Approximating the distribution of Greenwood's statistic

R.J.M.M. Does
Department of Medical Informatics and Statistics, University of Limburg, P.O. Box 616, 6200 MD Maastricht, The Netherlands

R. Helmers
Centre for Mathematics and Computer Science P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

C.A.J. Klaassen
Department of Mathematics, University of Leiden, P.O. Box 9512, 2300 RA Leiden, The Netherlands

Statistics based on uniform spacings are often used in goodness-of-fit problems. In this paper special attention is paid to the distribution of Greenwood's statistic. Although its asymptotic distribution is normal, the normal approximation is extremely bad, even for large sample sizes. It is shown that the Edgeworth expansion yields a considerably better approximation for the distribution of this statistic. Furthermore, an overview is given of the higher order asymptotics for the sum of functions of uniform spacings, of which Greenwood's statistic is a special case.

Key Words & Phrases: goodness-of-fit problems, uniform spacings, Edgeworth expansions, Monte Carlo study.

1. Introduction

On the basis of past experience or some specific characteristics of an experiment one often hypothesizes that the observations come from a specified distribution. When one tries to ascertain if the observations contradict this hypothesis, one is dealing with a goodness-of-fit testing problem. Let $X_1, X_2, ..., X_n$ be a random sample from a continuous distribution F_0. Then the probability integral transformation, $U_j = F_0(X_j)$, will transform $X_1, X_2, ..., X_n$ into $U_1, U_2, ..., U_n$ where the $U_j, j = 1,2,...,n$, are independent uniform $(0,1)$ random variables. Thus testing the null-hypothesis $H_0: X_1, X_2, ..., X_n$ i.i.d. F_0 can be reduced to testing the simple hypothesis $H_0: U_1, U_2, ..., U_n$ are independent uniform $(0,1)$ random variables. The resulting problem op testing uniformity and the related distribution theory have received considerable attention in the literature. A particular application of tests for uniformity is in checking the performance of random number generators which are used in simulation studies.

In this paper we consider a test statistic which is based on spacings. Under

the null-hypothesis these spacings are in fact uniform spacings. These are defined as the differences between consecutive observations from a uniform distribution: i.e. let $U_1, U_2, \ldots$ be a sequence of independent uniform (0,1) random variables, let $U_{1:n} \leqslant U_{2:n} \leqslant \cdots \leqslant U_{n:n}$ denote the ordered $U_1, \ldots, U_n$ and let $U_{0:n} = 0$ and $U_{n+1:n} = 1$, then the uniform spacings are defined by

$$D_{jn} = U_{j:n} - U_{j-1:n} \quad , \quad j = 1,2,\ldots,n+1. \tag{1}$$

Although discussed earlier by WHITWORTH (1887) and others the development of goodness-of-fit tests based on uniform spacings received its principal impetus from GREENWOOD (1946). To test the uniformity of the sample he proposed

$$G_n = \sum_{j=1}^{n+1} D_{jn}^2, \tag{2}$$

as a test statistic. In the literature there are many other statistics which have been suggested to provide tests based on a function of uniform spacings (cf. the last section of this paper). However the Greenwood statistic yields the locally most powerful one among such tests, against linear non-uniform alternatives (cf. WEISS, 1956).

This paper is organized as follows. In Section 2 we deal with the distribution of the Greenwood statistic. In this section we also present the results of a numerical investigation. Section 3 discusses higher order asymptotics for statistics which are the sum of functions of uniform spacings.

2. THE GREENWOOD STATISTIC

In Section 1 we introduced the Greenwood statistic as the sum of the squared lengths of the $n+1$ uniform spacings obtained when a unit interval is divided by n points at random (cf. (1) and (2)). However, the exact distribution of G_n under the null-hypothesis is not known in a manageable form for $n > 3$. For $n = 1$ and 2 GREENWOOD (1946) gives exact expressions for the corresponding distribution functions. GARDNER (1952) presents the distribution function of (2) for $n = 3$. Most recently KUMGANBAYEV and VOINOV (1986) found a method to obtain - at least in principle - the exact distribution function of the Greenwood statistic for any value of n. However it appears that the method is not easy to apply.

