moments of G are

$$\mu = \mathcal{E}(G) = \frac{(n+1)(n+1-m)}{m(n+2)}, \tag{2.1}$$

$$\sigma^2 = \mathcal{V}ar(G) = \frac{\{n(4m+2) - 6m^2 + 2m + 4\}(m+1)a_n}{3m^3}, \tag{2.2}$$

and, for $3m \leq n+3$,

$$\mu_3 = \frac{\{n^2p(m,3) - np(m,2) + p(m,1)\}(m+1)b_n}{5m^5}; \tag{2.3}$$

for $3m > n+3$ and $2m \leq n+2$, the term

$$\frac{2(3m-n-3)(3m-n-2)(3m-n-1)(3m-n)(3m-n+1)c_n}{15m^6} \tag{2.4}$$

is added to the value given in (2.3).

The asymptotic distribution theory for G is summarized in the following theorem.

THEOREM 1. *Let $m = m(n)$ be a sequence of integers in the range $1 \leq m \leq n$.*

(a) For all m we have $m^2G(m) = (n+1-m)^2G(n+1-m)$.
(b) If $m = o(n)$, then $(G-\mu)/\sigma$ has asymptotically a standard normal distribution.
(c) If $m/n \longrightarrow c \in (0, \frac{1}{2}]$, then G has asymptotically the distribution of

$$\sum_1^\infty \lambda_k(c)\omega_k,$$

where the ω_k are independent standard exponential variates and

$$\lambda_k(c) = \{1 - \cos(2\pi kc)\}/(\pi kc)^2.$$

The apparent distinction between (b) and (c) disappears if the limit in (c) is centered and standardized. Let $F_{m,n}$ be the cumulative distribution function of $(G - \mu)/\sigma$. Let H_c be the cumulative distribution function of $\{\sum \lambda_k(c)(\omega_k - 1)\}/\{\sum \lambda_k^2(c)\}^{\frac{1}{2}}$, and let $c = c(m,n) = m/(n+1)$. By considering a subsequence of any potential counterexample sequence along which $m/(n+1)$ converges we may prove the following corollary.

COROLLARY. *As $n \longrightarrow \infty$, $\sup\{|F_{m,n}(x) - H_c(x)|\} \longrightarrow 0$ where the supremum extends over all x and all m in the range $1 \leq m \leq (n+1)/2$.*

Proof of the theorem. Statement (a) is an elementary algebraic manipulation. Statement (b) is in Hall (1986). Hall also shows that if $m/n \longrightarrow c \in (0,1)$ then $Z = m^2G/(n+1)^2$ converges in distribution to $\int\{B(t+c) - B(t)\}^2\,dt$, where B is a Brownian bridge and we put $B(t+c) = B(t+c-1)$ for $t+c > 1$. Following Durbin (1973), this integral has the distribution of a sum of weighted chi-squareds. The weights are the eigenvalues of the integral equation

$$\lambda f(t) = \int f(s)\rho(s,t)\,ds, \tag{2.5}$$

where $\rho(s,t) = \mathit{Cov}\{B(s{+}c){-}B(s), B(t{+}c){-}B(t)\}$. Direct calculation with this covariance shows that any eigenfunction of (2.5) is periodic. Differentiating (2.5) twice, we see that

$$\lambda f''(s) = f(s+c) + f(s-c) - 2f(s).$$

Expanding f in a Fourier series in the family $\sin(2\pi ks)$, $\cos(2\pi ks)$, we find that the nonzero eigenvalues of (2.5) are $2\{1 - \cos(2\pi kc)\}/(2\pi k)^2$, each occurring with multiplicity 2. Since $G = Z/c^2$, statement (c) follows after some algebraic manipulation.

2.2. General Statistics, C.

Cressie (1976) suggests that the statistic $M = \sum \log\{(n+1)D_i/m\}$ will be sensitive to the presence of clusters in the data. Cheng and Stephens (1987) have shown that for alternative densities with sharp peaks which tend to produce such clusters this statistic is indeed more sensitive than G. This motivates consideration of the more general statistic C.

