Approximating the Distribution of the Scan Statistic Using Moments of the Number of Clumps

Fred W. Huffer
Department of Statistics
Florida State University
Tallahassee, FL 32306, U.S.A.
huffer@stat.fsu.edu

Chien-Tai Lin
Department of Mathematics
Tamkang University
Tamsui, Taiwan, R.O.C.
chien@sparc20.math.tku.edu.tw

August 8, 1995

Abstract

Let $X_1, X_2, \ldots, X_n$ be randomly distributed points on the unit interval. Let $N_{x,x+d}$ be the number of these points contained in the interval $(x, x+d)$. The scan statistic N_d is defined as the maximum number of points in a window of length d, that is, $N_d = \sup_x N_{x,x+d}$. This statistic is used to test for the presence of non-random clustering. We say that m points form an $m : d$ clump if these points are all contained in some interval of length d. Let Y denote the number of $m : d$ clumps. In this paper we show how to compute the lower order moments of Y, and we use these moments to obtain approximations and bounds for the distribution of the scan statistic N_d. Our approximations are based on using the "method of moments" to approximate the distribution of Y. We try two basic types of "method of moments" approximations, one involving a simple Markov chain model, and others using a variety of different compound Poisson approximations. Our results compare favorably with other approximations and bounds in the literature. In particular, our approximations MC2 and CPG2, which use only the first two moment of Y, do quite well and ought to be generally useful. In our work, we calculate the moments of Y using recursions given by Huffer (1988). We give explicit general formulas for the first two moments of Y and show how the computer programs of Lin (1993) may be used to calculate the third and fourth moments.

1 Introduction

Let $X_1, X_2, \ldots, X_n$ be n points independently drawn from a uniform distribution on the unit interval. The corresponding order statistics are $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. We say that an $m : d$ clump exists if there are m consecutive points all contained within an interval of length d. We let Y_I denote the number of $m : d$ clumps, i.e.

$$Y_I = \sum_{i=1}^{n-m+1} I\{X_{(i+m-1)} - X_{(i)} \leq d\}\,, \tag{1}$$

where for convenience we take $X_{(0)} = 0$ and $X_{(n+1)} = 1$.

Random points on a circle arise in some applications. We can also define the number of $m : d$ clumps in this situation in a manner similar to the above. Given $n+1$ random points $X_1, X_2, \ldots, X_{n+1}$ on a circle with unit circumference, we define the circular order statistics $X_{(1)}, X_{(2)}, \ldots, X_{(n+1)}$ to be these same points ordered in a clockwise fashion starting with $X_{(1)} \equiv X_1$. We define $X_{(j)} - X_{(i)}$ to be distance one must travel in a clockwise direction to go from $X_{(i)}$ to $X_{(j)}$. We can now define the number of $m : d$ clumps Y_C by

$$Y_C = \sum_{i=1}^{n+1} I\{X_{(i+m-1)} - X_{(i)} \leq d\}\,.$$

Here we are using addition mod $n+1$ in the subscripts, that is, we take $X_{(0)} = X_{(n+1)}$ and $X_{(i)} = X_{(n+1+i)}$. For remarks that pertain to both the interval and circle cases, we shall drop the subscripts I and C and just use Y to denote the number of $m : d$ clumps. Note that to obtain the closest correspondence between the formulas in the interval and circle cases, the number of random points on the circle is always $n+1$ whereas the number of points on the interval is n.

Let $N_{x,x+d}$ be the number of points X_i contained in the interval $(x, x+d)$. The scan statistic N_d is defined as the maximum number of points in a window of length d, that is, $N_d = \sup_x N_{x,x+d}$. This statistic is used to test for the presence of non-random clustering. Scan statistics have been applied in many fields and are described in Newell (1963), Naus (1965, 1966, 1979), Neff and Naus (1980), and Wallenstein and Neff (1987). Because of the complicated and impractical computation of the exact

Key words: Compound Poisson approximations; Markov Chain approximation; Spacings.

distribution of the scan statistic and the limited applicability of the asymptotic results, recent research in this area has mainly concentrated on finding approximations and bounds. See, e.g., Cressie (1980), Naus (1982), Gates and Westcott (1984), Berman and Eagleson (1985), Wallenstein and Neff (1987), Glaz (1989), Loader (1991), Glaz (1992), Roos (1993), and Glaz et al. (1994).

In this paper we are interested in the number of clumps Y mainly because it is closely related to the scan statistic N_d. In particular, $Y \geq 1$ if and only if $N_d \geq m$. This implies $P\{Y \geq 1\} = P\{N_d \geq m\}$. The number of clumps has some importance in its own right. Glaz (1993) suggests that Y (which he calls the multiple scan statistic) would be a reasonable statistic for testing uniformity; if the value of Y is much larger than expected, we reject the hypothesis of uniformity. We can use the moments of Y to approximate the distribution of Y and hence obtain approximate critical values for this test. The distribution of Y has been studied by several authors. Among them are Glaz and Naus (1983), Dembo and Karlin (1992), Roos (1993), and Glaz et al. (1994). Much of the recent research has been devoted to finding Poisson and compound Poisson approximations to the distribution of Y.

The goal of this paper is to approximate the distribution of the scan statistic by using the moments of the number of clumps Y. We work with the moments of Y because they are relatively easy to compute and lead naturally to approximations and bounds for the value of $P\{Y \geq 1\} = P\{N_d \geq m\}$. The paper will be organized as follows.

In section 2 we show how to use the moments of Y to compute bounds and approximations for $P\{Y \geq 1\}$. Section 2.1 presents upper and lower bounds for $P\{Y \geq 1\}$. Section 2.2 contains an approximation to $P\{Y \geq 1\}$ based on a simple two-state Markov chain model. This approximation uses only the first two moments of Y. Section 2.3 presents approximations to $P\{Y \geq 1\}$ based on a variety of different compound Poisson approximations to the distribution of Y. Tables are given to illustrate our bounds and approximations and compare them with others in the literature.

In Section 3 we describe how to compute the moments of Y. In section 3.2 we give explicit general expressions for the first and second moments of both Y_I and Y_C. For higher order moments, we use the formulas in Section 3.3 and the computer programs written by Lin (1993) to derive expressions valid for particular values of m. In Section

3.4 we illustrate this process by computing the third and fourth moments of Y_I when $m = 4$.

2 Approximations and Bounds

In the next section we will demonstrate how to compute the moments $\mu_i = EY^i$ for $1 \le i \le 4$. In this section we show how to use these moments to compute bounds and approximations for $P\{Y \ge 1\}$.

2.1 Bounds for $P\{Y \ge 1\}$

Let $p = P\{Y \ge 1\}$. Our approach to obtaining bounds for p is a simple one; to obtain an upper (lower) bound for p, we find a convenient random variable which is an upper (lower) bound for the indicator $I_{\{Y \ge 1\}}$, and then compute the expectation of this random variable. Let w denote the largest possible value of Y; for Y_I this value is $w = n - m + 1$. If ϕ and ψ are any two functions satisfying $\phi(0) \le 0 \le \psi(0)$ and $\phi(k) \le 1 \le \psi(k)$ for $k = 1, 2, \ldots, w$, then clearly $E\phi(Y) \le p \le E\psi(Y)$. When ϕ and ψ are polynomials of order at most 4, these expectations can be easily computed from the values $\mu_1, \mu_2, \mu_3, \mu_4$. Thus, by selecting appropriate polynomials ϕ and ψ, we can obtain bounds on p using the moments.

We have obtained excellent results with the following families of polynomials. For integer pairs (i, j) satisfying $1 \le i \le j - 2$ define

$$\phi(y; i, j) = 1 - \frac{(y - i)(y - i - 1)(y - j)(y - j - 1)}{i(i + 1)j(j + 1)}. \tag{2}$$

For integers i satisfying $2 \le i \le w - 2$ define

$$\psi(y; i) = 1 - \frac{(y - 1)(y - w)(y - i)(y - i - 1)}{wi(i + 1)}. \tag{3}$$

These polynomials are of order 4 and can easily be shown to satisfy the required conditions. Therefore,

$$\max_{i,j} E\phi(Y; i, j) \le p \le \min_{i} E\psi(Y; i). \tag{4}$$

The lower and upper bounds in (4) have been computed in numerous examples. In our tables they are labeled as LB and UB. The max and min in (4) are usually attained for fairly small values of i and j, so there is actually very little calculation involved

in implementing these bounds. The bounds we have just described are not necessarily the strongest possible. Given the first k moments of Y and no other information, the best possible upper and lower bounds for p can be obtained via linear programming (see Kwerel (1975), Prékopa (1988), Krauth (1991)). We have not pursued this more complicated route since we suspect that the resulting improvements in the bounds would be slight.

In judging the adequacy of the bounds and approximations we propose in this paper, we shall rely heavily on the work of Glaz (1989). For the scan statistic on the interval, let $P(m;d,n) = P\{N_d \geq m\} = P\{Y_I \geq 1\} = p$. Glaz (1989) gave extensive tables comparing various approximations and bounds for $P(m;d,n)$ over a broad range of values for m, d, and n. We display excerpts from his tables in our tables 3, 5, 7, 9, 10, 11, 12, 13. We shall compare our upper bound UB with that given in equation (2.9) of Glaz (1989); we refer to this bound as GUB in our tables. This bound was the best of the two upper bounds that Glaz studied. We find that UB and GUB are very close competitors. For small values of m, UB can sometimes be a substantial improvement over GUB. But for larger values of m, GUB tends to be slightly better.

Glaz (1989) also studied the performance of one lower bound which is given in his equation (2.12). He refers to this as the Kwerel lower bound because it is based on an inequality of Kwerel (1975). (The same inequality was also given by Dawson and Sankoff (1967).) For this reason we abbreviate this bound as KLB in our tables. In table 2 for $n = 25$ we see that our bound LB improves uniformly on KLB; in many cases the improvement is fairly dramatic. Because of the difficulty of the computations involved, Glaz did not report KLB in his tables for larger values of n. But it is not hard to show that LB will be strictly larger than KLB for all values of m, d, and n. One might expect this on general grounds. The bound KLB can be written in terms of the first two moments of Y_I and is, in fact, the best lower bound based on the first two moments. Any reasonable bound based on the first four moments (such as LB) ought to do better.

We note that when p is small (say $p < .10$) the bounds LB and UB are usually fairly tight. Thus, at commonly used significance levels like 0.05 or 0.01, the bounds LB and UB will often suffice for carrying out hypothesis tests using the scan statistic.

We cannot currently compute the bounds LB and UB for $m > 10$. This is because

of the difficulty of computing μ_4 (and to a lesser extent μ_3) by our current methods. Our approach (described in more detail in Section 3) involves obtaining, for each value of m, an exact algebraic expression for μ_4 which is valid for all n and d. The computer time and memory required to construct this expression increases with m and prevents us from using larger values of m. The advantage of our approach is that, once the expression for μ_4 has been obtained for some m, this expression can be stored and used to rapidly compute μ_4 to arbitrary precision for any values of n and d. The remarks of this paragraph also apply to those approximations to p developed in Section 2.3 which use the fourth moment.

2.2 A "Markov Chain" Approximation

We shall now develop some approximations to $p = P\{Y \geq 1\}$. If you are given values for the first k moments of Y, a natural way to approximate $P\{Y \geq 1\}$ is to find a "reasonable" discrete distribution which has the same first k moments as Y and use the probability given by this distribution as your approximation. This may be thought of as a "method of moments" approximation. In this section and the next, we shall present two different "method of moments" approximations. We give these approximations only for the interval case, that is, $Y = Y_I$. The development for the circle case $Y = Y_C$ would be similar.

Our first approximation is based on a simple two-state Markov chain model. This approximation will use only the first two moments of Y_I. For this reason we can compute this approximation even for very large values of m. We shall call this approximation MC2.

