15

Small Sample Asymptotics for Higher-Order Spacings

Riccardo Gatto and S. Rao Jammalamadaka

University of Bern, Bern, Switzerland
University of California, Santa Barbara, CA, USA

Abstract: In this chapter, we give conditional representations for families of statistics based on higher-order spacings and spacing frequencies. This allows us to compute accurate approximations to the distribution of such statistics, including tail probabilities and critical values. These results generalize those discussed in Gatto and Jammalamadaka (1999) and are essential in using such statistics in various testing contexts.

Keywords and phrases: Goodness-of-fit tests, nonparametric tests, rank tests, m-step spacings, m-step spacing frequencies, two-sample tests, Dirichlet, gamma, negative binomial distributions

15.1 Introduction

In this article, we provide some conditional representations that allow us to compute accurately the distribution of a large number of test statistics based on higher-order spacings and "spacing frequencies," following the ideas suggested in Gatto and Jammalamadaka (1999). The key point is that many important test statistics including the chi-square goodness-of-fit statistic, can be rewritten as conditional statistics, and the technique we develop here allows for very accurate approximations of their P-values, or in finding the critical values at a given level. Testing problems that were already considered by Gatto and Jammalamadaka (1999) included the two following classes of tests: (i) The class of tests based on simple spacings statistics, that is, based on the gaps between successive values of the ordered sample; and (ii) the class of tests based on the "spacing-frequencies", that is, the frequencies of one sample that fall in between the successive order statistics of the other sample, which includes many rank tests. We generalize (i) to tests based on higher-order spacings, or m-step

spacings, which are the gaps between order statistics and ones that are m steps away; and (ii) to tests based on higher-order spacing frequencies, which are the frequencies of one sample that fall in between the order statistic of the other sample that are m steps away. The reason to consider such tests is that they have higher asymptotic local powers, as demonstrated in Rao and Kuo (1984) for higher order spacings, and in Jammalamadaka and Schweitzer (1985) for higher-order spacing frequencies.

For convenience, we first review the "conditional saddlepoint approximation" that has been described in Gatto and Jammalamadaka (1999) which is the main tool for the proposed accurate approximations. The saddlepoint approximation is a well-known method of asymptotic analysis that allows us to approximate efficiently contour integrals of a general type. This method, also called the method of steepest descent, was brought into statistical use by Daniels (1954) and Lugannani and Rice (1980) for approximating the distribution of the sum of independent and identically distributed (i.i.d.) observations. The saddlepoint formula $P_n(t_1 \mid t_2)$ below enables us to find the P-values of a test statistic $T_{1n}(S_1, \ldots, S_n)$ based on the dependent quantities $S_1, \ldots, S_n$ which admit the conditional representation $T_n(S_1, \ldots, S_n) \sim T_{1n}(X_1, \ldots, X_n) \mid T_{2n}(X_1, \ldots, X_n) = t_2$, where "$\sim$" signifies the equivalence in distribution. Consider the independent random variables $X_1, \ldots, X_n$, and a statistic (T_{1n}, T_{2n}), $T_{1n} = T_{1n}(X_1, \ldots, X_n) \in \mathrm{R}$ and $T_{2n} = T_{2n}(X_1, \ldots, X_n) \in \mathrm{R}$, defined by

$$\sum_{i=1}^{n} \begin{pmatrix} \psi_{1i}(X_i, T_{1n}, T_{2n}) \\ \psi_{2i}(X_i, T_{2n}) \end{pmatrix} = 0.$$

The joint cumulant generating function of the sum of score functions ψ_{1i} and ψ_{2i} is given by

$$K_n(\lambda, t) = \sum_{i=1}^{n} \log \mathrm{E}[\exp\{\lambda_1 \psi_{1i}(X_i, t_1, t_2) + \lambda_2 \psi_{2i}(X_i, t_2)\}], \tag{15.1}$$

where $\lambda = (\lambda_1, \lambda_2)$ and $t = (t_1, t_2)$.

