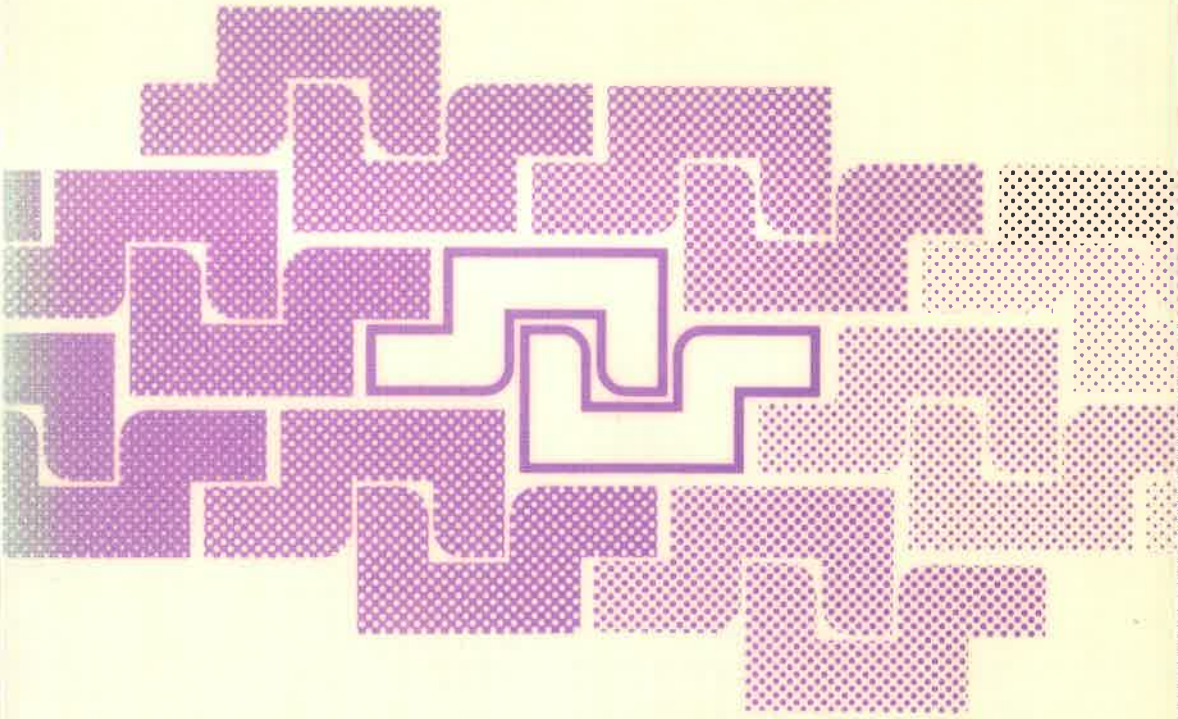

Statistical Extremes and Applications

edited by J. Tiago de Oliveira

NATO ASI Series

Series C: Mathematical and Physical Sciences Vol. 131

ON ORDERED UNIFORM SPACINGS FOR TESTING GOODNESS OF FIT

S. Rao Jammalamadaka

Univ. of California
Santa Barbara
U.S.A.

ABSTRACT. Tests for the goodness of fit problem, based on sample spacings i.e., observed distances between successive order statistics, have been used in the literature. This paper reviews some recent work on tests which make use of ordered spacings, like the largest spacing, sum of the k largest spacings and a test based on counting the number of "small" spacings where "small" is defined so as to optimize the large sample efficiency of the test.

1. INTRODUCTION

Many interesting statistical problems can be reduced to the following simple form: given some independent and identically distributed (i.i.d.) observations on $[0,1]$, test if they are uniformly distributed on the unit interval. Such problems include
(i) testing goodness of fit: given $X_1,\ldots,X_{n-1}$ i.i.d. from some cumulative distribution function (cdf) F, test if this F is a given (continuous) cdf F_0. By making the so called probability integral transformation $U_i=F_0(X_i)$, $i=1,\ldots,(n-1)$ on the data, this reduces to the interval $[0,1]$ and to testing uniformity.
(ii) testing for a Poisson process and/or exponentiality of interarrivals: given the times of occurrences of events in a finite interval, one would want to verify if these were generated by a Poisson process. From the property that a homogeneous Poisson process, suitably scaled, behaves like the uniform distribution, this problem is equivalent to testing uniformity on the unit interval.