In a situation where exact results are less tractable or not available, it is natural to derive asymptotic distributions. Hopefully these will give accurate approximations which can be used in establishing approximate critical regions for testing purposes. In 1947 Moran proved the asymptotic normality of G_n (cf. (2)). He also indicated that the tendency to normality is extremely slow. In Table 1 the values of skewness β_{1n} and kurtosis β_{2n} for some selected values of n are given.

TABLE I. The skewness β_{1n}, and kurtosis β_{2n} of Greenwood's statistic for some values of n.

n	β_{1n}	β_{2n}
5	1.587	6.827
10	1.706	8.351
20	1.584	8.201
30	1.437	7.493
50	1.218	6.378
70	1.073	5.673
100	0.926	5.026
150	0.775	4.439
250	0.613	3.909
500	0.440	3.473
1000	0.314	3.241
∞	0	3

Note that for a normal distribution the skewness vanishes and the kurtosis equals 3.

Better approximations for the distribution function of G_n may be obtained by using Edgeworth expansions (cf. (4)). In contrast to the normal approximation in which only the mean and variance play a role, the Edgeworth expansion involves the first four cumulants (moments). The third cumulant (the skewness) and fourth cumulant (the kurtosis minus 3) of G_n deviate from zero significantly as indicated in Table 1. This explains perhaps why the normal approximation is inadequate.

In DOES and HELMERS (1982) and DOES, HELMERS and KLAASSEN (1987) Edgeworth expansions with approximate cumulants for the sum of a function of uniform spacings with remainder $o(n^{-1})$ are established. For the special case of a quadratic function we can replace these approximate cumulants by their exact counterparts given in MORAN (1947: see also the corrigendum (1981) correcting an error in the formula for the third cumulant). In this way we arrive at an Edgeworth expansion of the distribution function F_n^* of the exactly standardized Greenwood statistic (cf. (2))

$$G_n^* = n^{-1/2}(n+2)(n+3)^{1/2}(n+4)^{1/2}G_n/2 - n^{-1/2}(n+3)^{1/2}(n+4)^{1/2},$$

which is given by

$$F_n^*(x) = \tilde{F}_n(x) + o(n^{-1}) \ , \ \text{as} \ n \to \infty, \tag{3}$$

where (Φ denotes the standard normal distribution function and ϕ its density),

$$\tilde{F}_n(x) = \Phi(x) - \phi(x)\{\beta_{1n}(x^2-1)/6 + (\beta_{2n}-3)(x^3-3x)/24 + \beta_{1n}^2(x^5-10x^3+15x)/72\}, \tag{4}$$

with (cf. MORAN (1974, 1981))

$$\beta_{1n} = \frac{(10n-4)(n+3)^{1/2}(n+4)^{1/2}}{n^{1/2}(n+5)(n+6)},$$

and

$$\beta_{2n} = \frac{(3n^3+303n^2+42n-24)(n+3)(n+4)}{n(n+5)(n+6)(n+7)(n+8)}.$$

The exact distribution function F_n^* is estimated by a Monte Carlo simulation based on 40,000 samples for $n = 10, 20, 30, 50, 70, 100, 150, 250$ and $x = -3.0(0.1)\ 3.0$. In Table II the Edgeworth expansion $\tilde{F}_n$, the normal approximation and the (estimated) exact distribution F_n^* are given for the above mentioned sample sizes and various values of the argument.