Step 1 Find $\alpha \in \mathrm{R}^2$ and $\beta \in \mathrm{R}$, solutions of the equations

$$\frac{\partial}{\partial \lambda} K_n(\lambda, t) = 0, \quad \frac{\partial}{\partial \lambda_2} K_n((0, \lambda_2), t) = 0.$$

Step 2 Define

$$K_n''(\lambda, t) = \frac{\partial^2}{\partial \lambda \partial \lambda^T} K_n(\lambda, t), \quad K_{2n}''(\lambda_2, t) = \frac{\partial^2}{\partial \lambda_2^2} K_n((0, \lambda_2), t),$$

$$s = \alpha_1 \left| \frac{\det(K_n''(\alpha, t))}{K_{2n}''(\beta, t)} \right|^{\frac{1}{2}}, \quad r = \mathrm{sgn}(\alpha_1)\{2[K_n((0, \beta), t) - K_n(\alpha, t)]\}^{\frac{1}{2}},$$

and

$$P_n(t_1 \mid t_2) = 1 - \Phi(r) + \phi(r)\left(\frac{1}{s} - \frac{1}{r}\right), \tag{15.2}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the standard normal density and distribution functions, and α_1 is the first element of α. Then, $\forall t_1$, t_2 and as $n \to \infty$,

$$\mathrm{P}[T_{1n} \geq t_1 \mid T_{2n} = t_2] = P_n(t_1 \mid t_2)\{1 + \mathrm{O}(n^{-1})\}. \tag{15.3}$$

Note that there is an asymptotically equivalent version of (15.2) which is given by

$$P_n^*(t_1 \mid t_2) = 1 - \Phi\left(r + \frac{1}{r}\log\left\{\frac{s}{r}\right\}\right), \tag{15.4}$$

and we refer to Example 15.2.2 for a numerical comparison.

The two steps given above allow one to approximate a tail probability or a P-value. If we are interested in quantiles or critical values, see Gatto (2001, Section 1) for an efficient algorithm for inverting this saddlepoint approximation.

15.2 Tests Based on Higher-Order Spacings

Statistics based on spacings play an important role in goodness-of-fit tests and in tests on hazard rates in the context of reliability; see Pyke (1965) for an excellent review. One-step spacings are the gaps between the successive ordered sample values and, more generally, m-step spacings are the gaps between m successive ordered sample values. One-step spacings are also very important with circular data, that is, when data are directions in two dimensions and are represented by angles. Indeed, one-step spacings are maximal invariant under changes of origin and sense of rotation. Except for one or two special cases, the exact distribution of such statistics based on uniform spacings is unknown. For most cases, the asymptotic distribution is known but it can be potentially misleading, especially when the sample size is moderate to small. Gatto and Jammalamadaka (1999, Section 3.1) derived saddlepoint approximations for test statistics based on uniform spacings. In this section, we generalize this result and provide saddlepoint approximations to test statistics based on higher-order or m-step uniform spacings. Tests based on such higher-order spacings are known to be more efficient as shown by Rao and Kuo (1984).

Consider $X_1, \ldots, X_{N-1}$ to be a sample of independent random variables from a given absolute continuous distribution F with support in R. The fundamental problem of goodness-of-fit, is to test $F = F_0$, where F_0 is specified. By the probability integral transform $U_i = F_0(X_i)$, $i = 1, \ldots, N-1$ the goodness-of-fit test is reduced to one of testing if $U_1, \ldots, U_{N-1}$ are uniformly distributed,

that is, to test the null hypothesis

$$\mathrm{H}_0 : F(u) = u, \ \forall u \in [0, 1).$$

Let $0 \leq U_{(1)} \leq \cdots \leq U_{(N-1)} \leq 1$, denote the ordered sample. The simple or one-step spacings $D_1, \ldots, D_N$ are the gaps between this ordered sample, viz.,

$$D_i = U_{(i)} - U_{(i-1)}, \ \ i = 1, \ldots, N,$$

where $U_{(0)} \stackrel{\text{def}}{=} 0$ and $U_{(N)} \stackrel{\text{def}}{=} 1$. More generally, the m-step disjoint spacings are the gaps between m successive values of the ordered sample. That is, denoting $\lfloor x \rfloor$ for the greatest integer less than or equal to x, for $M = \lfloor N/m \rfloor$,

$$D_{im}^{(m)} = U_{(im)} - U_{((i-1)m)}, \ \ i = 1, \ldots, M.$$

Let $h(\cdot)$ and $h_i(\cdot)$, $i = 1, \ldots, M$, be real-valued functions that satisfy some weak regularity conditions. Most spacings statistics can then be expressed as