J. Tiago de Oliveira (ed.), Statistical Extremes and Applications, 589–596.

(iii) testing for no preferred direction in circular data: A novel area of statistics is where the measurements are directions. Such directions in 2-dimensions can be represented as points on the perimeter of a unit circle and are referred to as the circular data. See for instance J.S. Rao [6]. One of the important questions here is whether the data is uniformly distributed (isotropic) or if indeed, there is a preferred direction. By cutting open the circle at anyone of the observations, this problem reduces to testing uniformity on $[0,1]$.

Among the several possible approaches to testing uniformity which include empirical distribution function methods and χ^2 methods, we would like to focus on those which utilize the spacings. Let $U_1,\ldots,U_{n-1}$ be $(n-1)$ i.i.d. random variables (r.v.'s) from a continuous cdf $F(u)$ on $[0,1]$. The null hypothesis of interest is

$$H_0: F(u) = u, \quad 0 \leq u \leq 1. \tag{1.1}$$

Define the order statistics

$$0 = U_{0,n} \leq U_{1,n} \leq \cdots \leq U_{n-1,n} \leq U_{n,n} = 1$$

and the spacings

$$Y_{i,n} = U_{i,n} - U_{i-1,n}, \quad i = 1,\ldots,n. \tag{1.2}$$

For notational convenience, we drop the second subscript n on the spacings and refer to them as $\{Y_i,\ i=1,\ldots,n\}$ which add upto 1. It is easy to see that the joint distribution of $(Y_1,\ldots,Y_{n-1})$ is a Dirichlet distribution with constant density $(n-1)!$ over $\{\underline{y}: y_i \geq 0, \sum_1^{n-1} y_i \leq 1\}$. Recall that a collection of r.v.'s $(Y_1,\ldots,Y_b)$ is said to have a Dirichlet distribution with parameters $(\nu_1,\ldots,\nu_b;\ \nu_{b+1})$, written as $D(\nu_1,\ldots,\nu_b;\ \nu_{b+1})$ if they have the joint density

$$\frac{\Gamma(\nu_1+\ldots+\nu_{b+1})}{\Gamma(\nu_1)\ldots\Gamma(\nu_{b+1})}\left(\prod_{i=1}^{b} y_i^{\nu_i-1}\right)(1-y_1-\ldots-y_b)^{\nu_{b+1}-1} \tag{1.3}$$

over the b-dimensional simplex $S_b = \{(y_1,\ldots,y_b): y_i \geq 0,$ $i = 1,\ldots,b,$ $\sum_1^b y_i \leq 1\}$ and zero outside S_b. For an elementary exposition on Dirichlet distributions and some properties, see Wilks [9], pp. 177-182 and for a more detailed discussion as well as tables, see Sobel et al [8]. One important property that we shall use later on, is that the marginal distributions of Dirichlet are also Dirichlet i.e., if $a < b$, then $(Y_1,\ldots,Y_a) \frown D(\nu_1,\ldots,\nu_a;$ $\nu_{a+1}+\ldots+\nu_{b+1})$ where "$\frown$" denotes "distributed as".

Coming back to uniform spacings, these are exchangeable with $E(Y_i) = \frac{1}{n}$ for all i under H_0. Thus tests of the form $\sum_{i=1}^n (Y_i - \frac{1}{n})^2$ and $\sum_{i=1}^n |Y_i - \frac{1}{n}|$ have been proposed and used to test H_0. See for instance Rao and Sethuraman [4], [5] for a unified treatment of the asymptotic theory and efficiencies. Simpler and more intuitive tests based on ordered spacings have also been used for this problem. Let

$$Y_{(1)} \leq \cdots \leq Y_{(n)} \tag{1.4}$$

be the ordered uniform spacings. Fisher [1] used the largest spacing $Y_{(n)}$ to construct a test of significance of the largest amplitude in harmonic analysis. J.S. Rao [3] defined the complement of the largest gap on the circle as the "circular range" since this is the smallest interval containing all the observations and proposed a test of uniformity on the circle based on it.