Let $w = n - m + 1$ denote the largest possible value of Y_I. In (1) we define Y_I as a sum of w indicator random variables. The approximation MC2 is based on the hope that this sequence of indicators behaves roughly like a two-state Markov chain. Let $\boldsymbol{P}$ be the transition matrix of a two-state Markov chain with off-diagonal entries $p_{01} = a$ and $p_{10} = b$. The stationary distribution for this chain is given by $\pi_0 = b/(a+b)$ and $\pi_1 = a/(a+b)$. Let $Z_1, Z_2, Z_3, \ldots$ be a Markov chain with transition matrix $\boldsymbol{P}$ which is started from the stationary distribution. Define $Y^* = \sum_{i=1}^{w} Z_i$. Routine calculations show that

$$P\{Y^* \geq 1\} \;=\; 1 - (1-\pi)\left(1 - \frac{\pi}{s}\right)^{w-1}, \tag{5}$$

$$EY^* = w\pi\,, \tag{6}$$

$$\operatorname{Var} Y^* = w\pi(1-\pi) + 2\pi(1-\pi)(s-1)(w - s(1-\phi))\,, \tag{7}$$

where for convenience we have introduced the quantities

$$\pi = \pi_1 = \frac{a}{a+b}, \quad s = \frac{1}{a+b}, \quad \text{and} \quad \phi = \left(1 - \frac{1}{s}\right)^w . \tag{8}$$

Setting the right hand sides of (6) and (7) equal to the mean $\mu = \mu_1$ and variance $\sigma^2 = \mu_2 - \mu_1^2$ of Y_I respectively, solving for π and s, and plugging these values into (5) leads to an approximation for p. The quantity ϕ is typically very small and neglecting it has little effect on the answer. If we drop ϕ from (7) it becomes a quadratic equation in s and the system of equations (6) and (7) can be solved in closed form as

$$\pi = \mu/w\,, \quad \text{and} \quad s = \frac{1}{2}\left(w + 1 - \sqrt{(w-1)^2 - 4c}\right) \tag{9}$$

where $c = (\sigma^2 - w\pi(1-\pi))/(2\pi(1-\pi))$. Using these values in (5) gives the approximation we have labeled as MC2 in our tables.

Considering the crudeness of the Markov chain model, the approximation MC2 does remarkably well. It is fairly accurate throughout the range of m, d, n, values we used; it does not seem to do badly anywhere. Glaz (1989) reviews various approximations to $P(m; d, n)$ and makes recommendations concerning their use. We have listed in our tables the three approximations recommended by Glaz. They are the approximations given by Glaz [1989, eq. (3.3)], Naus [1982, eq. (6.1)] and Wallenstein and Neff [1987, eq. (1)] which are labeled as `Glaz`, `Naus` and `WN` in our tables. The approximation MC2 compares well with these. On the whole, MC2 seems to do better than any of the other approximations for larger values of p, and its performance for small p is about as good as any of the others. Unfortunately, it is hard to draw firm conclusions because the values of $P(m; d, n)$ we use in our comparisons are only estimates obtained via simulation; they are (typically) not accurate beyond the third decimal place.

2.3 Compound Poisson Approximations

In this section we present some "method of moments" approximations using compound Poisson distributions. First, we shall define the compound Poisson (CP) distribution and motivate its use in this setting.

Let $Z_1, Z_2, Z_3, \ldots$ be a sequence of independent random variables with $Z_i \sim$ Poisson(λ_i) where $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \lambda_3, \ldots)$ satisfies $\lambda_i \geq 0$ for all i and $\sum_i \lambda_i < \infty$. Then $Y^* = \sum_i i Z_i$ has a compound Poisson distribution denoted by $Y^* \sim \mathrm{CP}(\boldsymbol{\lambda})$.

For points on the interval, the number of clumps may be written as $Y_I = \sum_{i=1}^{n-m+1} I_{B_i}$ where B_i denotes the event that $X_{(i+m-1)} - X_{(i)} \leq d$. When $d \ll m/n$, the probability $\beta \equiv P(B_i)$ is small. If the events B_i were independent, the distribution of Y_I would be approximately a Poisson distribution with a mean of $(n - m + 1)\beta$. Of course, the events are not independent; when $|i - j|$ is small, we expect the events B_i and B_j to be positively correlated, that is, we expect $P(B_i|B_j) > P(B_i)$ or perhaps even $P(B_i|B_j) \gg P(B_i)$. However, if n is large and $m/n \ll 1$, it is intuitively clear that B_i and B_j are very close to being independent whenever $|i - j| \geq m - 1$.

Problems exhibiting short range dependence and long range independence arise frequently in the applied probability literature. Poisson and compound Poisson approximations are frequently useful in the solution of these problems. The book by Aldous (1989) contains many examples and references. The "Poisson clumping heuristic" of Aldous may be applied in our situation. When n is large and $d \ll m/n \ll 1$, we expect the random set $\mathcal{S} = \{i : B_i \text{ occurs}\}$ to consist of groups of nearby values (because of the short range dependence) with the separation between groups determined, at least roughly, by a Poisson process (because of the long range independence). Thus $Y_I = |\mathcal{S}|$ should have approximately a compound Poisson distribution. (For any set A, we use $|A|$ to denote the number of elements in A.) In particular, we expect the distribution of Y_I to be well approximated by $\mathrm{CP}(\boldsymbol{\lambda})$ where λ_k is the expected number of groups of size k in $\mathcal{S}$.

If $Y^* \sim \mathrm{CP}(\boldsymbol{\lambda})$, it is immediate that

$$P\{Y^* \geq 1\} = 1 - \exp\{-\sum_{k=1}^{\infty} \lambda_k\}\,. \tag{10}$$

Our goal is to find $\boldsymbol{\lambda}$ such that the distribution of Y^* is close to that of Y_I, and then use (10) as our approximation to $p = P(m; d, n)$. Formula (10) suffices for the work in this paper. If one is interested in the entire distribution of Y_I, it is natural to use the approximation $P\{Y_I = j\} \approx P\{Y^* = j\}$. In this case, one needs a way to calculate the quantities $p_j = P\{Y^* = j\}$. A convenient way to do this is to start from

$p_0 = \exp(-\sum_i \lambda_i)$ and then for $j \geq 1$ use the recursion

$$jp_j = \sum_{i=1}^{j} (i\lambda_i) p_{j-i} \,. \tag{11}$$

Finding a good choice for $\boldsymbol{\lambda}$ is difficult. Empirical work on this problem has been done by Lin (1993). Some important theoretical work is that of Roos (1993). Theorem G of Roos (1993) and Theorem 1 of Glaz et al. (1994) show that if $n \to \infty$ and $d \to 0$ in such a way that EY_I remains fixed, then there is a sequence of compound Poisson random variables $Y_n^* \sim \mathrm{CP}(\boldsymbol{\lambda}_n)$ such that $\sup_B |P\{Y_I \in B\} - P\{Y_n^* \in B\}| = O(1/n)$. Roos (1993) gives an explicit description of the values $\boldsymbol{\lambda}_n$ used in this theorem. Unfortunately, it seems to be very difficult to compute these values unless m is quite small. Glaz et al. (1994) in their Equations (3.3)–(3.5) suggest a method of approximating $\boldsymbol{\lambda}$ and use this to approximate the distribution of Y_I.

In this paper we take an empirical approach to the choice of $\boldsymbol{\lambda}$. The approximations we have tried are of the following type: Given the first k moments of $Y = Y_I$, we choose $\boldsymbol{\lambda}$ to match these moments, that is, we find $\boldsymbol{\lambda}$ such that $Y^* \sim \mathrm{CP}(\boldsymbol{\lambda})$ has $E[(Y^*)^i] = EY^i$ for $1 \leq i \leq k$. There are many ways to choose $\boldsymbol{\lambda}$ which match the moments. We have worked mainly with two methods of doing this; the resulting approximations for p produced by these two methods we call APk and CPGk respectively. In method APk, we assume that $\lambda_i = 0$ for $i > k$ and we choose $\lambda_1, \lambda_2, \ldots, \lambda_k$ to match the moments. Note that AP1 is just a simple Poisson approximation; it is the same as the Poisson approximation given in (2.10) of Glaz et al. (1994). In method CPGk, we assume that the values λ_i decay geometrically starting with λ_{k-1}, that is $\lambda_i = \lambda_{k-1} r^{i-k+1}$ for $i \geq k$ where r satisfies $0 \leq r < 1$. We then choose $\lambda_1, \ldots, \lambda_{k-1}$ and r to match the given moments. The approximations CPGk are closely related to the compound Poisson approximation studied by Glaz et al. (1994). They propose a specific form for $\boldsymbol{\lambda}$ which has $\lambda_2, \ldots, \lambda_{m-1}$ decaying in a roughly geometric fashion and sets $\lambda_i = 0$ for $i > m-1$.

Let ξ_j denote the j^{th} cumulant of Y. The cumulants ξ_j are easily computed from the moments $\mu_\ell = EY^\ell$ using the relation

$$\mu_{\ell+1} = \sum_{k=0}^{\ell} \binom{\ell}{k} \mu_k \xi_{\ell+1-k} \,. \tag{12}$$

In this formula we take $\mu_0 = 1$. The approximations APk and CPGk are analytically convenient because the cumulants of Y^* are linear in $\boldsymbol{\lambda}$; it is straightforward to show

that if $Y^* \sim \mathrm{CP}(\boldsymbol{\lambda})$, then the cumulants ξ_j^* of Y^* are given by

$$\xi_j^* = \sum_{\ell=1}^{\infty} \ell^j \lambda_\ell. \tag{13}$$

The approximations APk and CPGk are implemented as the following series of steps. First, compute the first k moments of Y. Then use (12) to obtain the first k cumulants of Y. Next, equate the cumulants $\xi_j = \xi_j^*$ for $j = 1, \ldots, k$ and use (13) to solve for $\boldsymbol{\lambda}$. Finally, use (10) to compute the approximation for p.

The only one of these steps which requires elaboration is "solving for $\boldsymbol{\lambda}$". We must solve the system of k equations given by (13) with $j = 1, 2, \ldots, k$. For APk this step is very simple. Let $\boldsymbol{\xi}_k = (\xi_1, \xi_2, \ldots, \xi_k)'$ and $\boldsymbol{\lambda}_k = (\lambda_1, \lambda_2, \ldots, \lambda_k)'$. Define $\boldsymbol{B}_k = (b_{j\ell})$ to be the $k \times k$ matrix with entries $b_{j\ell} = \ell^j$. Then (13) becomes $\boldsymbol{\xi}_k = \boldsymbol{B}_k \boldsymbol{\lambda}_k$ so that $\boldsymbol{\lambda}_k = \boldsymbol{B}_k^{-1} \boldsymbol{\xi}_k$. Sometimes this produces a vector $\boldsymbol{\lambda}_k$ with one or more negative entries λ_i. When this happens, it means there does *not* exist a CP distribution of the assumed form ($\lambda_i = 0$ for $i > k$) with the given values for the first k moments. When $k = 1$, the system of equations (13) reduces to $\xi_1 = \lambda_1$, and (10) then leads to the approximation

$$\mathrm{AP1} = 1 - \exp\{-\xi_1\}\,. \tag{14}$$

For CPGk it is rather more difficult to "solve for $\boldsymbol{\lambda}$". The system of equations we must solve can be formulated in many ways. One convenient way is the following. We solve the system of k nonlinear equations

$$\xi_j = \sum_{\ell=1}^{k-2} \ell^j a_\ell + g_j(r) a_{k-1} \quad (j = 1, 2, \ldots, k) \tag{15}$$

for the k unknowns $a_1, \ldots, a_{k-1}, r$. Here we have introduced the functions

$$g_j(x) = \sum_{\ell=1}^{\infty} \ell^j x^\ell \tag{16}$$

and the auxiliary quantities $a_1, \ldots, a_{k-1}$ which are related to $\lambda_1, \ldots, \lambda_{k-1}$ via

$$\begin{aligned} \lambda_i &= a_i + r^i a_{k-1} \quad \text{for } i = 1, \ldots, k-2\,, \\ \lambda_{k-1} &= r^{k-1} a_{k-1}\,. \end{aligned} \tag{17}$$

The solution obtained is legitimate so long as $\lambda_i \geq 0$ for $i = 1, \ldots, k-1$ and $0 \leq r < 1$. Sometimes there does not exist a CP distribution of the assumed form ($\lambda_i = \lambda_{k-1} r^{i-k+1}$

for $i \geq k$) with the given values for the first k moments. In these cases, attempting to compute the approximation CPGk will lead to a non-legitimate solution of (15) in which at least one of these conditions is violated.

In our work we solve the system (15) using the procedure `fsolve` in the symbolic math package MAPLE. In doing this, it is useful to note that the functions $g_j(x)$ may be computed symbolically from $g_0(x) = x/(1-x)$ and the recursion

$$g_{j+1}(x) = x\frac{\partial}{\partial x} g_j(x)\,. \tag{18}$$

When $k = 2$, various simplifications occur and the solution to (15) has the simple closed form

$$\lambda_1 = \xi_1(1-r)^2 \quad \text{and} \quad r = \frac{(\xi_2/\xi_1) - 1}{(\xi_2/\xi_1) + 1}\,. \tag{19}$$

This solution is legitimate (leads to a valid CP distribution) whenever $\xi_2 \geq \xi_1$. Plugging the solution (19) into (10) leads to the approximation

$$\text{CPG2} = 1 - \exp\left\{-\frac{2\xi_1}{1 + (\xi_2/\xi_1)}\right\}\,. \tag{20}$$

Recall that ξ_1 and ξ_2 are just the mean and variance of Y.