TABLE II. Comparison of the (estimated) exact distribution function with the Edgeworth expansion and normal approximation.

x	-2.0	-1.5	-1.0	-0.5	0.0	0.5	1.0	1.5	2.0	2.5	3.0
F_{10}^*	.000	.004	.101	.349	.599	.769	.868	.924	.956	.974	.985
$\tilde{F}_{10}$	.000	.000	.109	.365	.613	.785	.890	.939	.946	.951	.968
F_{20}^*	.000	.010	.116	.345	.584	.759	.865	.926	.957	.975	.986
$\tilde{F}_{20}$	.000	.000	.104	.350	.605	.789	.896	.939	.945	.952	.969
F_{30}^*	.000	.016	.124	.341	.575	.753	.863	.925	.959	.978	.987
$\tilde{F}_{30}$	.000	.000	.110	.344	.596	.782	.890	.935	.946	.956	.972
F_{50}^*	.001	.025	.133	.338	.567	.745	.859	.922	.958	.976	.987
$\tilde{F}_{50}$	.000	.004	.120	.340	.581	.768	.880	.931	.949	.963	.979
F_{70}^*	.002	.029	.136	.339	.560	.741	.856	.923	.961	.980	.989
$\tilde{F}_{70}$	.000	.014	.128	.337	.571	.757	.872	.928	.952	.967	.982
F_{100}^*	.004	.034	.144	.338	.557	.736	.854	.925	.964	.982	.991
$\tilde{F}_{100}$	.000	.024	.135	.335	.562	.747	.865	.926	.955	.972	.985
F_{150}^*	.005	.039	.145	.329	.545	.724	.850	.922	.962	.982	.991
$\tilde{F}_{150}$	.000	.033	.142	.332	.552	.736	.858	.925	.958	.976	.989
F_{250}^*	.008	.047	.150	.330	.541	.724	.849	.925	.965	.984	.993
$\tilde{F}_{250}$	.005	.042	.148	.329	.541	.725	.852	.925	.962	.981	.992
Φ	.023	.067	.159	.309	.500	.692	.841	.933	.977	.994	.999

In Table III the accuracy of the two methods of approximation is illustrated. Let (cf. (3) and (4))

$$d(F_n^*, \tilde{F}_n) = \max_{-3 \leq x \leq 3} |F_n^*(x) - \tilde{F}_n(x)|$$

and

$$d(F_n^*, \Phi) = \max_{-3 \leq x \leq 3} |F_n^*(x) - \Phi(x)|$$

denote the maximum error in the region $|x| \leq 3$ when F_n^* is approximated by $\tilde{F}_n$ and by Φ, respectively.

TABLE III. Estimated maximum error in the region $-3 \leqslant x \leqslant 3$ when the (estimated) exact distribution is approximated by the Edgeworth expansion and normal distribution, respectively.

n	$d(F_n^*, \tilde{F}_n)$	$d(F_n^*, \Phi)$
10	.023	.101
20	.033	.087
30	.030	.076
50	.024	.068
70	.018	.061
100	.012	.058
150	.013	.046
250	.004	.042

The agreement between F_n^* and $\tilde{F}_n$ is reasonable. Already for $n = 10$-30 the accuracy of the Edgeworth expansion is within .035 of the (estimated) actual value. On the other hand the normal approximation is bad even for $n = 250$: in this case the maximum error in the region $-3.0 \leqslant x \leqslant 3.0$ is greater than .04. Hence for $n = 10$-30 the Edgeworth expansion performs already better than the normal approximation for $n = 250$!

EASTON and RONCHETTI (1986) have suggested a way of using Edgeworth expansions for the density of an arbitrary statistic to obtain a so-called saddlepoint approximation for it (cf. DANIELS (1987)). However this technique is beyond the scope of this paper.