2. SOME RESULTS ON EXACT DISTRIBUTIONS

For $b \leq (n-1)$, consider the b spacings $(Y_1,\ldots,Y_b)$ which have a $D(1,\ldots,1; (n-b))$ distribution. J.S. Rao and Sobel [7] define the following two b-dimensional incomplete Dirichlet integrals:

$$I_p^{(b)}(1,n) = \frac{(n-1)!}{(n-b-1)!} \int_0^p \cdots \int_0^p (1-\sum_1^b y_i)^{n-b-1} \prod_1^b dy_i \tag{2.1}$$

$$\{\sum_{1}^{b} y_i \leq 1\}$$

for $0 < p < 1$ and

$$J_p^{(b)}(1,n) = \frac{(n-1)!}{(n-b-1)!} \int_p^1 \cdots \int_p^1 (1 - \sum_1^b y_i)^{n-b-1} \prod_1^b dy_i \qquad (2.2)$$

$$\{\sum_{1}^{b} y_i \leq 1\}$$

for $0 < p < \frac{1}{b}$. These integrals represent respectively, $P(Y_i < p,$ $i=1,\ldots,b)$ ie., the maximum of these b spacings is less than p and $P(Y_i > p,\ i=1,\ldots,b)$ ie., the minimum of these b spacings exceeds p. Using recurrence relations on these, one can then obtain the distributions of ordered spacings and statistics based on these. For instance, it can be seen

$$I_p^{(b)}(1,n) = I_p^{(b-1)}(1,n) - (1-p)^{n-1} I_{p/(1-p)}^{(b-1)}(1,n) \qquad (2.3)$$

and by successive iterations, this reduces to

$$I_p^{(b)}(1,n) = \sum_{j=0}^{b} (-1)^j \binom{b}{j} \langle 1 - jp \rangle^{n-1}, \qquad (2.4)$$

where $\langle x \rangle = x$ if $x > 0$ and $=0$ if $x \leq 0$. It can be shown (refer Rao and Sobel [7]) that the joint distribution of the k largest spacings viz. $Y_{(n-k+1)}, \ldots, Y_{(n)}$ is given by an I-type integral

$$f_{Y_{(n-k+1)},\ldots,Y_{(n)}}(a_1,\ldots,a_k) = (n-1)_{P_k} \cdot n_{P_k} \cdot A_k^{n-1-k} I_{P_k}^{(n-k)}(1,n-k-1) \qquad (2.5)$$

$$= (n-1)_{P_k} \cdot n_{P_k} \cdot \sum_{j=0}^{\infty} (-1)^j \binom{n-k}{j} \langle A_k - ja_k \rangle^{n-k}$$

for $0 < a_1 < \ldots < a_k < 1$, $\sum_1^k a_i \leq 1$, and with the notation

$A_k = (1 - \sum_{1}^{k} a_i)$ and $p_k = a_k/A_k$. Starting from (2.5) one can obtain the distribution of the k^{th} largest spacing $Y_{(n-k+1)}$ or the sum of the k largest gaps $\sum_{i=n-k+1}^{n} Y_{(i)}$. The density functions of these two statistics are

$$f_{Y_{(n-k+1)}}(x) = (n-1).k.\binom{n}{k} \sum_{j=0}^{n-k} (-1)^j \binom{n-k}{j} \langle 1-(j+k)x \rangle^{n-2} \quad (2.6)$$

for $0 < x < \frac{1}{k}$ and

$$f_S(s) = n!(n-1) \sum_{j=q}^{n} \frac{(-1)^{k-j+1}}{(j-k)!(j-k)^{k-1}.k!k^{n-k-1}(n-j)!} \langle js-k \rangle^{n-2} \quad (2.7)$$

for $k/q < s \leq k/(q-1)$, $q=k+1,\ldots,n$. Analogous results for the smallest spacings may be obtained by using the J-function in (2.2) or more easily by using the complementary nature of these, namely that the k^{th} largest is the $(n-k+1)^{th}$ smallest and that the sum of the k smallest spacings is equal to the complement of the sum of the $(n-k)$ largest spacings. These and other details including fuller derivations may be found in Rao and Sobel [7] which streamlines and unifies the distribution theory, also derived by other methods in the literature before.