The approximations APk and CPGk were chosen partly on the basis of analytical convenience and partly on the basis of intuition and simulation work. When p is small and m is not too large, we thought it would be very rare for $\mathcal{S}$ (defined in the paragraph before equation (10)) to contain any large groups. This was the motivation underlying the approximation APk; if groups of more than k values in $\mathcal{S}$ are extremely rare, a CP approximation which assumes $\lambda_i = 0$ for $i > k$ ought to do well. In simulations involving larger values of p or m, we observed that large groups did occur in $\mathcal{S}$ and that the frequency distribution of the group sizes seemed to tail off in a roughly geometric fashion. This was the motivation for the approximations CPGk.

In our tables, we list values for the approximations AP1–AP4, CPG2 and CPG4. In some cases, the approximations AP2–AP4 and CPG4 cannot be computed because there exists no CP distribution having the required form and the given moments. When this happens, we place an asterisk in the table. On the whole, the approximations APk did not perform well. The approximation AP1 is always larger than p and is often far from p. However, AP1 is very easily computed, so it might be useful as a rough approximation to p. The approximation AP2 tends to be smaller than p, but is usually

a considerable improvement on AP1. The approximations AP2–AP4 often cannot be computed; they generally fail to exist when m or p is large. However, when AP3 and AP4 do exist, they are usually good approximations for p. Approximations CPG2 and CPG4 seem to be very reliable. CPG2 is very similar to MC2, but MC2 does a little better when n is small or p is close to 1. For small n, there are many cases in which CPG4 cannot be computed. But for large n, we were able to compute CPG4 except in a few cases with p close 1. When CPG4 exists, it is usually better than both CPG2 and MC2. The overall performance of CPG4 also seems to be superior to that of the approximations `Glaz`, `Naus`, and `WN` listed in our tables. However, our comparisons have so far been limited to $m \leq 10$, so we cannot draw any definite conclusions yet.

3 The Moments of Y

3.1 Notation

In our calculation of the moments, we rely extensively on properties of spacings. The spacings $S_1, S_2, \ldots, S_{n+1}$ are simply the lengths of the spaces between consecutive order statistics $X_{(i)}$. More precisely, for n random points on the unit interval (or $n+1$ points on the circle), we define $S_i = X_{(i)} - X_{(i-1)}$ for $1 \leq i \leq n+1$. (Reminder: for points on the interval we are taking $X_{(0)} = 0$ and $X_{(n+1)} = 1$.) Huffer (1988) and Lin (1993) have developed a general approach to calculating probabilities involving several linear combinations of spacings. Lin (1993) has devised algorithms and written computer programs which implement this approach. By re-expressing Y in terms of spacings, we are able to use these results and computer programs to compute the moments of Y.

The following notation is useful for representing sums of spacings: For any $\triangle \subset \{1, 2, \ldots, n+1\}$, define

$$S(\triangle) = \sum_{i \in \triangle} S_i \,.$$

Let δ_i denote the particular subset $\{i+1, i+2, \ldots, i+m-1\}$ so that we may write $X_{(i+m-1)} - X_{(i)} = S(\delta_i)$. As in section 2.3 we define the event $B_i = \{X_{(i+m-1)} - X_{(i)} \leq d\} = \{S(\delta_i) \leq d\}$. Then we have

$$Y = \sum_i I_{B_i}$$

with the range of summation being $1 \leq i \leq n-m+1$ for Y_I, and $1 \leq i \leq n+1$ for Y_C . Now define $P_i = P(B_i)$, $P_{i,j} = P(B_i \cap B_j)$, $P_{i,j,k} = P(B_i \cap B_j \cap B_k)$, etc. The

moments of Y can be easily expressed in terms of these quantities. For example the third moment is just

$$EY^3 = \sum_i \sum_j \sum_k P_{i,j,k} \,. \tag{21}$$

The results of Huffer (1988) and Lin (1993) can be used to evaluate quantities like $P_{i,j,k} = P\{S(\delta_i) \le d, S(\delta_j) \le d, S(\delta_k) \le d\}$. We have used these results to find convenient closed form expressions for $EY = \sum_i P_i$ and $EY^2 = \sum_{i,j} P_{i,j}$. These formulas are given in section 3.2 below. The derivation of these formulas is presented in Appendix B.

For the third and fourth moments we have been less ambitious. Instead of giving general formulas for EY^3 and EY^4, we have devised a procedure which uses the computer programs of Lin (1993) to produce formulas valid only for a given value of m. These formulas consist of piecewise polynomials in the argument d. To illustrate the final product of our methods, in Section 3.4 we give the explicit formulas for EY_I^3 and EY_I^4 when $m = 4$.

Essentially, our approach is to use the programs of Lin (1993) to find exact expressions for the terms $P_{i,j,k}$ occurring in equation (21), and then add these various terms together to obtain an exact expression for EY^3. Implementing this exactly as stated above would be highly inefficient because many of the terms $P_{i,j,k}$ are identical and the computer program would waste time computing the same quantities over and over. The reason so many terms are identical is that the spacings $S_1, S_2, \ldots, S_{n+1}$ are exchangeable random variables. Thus, the joint distribution of $(S(\delta_i), S(\delta_j), S(\delta_k))$ depends on i, j, k only through the pattern of overlaps among the sets $\delta_i, \delta_j, \delta_k$, that is, the distribution depends only on the values $|\delta_i \cap \delta_j|$, $|\delta_i \cap \delta_k|$, $|\delta_j \cap \delta_k|$, and $|\delta_i \cap \delta_j \cap \delta_k|$. After combining identical terms in (21) and the corresponding formula for EY^4, we obtain the "reduced" expressions reported in section 3.3. It is these expressions which are then evaluated by the computer programs. The process of going from the original formula (such as (21)) to the "reduced" formula is basically one of counting identical terms and is accomplished by elementary combinatorics. The derivations for the formulas in section 3.3 are given in Appendix C.

The formulas in Sections 3.2 and 3.3 are complicated, but it is important to note that the complexity does *not* increase with the sample size n. The number of terms in these formulas depends only on the value of m. Thus, the computer time involved in

calculating the moments will be roughly the same for all n.

3.2 Formulas for the First and Second Moments

In this section, we state formulas for the first and second moments of the number of the clumps. The proofs are given in Appendix B.

The following notation will be used in both interval and circular cases. For fixed n and d define

$$G(i) = \sum_{j=0}^{i} \binom{n}{j} d^j (1-d)^{n-j} \tag{22}$$

and

$$F(i,j) = \sum_{k=0}^{i} \sum_{l=0}^{j} \binom{n}{k\ ,\ l} d^{k+l} (1-2d)_+^{n-k-l} \,. \tag{23}$$

The values $G(i)$ are cumulative binomial probabilities, and $F(i,j)$ are cumulative trinomial probabilities.

Interval Case

$$E(Y_I) = (n-m+1)[1-G(m-2)] \,. \tag{24}$$

For $n \geq 2(m-1)$,

$$\begin{aligned} E(Y_I^2) &= E(Y_I) + (n-m+1)(n-m)[1-2G(m-2)] \\ &+ 4 \sum_{i=0}^{m-3} (m-i-2)[n-m-(m-i-1)(m-i-3)/2]G(i) \\ &- 2 \sum_{i=0}^{m-3} \sum_{j=0}^{m-3} [(n-2m+3)-(m-i-3)(m-j-3)]F(i,j) \\ &+ (n-2m+3)(n-2m+2)F(m-2,m-2) \,. \end{aligned} \tag{25}$$

Glaz and Naus (1983) study the number of clumps for random points on the unit interval. They derive the expectation, variance and approximate distribution of Y_I in terms of the quantities P_i and $P_{i,j}$ and give exact expressions for P_i and $P_{i,j}$. The formula they give for EY_I^2 is more complicated than ours and the number of terms in their formula grows with the sample size n. This makes their formula difficult to use for large n.

Circular Case

$$E(Y_C) = (n+1)[1 - G(m-2)]\,.$$

For $n \geq 2(m-2)$,

$$\begin{aligned} E(Y_C^2) &= E(Y_C) + n(n+1)[1 - 2G(m-2)] \\ &+ 4(n+1)\sum_{i=0}^{m-3}(m-i-2)G(i) - 2(n+1)\sum_{i=0}^{m-3}\sum_{j=0}^{m-3}F(i,j) \\ &+ (n+1)(n-2m+4)F(m-2,m-2)\,. \end{aligned} \tag{26}$$

3.3 Expressions for the Third and Fourth Moments

Interval Case

The expressions for $E(Y_I^3)$ and $E(Y_I^4)$ given below are valid for all n so long as empty sums (where the lower limit exceeds the upper limit) are taken to be zero, and binomial coefficients $\binom{a}{b}$ are defined to be zero whenever $a < b$. We also use the notation $(x)_+ = \max(x, 0)$.

$$\begin{aligned} E(Y_I^3) &= E(Y_I) + 6\sum_{i=1}^{m-2}(n-m-i+1)_+ P_{0,i} + 6\binom{n-2m+3}{2}P_{0,m-1} \\ &+ 6\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}(n-m-j+1)_+ P_{0,i,j} + 12\sum_{i=1}^{m-2}\binom{n-2m-i+3}{2}P_{0,i,i+m-1} \\ &+ 6\binom{n-3m+5}{3}P_{0,m-1,2m-2}\,. \end{aligned} \tag{27}$$

$$\begin{aligned} E(Y_I^4) &= E(Y_I) + 14\sum_{i=1}^{m-2}(n-m-i+1)_+ P_{0,i} + 14\binom{n-2m+3}{2}P_{0,m-1} \\ &+ 36\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}(n-m-j+1)_+ P_{0,i,j} + 72\sum_{i=1}^{m-2}\binom{n-2m-i+3}{2}P_{0,i,i+m-1} \\ &+ 36\binom{n-3m+5}{3}P_{0,m-1,2m-2} + 24\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}\sum_{k=j+1}^{j+m-2}(n-m+1-k)_+ P_{0,i,j,k} \\ &+ 48\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}\binom{n-2m-j+3}{2}P_{0,i,j,j+m-1} \\ &+ 72\sum_{i=1}^{m-2}\binom{n-3m-i+5}{3}P_{0,i,i+m-1,i+2m-2} \end{aligned}$$

$$+ 24 \sum_{i=1}^{m-2} \sum_{j=1}^{m-2} \binom{n-2m-i-j+3}{2} P_{0,i,i+m-1,i+m-1+j}$$
$$+ 24 \binom{n-4m+7}{4} P_{0,m-1,2m-2,3m-3} \,. \tag{28}$$

Circular Case

In the circular case, it is possible for the spacings involved in quantities like $P_{i,j,k}$ and $P_{i,j,k,\ell}$ to "wrap around" the circle. The formulas we give below are valid provided there is no "wrapping around" and the conditions on n ensure this does not occur.

