43. Scan Statistics

Section 43.1 introduces the concept of scan statistics and overviews types used to localize unusual clusters in continuous time or space, in sequences of trials or on a lattice. Section 43.2 focuses on scan statistics in one dimension. Sections 43.2.2 and 43.2.3 deal with clusters of events in continuous time. Sections 43.2.4 and 43.2.5 deal with success clusters in a sequence of discrete binary (s-f) trials. Sections 43.2.6 and 43.2.7 deal with the case where events occur in continuous time, but where we can only scan a discrete set of positions. Different approaches are used to review data when looking for clusters (the retrospective case in Sects. 43.2.2, 43.2.5, 43.2.6), and for ongoing surveillance that monitors unusual clusters (the prospective case in Sects. 43.2.2, 43.2.3, 43.2.7). Section 43.2.7 describes statistics used to scan for clustering on a circle (are certain times of the day or year more likely to have accidents?). Section 43.3 describes statistics used to scan continuous space or a two-dimensional lattice for unusual clusters. Sections 43.2 and 43.3 focus on how unusual the largest number of events within a scanning window is. Section 43.4.1 deals with scanning for unusually sparse regions. In some cases the researcher is more interested in the number of clusters, rather than the size of the largest or smallest, and Sect. 43.4.2 describes results useful for this case. The double-scan statistic of Sect. 43.4.3 allows the researcher to test for unusual simultaneous or lagged clustering of two different types of events. Section 43.4.4 describes scan statistics that can be used on data with a complex structure.

43.1 Overview

During design, monitoring, or analysis work, engineers and other scientists often need to take into account unusually large *clusters* of events in time or space. Mechanical engineers design system capacity to provide reliability to pipeline systems. Telecommunication experts seek to avoid outage caused by too many mobiles transmitting within the same area served by a single base station. Quality control experts monitor for clusters of defectives. Epidemiologists investigate *hotspots* of cancer cases, and carry out syndrome surveillance to monitor for bioterrorism attacks. Computer scientists base an information flow control mechanism on a large enough number of information packets within a temporal *sliding window*. Astronomers scan for muon clusters. Electrical engineers build in multiple redundancies to improve reliability, and use clusters of successes

as a criteria for start-up reliability. Molecular biologists search for clusters of specific types of patterns in protein or DNA sequences to focus on regions with important biologic functions. These scientists seek to determine which clusters are unlikely to occur by chance. The distributions given by various cluster statistics are tools that answer this question.

Scan statistics measure an unusually large cluster of events in time and or space. Given events distributed over a time period $(0, T)$, S_w is the largest number of events in any subinterval of length w. S_w is called the (temporal) *scan statistic*, from the viewpoint that one scans the time period $(0, T)$ with a window of size w, and finds the maximum cluster of points. A simple example illustrates this.

Example 1: A public health officer reviewing the records for a nursing home observed 60 deaths over the five-year period from January 1, 2000 through December 31, 2004. This was about average for such facilities. However, the officer observed that in the one-year period between April 1, 2003 to March 31, 2004 there were 23 deaths. Given that the 60 dates of death were independent, and occurred at random over the five-year period, how likely is it that there would be any one year period with 23 or more deaths? That is, if we scan the $T = 5$ year period with a $w = 1$ year period, how likely is it that the scan statistic $S_w \geq 23$? Note that the officer did not just divide the five-year period up into five calendar years and look at the calendar year with the largest number of deaths. In fact, the cluster observed did not occur in a calendar year. The scan statistic takes into account that the officer looked at a very large number of overlapping one-year periods. We see below that, even taking into account these multiple comparisons, the cluster of 23 deaths in a one-year period is fairly unusual. $P(S_w \geq 23, \text{ given } N = 60, w/T = 0.2) < 0.03$.