3. A TEST BASED ON THE NUMBER OF "SMALL" SPACINGS

To detect certain clustering alternatives one may simply count the number of "small" spacings and reject the hypothesis of uniformity if there are too many "small" spacings. This type of a statistic was investigated in Puri, Rao and Yoon [2]. Recall that an average spacing is of the order of $\frac{1}{n}$ under H_0. Therefore one may define for $0 < \delta < 1$, a spacing as "small" if it is smaller than $\delta_n = \delta/n$. Let

$$R_n = R_n(\delta) = \text{number of } Y_i \leq \delta/n \,. \quad (3.1)$$

The following results are quoted from Puri, Rao and Yoon (Ibid).

Theorem 1 (Exact null distribution). Under the hypothesis of uniformity, the probability mass function of R_n defined in (3.1) is given by

$$P(R_n = k) = \binom{n}{k} \sum_{j=0}^{k} (-1)^j \binom{k}{j} \langle 1-(n-k+j)\delta_n \rangle^{n-1} \qquad (3.2)$$

for $k=0,1,\ldots,n-1$.

Theorem 2 (Asymptotic distribution under close alternatives). When $U_1,\ldots,U_{n-1}$ are i.i.d. from the sequence of close alternatives densities

$$a_n(x) = 1 + \frac{\ell(x)}{n^{1/4}} , \quad 0 \leq x \leq 1 \qquad (3.3)$$

then $\sqrt{n}\left(\frac{R_n}{n} - G_n(\delta) \right)$ is asymptotically $N(0,\sigma^2)$ where

$$G_n(x) = (1-e^{-x}) + \frac{e^{-x}}{\sqrt{n}} \left(x - \frac{x^2}{2}\right) \left(\int_0^1 \ell^2(p)dp \right), \quad x \geq 0$$

and

$$\sigma^2 = e^{-\delta}(1-e^{-\delta}-\delta^2 e^{-\delta}) .$$

The asymptotic null distribution of R_n is obtained from Theorem 2 by putting $\ell(x) \equiv 0, 0 \leq x \leq 1$, which affects only the mean. This type of a result allows one to compute the asymptotic relative efficiency (Pitman efficiency), which is the reciprocal of the sample size required for the test to attain a specified power. Because of the nature of the alternatives (3.3), the Pitman efficiency for $R_n(\delta)$ is given by the expression

$$\left(\frac{\mu_\Delta}{\sigma}\right)^4 = \frac{\left(\int_0^1 \ell^2(p)dp\right)^4 \left(\delta - \frac{\delta^2}{2}\right)^4}{(e^\delta - 1 - \delta^2)^2} \qquad (3.4)$$

where μ_Δ denotes the change in the mean under the alternatives from that under the null hypothesis. By numerical evaluations, it can be seen that among the several possible definitions of

"small" ie. choices of δ, the value $\delta = 0.7379$ yields the largest asymptotic efficiency. Thus, an optimal definition of a "small" spacing is when the spacing is about 73.79% smaller than the average value $1/n$. One may use this as the definition of "small" spacing and test the hypothesis (1.1) using Theorem 1 for the null distribution.

An illustrative example. Suppose a fire station received 20 calls on a particular day and we wish to test if these are uniformly distributed over the entire day or they tend to cluster around some particular time of the day. Suppose the calls are received at the following times:

1.10, 4.30, 6.00, 6.10, 7.00, 8.00, 8.30, 8.45, 9.30, 10.05, 13.00, 14.10, 16.00, 17.50, 19.30, 21.15, 22.00, 22.15, 23.00, 23.30.