For $n \geq 3(m-2)$,

$$E(Y_C^3) = E(Y_C) + 6(n+1) \sum_{i=1}^{m-2} P_{0,i} + 3(n+1)(n-2m+4) P_{0,m-1}$$
$$+ 6(n+1) \sum_{i=1}^{m-2} \sum_{j=i+1}^{i+m-2} P_{0,i,j} + 6(n+1) \sum_{i=1}^{m-2} (n-2m+4-i) P_{0,i,i+m-1}$$
$$+ (n+1)(n-3m+6)(n-3m+5) P_{0,m-1,2m-2} \,. \tag{29}$$

For $n \geq 4(m-2)$,

$$E(Y_C^4) = E(Y_C) + 14(n+1) \sum_{i=1}^{m-2} P_{0,i} + 7(n+1)(n-2m+4) P_{0,m-1}$$
$$+ 36(n+1) \sum_{i=1}^{m-2} \sum_{j=i+1}^{i+m-2} P_{0,i,j} + 36(n+1) \sum_{i=1}^{m-2} (n-2m+4-i) P_{0,i,i+m-1}$$
$$+ 6(n+1)(n-3m-6)(n-3m+5) P_{0,m-1,2m-2}$$
$$+ 24(n+1) \sum_{i=1}^{m-2} \sum_{j=i+1}^{i+m-2} \sum_{k=j+1}^{j+m-2} P_{0,i,j,k}$$
$$+ 24(n+1) \sum_{i=1}^{m-2} \sum_{j=i+1}^{i+m-2} (n-2m+4-j) P_{0,i,j,j+m-1}$$
$$+ 12(n+1) \sum_{i=1}^{m-2} (n-3m+6-i)(n-3m+5-i) P_{0,i,i+m-1,i+2m-2}$$
$$+ 12(n+1) \sum_{i=1}^{m-2} \sum_{j=1}^{m-2} (n-2m+4-i-j) P_{0,i,i+m-1,i+m-1+j}$$
$$+ (n+1)(n-4m+8)(n-4m+7)(n-4m+6) P_{0,m-1,2m-2,3m-3} \,. \tag{30}$$

3.4 Example

To illustrate the results of the preceding section, we shall give explicit formulas for EY_I^3 and EY_I^4 when $m = 4$. These formulas are valid for all n and d. The formulas

are written in terms of functions b and R which are defined for integers j, $k \geq 0$ by

$$b(j,k) = \begin{cases} \binom{n-j}{k} & \text{for } n \geq j+k\,, \\ 0 & \text{for } n < j+k\,. \end{cases} \tag{31}$$

and

$$R(j,k) = \begin{cases} \binom{n}{j} d^j (1-kd)^{n-j} & \text{for } kd < 1\,, \\ 0 & \text{for } kd \geq 1\,. \end{cases} \tag{32}$$

Substituting the value $m = 4$ into (27), we find that $E(Y_I^3)$ is equal to the following sum:

$$\begin{aligned} & b(3,1)P_0 + 6b(4,1)P_{0,1} + 6b(5,1)P_{0,2} + 6b(5,2)P_{0,3} + 6b(5,1)P_{0,1,2} + 6b(6,1)P_{0,1,3} \\ + \quad & 6b(6,1)P_{0,2,3} + 6b(7,1)P_{0,2,4} + 12b(6,2)P_{0,1,4} + 12b(7,2)P_{0,2,5} + 6b(7,3)P_{0,3,6}. \end{aligned}$$

The program of Lin (1993) has been designed so that it can be conveniently used to evaluate sums of this form. It can also handle a broad array of other problems involving linear combinations of spacings. The output of Lin's program is an algebraic expression written in a form suitable for reading by the symbolic math package MAPLE. In this example we obtain the following expression for $E(Y_I^3)$:

```
 (+6*b(7,3)+30*b(7,2)+61*b(7,1)+37*b(7,0)+19*b(6,0)+7*b(5,0)+1*b(4,0))
+(-18*b(7,3)-12*b(7,2)+29*b(7,1)+23*b(7,0)-1*b(6,0)-13*b(5,0)
                                              -1*b(4,0))*R(0,1)
+(-18*b(7,3)-60*b(7,2)-73*b(7,1)-31*b(7,0)-7*b(6,0)-1*b(5,0)
                                              -1*b(4,0))*R(1,1)
+(-18*b(7,3)-84*b(7,2)-157*b(7,1)-91*b(7,0)-43*b(6,0)-13*b(5,0)
                                              -1*b(4,0))*R(2,1)
+(+18*b(7,3)-66*b(7,2)-54*b(7,1)-60*b(7,0)-18*b(6,0)+6*b(5,0))*R(0,2)
+(+36*b(7,3)-36*b(7,2)-78*b(7,1)-66*b(7,0)-30*b(6,0))*R(1,2)
+(+72*b(7,3)+72*b(7,2)+18*b(7,1)-18*b(7,0)-18*b(6,0))*R(2,2)
+(+108*b(7,3)+252*b(7,2)+174*b(7,1)+48*b(7,0))*R(3,2)
+(+108*b(7,3)+324*b(7,2)+282*b(7,1)+72*b(7,0))*R(4,2)
+(-6*b(7,3)+48*b(7,2)-36*b(7,1))*R(0,3)
+(-18*b(7,3)+96*b(7,2)-36*b(7,1))*R(1,3)
+(-54*b(7,3)+168*b(7,2)-24*b(7,1))*R(2,3)
```

```
+(-144*b(7,3)+216*b(7,2))*R(3,3)
+(-324*b(7,3)+144*b(7,2))*R(4,3)
+(-540*b(7,3))*R(5,3)
+(-540*b(7,3))*R(6,3)
```

This expression can now be numerically evaluated by MAPLE for various values of n and d. Some simplifications of this formula are possible. For example, the leading term

```
(+6*b(7,3)+30*b(7,2)+61*b(7,1)+37*b(7,0)+19*b(6,0)+7*b(5,0)+1*b(4,0))
```

can be shown to be simply $(n-3)_+^3$. However, we do not know a good general way to recognize such simplifications.

The fourth moment is obtained in a similar fashion. Substituting $m = 4$ into (28), we find that $E(Y_I^4)$ is equal to

$$\begin{aligned}
& b(3,1)P_0 + 14b(4,1)P_{0,1} + 14b(5,1)P_{0,2} + 14b(5,2)P_{0,3} \\
+ \;& 36b(5,1)P_{0,1,2} + 36b(6,1)P_{0,1,3} + 36b(6,1)P_{0,2,3} + 36b(7,1)P_{0,2,4} \\
+ \;& 72b(6,2)P_{0,1,4} + 72b(7,2)P_{0,2,5} + 36b(7,3)P_{0,3,6} \\
+ \;& 24b(6,1)P_{0,1,2,3} + 24b(7,1)P_{0,1,2,4} + 24b(7,1)P_{0,1,3,4} + 24b(8,1)P_{0,1,3,5} \\
+ \;& 24b(7,1)P_{0,2,3,4} + 24b(8,1)P_{0,2,3,5} + 24b(8,1)P_{0,2,4,5} + 24b(9,1)P_{0,2,4,6} \\
+ \;& 48b(7,2)P_{0,1,2,5} + 48b(8,2)P_{0,1,3,6} + 48b(8,2)P_{0,2,3,6} + 48b(9,2)P_{0,2,4,7} \\
+ \;& 72b(8,3)P_{0,1,4,7} + 72b(9,3)P_{0,2,5,8} + 24b(7,2)P_{0,1,4,5} + 24b(8,2)P_{0,1,4,6} \\
+ \;& 24b(8,2)P_{0,2,5,6} + 24b(9,2)P_{0,2,5,7} + 24b(9,4)P_{0,3,6,9}\,.
\end{aligned}$$

which can be re-expressed as the following algebraic expression:

```
  (+24*b(9,4)+180*b(9,3)+590*b(9,2)+1105*b(9,1)+671*b(9,0)
      +369*b(8,0)+175*b(7,0)+65*b(6,0)+15*b(5,0)+1*b(4,0))
 +(-96*b(9,4)-252*b(9,3)+92*b(9,2)+923*b(9,1)+675*b(9,0)
      +367*b(8,0)+95*b(7,0)-45*b(6,0)-29*b(5,0)-1*b(4,0))*R(0,1)
 +(-96*b(9,4)-540*b(9,3)-1252*b(9,2)-1521*b(9,1)-713*b(9,0)
      -253*b(8,0)-45*b(7,0)+7*b(6,0)-1*b(5,0)-1*b(4,0))*R(1,1)
 +(-96*b(9,4)-684*b(9,3)-2116*b(9,2)-3709*b(9,1)-2181*b(9,0)
      -1145*b(8,0)-505*b(7,0)-165*b(6,0)-29*b(5,0)-1*b(4,0))*R(2,1)
 +(+144*b(9,4)-324*b(9,3)-1306*b(9,2)-864*b(9,1)-434*b(9,0)
```

```
     -616*b(8,0)-270*b(7,0)-20*b(6,0)+14*b(5,0))*R(0,2)
+(+288*b(9,4)+216*b(9,3)-1220*b(9,2)-2596*b(9,1)-1448*b(9,0)
     -852*b(8,0)-400*b(7,0)-92*b(6,0))*R(1,2)
+(+576*b(9,4)+1728*b(9,3)+1496*b(9,2)-56*b(9,1)-400*b(9,0)
     -168*b(8,0)-200*b(7,0)-64*b(6,0))*R(2,2)
+(+864*b(9,4)+4104*b(9,3)+7668*b(9,2)+7164*b(9,1)
     +2736*b(9,0)+684*b(8,0)+156*b(7,0))*R(3,2)
+(+864*b(9,4)+4968*b(9,3)+11556*b(9,2)+13812*b(9,1)
     +6360*b(9,0)+2148*b(8,0)+348*b(7,0))*R(4,2)
+(-96*b(9,4)+684*b(9,3)-24*b(9,2)-876*b(9,1)-936*b(9,0)
     -120*b(8,0))*R(0,3)
+(-288*b(9,4)+1188*b(9,3)+1176*b(9,2)-180*b(9,1)-744*b(9,0)
     -120*b(8,0))*R(1,3)
+(-864*b(9,4)+1404*b(9,3)+3528*b(9,2)+2124*b(9,1)-144*b(8,0))*R(2,3)
+(-2304*b(9,4)-864*b(9,3)+4512*b(9,2)+5088*b(9,1)+1392*b(9,0))*R(3,3)
+(-5184*b(9,4)-9720*b(9,3)-3360*b(9,2)+2568*b(9,1)+1344*b(9,0))*R(4,3)
+(-8640*b(9,4)-24840*b(9,3)-20160*b(9,2)-5640*b(9,1))*R(5,3)
+(-8640*b(9,4)-29160*b(9,3)-27360*b(9,2)-7560*b(9,1))*R(6,3)
+(+24*b(9,4)-288*b(9,3)+648*b(9,2)-288*b(9,1)+24*b(9,0))*R(0,4)
+(+96*b(9,4)-864*b(9,3)+1296*b(9,2)-288*b(9,1))*R(1,4)
+(+384*b(9,4)-2448*b(9,3)+2112*b(9,2)-144*b(9,1))*R(2,4)
+(+1440*b(9,4)-6048*b(9,3)+2448*b(9,2))*R(3,4)
+(+4896*b(9,4)-12096*b(9,3)+1584*b(9,2))*R(4,4)
+(+14400*b(9,4)-17280*b(9,3))*R(5,4)
+(+34560*b(9,4)-12960*b(9,3))*R(6,4)
+(+60480*b(9,4))*R(7,4)
+(+60480*b(9,4))*R(8,4) .
```

This expression appears rather bulky, but it is easily handled by MAPLE, and computations using this formula are fast and accurate.

REFERENCES

Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York.

Berman, M., and Eagleson, G. K. (1985). A useful upper bound for the tail probabilities of the scan statistic when the sample size is large. *J. Amer. Statist. Assoc.* **80**, 886–889.

Cressie, N. (1980). The asymptotic distribution of the scan statistic under uniformity. *Ann. Prob.* **8**, 828–840.

Dawson, D. A., and Sankoff, D. (1967) An inequality for probabilities. *Proceedings of the American Mathematical Society.* **18**, 504–507.

Dembo, A., and Karlin, S. (1992). Poisson approximation for r-scan processes. *Ann. Appl. Prob.* **2**, 329–357.

Gates, D. J., and Westcott, M. (1984). On the distributions of scan statistics. *J. Amer. Statist. Assoc.* **79**, 423–429.

Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic. *J. Amer. Statist. Assoc.* **84**, 560–566.

Glaz, J. (1992). Approximations for tail probabilities and moments of the scan statistic. *Comput. Statist. Data Anal.* **14**, 213–227.

Glaz, J. (1993). Approximations for the tail probabilities and moments of the scan statistic. *Statist. in Med.* **12**, 1845–1852.

Glaz, J., and Naus, J. (1983). Multiple clusters on the line. *Commun. Statis.–Theor. Meth.* **12**, 1961–1986.

Glaz, J., Naus, J., Roos, M., and Wallenstein, S. (1994). Poisson approximations for the distribution and moments of ordered m-spacings. *J. Appl. Prob.* **31A**, 271–281.

Huffer, F. (1988). Divided differences and the joint distribution of linear combinations of spacings. *J. Appl. Prob.* **25**, 346–354.

Krauth, J. (1991). Lower bounds for the tail probabilities of the scan statistic. In *Classif., Data Anal., Knowledge Organiz: Models Meth. with Appl.* (Hans-Hermann Bock and Peter Ihm, eds.) Springer-Verlag, New York. 61–67.

Kwerel, S. M. (1975). Most stringent bounds on aggregated probabilities of partially specified dependent probability systems. *J. Amer. Statist. Assoc.* **70**, 472–479.

Lin, C. T. (1993). The computation of probabilities which involve spacings, with applications to the scan statistic. Ph.D. Dissertation, Department of Statistics, Florida State University.

Loader, C. R. (1991). Large-deviation approximations to the distribution of scan statistics. *Adv. Appl. Prob.* **23**, 751–771.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line. *J. Amer. Statist. Assoc.* **60**, 532–538.