$$T_n^* = \sum_{i=1}^{M} h_i(MD_{im}^{(m)}), \tag{15.5}$$

which is not symmetric in the spacings, or as

$$T_n = \frac{1}{M} \sum_{i=1}^{M} h(MD_{im}^{(m)}), \tag{15.6}$$

which is symmetric in the spacings. Sethuraman and Rao (1970) and Rao and Sethuraman (1975) showed that the class of symmetric tests (15.6) based on one-step spacings cannot discriminate alternatives converging to the null hypothesis at asymptotic rates faster than $N^{-1/4}$, which is a drawback when compared, for example, to the Kolmogorov-Smirnov test. Del Pino (1979) showed that tests based on m-step spacings, $m > 1$, have better asymptotic efficiencies than tests based on one-step spacings. Typical examples of symmetric test statistics (15.6) are obtained with

$$h(x) = \log x, \ |x - 1|, \ x^a,$$

$a > -1/2$ and $\neq 0$ or 1. The first two functions lead to the Rao and the log higher-order test statistics and they will be developed in Examples 15.2.1 and 15.2.2 below. The last function for $a = 2$ leads to the Greenwood higher-order test statistic and will be developed in Example 15.2.3. It has maximum asymptotic relative efficiency among symmetric m-step spacings statistics, is asymptotically more efficient that the one-step Greenwood statistic, and indeed the efficiency grows with m; see Table 2 in Rao and Kuo (1984).

The exact distribution of spacings statistics is unknown in most cases and it is common practice to rely on the limiting normal distribution, which does however not guarantee sufficient accuracy, if we have a sample of small to moderate size, or if we are interested in small tail probabilities. If a higher accuracy is desired, the conditional saddlepoint approximation can be applied with the following conditional representation of the m-step spacings. If $Y_1, \ldots, Y_M$ are independent Gamma(m, b) random variables with density $\{b^m/\Gamma(m)\}y^{m-1}e^{-by}$, $y \geq 0$, then, under H_0 and $\forall b > 0$,

$$(MD_{1\cdot m}^{(m)}, \ldots, MD_{M\cdot m}^{(m)}) \sim \left\{ (Y_1, \ldots, Y_M) \mid \sum_{i=1}^{M} Y_i = M \right\}. \tag{15.7}$$

The equivalence in (15.7) is easy to justify; see, for example, Wilks (1962, Section 7.7). Thus $(D_{1\cdot m}^{(m)}, \ldots, D_{(M-1)\cdot m}^{(m)}) \sim$ Dirichlet$(m, \ldots, m; m)$, and these m-spacings admit the conditional Gamma representation (15.7). This conditional representation together with the computational steps given in Section 15.1 allow us to compute a saddlepoint approximation for the distribution of symmetric and asymmetric test statistics based on m-step spacings. The particular case $m = 1$ in (15.7) corresponds to the exponential representation of simple spacings, and using this, Gatto and Jammalamadaka (1999, Section 3.1) developed four examples with one-step spacing statistics: the Rao spacings test, the log spacings test, the Greenwood spacings test, and the locally most powerful spacings test given by $h_i(ND_i) = \Phi^{(-1)}(\frac{i}{N+1})ND_i$. Saddlepoint approximations were computed for these four examples with sample sizes as low as $N = 3$, and they showed a very high accuracy, even for small tail probabilities. By means of this new conditional representation, we provide some further examples for the case of higher-order spacings.

Example 15.2.1 (The Rao higher-order spacings test) In order to apply Steps 1 and 2 of the saddlepoint approximation in Section 15.1, we must determine the joint cumulant generating function of the score functions

$$\begin{aligned} \psi_{1i}(x, t_1) &= \begin{cases} (1 - x - t_1), & \text{if } x \in [0, 1), \\ (x - 1 - t_1), & \text{if } x \in [1, \infty), \end{cases} \\ \psi_{2i}(x, t_2) &= x - t_2, \end{aligned}$$

with $\psi_{ji} = \psi_j$, $i = 1, \ldots, n$, $j = 1, 2$. With some algebraic computations, we can see that, for $b = m$ and $t_2 = 1$, this cumulant generating function has the form

$$\begin{aligned} K_M((\lambda_1, \lambda_2), (t_1, 1)) = M\Big[& m \log m - m \log(m + \lambda_1 - \lambda_2) + \lambda_1(1 - t_1) - \lambda_2 \\ & + \log\Big\{ P(m, m + \lambda_1 - \lambda_2) + (\frac{m + \lambda_1 - \lambda_2}{m - \lambda_1 - \lambda_2})^m e^{-2\lambda_1} \\ & [1 - P(m, m - \lambda_1 - \lambda_2)] \Big\} \Big], \end{aligned}$$