Instead of considering a distribution function one might be interested in its inverse function, i.e. its percentage points. In recent years there has been a revival of interest in the percentage points of the Greenwood statistic. BURROWS (1979), using recursion and numerical integration, produced a table of *exact* percentage points of G_n for $n = 2$ (1) 10. CURRIE (1981) extended the tabulation of Burrows (1979) up to sample size 20. HILL (1979; see also his corrigendum (1981)) fitted Johnson curves and lognormal curves. STEPHENS (1981) approximated percentage points for the Greenwood statistic for various sample sizes n and values of the level α by fitting Pearson curves to the first four moments. A comparison of Pearson curve points and those given by Johnson curves or by lognormal curves fitted to the distribution has been given by CURRIE (1981), for n up to 20. It shows that all three methods give very similar results for $n = 20$. In DOES, HELMERS and KLAASSEN (1984) the exact percentage points of Greenwood's statistic are approximated by Cornish-Fisher expansions. Their numerical results indicate that the Cornish-Fisher approximation behaves quite satisfactory for sample sizes $n \geqslant 12$ although its performance is inferior to that of Pearson curves approximation (cf. STEPHENS (1981)). It should be noted that the approximate percentage points based on Cornish-Fisher expansions can be computed for any value of α and n. On the other hand to apply the Pearson curves method of STEPHENS (1981) one has either to rely on interpolation for values α and n different from those occuring

in his paper or to extend his table.

3. Higher order asymptotics for the sum of functions of uniform spacings

Let $g:[0,\infty)\to\mathbb{R}$, be a measurable nonlinear function and define statistics T_n by

$$T_n = \sum_{j=1}^{n+1} g((n+1)D_{jn}),\ n\geqslant 1. \tag{5}$$

Statistics of this form can be used for testing uniformity. For the special case with $g(x)=x^2, T_n$ reduces to $(n+1)^2G_n$ (cf. (2)). However, there are several other statistical applications of uniform spacings. The first study of uniform spacings was concerned with the randomness of a series of events and was motivated by the fact that the intervals between successive events of a Poisson process, conditioned on the number of events in a specified interval, are distributed like uniform spacings (cf. Whitworth (1887) and also the paper of Steutel (1967) in this journal). Furthermore, functions of uniform spacings are used in time series analysis for testing for null correlation, studies of empirical processes, density and regression estimation and coverage problems in which one considers typically intervals of fixed length centered at the U_j's and where one is interested in the sample size required to cover $(0,1)$ k times. A more or less complete account of the statistical applications of uniform spacings can be found in the review papers of Pyke (1965, 1972), and Deheuvels (1985).

In the literature there are many statistics of the form (5) which have been proposed to provide tests based on uniform spacings. A few examples are: $g(x)=x^2$, suggested by Greenwood (1946), $g(x)=|x-1|$ suggested by M.G. Kendall in the discussion of Greenwood (1946), $g(x)=x^r, r>0, r\neq 1$ proposed by Kimball (1950), $g(x)=\log x$ and $g(x)=x^{-1}$ both suggested by Darling (1953). In a paper of Sethuraman and Rao (1970) a unified treatment is presented of the computation of (asymptotic relative) Pitman efficiencies of tests based on the sum of a function of uniform spacings. Though Chibisov (1961) shows that the efficiency of any test symmetric in the spacings is zero relative to the Kolmogorov-Smirnov test, it is still useful to know about the efficiency of one symmetric spacings test relative to another. It must be pointed out that the negative result about the use of spacings concerns only a very specific type of local alternatives for which tests based on spacings are inefficient (cf. Deheuvels (1985)). Sethuraman and Rao (1970) show that among a large class of symmetric tests of the form (5) the test as proposed by Greenwood (1946) has maximum efficacy.

The first general attempt to derive limit theorems for statistics of the form (5) was made by Darling (1953). He derived a formula for the characteristic function of T_n, from which he was able to obtain the limit distribution for several special functions g. Le Cam (1958) presented a powerful technique for proving first order limit theorems using a well-known characterization for

uniform spacings: Note that interarrival times $Y_j, j=1,2,...,$ of a Poisson process are exponentially distributed and that given the $(n+1)$-th point $Q_{n+1}(=\Sigma_{j=1}^{n+1} Y_j)$ in this Poisson process, the first n points are distributed like the order statistics of n iid observations from a uniform distribution on the interval $(0, Q_{n+1})$. Consequently, we have

$$\mathcal{L}(D_{1n}, D_{2n}, ..., D_{n+1,n}) = \mathcal{L}(\frac{Y_1}{Q_{n+1}}, \frac{Y_2}{Q_{n+1}}, \ldots, \frac{Y_{n+1}}{Q_{n+1}})$$

and hence

$$\mathcal{L}((n+1)D_{1n}, (n+1)D_{2n}, ..., (n+1)D_{n+1,n}) =$$
$$\mathcal{L}(Y_1, Y_2, ..., Y_{n+1} | Q_{n+1} = n+1).$$