—— (1966). A power comparison of two tests of nonrandom clustering. *Technometrics* **8**, 493–517.

—— (1979). An indexed bibliography of clusters, clumps and coincidences. *Internat. Statist. Rev.* **47**, 47–78.

—— (1982). Approximations for distributions of scan statistics. *J. Amer. Statist. Assoc.* **77**, 177–183.

Neff, N. D., and Naus, J. I. (1980). The distribution of the size of the maximum cluster of points on a line. *IMS Series of Selected Tables in Mathematical Statistics* **6**. Providence, RI: American Mathematical Society.

Newell, G. F. (1963). Distribution for the smallest distance between any pair of kth nearest neighbor random points on a line. *Procceedings of Symposium on Time Series Analysis*, Brown University, John Wiley & Sons, New York, 89–103.

Prékopa, A. (1988). Boole-Bonferroni inequalities and linear programming. *Operations Research* **36**, 145–162.

Roos, M. (1993). Compound Poisson approximations for the number of extreme spacings. *Adv. Appl. Prob.* **25**, 847–874.

Wallenstein, S., and Neff, N. (1987). An approximation for the distribution of the scan statistic. *Statistics in Medicine* **6**, 197–207.

Appendices

A Tables

Tables 1, 2, 4, 6, 8 give examples of all the bounds and approximations investigated in this paper. Recall that LB and UB are defined in Section 2.1, MC2 is defined in Section 2.2, and AP1–AP4, CPG2 and CPG4 are defined Section 2.3. These tables consider a variety of sample sizes n, but are restricted to $m \leq 10$. For small values of n (say, $n \leq 20$) there is no need for approximation; exact values of $P(m; d, n)$ may be computed or found in the tables of Neff and Naus (1980). We have included Table 1, which uses $n = 20$, in order to have one table in which our approximations are compared with the exact values. In this table the numbers in the column labeled `Prob` are the exact values of $P(m; d, n)$ computed using the programs of Lin (1993). In Tables 2, 4, 6, 8 the values in the `Prob` column are estimates based on simulations with 1,000,000 trials. There are two exceptional cases (indicated by †) in Table 2 where the values given are exact.

Tables 3, 5, 7, 9 are (for the most part) excerpts from the tables of Glaz (1989). Glaz (1992) refines the methods of Glaz (1989), and we have used values from this later paper when they are available. In Table 5 the values marked ‡ are from Glaz (1992). The column labeled `GCP` in Table 5 is taken from Glaz et al. (1994) and gives values for the compound Poisson approximation described in Equations (3.3)–(3.5) of this paper. The columns labeled `Glaz`, `Naus`, `WN` are the approximations given by Glaz [1989, eq. (3.3)], Naus [1982, eq. (6.1)] and Wallenstein and Neff [1987, eq. (1)]. The lower bound `KLB` and upper bound `GUB` are taken from Equations (2.12) and (2.9) of Glaz (1989). (Glaz (1989) only gives `KLB` for $n = 25$, so this column is missing from most of our tables.) The values in column `Prob` are the same values used in our Tables 2, 4, 6, 8. For easy comparison, we have chosen the same values of n, m, d used by Glaz (1989), and have constructed Tables 2, 4, 6, 8 so that they exactly parallel Tables 3, 5, 7, 9.

The approximations MC2 and CPG2 use only the first two moments of Y_I and can easily be computed for larger values of m. In Tables 10, 11, 12, 13, we study the

performance of MC2 and CPG2 for $m > 10$. For comparison we include in our tables the values of `Glaz`, `Naus`, `WN`, `KLB`, `GUB` described above. These values were taken from Glaz (1989). In Tables 10, 11, 12, 13, the values of `Prob` were also taken from Glaz (1989); they are estimates from simulations with 20,000 trials.

Table 1: Approximations to $P(m; d, n)$ for $n = 20$

m	d	Prob	AP1	AP2	CPG2	MC2	AP3	AP4	CPG4	LB	UB
3	.005	.07428	.07737	.07420	.07433	.07432	.07428	.07428	.07428	.07428	.07430
3	.01	.24801	.26175	.24788	.24869	.24845	.24797	$*$	$*$	.24788	.24894
3	.05	.99619	.99139	$*$	$*$	.99637	$*$	$*$	$*$	.96758	> 1
4	.01	.01567	.01692	.01558	.01568	.01568	.01568	.01567	.01567	.01567	.01568
4	.05	.65524	.72286	.65652	.66692	.66179	$*$	$*$	$*$	.62642	.70729
4	.10	.99770	.99588	$*$	$*$	.99789	$*$	$*$	$*$	.97189	> 1
5	.01	.00064	.00068	.00064	.00064	.00064	.00064	.00064	.00064	.00064	.00064
5	.05	.17752	.22464	.16585	.17934	.17868	.17919	.17753	.17772	.17589	.18360
5	.10	.80160	.88084	.80573	.82268	.81166	$*$	$*$	$*$	.73514	.92046
6	.05	.02984	.03787	.02778	.02994	.02991	.03028	.02976	.02985	.02972	.03023
6	.10	.35340	.47671	.30464	.36241	.35803	.35877	.35743	.35746	.33719	.37929
6	.15	.83405	.92210	.83407	.86040	.84419	$*$	$*$	$*$	.76148	.99280
7	.05	.00377	.00460	.00359	.00377	.00377	.00380	.00376	.00377	.00376	.00379
7	.10	.10091	.14576	.08109	.10208	.10159	$*$	$*$	.10135	.09907	.10544
7	.15	.43985	.61027	$*$	.45594	.44654	.44683	$*$	$*$	.40217	.48870
7	.20	.83381	.93550	.82161	.86455	.84238	$*$	$*$	$*$	.75603	.99336
8	.05	.00038	.00044	.00036	.00038	.00038	.00038	.00038	.00038	.00038	.00038
8	.10	.02182	.03054	.01846	.02193	.02188	$*$	$*$	.02186	.02161	.02256
8	.15	.15982	.24810	$*$	.16315	.16146	$*$	$*$	.16215	.15304	.17047
8	.20	.47596	.67600	$*$	.49725	.48287	$*$	$*$	$*$	.42550	.53459
9	.10	.00376	.00498	.00337	.00376	.00376	.00386	$*$	.00376	.00373	.00382
9	.15	.04479	.06859	$*$	.04528	.04507	$*$	$*$	.04506	.04387	.04679
9	.20	.19701	.32003	$*$	.20289	.19952	$*$	$*$	.20303	.18527	.21412
9	.25	.48419	.70528	$*$	.50940	.49047	$*$	$*$	$*$	.43407	.54629
10	.10	.00052	.00066	.00049	.00052	.00052	.00053	.00052	.00052	.00052	.00053
10	.15	.01008	.01451	.00822	.01013	.01011	$*$	$*$	.01010	.00997	.01042
10	.20	.06410	.10399	$*$	.06523	.06469	$*$	$*$	.06486	.06191	.06748
10	.25	.21622	.36248	$*$	.22464	.21946	$*$	$*$	$*$	.20047	.23671

Note: P(m,d,n) (Prob) is exact.

$*$ indicates the approximation is not defined for this case.

Table 2: Approximations to $P(m; d, n)$ for $n = 25$

m	d	Prob	AP1	AP2	CPG2	MC2	AP3	AP4	CPG4	LB	UB
3	.01	.42507†	.44703	.42539	.42673	.42613	$*$	$*$	$*$	.42326	.43311
4	.01	.03825	.04201	.03780	.03819	.03818	.03817	.03815	.03815	.03814	.03819
4	.05	.90919	.93897	.91691	.91942	.91499	$*$	$*$	$*$	.84476	> 1
5	.05	.41014	.51125	.38308	.41736	.41445	.41103	$*$	$*$	.39565	.43833
5	.10	.98544	.99302	.99010	.99032	.98791	$*$	$*$	$*$	.93264	> 1
6	.05	.09916	.13351	.08743	.09990	.09968	.10246	$*$	.09941	.09817	.10317
6	.10	.72834	.85912	.68723	.75167	.73965	.70834	$*$	$*$	.65721	.86571
7	.05	.01714	.02278	.01562	.01735	.01734	.01780	$*$	.01734	.01722	.01772
7	.20	.99811	.99931	.99913	.99914	.99814	$*$	$*$	$*$	.96774	> 1
8	.10	.10132	.15682	$*$	.10239	.10193	$*$	$*$	.10222	.09770	.10819
9	.10	.02527	.03771	$*$	.02521	.02516	$*$	$*$	.02523	.02464	.02631
10	.20	.31509†	.52687	$*$	.32620	.31893	$*$	$*$	$*$	.27775	.36342

Note: P(m,d,n) (Prob) was estimated from 1,000,000 trials except for the values indicated by † which are exact. $*$ indicates the approximation is not defined for this case.

Table 3: Approximations to $P(m; d, n)$ for $n = 25$

m	d	Prob	Glaz	Naus	WN	KLB	GUB
3	.01	.42507†	.424	.402	.509	.379	.510
4	.01	.03825	.038	.038	.039	.038	.039
4	.05	.90919	.879	.827	> 1	.697	> 1
5	.05	.41014	.391	.377	.455	.314	.456
5	.10	.98544	.953	.925	> 1	.815	> 1
6	.05	.09916	.097	.096	.101	.081	.101
6	.10	.72834	.667	.646	.880	.527	.894
7	.05	.01714	.017	.017	.017	.015	.017
7	.20	.99811	.969	.970	> 1	.875	> 1
8	.10	.10132	.099	.099	.102	.076	.102
9	.10	.02527	.025	.025	.025	.019	.025
10	.20	.31509†	.298	.299	.317	.216	.324

These approximations and bounds are Glaz (1989, eq. 3.3), Naus (1982, eq. 6.1), Wallenstein and Neff (WN) (1987, eq. 1), Glaz (KLB) (1989, eq. 2.12), and Glaz (GUB) (1989, eq. 2.9). All the values are taken from Glaz's (1989) Table 1.

Table 4: Approximations to $P(m;d,n)$ for $n = 100$

m	d	Prob	AP1	AP2	CPG2	MC2	AP3	AP4	CPG4	LB	UB
3	.001	.35218	.36525	.35139	.35203	.35196	.35180	.35180	.35180	.35119	.35554
3	.005	.99977	.99985	.99982	.99982	.99981	$*$	$*$	$*$	.97852	> 1
4	.001	.01373	.01448	.01378	.01381	.01381	.01382	.01381	.01381	.01381	.01382
4	.005	.68032	.74538	.67054	.68372	.68287	.68127	.68092	.68093	.63779	.77320
4	.01	.99823	.99955	.99838	.99864	.99851	$*$	$*$	$*$	.95941	> 1
5	.005	.12481	.14840	.12020	.12502	.12499	.12575	.12479	.12493	.12433	.12728
5	.01	.72584	.82863	.68908	.73249	.73079	.73218	.72763	.72810	.65097	.88299
6	.01	.21309	.27825	.18820	.21311	.21293	$*$	$*$	.21296	.20797	.22548
7	.01	.03762	.04900	.03339	.03723	.03723	.03856	$*$	.03734	.03703	.03868
8	.01	.00523	.00659	.00475	.00515	.00515	.00529	$*$	.00517	.00514	.00529
9	.05	.99744	.99999	$*$	.99899	.99846	$*$	$*$	$*$	.94566	> 1
10	.05	.92437	.99679	$*$	.93914	.93303	$*$	$*$	$*$	.80081	> 1

Note: P(m,d,n) (Prob) was estimated from 1,000,000 trials.

$*$ indicates the approximation is not defined for this case.

Table 5: Approximations to $P(m;d,n)$ for $n = 100$

m	d	Prob	Glaz	Naus	WN	GUB	GCP
3	.001	.35218	.351	.347	.426	.426	
3	.005	.99977	.999	.998	> 1	> 1	
4	.001	.01373	.014	.014	.014	.014	
4	.005	.68032	.670	.657	> 1	> 1	
4	.01	.99823	.997	.992	> 1	> 1	.9967
5	.005	.12481	.124	.124	.131	.132	
5	.01	.72584	.710‡	.694	> 1	> 1‡	.7144
6	.01	.21309	.210‡	.208	.232	.232‡	.2123
7	.01	.03762	.037‡	.037	.038	.038‡	.0376
8	.01	.00523	.0052‡			.0052‡	.0052
9	.05	.99744	.983	.979	> 1	> 1	
10	.05	.92437	.888‡	.863	> 1	> 1	

All the approximated values are taken from Glaz's (1989) Table 2 except GCP which is taken from Table 1 (3.4) of Glaz ed al. (1994) and values indicated ‡ which are taken from Glaz's (1992) Table 1.