For general ϕ we present a slight variation of Theorem 2 of Hall (1986) specialized to the case of the null hypothesis. If $m(n)/\log n \longrightarrow \infty$, then $\sup\{|(n+1)D_i/m - 1|;\ 0 \leq i \leq n\}$ tends to 0 in probability. Fix an integer $r \geq 2$, and assume m^{r-1}/n is bounded away from 0. If ϕ admits a Taylor expansion at 1 of the form $\phi(x) = \sum_1^r \phi^{(j)}(1)(x-1)^j + o(|x-1|^r)$, then we may write

$$C = \sum_2^r \phi^{(j)}(1)S_j/j! + \sum \left(\frac{(n+1)D_i}{m-1}\right)^r \epsilon_i,$$

where $S_j = \sum\{(n+1)D_i/m - 1\}^j$ and $\max\{|\epsilon_i|;\ 1 \leq i \leq n\} \longrightarrow 0$ in probability. Let $\mu_j = \mathcal{E}(S_j)$. The proof given by Hall then shows that

$$\frac{1}{\sigma}\left(C - \sum_2^r \frac{\mu_j\phi^{(j)}(1)}{j!} - \phi''(1)(G - \mu)\right) \longrightarrow 0$$

in probability, where μ and σ are the mean and standard deviation of G.

Thus, provided that ϕ has r derivatives at 1, that $\phi''(1) \neq 0$ and that m^{r-1}/n is bounded away from 0, the statistic $\{C - \sum_2^r \mu_j \phi^{(j)}(1)/j!\}/\{\phi''(1)\sigma\}$ has asymptotically the same distribution as $(G - \mu)/\sigma$; the latter distribution is given in the corollary above.

This result gives centering constants which depend not only on m and n but also on r. In principle it seems undesirable to have an approximation to the distribution of a statistic depend not only on the statistic but also on what statistic might have been used for a different sample size. The problem does not arise for G, since the centering constant in the corollary is the exact mean of G. Following a suggestion from a referee, we made some numerical calculations of the effect of using $r = 4$ versus $r = 6$ for the statistic M with $n = 50$, 100, and 200 and m running from 1 to $n/2$. We found that for m larger than 5 to 7 use of $r = 4$ was adequate, while smaller values of m required r to be at least 6. No general formula for a useful value of r as a function of m and n is known to us.

Although we do not have a complete solution for this problem, we are able to deal with the most common special case of C other than G, namely, M, the high-order-spacing version of Moran's statistic. Cressie (1976) gives formulae for the exact mean and variance of $M - (n+1)\log\{(n+1)/m\} = \sum_0^n \log D_i$ and proves that if $m^3/n \longrightarrow 0$, then $\{M - \mathcal{E}(M)\}/\mathcal{V}ar^{\frac{1}{2}}(M)$ is asymptotically standard normal. When m^3/n is bounded away from 0 it is possible to prove, using asymptotic expansions for the digamma function, that $\{\sum_2^4 (-1)^{j-1}\mu_j/j - \mathcal{E}(M)\}/\sigma \longrightarrow 0$. Therefore the corollary holds with $\{M - \mathcal{E}(M)\}/\mathcal{V}ar^{\frac{1}{2}}(M)$ replacing $(G - \mu)/\sigma$.

3. NUMERICAL RESULTS

The results of Section 2 permit tabulation of critical points from the limiting distribution of G. To test the null hypothesis that the underlying distribution is uniform on the unit interval or the circle of unit circumference, compute $(G - \mu)/\sigma$ from (2.1), (2.2), and compute $c = m/(n+1)$. If $c > \frac{1}{3}$, use this value to enter Table 1. If $c < \frac{1}{3}$, the approximation may be improved by adjusting for the exact skewness of the statistic as follows. Compute the standardized skewness $\gamma_1 = \mu_3/\sigma^3$ from (2.3, 2.4), and use this value to enter Table 1. If $m/(n+1) > \frac{1}{2}$, let $m^* = n+1-m$, and compute $G(m^*) = m^2 G(m)/(m^*)^2$, and proceed as above with $G(m^*)$. To interpolate between columns in the table use linear interpolation in the logarithm of the upper-tail probability. To interpolate between rows use linear interpolation in γ_1; this is justified for c near 0 by a Cornish-Fisher expansion of the critical points of T.