From this it it follows that

$$\mathcal{L}(T_n) = \mathcal{L}(W_n | S_n = 0), \tag{6}$$

where

$$W_n = \sum_{j=1}^{n+1} g(Y_j), \text{ and } S_n = \sum_{j=1}^{n+1} (Y_j - 1),$$

i.e., T_n has the same distribution as a sum of independent random variables given another sum of independent random variables. If g would be linear then T_n would be degenerate. A survey of the general area of first order limit theorems may be found in the just mentioned review papers of Pyke and Deheuvels.

In this journal two review papers about higher order asymptotics have been published namely Albers (1975) and Does (1984). For an introduction to this subject the reader is referred to these papers. The first question that is discussed in second order asymptotics is that of the rate of convergence to the (normal) limit. According to Pyke (1972) such a study is of interest. With the aid of the characterization (6), Does and Klaassen (1984a, 1984b) proved Berry-Esseen bounds of the order $n^{-1/2}$ for the normal approximation for statistics based on uniform spacings under natural moment assumptions.

The next step is to go beyond Berry-Esseen bounds and investigate higher order terms of the distribution function of the statistic T_n defined in (5). Such results are called Edgeworth expansions and are of interest for several reasons (cf. Bickel (1974)).

In Does and Helmers (1982) Edgeworth expansions were established for statistics of the form (5) under a natural moment assumption and an integrability condition on the simultaneous characteristic function of $(Y-1, g(Y))$ (cf. (6)) where Y is an exponential random variable with expectation 1. In Does, Helmers and Klaassen (1987) it is shown that the latter integrability condition can be replaced by a much weaker and more natural Cramér-type condition (cf. Feller (1971), Chapter XVI). It is shown in the latter paper that the Cramér-type condition holds under an easily verifiable and mild assumption on the function g : if $(c,d) \subset (0,\infty)$ is an open interval on which g is almost

everywhere differentiable with derivative g' such that g' is not essentially constant on (c,d) then the Cramér-type condition holds. Furthermore an indication is given how to generalize the results to functions g_{jn}; i.e. functions depending on the $j-th$ spacing and sample size n. A Berry-Esseen theorem for this more general case was proved in DOES and KLAASSEN (1984b).

ACKNOWLEDGEMENTS

The authors are grateful to the editor and the referees for their comments and suggestions. They also thank Mr. M.P.E. Janssen for carrying out the computations leading to the Tables.

REFERENCES

ALBERS, W. (1975), Efficiency and deficiency considerations in the symmetry problem, *Statistica Neerlandica 29,* 81-92.

BICKEL, P.J. (1974), Edgeworth expansions in nonparametric statistics, *Annals of Statistics 2,* 1-20.

BURROWS, P.M. (1979), Selected percentage points of Greenwood's statistic, *Journal of the Royal Statistical Society Series A 142,* 256-258.

CHIBISOV, D.M. (1961), On the tests of fit based on sample spacings, *Theory of Probability and Its Applications 6,* 325-329.

CURRIE, I.D. (1981), Further percentage points of Greenwood's statistic, *Journal of the Royal Statistical Society Series A 144,* 360-363.

DANIELS, H.E. (1987), Tail probability approximations, *International Statistical Review 55,* 37-48.

DARLING, D.A. (1953), On a class of problems related to the random division of an interval, *Annals of Mathematical Statistics 24,* 239-253.

DEHEUVELS, P. (1985), Spacings and applications, in: F. Konecny, J. Mogyoródi and W. Wertz (eds), *Proceedings of the 4th Pannonian Symposium on Mathematical Statistics,* Reidel, Dordrecht, 1-30.