Table 6: Approximations to $P(m;d,n)$ for $n = 500$

m	d	Prob	AP1	AP2	CPG2	MC2	AP3	AP4	CPG4	LB	UB
4	.001	.99735	.99919	.99682	.99745	.99742	.99734	.99733	.99733	.95048	> 1
5	.001	.50888	.57725	.49307	.50886	.50875	.51141	.50812	.50863	.49128	.55021
6	.001	.06873	.08042	.06604	.06830	.06830	.06880	.06829	.06837	.06820	.06915
8	.005	.97756	.99897	$*$	.97876	.97825	$*$	$*$	.98007	.87547	> 1
9	.005	.68108	.86927	$*$	.67903	.67836	$*$	$*$	.68584	.57121	.94967
10	.005	.27074	.41673	$*$	.26699	.26688	$*$	$*$	.27148	.23490	.32096

Note: P(m,d,n) (Prob) was estimated from 1,000,000 trials.

$*$ indicates the approximation is not defined for this case.

Table 7: Approximations to $P(m;d,n)$ for $n = 500$

m	d	Prob	Glaz	Naus	WN	GUB
4	.001	.99735	.997	.996	> 1	> 1
5	.001	.50888	.505	.503	.700	.700
6	.001	.06873	.069	.068	.071	.071
8	.005	.97756	.970	.968	> 1	> 1
9	.005	.68108	.670	.665	> 1	> 1
10	.005	.27074	.268	.266	.309	.311

All the approximated values are taken from Glaz's (1989) Table 3.

Table 8: Approximations to $P(m;d,n)$ for $n = 1000$

m	d	Prob	AP1	AP2	CPG2	MC2	AP3	AP4	CPG4	LB	UB
6	.001	.92690	.97318	.89412	.92743	.92724	$*$	$*$	.92735	.82754	> 1
7	.001	.35264	.44264	.31704	.35192	.35188	$*$	$*$	.35298	.33922	.38489
8	.001	.06081	.07817	.05504	.06049	.06049	.06256	$*$	.06077	.06026	.06302
9	.001	.00767	.00990	.00727	.00783	.00783	.00803	$*$	.00786	.00782	.00804

Note: P(m,d,n) (Prob) was estimated from 1,000,000 trials.

$*$ indicates the approximation is not defined for this case.

Table 9: Approximations to $P(m; d, n)$ for $n = 1000$

m	d	Prob	Glaz	Naus	WN	GUB
6	.001	.92690	.923	.921	> 1	> 1
7	.001	.35264	.351	.341	.432	.432
8	.001	.06081	.061	.061	.063	.063
9	.001	.00767	.0079	.0078	.0079	.0079

All the approximated values are taken from Glaz's (1989) Table 4.

Table 10: Approximations to $P(m; d, n)$ for $n = 25$

m	d	Prob	CPG2	MC2	Glaz	Naus	WN	KLB	GUB
12	.20	.043	.04335	.04308	.043	.043	.043	.031	.043

Note: Prob, Glaz, Naus, WN, KLB, GUB are taken from Glaz's (1989) Table 1.

Table 11: Approximations to $P(m; d, n)$ for $n = 100$

m	d	Prob	CPG2	MC2	Glaz	Naus	WN	GUB
12	.05	.353	.35903	.35712	.345	.338	.399	.413
14	.05	.060	.05761	.05756	.059	.058	.060	.061
14	.10	.999	.99983	.99953	.984	.981	> 1	> 1
16	.10	.858	.88372	.87070	.800	.783	> 1	> 1
18	.10	.408	.41394	.40925	.401	.383	.449	.494
20	.10	.116	.11460	.11420	.121	.115	.120	.128
22	.10	.025	.02307	.02304	.025	.024	.024	.025
26	.20	.900	.92391	.90099	.839	.830	> 1	> 1
28	.20	.585	.59467	.57744	.569	.542	.619	.775

Note: Prob, Glaz, Naus, WN, GUB are taken from Glaz's (1989) Table 2.

Table 12: Approximations to $P(m; d, n)$ for $n = 500$

m	d	Prob	CPG2	MC2	Glaz	Naus	WN	GUB
11	.01	.998	.99870	.99854	.996	.995	> 1	> 1
12	.005	.019	.01753	.01753	.018	.018	.018	.018
12	.01	.935	.93285	.93142	.918	.910	> 1	> 1
13	.01	.665	.65461	.65349	.651	.640	> 1	> 1
14	.01	.341	.32922	.32893	.336	.329	.397	.408
15	.01	.135	.13208	.13204	.137	.135	.144	.148
16	.01	.048	.04611	.04610	.048	.048	.049	.049

Note: Prob, Glaz, Naus, WN, GUB are taken from Glaz's (1989) Table 3.

Table 13: Approximations to $P(m; d, n)$ for $n = 1000$

m	d	Prob	CPG2	MC2	Glaz	Naus	WN	GUB
12	.005	.996	.99603	.99585	.994	.992	> 1	> 1
13	.005	.891	.88777	.88700	.885	.877	> 1	> 1
14	.005	.575	.56259	.56218	.572	.562	.822	.846
15	.005	.267	.25601	.25593	.266	.261	.302	.308
16	.005	.098	.09471	.09470	.099	.098	.103	.104
17	.005	.032	.03095	.03095	.033	.032	.033	.034
18	.005	.010	.00930	.00930	.0097	.0096	.0097	.0098
19	.005	.0025	.00261	.00261	.0027	.0027	.0027	.0027

Note: Prob, Glaz, Naus, WN, GUB are taken from Glaz's (1989) Table 4.

B Derivation of the First and Second Moments

Before beginning the proofs, we need some preliminaries. First, we note the following facts:

$$\begin{aligned} P_i &= P_0 \quad \text{for all } i\,, \\ P_{i,i+j} &= P_{0,j} \quad \text{for all } i,j\,, \\ P_{i,i+j} &= P_{0,m-1} \quad \text{for } j \geq m-1\,. \end{aligned} \tag{33}$$

Here i and j are positive integers which satisfy certain obvious restrictions, for example, for random points on the interval, $P_{i,j}$ is not defined except when both i and j are less than $n-m+3$. These facts are simple consequences of the exchangeability of the spacings. In particular, the condition $j \geq m-1$ in the third fact ensures that B_i and B_{i+j} involve disjoint sets of spacings.

It is also convenient to introduce the auxiliary quantity $Q(i,j,k)$ defined as follows. Let $\triangle_1, \triangle_2, \triangle_3$ be disjoint subsets of $\{1,2,\ldots,n+1\}$ having sizes $|\triangle_1| = i, |\triangle_2| = j, |\triangle_3| = k$. Define

$$Q(i,j,k) = P\{S(\triangle_1) + S(\triangle_3) > d, S(\triangle_2) + S(\triangle_3) > d\}\,. \tag{34}$$

The values i,j,k can be zero in which case we take $S(\emptyset) = 0$. This definition implicitly relies upon the exchangeability of the spacings.

The quantity $Q(i,j,k)$ is useful mainly because it satisfies the following recursion: When i,j and k are positive,

$$Q(i,j,k) = Q(i-1,j,k) + Q(i,j-1,k) - Q(i,j,k-1)\,. \tag{35}$$

This is proved in Huffer (1988), see equation (10). Given values for the boundary terms $Q(0,j,k), Q(i,0,k)$ and $Q(i,j,0)$, this recursion completely determines Q. The boundary terms are easily evaluated using elementary properties of spacings. In terms of the functions G and F defined in (22) and (23), these boundary terms are

$$Q(0,j,k) = Q(i,0,k) = P\{S_1 + \cdots + S_k > d\} = G(k-1)\,, \tag{36}$$

$$Q(i,j,0) = P\{S_1 + \cdots + S_i > d, S_{i+1} + \cdots + S_{i+j} > d\} = F(i-1,j-1)\,. \tag{37}$$

We shall need the following two facts in our arguments. They are derived using the recursion and boundary terms given above. We shall defer the proofs of these facts until the end of this section.

Lemma 1 *For $n \geq 2(x-1)$,*

$$\sum_{k=1}^{x-1} Q(k,k,x-k) = 2\sum_{i=0}^{x-2}(x-1-i)G(i) - \sum_{i=0}^{x-2}\sum_{j=0}^{x-2} F(i,j).$$

Lemma 2 *For $n \geq 2(x-1)$,*

$$\begin{aligned}\sum_{k=1}^{x-1} kQ(k,k,x-k) &= \sum_{i=0}^{x-2}[(x-i-1)^3 + (x-i-1)]G(i) \\ &\quad - \sum_{i=0}^{x-2}\sum_{j=0}^{x-2}[(x-i-2)(x-j-2) + (x-1)]F(i,j).\end{aligned}$$

Finally, we note that the quantities P_i and $P_{i,j}$ are easily re-expressed in terms of F, G, and Q. It is clear that

$$P_0 = P\{S(\delta_0) \leq d\} = 1 - P\{S(\delta_0) > d\} = 1 - G(m-2)\,. \tag{38}$$

Also, by switching to complementary events and using (34) we see that

$$\begin{aligned}P_{0,k} &= P\{S(\delta_0) \leq d, S(\delta_k) \leq d\} \\ &= 1 - P\{S(\delta_0) > d\} - P\{S(\delta_k) > d\} + P\{S(\delta_0) > d, S(\delta_k) > d\} \\ &= 1 - 2G(m-2) + Q(k,k,m-1-k) \end{aligned} \tag{39}$$

for $k \leq m-1$. When $k = m-1$, fact (37) implies

$$P_{0,m-1} = 1 - 2G(m-2) + F(m-2, m-2)\,. \tag{40}$$

Interval Case

We may now proceed with the proofs. Using (33) and (38) it is immediate that

$$E(Y_I) = \sum_{i=1}^{n-m+1} P_i = (n-m+1)P_0 = (n-m+1)[1 - G(m-2)].$$

Noting that $P_{i,i} = P_i$, the second moment of Y_I can be written as

$$E(Y_I^2) = \sum_{i=1}^{n-m+1}\sum_{j=1}^{n-m+1} P_{i,j} = E(Y_I) + 2\sum_{i<j} P_{i,j}\,. \tag{41}$$

For any $k > 0$, the number of pairs (i,j) with $j - i = k$ is equal to $n - m - k + 1$. A short combinatorial argument shows that the number of pairs (i,j) with $j - i \geq m-1$ is equal to $\binom{n-2m+3}{2}$. Thus using (33) leads to

$$EY_I^2 = EY_I + 2\sum_{k=1}^{m-2}(n-m-k+1)P_{0,k} + 2\binom{n-2m+3}{2}P_{0,m-1}\,. \tag{42}$$

Now substituting for $P_{0,k}$ and $P_{0,m-1}$ using (39) and (40), and then doing some manipulations we obtain

$$\begin{aligned} EY_I^2 &= EY_I + (n-m+1)(n-m)(1-2G(m-2)) \\ &+ (n-2m+3)(n-2m+2)F(m-2,m-2) \\ &+ 2(n-m+1)\sum_{k=1}^{m-2} Q(k,k,m-k-1) - 2\sum_{k=1}^{m-2} kQ(k,k,m-k-1)\,. \end{aligned}$$

The remaining sums can be evaluated using Lemmas 1 and 2 with $x = m-1$. This leads to the desired result given in (25).

Circular Case

The arguments are very similar to those in the interval case. Using (33) and (38) it is immediate that

$$E(Y_C) = \sum_{i=1}^{n+1} P_i = (n+1)[1-G(m-2)]\,.$$

The second moment can be written as

$$E(Y_C^2) = \sum_{i=1}^{n+1}\sum_{j=1}^{n+1} P_{i,j} = E(Y_C) + \sum_{i\neq j} P_{i,j}\,.$$

Employing (33), we may combine equal terms to obtain

$$= E(Y_C) + 2(n+1)\sum_{k=1}^{m-2} P_{0,k} + (n+1)(n-2m+4)P_{0,m-1}\,. \tag{43}$$

This expression is valid only for $n \geq 2(m-2)$.

When $n \geq 2(m-2)$, the coefficient $(n+1)(n-2m+4)$ in (43) is the number of ways to select two disjoint sets δ_i and δ_j of $m-1$ consecutive spacings. It is obtained by the following argument. Since we have $n+1$ spacings on the circle, we have $n+1$ ways to select the first set δ_i. Once we have chosen the first set of spacings, we have only $(n+1)-(m-1)$ spacings left. The number of ways to choose $m-1$ consecutive spacings from these is $(n+1)-(m-1)-(m-2) = (n-2m+4)$.