There is a large volume of literature and much current research being done on scan statistics. Three recent books [43.1–3] summarize and reference many results. Chapters 9 through 12 of [43.1] deals with scan statistics in a sequence of trials. There are useful engineering applications to reliability in consecutive systems, quality control, matching in genetic sequences, sooner and later waiting time problems and 556 references on runs and scans. Recent advances from articles by researchers are detailed in [43.2] together with applications. In [43.3], the first six chapters systematically show many applications with useful simple formulae for scan statistics, and are aimed at practitioners; the remaining 12 chapters develop the theory and methodology of scan statistics, and this is followed by a bibliography of over 600 references. In this review we draw heavily on these references, particularly [43.3], and subsequent research to give an overview of scan statistics, and highlight many results that have proved useful in many scientific applications. Scan statistics have been developed and applied for a variety of temporal and spatial scenarios. Time can be viewed as continuous or a discrete sequence of trials or time periods. Space can be viewed as continuous, or as a discrete grid of points at which events can occur. In two dimensions, results have been derived for square, rectangular, circular, triangular, and other shapes of scanning windows. The scan statistic S_w is the largest number of points in any window (for a fixed window size, shape, and orientation; or for a range of window sizes). The two-dimensional regions scanned include rectangles, the surface of a sphere, and more irregularly shaped geographical areas. In certain applications the events can only occur naturally at a discrete set of points in space. In other applications the underlying events can occur anywhere in space, but the method of observation limits events to a grid or lattice of points. Scan distributions have been derived for uniform, Poisson, and other distributed points in continuous space, and binomial, hypergeometric, and other distributions on two-dimensional grids. In the next section we discuss temporal scenarios.

43.2 Temporal Scenarios

One aspect of the scenario is whether its view is retrospective or prospective. A researcher might be reviewing events over some past time period of length T. The events might be a call for service, a reported cancer case, or an unacceptable item from an assembly line. In the *retrospective* case, the total number of events in the review period is a known number, N. The retrospective scan statistic analysis will typically be conditioned on N, a fixed known number, and in this case is referred to as either the *retrospective* or *conditional* case.

In other applications, the scientist uses scan statistics *prospectively* either to design a system's capacity to handle clustered demands, or to set up a monitoring system that will sound an alarm when an unusual cluster occurs. System capacity can be designed to give a specified small probability of overload within some future

period of operations of length T. Scan monitoring systems can be similarly designed so that (provided the process is "in control") there is, for example, only a 1% chance of a false alarm within a year; this is equivalent to saying that there is a 99% chance that the waiting time until a false alarm is greater than a year. Note that the total number of events, N, in time T is not known at either the system design time, or at the time an alarm is to be sounded. In the prospective case, the distribution of scan statistics cannot be conditioned on N. However, we often have information on the expected number of events in $(0, T)$. The prospective case is referred to as either the *prospective* or the *unconditional* case.

For each of the retrospective and prospective views, scan statistic distributions have been developed for continuous and for several discrete time scenarios. For example, the starting time of a hospital emergency room admission might be recorded to the nearest minute, and whether the patient had a particular syndrome may be recorded for each admission. For the event "admission of patient with syndrome," the scan statistic might be based on reported admission times, where time is viewed as a continuum. The *continuous scan statistic* is the maximum number of events in a window of length w that scans the time period $(0, T)$. In the continuous scenario, the times of occurrence of events are reported and for each time t in the review period, $(w \le t \le T)$, one knows the observed number of events $Y_w(t)$ and the expected number of events $E_w(t)$ in the subinterval $[t-w, t)$.

Alternatively, the analyst may only have a sequential list of patients available, and may only know whether or not each has the syndrome. In this case, the data is in the form of a discrete sequence of binary trials, and a *discrete* case scan statistic will be used. The data is viewed as a sequence of T trials, where for each trial whether or not an event has occurred is recorded; the discrete scan statistic is the maximum number of events in any w consecutive trials. For $t = w, w+1, \ldots, T$, $Y_w(t)$ and $E_w(t)$ are the observed and expected number of events within the w consecutive trials, $t-w+1, t-w+2, \ldots, t$.

In other cases, the reported data may only give hourly summary counts of patients with the syndrome, and the researcher may be keeping a moving sum of the number of such patients in the past six hours. In this discrete case we might use the *ratchet scan statistic*. In the ratchet scenario, time is divided into T disjoint intervals (hours, days, or weeks) and the reported data consists of the number of events in each interval. For $t = w, w+1, \ldots, T$, $Y_w(t)$ and $E_w(t)$ are the observed and expected number of events within the w consecutive intervals, $t-w+1, t-w+2, \ldots, t$.