where $P(m,x) = 1 - e^{-x}\sum_{i=1}^{m-1} x^j/j!$, $m = 1, 2, \ldots$, and $x \in \mathrm{R}$. The derivatives of $K_M((\lambda_1, \lambda_2), (t_1, 1))$ with respect to λ_1 and λ_2 can be obtained by automatic symbolic computation (e.g., with *Maple*). The advantage of choosing $b = m$ as scale parameter in the conditional Gamma representation is that the expectation of the sample mean of the Gamma random variables becomes one, and hence the "conditional saddlepoint equation," that is, the second equation in Step 1, has the trivial solution $\beta = 0$. Furthermore, $\beta = 0$ leads to $K''_{2M}(\beta, t) = M\mathrm{Var}(Y_1) = M/m$ and to $K_M((0,\beta),t) = 0$ in the formulas of s and r in Step 2.

Example 15.2.2 (The log higher-order spacings test) The choice of the score function $h(x) = \log x$ in (15.6) was proposed by Darling (1953) and it maximizes Bahadur efficiency; see Zhou and Jammalamadaka (1989). For the case $b = m$ and $t_2 = 1$, the joint cumulant generating function in (15.1) is given by

$$K_M((\lambda_1, \lambda_2), (t_1, 1)) = M\Big[-\lambda_1 t_1 - \lambda_2 + m\log m - (\lambda_1 + m)\log(m - \lambda_2) + \log\frac{\Gamma(\lambda_1 + m)}{\Gamma(m)}\Big]$$

provided that $\lambda_1 > -m$ and $\lambda_2 < m$. The second derivatives of $K_M((\lambda_1, \lambda_2), (t_1, 1))$ with respect to λ_1 and λ_2 are the following:

$$\partial^2 K_M((\lambda_1, \lambda_2), (t_1, 1))/(\partial\lambda_1)^2 = \Psi(1, \lambda_1 + m),$$

$$\partial^2 K_M((\lambda_1, \lambda_2), (t_1, 1))/(\partial\lambda_1\partial\lambda_2) = (m - \lambda_2)^{-1},$$

and

$$\partial^2 K_M((\lambda_1, \lambda_2), (t_1, 1))/(\partial\lambda_2)^2 = \frac{\lambda_1 + m}{(m - \lambda_2)^2},$$

where $\Psi(z) = \Gamma'(z)/\Gamma(z)$ is the digamma function and $\Psi(z,n) = (d/dz)^n\Psi(z)$ is the polygamma function, with $\Re\{z\} > 0$ and $n \in \mathcal{N}$. The first derivatives are not necessary because the saddlepoint equation can be efficiently solved by a minimization routine such as *Matlab*'s routine `fminsearch`. In this example, we consider $N = 6$ and $m = 2$, yielding the very small number of summands or effective sample size $M = 3$. The numerical results are displayed in Figure 15.1 in terms of absolute errors $\mid \mathrm{P_{MC}} - \mathrm{P_{SP}} \mid$ and relative absolute error $\mid \mathrm{P_{MC}} - \mathrm{P_{SP}} \mid / \min\{\mathrm{P_{MC}}, 1 - \mathrm{P_{MC}}\}$, where $\mathrm{P_{MC}}$ and $\mathrm{P_{SP}}$ denote the distribution of the test statistic obtained by the 10^6 Monte Carlo simulated values of the test statistic and by the saddlepoint approximation in the Lugannani and Rice form in (15.2), or in its asymptotic equivalent version in (15.4), sometimes referred to as "Barndorff-Nielsen formula."

From Figure 15.1, we can see that the saddlepoint approximation has a small relative error over the whole domain of the distribution, and therefore is uniformly accurate. The Lugannani and Rice version in (15.2) has all relative errors below 10 %, and it appears substantially more accurate than its asymptotic equivalent formula in (15.4). For this test of uniformity, the small left tail probabilities are the most important. Note that the small increment of relative errors at both ends of the domains is not necessarily due to an inaccuracy of the saddlepoint approximation, because it is based on very few simulated values. (A further analysis based on importance sampling would provide a more reliable comparison.) The domain of the distribution is $(-\infty, 0)$ (all approximated distributions are almost zero at the left of -1), and the density function has a negative skewness.

Matlab programs for the computation of this saddlepoint approximation can be found at the address `http://www.stat.unibe.ch/~gatto`.