The optimal definition of "small" spacing is when it is smaller than $\delta_n = (0.7379)\frac{24}{20}$ hrs. $\simeq$ 53 mts. The observed R_n, the number of small spacings, for this data is, 10. From Theorem 1, it can be checked that this value of $R_n = 10$ is not significant even $\alpha = 0.10$. Therefore, it may be concluded that there is no reason to reject the hypothesis of randomness of these calls in time.

REFERENCES

Fisher, R.A. (1929), "Tests of Significance in Harmonic Analysis". Proc. Roy. Soc. Ser. A. 125, pp. 54-59.

Puri, M.L., Rao, J.S. and Yoon, Y. (1979), "A Simple Test for Goodness of Fit Based on Spacings with some Efficiency Comparisons", in Contributions to Statistics (J. Jurecková, Ed.) Academia (Prague), pp. 197-209.

Rao, J.S. (1969), "Some Contributions to the Analysis of Circular Data". Unpublished Ph. D. thesis, Indian Statistical Inst., Calcutta.

Rao, J.S. and Sethuraman, J. (1970), "Pitman Efficiencies of Tests based on Spacings", in Nonparametric Techniques in Statistical Inference (Ed. M.L. Puri) Cambridge Univ. Press, pp. 405-415.

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, pp. 299-313.

Rao, J.S. (1976), "Some Tests based on Arc Lengths for the Circle". Sankhyá, Ser. B. 38, pp. 329-338.

function F_o. It is intuitively clear that no information is lost if $z_1 \leq \ldots \leq z_m$ is replaced by $y_1 \leq \ldots \leq y_{m-1} \leq z_m$ where, given z_m, $y_1, \ldots, y_{m-1}$ is distributed as the ordered sample of m-1 independent random variables which are distributed according to the appropriately defined restriction of F_o to $(0, z_m)$. In the next step this procedure is also applied when the original sample has a common distribution function F which is close - in a sense to be described later - to F_o; thus, the values $y_1 \leq \ldots \leq y_{m-1}$ are still generated according to the restriction of the distribution function F_o being known to the statistician (see also Weiss [8], page 796). It can be proved that for every critical function ψ the power function of $\psi(y_1, \ldots, y_{m-1}, z_m)$ is an approximation to the power function of $\psi(z_1, \ldots, z_m)$ if F is close to F_o. These ideas can be expressed in a mathematical model by using the conditional distribution $K(z_m, \cdot)$ of $z_1 \leq \ldots \leq z_m$ given z_m where F_o is the actual distribution function. The final outcomes $y_1 \leq \ldots \leq y_{m-1} \leq z_m$ of the random experiment are governed by the distribution which is induced by z_m and the Markov kernel K. By making use of the Fubini theorem for the distributions of Markov kernels it can easily be shown that the test procedure based on $\psi(y_1, \ldots, y_{m-1}, z_m)$ is equivalent - as far as power functions are concerned - to the test procedure based on $\tilde{\psi}(z_m)$ where $\tilde{\psi} = \int \psi(x) K(\cdot, dx)$. Thus, it is not necessary that the statistician consults his random number generator to obtain the values $y_1, \ldots, y_{m-1}$ as described above.

One has to be cautious when applying this method to other cases. As a second example we mention the case where z_m is omitted from $z_1 \leq \ldots \leq z_m$. It can be shown by examples that there are cases where, roughly speaking, all the information which is of interest for us concerning the unknown distribution function F is carried by z_m; thus, omitting z_m from $z_1, \ldots, z_m$ can possibly result in the loss of all power within a particular testing problem.

Moreover, we indicate the possibility of making the statistical inference within an approximate, simplified model. For this reason we study again the case that the ordered values $z_1 \leq \ldots \leq z_{m-1}$ are removed from $z_1 \leq \ldots \leq z_m$. If the underlying distribution function belongs to some domain of attraction of an extreme value distribution then it is well known that for m being fixed the standardized distributions of z_m and $(z_1, \ldots, z_m)$ converge in