The coefficient $2(n+1)$ in (43) is the number of ways to choose two sets δ_i and δ_j of $m-1$ consecutive spacings having a given number $k(\leq m-2)$ of spacings in common. There are $n+1$ ways to choose δ_i. Given δ_i, there are two ways to choose δ_j.

Substituting (39) and (40) for $P_{0,k}$ and $P_{0,m-1}$ in (43) and then doing some manipulations we obtain

$$E(Y_C^2) \;=\; E(Y_C) + n(n+1)(1-2G(m-2)) + (n+1)(n-2m+4)F(m-2,m-2)$$

$$+ 2(n+1) \sum_{k=1}^{m-2} Q(k,k,m-k-1)\,.$$

Now applying Lemma 1 with $x = m-1$ gives us the result (26).

Proving the Lemmas

Our proofs of lemmas 1 and 2 will be based on the following fact.

Lemma 3 *If*

$$S(x,y) = f(x,y) + S(x-1,y) + S(x,y-1) - S(x-1,y-1)$$

for integers $x > c, y > c$, *and* $S(x,c) = S(c,y) = 0$ *for* $x \geq c, y \geq c$, *then*

$$S(x,y) = \sum_{i=c+1}^{x} \sum_{j=c+1}^{y} f(i,j).$$

This fact has an easy induction proof (which we omit). Note that the relation

$$S(x,y) - S(x-1,y) - S(x,y-1) + S(x-1,y-1) = f(x,y)$$

is a discrete version of

$$\frac{\partial^2 S}{\partial x \partial y} = f(x,y) \text{ with solution } S(x,y) = \int_c^x du \int_c^y dv\, f(u,v)\,.$$

Proof of Lemma 1:

Define

$$S(x,y) = \sum_{k=1}^{x \wedge y - 1} Q(x-k, y-k, k),$$

where $x \wedge y = \min(x,y)$. This definition is motivated by the fact that

$$S(x,x) = \sum_{k=1}^{x-1} Q(x-k, x-k, k) = \sum_{k=1}^{x-1} Q(k,k,x-k).$$

Write $z \equiv (x \wedge y) - 1$. The recursion (35) implies

$$\begin{aligned} S(x,y) &= \sum_{k=1}^{z} Q(x-1-k, y-k, k) + \sum_{k=1}^{z} Q(x-k, y-1-k, k) \\ &\quad - \sum_{k=1}^{z} Q(x-k, y-k, k-1) \\ &= I(x \leq y) Q(0, y-x+1, x-1) + S(x-1,y) \\ &\quad + I(x \geq y) Q(x-y+1, 0, y-1) + S(x,y-1) \\ &\quad - Q(x-1, y-1, 0) - S(x-1, y-1) \end{aligned}$$

so long as we take $S(1,y) = S(x,1) = 0$,

$$= f(x,y) + S(x-1,y) + S(x,y-1) - S(x-1,y-1)$$

with

$$f(x,y) = (1+\delta_{xy})G(x \wedge y - 2) - F(x-2,y-2). \tag{44}$$

Therefore,

$$S(x,y) = \sum_{i=2}^{x}\sum_{j=2}^{y} f(i,j)$$

and with a little manipulation we see that

$$\begin{aligned} S(x,x) &= \sum_{i=0}^{x-2} G(i) + \sum_{i=0}^{x-2}(2x-3-2i)G(i) - \sum_{i=0}^{x-2}\sum_{j=0}^{x-2} F(i,j) \\ &= 2\sum_{i=0}^{x-2}(x-1-i)G(i) - \sum_{i=0}^{x-2}\sum_{j=0}^{x-2} F(i,j). \end{aligned}$$

Proof of Lemma 2

Define

$$S^*(x,y) = \sum_{k=1}^{x\wedge y-1} (x \wedge y - k)Q(x-k,y-k,k).$$

This definition is motivated by the fact that

$$S^*(x,x) = \sum_{k=1}^{x-1}(x-k)Q(x-k,x-k,k) = \sum_{k=1}^{x-1} kQ(k,k,x-k).$$

Write $z \equiv (x \wedge y) - 1$ again. The recursion (35) implies

$$\begin{aligned} S^*(x,y) &= \sum_{k=1}^{z}(x \wedge y - k)Q(x-1-k,y-k,k) \\ &\quad + \sum_{k=1}^{z}(x \wedge y - k)Q(x-k,y-1-k,k) \\ &\quad - \sum_{k=1}^{z}(x \wedge y - k)Q(x-k,y-k,k-1) \\ &= I(x \le y)\,[Q(0,y-x+1,x-1) + S(x-1,y)] + S^*(x-1,y) \\ &\quad + I(x \ge y)\,[Q(x-y+1,0,y-1) + S(x,y-1)] + S^*(x,y-1) \\ &\quad - (x \wedge y - 1)Q(x-1,y-1,0) - S^*(x-1,y-1) \end{aligned}$$

so long as we take $S^*(1,y) = S^*(x,1) = 0$,

$$= f^*(x,y) + S^*(x-1,y) + S^*(x,y-1) - S^*(x-1,y-1)$$

with

$$\begin{aligned}
f^*(x,y) &= (1+\delta_{xy})G(x\wedge y-2)+I(x\le y)S(x-1,y)\\
&\quad +I(x\ge y)S(x,y-1)-(x\wedge y-1)F(x-2,y-2)\\
&= (1+\delta_{xy})G(x\wedge y-2)+I(x\le y)\left[S(x-1,y)-S(x,y)\right]\\
&\quad +I(x\ge y)\left[S(x,y-1)-S(x,y)\right]+(1+\delta_{xy})S(x,y)\\
&\quad -(x\wedge y-1)F(x-2,y-2)\\
&= f(x,y)+I(x\le y)\left[S(x-1,y)-S(x,y)\right]\\
&\quad +I(x\ge y)\left[S(x,y-1)-S(x,y)\right]+(1+\delta_{xy})S(x,y)\\
&\quad -(x\wedge y-2)F(x-2,y-2).
\end{aligned}$$

Hence, by Lemma 3 we have

$$\begin{aligned}
S^*(x,x) &= \sum_{i=2}^{x}\sum_{j=2}^{x} f^*(i,j)\\
&= S(x,x)-2\sum_{i=2}^{x}S(i,i)+\sum_{i=2}^{x}\sum_{j=2}^{x}S(i,j)+\sum_{i=2}^{x}S(i,i)\\
&\quad -\sum_{i=2}^{x}\sum_{j=2}^{x}(i\wedge j-2)F(i-2,j-2)\\
&= -\sum_{i=2}^{x-1}S(i,i)+\sum_{i=2}^{x}\sum_{j=2}^{x}S(i,j)-\sum_{i=2}^{x}\sum_{j=2}^{x}(i\wedge j-2)F(i-2,j-2)\\
&= -\sum_{i=2}^{x-1}\sum_{k=2}^{i}\sum_{h=2}^{i}f(k,h)+\sum_{i=2}^{x}\sum_{j=2}^{x}\sum_{k=2}^{i}\sum_{h=2}^{j}f(k,h)\\
&\quad -\sum_{i=2}^{x}\sum_{j=2}^{x}(i\wedge j-2)F(i-2,j-2).
\end{aligned}$$

Now expand $f(k,h)$ using the definition (44). Collecting the terms which involve F, we have

$$\begin{aligned}
&\sum_{i=2}^{x-1}\sum_{k=0}^{i-2}\sum_{h=0}^{i-2}F(k,h)-\sum_{i=2}^{x}\sum_{j=2}^{x}\sum_{k=0}^{i-2}\sum_{h=0}^{j-2}F(k,h)-\sum_{i=2}^{x}\sum_{j=2}^{x}(i\wedge j-2)F(i-2,j-2)\\
=\ &\sum_{k=0}^{x-3}\sum_{h=0}^{x-3}(x-h\vee k-2)F(k,h)-\sum_{k=0}^{x-2}\sum_{h=0}^{x-2}(x-k-1)(x-h-1)F(k,h)\\
&-\sum_{k=0}^{x-2}\sum_{h=0}^{x-2}(h\wedge k)F(k,h)\\
=\ &-\sum_{k=0}^{x-2}\sum_{h=0}^{x-2}\left[(x-k-1)(x-h-1)+h\wedge k-x+h\vee k+2\right]F(k,h)
\end{aligned}$$

$$= \quad -\sum_{k=0}^{x-2}\sum_{h=0}^{x-2}\left[(x-k-2)(x-h-2)+x-1\right]F(k,h).$$

The notation $h \vee k \equiv \max(h,k)$ was used in the above. Collecting the terms which involve G, we have

$$\begin{aligned}
& -\sum_{i=2}^{x-1}\sum_{k=2}^{i}\sum_{h=2}^{i}(1+\delta_{kh})G(h\wedge k-2)+\sum_{i=2}^{x}\sum_{j=2}^{x}\sum_{k=2}^{i}\sum_{h=2}^{j}(1+\delta_{kh})G(h\wedge k-2) \\
= \quad & -2\sum_{i=2}^{x-1}\sum_{k=0}^{i-2}(i-1-k)G(k)+\sum_{i=2}^{x}\sum_{j=2}^{x}\sum_{k=2}^{i}\sum_{h=2}^{j}I(k\le h)G(k-2) \\
& +\sum_{i=2}^{x}\sum_{j=2}^{x}\sum_{k=2}^{i}\sum_{h=2}^{j}I(h\le k)G(h-2) \\
= \quad & -2\sum_{k=0}^{x-3}\sum_{i=k+2}^{x-1}(i-1-k)G(k)+2\sum_{k=2}^{x}\sum_{h=k}^{x}\sum_{i=k}^{x}\sum_{j=h}^{x}G(k-2) \\
= \quad & -\sum_{k=0}^{x-2}(x-k-2)(x-k-1)G(k)+\sum_{k=0}^{x-2}(x-k-1)^2(x-k)G(k) \\
= \quad & \sum_{k=0}^{x-2}\left[(x-k-1)^3+(x-k-1)\right]G(k).
\end{aligned}$$

Therefore,

$$\begin{aligned}
S^*(x,x) \quad = \quad & \sum_{i=0}^{x-2}[(x-i-1)^3+(x-i-1)]G(i) \\
& -\sum_{i=0}^{x-2}\sum_{j=0}^{x-2}[(x-i-2)(x-j-2)+(x-1)]F(i,j).
\end{aligned}$$

C Expressions for Third and Fourth Moments

In this section we derive the expressions given in Section 3.3 for the third and fourth moments of Y_I and Y_C.

C.1 The Third Moments of Y

Interval Case: From the definition of Y_I we obtain

$$E(Y_I^3) = \sum_{i=1}^{n-m+1}\sum_{j=1}^{n-m+1}\sum_{k=1}^{n-m+1} P_{i,j,k}\,.$$

Grouping together terms according to the number of distinct values in the 3-tuple (i,j,k) and then rewriting the sums so that the indices i,j,k are ordered (that is,

$i < j < k$) leads to

$$= E(Y_I) + 2 \cdot \frac{3!}{1!2!} \sum_{i=1}^{n-m} \sum_{j=i+1}^{n-m+1} P_{i,j} + 3! \sum_{i=1}^{n-m-1} \sum_{j=i+1}^{n-m} \sum_{k=j+1}^{n-m+1} P_{i,j,k} \,. \tag{45}$$

Because the spacings are exchangeable, many of the terms in (45) are equal. The argument which follows is really nothing more than counting and combining the similar terms. In this argument we shall freely use the exchangeability property, frequently without explicit mention.

The first sum in (45) is the same as that in (41). Thus the argument leading to (42) gives us

$$\begin{aligned} = \quad & E(Y_I) + 6 \sum_{i=1}^{m-2} (n-m-i+1) P_{0,i} + 6 \binom{n-2m+3}{2} P_{0,m-1} \\ & + 6 \sum_{i=1}^{n-m-1} \sum_{j=i+1}^{n-m} \sum_{k=j+1}^{n-m+1} P_{i,j,k} \,. \end{aligned} \tag{46}$$

The triple summation in (46) can be rewritten as

$$\sum_{i,j,k} P_{i,j,k},$$

where $1 \le i < j < k \le n-m+1$. Making the change of variables $r = j - i$ and $s = k - j$ gives us

$$= \sum_{i,r,s} P_{i,i+r,i+r+s},$$

where $i, r, s \ge 1$ and $i + r + s \le n - m + 1$. Using exchangeability to simplify the above equation yields

$$= \sum_{i,r,s} P_{0,r,r+s} \,.$$

Summing over i leads to

$$= \sum_{r,s} (n-m+1-r-s) P_{0,r,r+s}, \tag{47}$$

where $r, s \ge 1$ and $r + s \le n - m$.