Example 15.2.3 (The Greenwood higher-order spacings test) The choice of the score function $h(x) = x^2$ in (15.6) defines the Greenwood test statistic. The joint cumulant generating function (15.1) for $b = m$ and $t_2 = 1$ is given by

$$\begin{aligned} K_M((\lambda_1, \lambda_2), (t_1, 1)) &= M \Bigg[m \log 2 + m \log m - \lambda_1 t_1 - \lambda_2 - \frac{(m - \lambda_2)^2}{4\lambda_1} \\ &\quad - \frac{m}{2} \log(-\lambda_1) + (m-1) \log(m - \lambda_2) \\ &\quad + \log \sum_{j=0}^{m-1} (-1)^{j+m-1} \left(\frac{2}{m - \lambda_2} \right)^j \frac{\Gamma(\frac{j+1}{2}, -\frac{(m-\lambda_2)^2}{4\lambda_1})}{\Gamma(j+1)\Gamma(m-j)} \Bigg] \end{aligned}$$

provided that $\lambda_1 < 0$ and $\lambda_2 < m$, and where $\Gamma(a, x) = \int_x^\infty e^{-t} t^{a-1} dt$ is the incomplete Gamma function.

15.3 Tests Based on Higher-Order Spacing-Frequencies

Consider a first sample of $(N-1)$ independent random variables $X_1, \ldots, X_{N-1}$, with underlying absolute continuous distribution F defined on $A \subset \mathrm{R}$, and a second sample of n independent random variables $Y_1, \ldots, Y_n$, with underlying absolute continuous distribution G, also defined on $A \subset \mathrm{R}$. The general two-sample problem is to test the null hypothesis H_0: $F = G$. Define the random

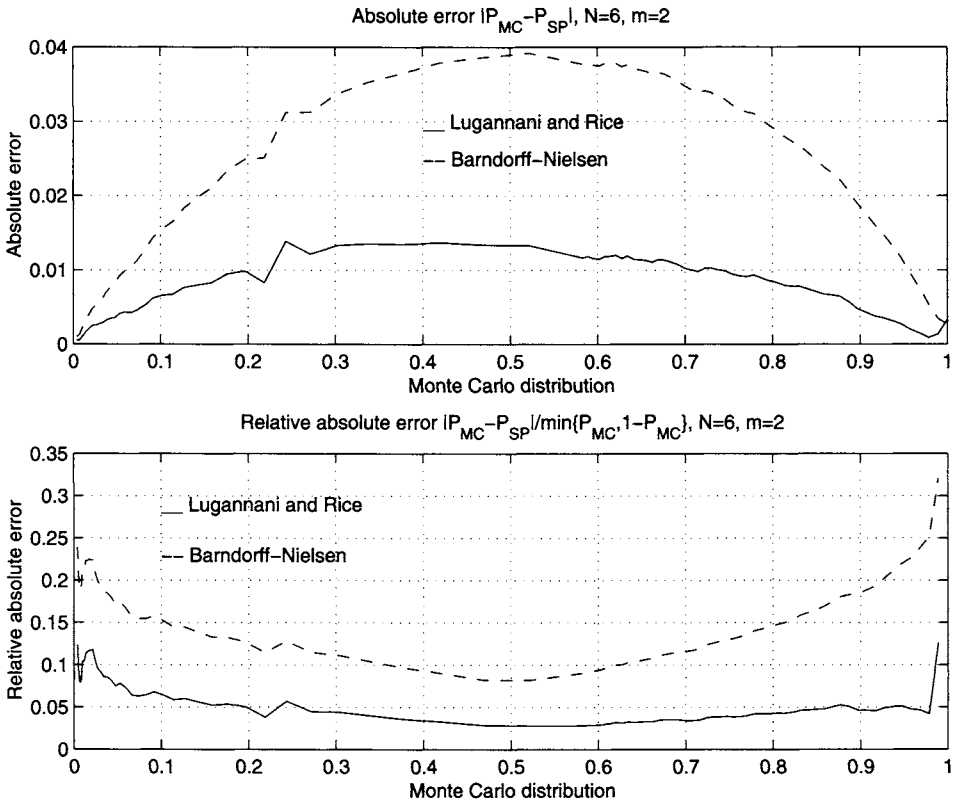

Figure 15.1: Saddlepoint and Monte Carlo approximations to the distribution of the log higher-order spacings statistic, $N = 6$, $m = 2$ and $M = 3$. Upper figure: absolute error $\mid P_{MC} - P_{SP} \mid$. Lower figure: relative absolute error $\mid P_{MC} - P_{SP} \mid / \min\{P_{MC}, 1 - P_{MC}\}$. P_{MC}: Monte Carlo approximation to the distribution. P_{SP}: saddlepoint approximations to the distribution. Solid line: Lugannani and Rice approximation in (15.2). Dashed line: Barndorff-Nielsen approximation in (15.4)