The values in Table 1 are approximations to the critical points for the distribution of

$$T = \frac{\sum_1^\infty \lambda_k(c)(\omega_k - 1)}{\{\sum_1^\infty \lambda_k^2(c)\}^{\frac{1}{2}}}.$$

For $c \geq 0.1$ the points in Table 1 are critical points for the truncated sum

$$T^* = \frac{\sum_1^{100} \lambda_k(c)(\omega_k - 1)}{\left\{\sum_1^{100} \lambda_k^2(c)\right\}^{\frac{1}{2}}}.$$

The truncation causes no problems, since

$$\mathcal{V}ar\left\{\sum_{101}^{\infty} \lambda_k(c)(\omega_k - 1)\right\}$$

TABLE 1: Critical values for $(G - \mu)/\sigma$.[a]

		Upper-tail level of significance								
c	γ_1	50	25	15	10	5	2.5	1	0.5	0.1
0.5	1.960	−0.297	0.391	0.898	1.301	1.989	2.677	3.587	4.275	5.873
0.4	1.955	−0.296	0.392	0.898	1.301	1.988	2.675	3.585	4.275	5.869
0.3	1.744	−0.257	0.430	0.918	1.303	1.958	2.611	3.475	4.128	5.645
0.2	1.389	−0.213	0.491	0.963	1.323	1.922	2.507	3.269	3.840	5.156
0.1	0.944	−0.152	0.560	1.005	1.333	1.861	2.360	2.992	3.455	4.495
0.02	0.408	−0.068	0.632	1.033	1.316	1.752	2.146	2.623	2.960	3.688
0.01	0.287	−0.048	0.646	1.036	1.308	1.722	2.092	2.536	2.847	3.508
0.001	0.090	−0.015	0.666	1.037	1.291	1.670	2.002	2.393	2.661	3.220
0	0	0.000	0.675	1.036	1.282	1.645	1.960	2.326	2.576	3.090

[a]See the text for interpolation techniques.

is very small in this case. The distribution function of T^* is then available in closed form; see Johnson and Kotz (1970, p. 222). For c <0.1 the eigenvalues $\lambda_k(c)$ fall off too slowly with increasing k to neglect the truncated terms. Instead we truncated the sum at 40 terms and replaced $\sum_{41}^{\infty} \lambda_k(c)(\omega_k - 1)$ by $a(\chi^2_\nu - \nu)$, where a and ν are chosen to match the variance and skewness of the truncated terms. Such a choice is feasible because the cumulants of T can be computed analytically. Indeed, the rth cumulant, κ_r, of T is $\kappa_r = (r-1)! \sum_1^{\infty} \lambda_k^r(c)$. The latter sums may be evaluated using trigonometric identities and the relation, for $0 \le x \le 1$,

$$\sum_{k=1}^{\infty} \cos(2\pi k x)\, k^{-2n} = \frac{(-1)^{n-1}(2\pi)^{2n} B_{2n}(x)}{2(2n)!},$$

where $B_k(x)$ is the kth Bernoulli polynomial; see Abramowitz and Stegun (1965, p. 804 ff.) for details on these polynomials. In particular, $\kappa_1 = (1-c)/c$, $\kappa_2 = (4/3 - 2c)/c$, and for $c \le \frac{1}{3}$, $\kappa_3 = (22/5 - 8c)/c$, while for $c > \frac{1}{3}$, $\kappa_3 = (\frac{22}{5} - 8c)/c + 2(3c-1)^5/(15c^6)$.

For $c < 0.1$ the points in Table 1 were then computed, following Durbin and Knott (1973), by numerical inversion of the characteristic function of

$$T^* = \frac{\sum_1^{40} \lambda_k(c)(\omega_k - 1) + a(\chi^2_\nu - \nu)}{\left\{\sum_1^{40} \lambda_k^2(c) + 2a^2\nu\right\}^{\frac{1}{2}}}.$$

The two approximations were compared at $c = 0.1$ and agree to all the decimal places given in the table.