If $n \ge 3(m-1)$, the summation (47) can be separated into four parts. The sum is first broken into two parts depending on whether or not δ_0 intersects δ_r. Each of these parts is in turn broken in two depending on whether or not δ_r intersects δ_{r+s}. The resulting four parts can be written in short as

$$\sum_{r,s} = \sum_{\substack{r \le m-2 \\ s \le m-2}} + \sum_{\substack{r \le m-2 \\ s \ge m-1}} + \sum_{\substack{r \ge m-1 \\ s \le m-2}} + \sum_{\substack{r \ge m-1 \\ s \ge m-1}} \,.$$

The second and third sums are the same by symmetry, so that we have

$$= \sum_{r,s \le m-2} + 2 \sum_{\substack{r \le m-2 \\ s \ge m-1}} + \sum_{r,s \ge m-1} ,$$

where $r, s \ge 1$ and $r + s \le n - m$.

We now examine these three terms. For convenience, we shall call them (a), (b) and (c).

$$(a) = \sum_{r=1}^{m-2} \sum_{s=1}^{m-2} (n - m + 1 - r - s) P_{0,r,r+s} = \sum_{r=1}^{m-2} \sum_{t=r+1}^{r+m-2} (n - m + 1 - t) P_{0,r,t} .$$

$$(b) = \sum_{r=1}^{m-2} \sum_{s=m-1}^{n-m-r} (n - m + 1 - r - s) P_{0,r,r+s} = \sum_{r=1}^{m-2} \sum_{t=r+m-1}^{n-m} (n - m + 1 - t) P_{0,r,t} .$$

Because δ_r and δ_t are disjoint, the above sum can be simplified as

$$= \sum_{r=1}^{m-2} \binom{n - 2m - r + 3}{2} P_{0,r,r+m-1} .$$

Finally, we have

$$(c) = \sum_{r,s \ge m-1} (n - m + 1 - r - s) P_{0,r,r+s},$$

where $r + s \le n - m$. Because δ_0, δ_r and δ_{r+s} are mutually disjoint, the above expression can be rewritten as

$$= \binom{n - 3m + 5}{3} P_{0,m-1,2m-2} .$$

Combining these intermediate terms and putting them back in (45) we get the result (27).

Circular Case: The argument for (45) when applied to Y_C leads to

$$E(Y_C^3) = E(Y_C) + 6 \sum_{i=1}^{n} \sum_{j=i+1}^{n+1} P_{i,j} + 6 \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n+1} P_{i,j,k} . \tag{48}$$

The sums in (48) can be handled in much the same way as in the derivations of (43) and (27). We shall not give the details, but will just comment on parts of the argument.

In the circular case the $n + 1$ spacings may be viewed as being arranged around a circle. The factors of $n + 1$ which occur repeatedly in (29) all come from the fact that

any collection of sets such as $\delta_0, \delta_i, \delta_j$ can be "rotated" to begin at any of the $n+1$ positions on the circle.

The coefficient $(n+1)(n-2m+4-i)$ in the fifth term of (29) is the number of ways to select three distinct sets $\delta_j, \delta_k, \delta_\ell$ so that two of the sets have a given amount of overlap and the remaining set is disjoint from these two. (Note that here we do not consider different orderings of the sets to be distinct.) This coefficient can be derived by extending the argument given for $E(Y_C^2)$. Once we have chosen the first set δ_i, there is only one choice for the overlapping set δ_j. After δ_i and δ_j are chosen, there are $(n+1)-(m-1)-i$ spacings left. Therefore, the number of ways to choose $m-1$ consecutive spacings for δ_k from these is $(n+1)-(m-1)-i-(m-2) = n-2m+4-i$.

The value $(n+1)(n-3m+6)(n-3m+5)/6$ is the number of ways to select three disjoint sets of $m-1$ consecutive spacings when the ordering of the sets does not count. After choosing the first set (in $n+1$ ways), there are $(n+1)-(m-1) = n-m+2$ consecutive spacings remaining. Using the lemma stated below, the number of ways to choose two nonoverlapping sets of $m-1$ spacings from these remaining spacings is $\binom{(n-m+2)-2(m-1)+2}{2}$. We must now divide by 3, since any of our three sets could have been designated the "first". Thus, our three disjoint sets can be selected in $(n+1)\binom{n-3m+6}{2} \times \frac{1}{3} = (n+1)(n-3m+6)(n-3m+5)/6$ ways. When we plug this into equation (48), the 1/6 cancels the 6 and we get the coefficient of the last term in (29).

Lemma 4 *Let $\mathcal{L} = \{1, 2, \ldots, L\}$. Suppose you wish to select k disjoint subsets $R_1, R_2, \ldots, R_k$ from $\mathcal{L}$ having given cardinalities $|R_i| = r_i$ for $1 \le i \le k$. If each set must consist of consecutive integers, and the sets $R_1, R_2, \ldots, R_k$ must be arranged from left to right, then the number of ways this can be done is*

$$\binom{L-(r_1+r_2+\cdots+r_k)+k}{k}.$$

Proof: For $2 \le i \le k$, let g_i be the size of the gap between sets R_{i-1} and R_i. Let g_1 be the number of integers (in $\mathcal{L}$) to the left of R_1, and g_{k+1} be the number of integers to the right of R_k. Every choice of $R_1, \ldots, R_k$ corresponds uniquely to a choice of gaps $g_1, \ldots, g_{k+1}$ satisfying $g_i \ge 0$ for all i and $\sum_i g_i = L-(r_1+\cdots+r_k)$. Counting the number of ways to choose the gaps $g_1, \ldots, g_{k+1}$ is a well known elementary combinatorics problem whose answer is given in Lemma 4.

C.2 The Fourth Moments of Y

The derivation of the fourth moment of Y is very similar to that of the third moment. Hence we shall just remark on parts of the proof.

Interval Case

$$\begin{aligned} E(Y_I^4) &= E(Y_I) + \left(\frac{4!}{2!2!} + 2\frac{4!}{1!3!}\right) \sum_{i=1}^{n-m} \sum_{j=i+1}^{n-m+1} P_{i,j} + 3\frac{4!}{1!1!2!} \sum_{i=1}^{n-m-1} \sum_{j=i+1}^{n-m} \sum_{k=j+1}^{n-m+1} P_{i,j,k} \\ &\quad + 4! \sum_{i=1}^{n-m-2} \sum_{j=i+1}^{n-m-1} \sum_{k=j+1}^{n-m} \sum_{\ell=k+1}^{n-m+1} P_{i,j,k,\ell} \,. \end{aligned} \tag{49}$$

For $n \geq 4(m-1)$, we can rewrite the last summation in (49) as

$$= \sum_{r,s,t} (n-m+1-r-s-t) P_{0,r,r+s,r+s+t},$$

where $r, s, t \geq 1$ and $r+s+t \leq n-m$. This can be separated into eight parts which can be written in short as

$$\sum_{r,s,t} = \sum_{r,s,t \leq m-2} + \sum_{\substack{r,s \leq m-2 \\ t \geq m-1}} + \sum_{\substack{r,t \leq m-2 \\ s \geq m-1}} + \sum_{\substack{r \leq m-2 \\ s,t \geq m-1}} + \sum_{\substack{r \geq m-1 \\ s,t \leq m-2}} + \sum_{\substack{r,t \geq m-1 \\ s \leq m-2}} + \sum_{\substack{r,s \geq m-1 \\ t \leq m-2}} + \sum_{r,s,t \geq m-1} .$$

It is easy to see that

$$\sum_{\substack{r,s \leq m-2 \\ t \geq m-1}} = \sum_{\substack{r \geq m-1 \\ s,t \leq m-2}} \quad \text{and} \quad \sum_{\substack{r \leq m-2 \\ s,t \geq m-1}} = \sum_{\substack{r,t \geq m-1 \\ s \leq m-2}} = \sum_{\substack{r,s \geq m-1 \\ t \leq m-2}} .$$

Combining these equal terms together we get

$$\sum_{r,s,t} = \sum_{r,s,t \leq m-2} + 2 \sum_{\substack{r,s \leq m-2 \\ t \geq m-1}} + 3 \sum_{\substack{r \leq m-2 \\ s,t \geq m-1}} + \sum_{\substack{r,t \leq m-2 \\ s \geq m-1}} + \sum_{r,s,t \geq m-1} ,$$

where $r, s, t \geq 1$ and $r+s+t \leq n-m$.

Following the derivation of $E(Y_I^3)$, we may rewrite these sums as

$$\begin{aligned} &= \sum_{r=1}^{m-2} \sum_{w=r+1}^{r+m-2} \sum_{u=w+1}^{w+m-2} (n-m+1-u) P_{0,r,w,u} \\ &\quad + 2 \sum_{r=1}^{m-2} \sum_{w=r+1}^{r+m-2} \binom{n-2m-w+3}{2} P_{0,r,w,w+m-1} \\ &\quad + 3 \sum_{r=1}^{m-2} \binom{n-3m-r+5}{3} P_{0,r,r+m-1,r+2m-2} \\ &\quad + \sum_{r=1}^{m-2} \sum_{t=1}^{m-2} \binom{n-2m-r-t+3}{2} P_{0,r,r+m-1,r+m-1+t} \\ &\quad + \binom{n-4m+7}{4} P_{0,m-1,2m-2,3m-3} \,. \end{aligned}$$

(The binomial coefficients occurring above can also be obtained by repeated applications of Lemma 4.) Putting all these terms back in (49), we obtain the result (28).

Circular Case From the definition of Y_C we obtain

$$E(Y_C^4) = E(Y_C) + \left(\frac{4!}{2!2!} + 2\frac{4!}{1!3!}\right)\sum_{i=1}^{n}\sum_{j=i+1}^{n+1} P_{i,j} + 3\frac{4!}{1!1!2!}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sum_{k=j+1}^{n+1} P_{i,j,k}$$
$$+4!\sum_{i=1}^{n-2}\sum_{j=i+1}^{n-1}\sum_{k=j+1}^{n}\sum_{\ell=k+1}^{n+1} P_{i,j,k,\ell}\,. \tag{50}$$

For $n \geq 4(m-2)$, the last summation in (50) can be broken down into cases with each case then being simplified as in the derivation of $E(Y_C^3)$. Doing this, the last summation can be rewritten as

$$(n+1)\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}\sum_{k=j+1}^{j+m-2} P_{0,i,j,k} + (n+1)\sum_{i=1}^{m-2}\sum_{j=i+1}^{i+m-2}(n-2m+4-j)P_{0,i,j,j+m-1}$$
$$+(n+1)\sum_{i=1}^{m-2}(n-3m+6-i)(n-3m+5-i)/2\cdot P_{0,i,i+m-1,i+2m-2}$$
$$+(n+1)\sum_{i=1}^{m-2}\sum_{j=1}^{m-2}(n-2m+4-i-j)/2\cdot P_{0,i,i+m-1,i+m-1+j}$$
$$+(n+1)(n-4m+8)(n-4m+7)(n-4m+6)/24\cdot P_{0,m-1,2m-2,3m-3}\,. \tag{51}$$

The first three terms in (51) can be obtained by arguments very similar to those needed in (29).

The coefficient $(n+1)(n-2m+4-i-j)/2$ in the fourth term is the number of ways to select four sets $\delta_r, \delta_s, \delta_t, \delta_u$ of $m-1$ consecutive spacings such that: δ_r and δ_s have a given overlap, δ_t and δ_u have a given overlap, and $\delta_r \cup \delta_s$ is disjoint from $\delta_t \cup \delta_u$. After selecting the set $\delta_r \cup \delta_s$ (in $n+1$ ways), we have $(n+1)-(m-1)-i$ spacings left. The number of ways to choose $\delta_t \cup \delta_u$ from these is $(n+1)-(m-1)-i-(m-1)-j+1 = (n-2m+4-i-j)$. We must now divide by two since either $\delta_r \cup \delta_s$ or $\delta_t \cup \delta_u$ could have been chosen first.

The coefficient $(n+1)(n-4m+8)(n-4m+7)(n-4m+6)/24$ is the number of ways to select four disjoint sets of $m-1$ consecutive spacings (without taking order into consideration). The first set may be chosen in $n+1$ ways. According to Lemma 4, the remaining three sets can then chosen in $(n-4m+8)(n-4m+7)(n-4m+6)/6$ ways. Now we divide by 4 since any of the four sets could have been designated as the "first". This gives us the desired coefficient.

Combining the equation (51) with the results in $E(Y_C^2)$ and $E(Y_C^3)$ yields the final expression (30).