variables

$$S_j = \sum_{i=1}^{n} I\{Y_i \in [X_{(j-1)}, X_{(j)})\}, \quad j = 1, \ldots, N,$$

where for convenience, we take $X_{(0)} \stackrel{\text{def}}{=} \inf\{A\}$ and $X_{(N)} \stackrel{\text{def}}{=} \sup\{A\}$. The numbers $\{S_1, \ldots, S_N\}$ are called the spacing frequencies becaise they correspond to the frequencies or counts of the $\{Y_i\}$ that fall in between successive $\{X_{(j)}\}$. In fact, if $R(X_{(k)})$ denotes the rank of the kth largest $\{X_j\}$ in the combined sample, $k = 1, \ldots, N$, it is easily seen that $R(X_{(k)}) = \sum_{j=1}^{k}(S_j + 1)$, or, $S_k = R(X_{(k)}) - R(X_{(k-1)}) - 1$, $k = 1, \ldots, N$, so that the $\{S_j\}$ are also the "rank differences."

Let $h(\cdot)$ and $h_j(\cdot)$, $j = 1, \ldots, N$, be real-valued functions satisfying certain regularity conditions. Holst and Rao (1980) consider statistics of the form $N^{-1/2}\sum_{j=1}^{N} h_j(S_j)$ and $N^{-1/2}\sum_{j=1}^{N} h(S_j)$ and their asymptotic properties when both N and n tend to infinity; formally, through nondecreasing sequences of positive integers $\{N_\nu\}$ and $\{n_\nu\}$ such that, as $\nu \to \infty$,

$$N_\nu \to \infty,\ n_\nu \to \infty \text{ and } \frac{N_\nu}{n_\nu} = \rho_\nu \to \rho,\ 0 < \rho < \infty.$$

Specifically, they show that if $V_1, \ldots, V_N$ are independent geometric random variables with probability distribution function

$$\mathrm{P}[V_1 = k] = \{\rho/(\rho+1)\}^k \cdot 1/(\rho+1), \quad k = 0, 1, 2, \ldots, \tag{15.8}$$

then, under H_0,

$$\sum_{j=1}^{N} h_j(S_j) \xrightarrow{\mathcal{D}} \mathcal{N}(\mu, \sigma^2), \tag{15.9}$$

where $\mu = \mathrm{E}[\sum_{j=1}^{N} h_j(V_j)]$ and $\sigma^2 = \mathrm{Var}\left(\sum_{j=1}^{N} h_j(V_j) - \beta\sum_{i=j}^{N} V_j\right)$ in which β is the regression coefficient given by

$$\beta = \mathrm{Cov}\left(\sum_{j=1}^{N} h_j(V_j), \sum_{j=1}^{N} V_j\right) \Big/ \mathrm{Var}\left(\sum_{j=1}^{N} V_j\right).$$

As we stated already, the asymptotic efficiencies are improved by considering the corresponding higher-order spacings. Therefore, we now consider the more general case. For $m \geq 1$, denote $M = \lfloor N/m \rfloor$, and define the "nonoverlapping" or disjoint mth order spacing-frequencies

$$S_{k\cdot m}^{(m)} = \sum_{j=0}^{m-1} S_{k\cdot m+j} = \sum_{k=1}^{M} I\{Y_j \in [X_{(k\cdot m-1)}, X_{(k\cdot m+m-1)})\}, \quad k = 1, \ldots, M-1,$$

where we take $S_k^{(m)} = S_{k-M}^{(m)}$ for $k > M$ circularly, for convenience. Let $h(\cdot)$ and $h_j(\cdot)$, $j = 1, \ldots, N$, be real-valued functions satisfying certain regularity

conditions [see Assumption (A), in Jammalamadaka and Schweitzer (1985)], and define the general classes of test statistics

$$T_\nu^* = \sum_{j=1}^{M} h_j(S_{j\cdot m}^{(m)}), \tag{15.10}$$

and

$$T_\nu = \sum_{j=1}^{M} h(S_{j\cdot m}^{(m)}), \tag{15.11}$$

which represent, respectively, the nonsymmetric and the symmetric test statistics based on such higher-order spacing frequencies. Jammalamadaka and Schweitzer (1985) discuss the asymptotic normality of such statistics (and indeed, more general ones based on the "overlapping" mth-order spacing frequencies) both under the null hypothesis, as well as under close alternatives.