The quality of the asymptotic approximation suggested here was studied in a Monte Carlo experiment. For sample sizes $n = 40$ and 100 and spacing orders $m = 1, 5, 20$, and (for $n = 100$ only) $m = 40$ we generated 10,000 Monte Carlo values of G. For various significance levels the number of rejections using critical points derived from an extended version of Table 1 is recorded in Table 2. We also evaluated the normal approximation to the distribution of $(G - \mu)/\sigma$. Except for $m = 1$ this approximation is very poor; for $m = 1$ Table 2 also presents the number of Monte Carlo samples rejected using this normal approximation. The poor performance of the normal approximation

TABLE 2: Number of rejected Monte Carlo samples in 10,000 trials.

	Normal approx.	Sum-of-weighted-chi-squareds approximation			
Upper-tail area	$m = 1$	$m = 1$	5	20	40
		$n = 40$			
0.99	10000	9707	9767	9999	
0.95	9912	9358	9368	9743	
0.75	7338	7593	7496	7382	
0.50	4240	5095	5073	4831	
0.25	2048	2472	2486	2524	
0.05	644	470	488	494	
0.01	294	98	105	95	
		$n = 100$			
0.99	9998	9812	9847	9868	9943
0.95	9787	9408	9459	9435	9604
0.75	7348	7545	7475	7546	7492
0.50	4317	5006	5012	5077	4987
0.25	2154	2434	2469	2535	2540
0.05	618	460	475	506	505
0.01	255	110	97	100	107

conforms with the findings of Stephens (1981), who gives accurate finite-sample points for the case $m = 1$.

The results show that the asymptotic sum-of-weighted-chi-squareds approximation is good for n of 40 or more and for all m. The normal approximation does not appear to work as well even in the case $m = 1$. The approximations are best for points near the 5% level and worst in the lower tail.

In Section 2.2 we indicated that for sufficiently smooth ϕ the statistic C can be referred to Table 1 provided exact expressions for the mean and variance are available. In particular, Cressie (1976) gives the mean and variance of M. Since the exact skewness of the statistic is not available, the value $c = m/(n+1)$ must be used to enter the table; as a result the asymptotic approximation cannot be expected to be as accurate.

ACKNOWLEDGMENT

This work was completed in part while the first author was visiting at the University of British Columbia and Simon Fraser University. The authors gratefully acknowledge financial support from the National Sciences Foundation, the SIAM Institute for Mathematics in Society, and the Natural Sciences and Engineering Research Council of Canada. We thank Professor M.A. Stephens for many useful conversations, and anonymous referees whose suggestions resulted in substantial improvements to the present work.

REFERENCES

Abramowitz, M., and Stegun, I.M., *eds.* (1965). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (National Bureau of Standards). Dover, New York.

Cheng, R.C.H., and Stephens, M.A. (1987). A comparison of symmetric spacings tests and the Cramer–von Mises statistic when the alternative is not smooth. Technical Report, Department of Mathematics and Statistics, Simon Fraser University.

Cressie, N. (1976). On the logarithms of high order spacings. *Biometrika*, 63, 343–355.

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). *Distribution Theory for Tests Based on the Sample Distribution Function*. Regional Conf. Ser. in Appl. Math., 9. SIAM, Philadelphia.

Durbin, J., and Knott, M. (1973). Components of Cramer–Von Mises statistics, I. *J. Roy. Statist. Soc. Ser. B*, 34, 298–307.

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

Guttorp, P., and Lockhart, R. (1988). On the asymptotic distribution of quadratic forms in uniform order statistics. *Ann. Statist.*, 16, 433–449.

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

Holst, L. (1979). Asymptotic normality of sum-functions of spacings. *Ann. Probab.*, 7, 1066–1072.

Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika*, 48, 419–426.

Johnson, N.L., and Kotz, S. (1970). *Distributions in Statistics. Volume 1: Continuous Univariate Distributions*. Houghton Mifflin, Boston.

Moran, P.A.P. (1947). The random division of an interval—part I. *J. Roy. Statist. Soc. Ser. B*, 9, 92–98.

Stephens, M.A. (1981). Further percentage points for Greenwood's statistic. *J. Roy. Statist. Soc. Ser. A*, 144, 364–366.

Received 26 October 1987
Revised 4 July 1989
Accepted 15 August 1989

Department of Statistics, GN-22
University of Washington
Seattle, Washington U.S.A.

Department of Mathematics and Statistics
Simon Fraser University
Burnaby, B.C. V5A 1S6