DOES, R.J.M.M. (1984), The asymptotic behavior of simple linear rank statistics, *Statistica Neerlandica 38,* 109-130.

DOES, R.J.M.M. and R. HELMERS (1982), Edgeworth expansions for functions of uniform spacings, in: B.V. Gnedenko, M.L. Puri and I. Vincze (eds), Colloquia Mathematica Societatis János Bolyai Volume 32: *Nonparametric Statistical Inference,* North-Holland, Amsterdam, 203-212.

DOES, R.J.M.M., R. HELMERS and C.A.J. KLAASSEN (1984), Approximating the percentage points of Greenwood's statistic with Cornish-Fisher expansions, Centre for Mathematics and Computer Science Report MS-R8405, Amsterdam.

DOES, R.J.M.M., R. HELMERS and C.A.J. KLAASSEN (1987), On the Edgeworth expansion for the sum of a function of uniform spacings, *Journal of Statistical Planning and Inference 17,* 149-157.

DOES, R.J.M.M. and C.A.J. KLAASSEN (1984a), The Berry-Esseen theorem for functions of uniform spacings, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 65,* 461-471.

DOES, R.J.M.M. and C.A.J. KLAASSEN (1984b), Second order asymptotics for statistics based on uniform spacings, in: P. Mandl and M. Hušková (eds), *Asymptotic Statistics 2,* North-Holland, Amsterdam, 231-239.

EASTON, G.S. and E. RONCHETTI (1986), General saddlepoint approximations with applications to *L* statistics, *Journal of the American Statistical Association 81,* 420-430.

FELLER, W. (1971), *An Introduction to Probability Theory and Its Applications,* Vol. II, 2nd edition, Wiley, New York.

GARDNER, A. (1952), Greenwood's "Problem of intervals": an exact solution for $n = 3$, *Journal of the Royal Statistical Society Series B 14,* 135-139.

GREENWOOD, M. (1946), The statistical study of infectious diseases, *Journal of the Royal Statistical Society Series A 109,* 85-110.

HILL, I.D. (1979), Approximating the distribution of Greenwood's statistic with Johnson distributions, *Journal of the Royal Statistical Society Series A 142,* 378-380, Corrigendum (1981), *Journal of the Royal Statistical Society Series A 144,* 388.

KIMBALL, B.F. (1950), On the asymptotic distribution of the sum of powers of unit frequency differences, *Annals of Mathematical Statistics 21,* 263-271.

KUMGANBAYEV, M. and V.G. VOINOV (1986), On a derivation of probability density functions. The pdf of Greenwood's statistic, Preprint, submitted to the Proceedings of the Tenth Prague Conference on Information Theory, Prague.

LE CAM, L. (1958), Un théorème sur la division d'un intervalle par des points pris au hasard, *Publications de l'Institut de Statistique de l'Université de Paris 7,* 7-16.

MORAN, P.A.P. (1947), The random division of an interval, *Journal of the Royal Statistical Society Series B 9,* 92-98, Corrigendum (1981), *Journal of the Royal Statistical Society Series A 144,* 388.

PYKE, R. (1965), Spacings, *Journal of the Royal Statistical Society Series B 27,* 395-449.

PYKE, R. (1972), Spacings revisited, *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability 1,* 417-427.

SETHURAMAN, J. and J.S. RAO (1970), Pitman efficiencies of tests based on spacings, in: M.L. Puri (ed.), *Nonparametric Statistical Inference,* University Press, Cambridge, 405-415.

STEPHENS, M.A. (1981), Further percentage points for Greenwood's statistic, *Journal of the Royal Statistical Society Series A 144,* 364-366.

STEUTEL, F.W. (1967), Random division of an interval, *Statistica Neerlandica 21,* 231-244.

WEISS, L. (1956), A certain class of tests of fit, *Annals of Mathematical Statistics 27,* 1165-1170.

WHITWORTH, W.A. (1887), *Choice and Chance,* University Press, Cambridge.

Received January 1988.