The following optimality result has been proved there; see Theorem 3.2 in Jammalamadaka and Schweitzer (1985) for further details. Consider $\{G_N\}$, a smooth sequence of distribution functions converging towards F, as $N \to \infty$. It turns out that the asymptotically most powerful test for the null hypothesis H_0 against the sequence of simple alternatives

$$\mathrm{A}_N : G = G_N$$

is to reject H_0 when

$$\sum_{j=1}^{M} l\left(\frac{j}{M+1}\right) S_{j\cdot m}^{(m)} > c, \tag{15.12}$$

where $l(\cdot)$ is the derivative of $L(u) = \lim_{N\to\infty} N^{\frac{1}{2}}[G_N(F^{(-1)}(u)) - u]$, $0 \le u \le 1$. However, such linear combinations of higher-order spacing frequencies in $\{S_{j\cdot m}^{(m)}\}$ are equivalent to linear combinations in one-step spacing frequencies S_j, already discussed in Gatto and Jammalamadaka (1999, Section 4) and need no further elaboration.

However, among the class of symmetric tests, there is reason to consider higher-order spacing frequencies. It is shown there that the sum of squares, leading to the statistic

$$\sum_{j=1}^{M} (S_{j\cdot m}^{(m)})^2, \tag{15.13}$$

is the optimal choice among all such symmetric nonoverlapping statistics. When $m = 1$, this has been introduced by Dixon (1940) and has been shown to be locally most powerful by Holst and Rao (1980) among such tests based on one-step spacing-frequencies.

For the more general statistics based on the mth-order spacing frequencies, consider the independent random variables $\eta_1, \ldots, \eta_M$ with the same negative binomial distribution with parameters m and $\rho/(1+\rho)$, viz.,

$$\mathrm{P}[\eta_1 = j] = \binom{m+j-1}{j} \left(\frac{1}{1+\rho}\right)^j \left(\frac{\rho}{1+\rho}\right)^m, \quad j = 0, 1, \ldots . \tag{15.14}$$

A moment's reflection shows that these negative binomial random variables arise by taking sums of the independent geometric random variables m at a time, corresponding to one-step spacing frequencies. It can be verified that under H_0, the mth-order spacing frequencies have the same distribution as independent negative binomial random variables conditioned to sum up to n, that is, if $\eta_1, \ldots, \eta_M$ are i.i.d. with probability function (15.14), then $\forall p \in (0,1)$, it can be checked that

$$\{S_1^{(m)}, \ldots, S_M^{(m)}\} \sim \{\eta_1, \ldots, \eta_M\} \mid \sum_{j=1}^{M} \eta_j = n.$$

To illustrate the power of our conditional approach through which accurate saddlepoint approximations can be obtained, we quote a simple result for symmetric statistics based on nonoverlapping mth-order spacing frequencies, which is a consequence of the results of Jammalamadaka and Schweitzer (1985, Theorem 4.2).

Proposition 15.3.1 *Under* H_0, *if* $\eta \sim \eta_1$,

$$M^{-1/2} \sum_{j=1}^{M} \{h(S_{j\cdot m}^{(m)}) - \mathrm{E}[h(\eta)]\} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \tag{15.15}$$

where

$$\sigma^2 = \mathrm{Var}(h(\eta)) - \frac{\rho^2}{1+\rho}(\mathrm{Cov}^2(h(\eta), \eta).$$

The same conditioning idea used for obtaining the first-order approximation in (15.15) can be exploited for the construction of our saddlepoint approximation. By defining

$$T_{1\nu}^* = \sum_{j=1}^{M} h_j(\eta_j), \quad T_{1\nu} = \frac{1}{M} \sum_{j=1}^{M} h(\eta_j) \quad \text{and} \quad T_{2\nu} = \frac{1}{M} \sum_{j=1}^{M} \eta_j,$$

the conditional distributions of $(T_{1\nu}^* \mid T_{2\nu} = 1)$ and $(T_{1\nu} \mid T_{2\nu} = 1)$ can be approximated again by Steps 1 and 2 of Section 15.1 and with the result below. These approximations are also accurate approximations to the distributions of T_ν^* and T_ν in (15.10) and (15.11), respectively. The following results, which can be proved by direct verification from our general results, show how one can find saddlepoint approximations for statistics in (15.12) and (15.13). Numerical evaluations are somewhat straightforward and are omitted.

Proposition 15.3.2 *The joint cumulant generating function on (15.1) for the test statistic (15.12) is given by*

$$K_M((\lambda_1,\lambda_2),(t_1,t_2)) = -\lambda_1 t_1 - M\lambda_2 t_2 + mM\log(1-p) - m\sum_{j=1}^{M}\log\left[1-p\exp\left\{\lambda_1 l\left(\frac{j}{M+1}\right)+\lambda_2\right\}\right],$$

where $0<p<1$ *and* $\lambda_1 l(j/(M+1))+\lambda_2 < -\log p$, *for* $j=1,\ldots,M$.

Proposition 15.3.3 *The joint cumulant generating function in (15.1) for the test statistic (15.13) is given by*

$$K_M((\lambda_1,\lambda_2),(t_1,t_2)) = M\Big[-\lambda_1 t_1 - \lambda_2 t_2 + m\log(1-p) - m\log\{1-pe^{\lambda_2}\}+\kappa(\lambda_1)\Big],$$

where $\kappa(\lambda_1) = \log \mathrm{E}[e^{\lambda_1 J^2}]$, J *is a negative binomial random variable with parameters* m *and* $1-pe^{\lambda_2}$, $0<p<1$, $\lambda_1<0$ *and* $\lambda_2 < -\log p$.

15.4 Conclusion

In this discussion, we develop accurate approximations valid for small to moderate sample sizes, for the distributions of statistics based on higher order spacings, and higher-order spacing frequencies, whose exact distributions are unavailable and asymptotics are quite inaccurate.

Acknowledgements. The research of R. Gatto was supported by the Swiss National Science Foundation.

References

1. Daniels, H. E. (1954). Saddlepoint approximations in statistics, *The Annals of Mathematical Statistics*, **25**, 631–650.

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

3. Del Pino, G. E. (1979). On the asymptotic distributionof k-spacings with applications to goodness-of-fit tests, *The Annals of Statistics*, **7**, 1058–1065.

4. Dixon, W. J. (1940). A criterion for testing the hypothesis that two samples are from the same population, *Annals of Mathematical Statistics*, **11**, 199–204.

5. Gatto, R. (2001). Symbolic computation for approximating the distributions of some families of one and two-sample nonparametric test statistics, *Statistics and Computing*, **11**, 449–455.

6. Gatto, R., and Jammalamadaka, S. R. (1999). A conditional saddlepoint approximation for testing problems, *Journal of the American Statistical Association*, **94**, 533–541.

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

8. Holst, L., and Rao, J. S. (1981). Asymptotic spacings theory with applications to the two-sample problem, *The Canadian Journal of Statistics*, **9**, 79–89.

9. Lugannani, R., and Rice, S. (1980). Saddle point approximation for the distribution of the sum of independent random variables, *Advances in Applied Probability*, **12**, 475–490.

10. Jammalamadaka, S. R., and Schweitzer, R. L. (1985). On tests for the two-sample problem based on higher order spacing-frequencies, In *Statistical Theory and Data Analysis* (Ed., K. Matusita), pp. 583–618, North-Holland, Amsterdam.

11. Pyke, R. (1965). Spacings, *The Journal of the Royal Statistical Society, Series B*, **27**, 395–449.

12. Rao, J. S. (1976). Some tests based on arc-lengths for the circle, *Sankhyā, Series B*, **4**, 329–338.

13. Rao, J. S., and Kuo, M. (1984). Asymptotic results on the Greenwood statistic and some of its generalizations, *The Journal of the Royal Statistical Society, Series B*, **46**, 228–237.

14. Rao, J. S., and Sethuraman, J. (1975). Weak convergence of empirical distribution functions of random variables subject to perturbations and scale factors, *The Annals of Statistics*, **3**, 299–313.

15. Sethuraman, J., and Rao, J. S. (1970). Pitmann efficiencies of tests based on spacings, In *Nonparametric Techniques in Statistical Inference* (Ed., M. L. Puri), Cambridge University Press, Cambridge.

16. Wilks, S. S. (1962). *Mathematical Statistics*, John Wiley & Sons, New York.

17. Zhou, X., and Jammalamadaka, S. R. (1989). Bahadur efficiencies of spacings test for goodness of fit, *Annals of the Institute of Statistical Mathematics*, **41**, 541–553.