Let S_w denote the scan statistic, $\max_t[Y_w(t)]$. For several important models for the above scenarios, exact formulae, approximations, and bounds are available for the distribution of S_w and related statistics. The following sections detail some of the most useful formula for the case where $E_w(t) = E_w$, a constant.

43.2.1 The Continuous Retrospective Case

The completely at random (constant background) model for this case is where N points (number of events) are independent uniform random variables over $(0, T)$.

$P(S_w \ge k | E_w, T)$ only depends on k, $N = (T/w)E_w$, and the ratio w/T. Choosing the units of measurement to make $T = 1$ simplifies the notation. A related scan statistic is the *minimum $(k-1)$th order gap*, W_k, the length of the smallest subinterval that contains k points. W_{k+1} is also referred to as the smallest *k-th-nearest neighbor distance* among the N points. The statistics are related by

$$P(S_w \ge k) = P(W_k \le w) \tag{43.1}$$

Now denote the common probability $P(k; N, w)$. W_N is the sample range, and W_2 is the smallest gap between any pair of points.

$$\begin{aligned} P(2; N, w) &= 1 - [1-(N-1)w]^N \,; \\ &\quad \text{for} \quad 0 \le w \le 1/(N-1) \\ &= 1 \,; \\ &\quad \text{for} \quad 1/(N-1) \le w \le 1 \,. \end{aligned} \tag{43.2}$$

$$\begin{aligned} P(N; N, w) &= N w^{N-1} - (N-1) w^N \,. \\ &\quad \text{for} \quad 0 \le w \le 1 \,. \end{aligned} \tag{43.3}$$

For a given k and N, the expressions for $P(k; N, w)$ are piecewise polynomials in w with different polynomials for different ranges of w. A direct integration approach can be used to derive the piecewise polynomials for a few simple cases, but it becomes overly complex in general. An alternative combinatorial approach is used by [43.4] to derive the piecewise polynomials for $k > N/2$, with one polynomial for $w \le 0.5$, and another polynomial for $w > 0.5$. For $k > N/2$, $0 \le w \le 0.5$, the formula is particularly simple,

$$\begin{aligned} P(k; N, w) &= [(k - E_w)(1/w) + 1] \\ &\quad \times P(Y_t = k) + 2P(Y_t > k) \,, \\ &= \left[\left(k w^{-1} - \mathrm{N} + 1\right)\right. \\ &\quad \left. \times P(Y_t = k)\right] + 2P(Y_t \ge k) \,, \end{aligned} \tag{43.4}$$

where E_w is $N(w/T)$ and $P(Y_t = k)$ is the binomial probability $b(k; N, w)$

$$b(k; N, w) = \binom{N}{k} w^k (1-w)^{N-k} . \quad (43.5)$$

For $w > 0.5, k > (N+1)/2$, there are some additional terms involving binomials and cumulative binomial terms. This leaves the case $k \le N/2$ when $w \le 0.5$. Below we discuss exact results, tabled values, approximations and bounds that can be used to compute values for some cases. However, for hypothesis testing, [43.5] uses (43.4) as an accurate approximation for small to moderate (< 0.10, and even larger) $P(k; N, w)$ when $k \le N/2, w \le 0.5$.

In Example 1, there were 60 dates of death over a five-year period, and a one-year period with 23 or more deaths. Assuming that the 60 deaths were randomly distributed over the five-year period, $P(S_1 \ge 23|T = 5) = P(23; 60, 0.2) = 0.029$, obtained by approximating using (43.4).

For certain applications one may seek to evaluate large values of $P(k; N, w)$, where the Wallenstein–Neff approximation may not be sufficiently accurate. In [43.3], Chapt. 8 discusses exact formulae, Chapt. 9 bounds, Chapt. 10 approximations, and Chapt. 2 details of the application of $P(k; N, w)$ to the continuous conditional (retrospective) case. We now give an overview of the types of exact results, other more accurate approximations, and bounds for $P(k; N, w)$.

Exact Results for $P(k; N, w)$

A general expression for $P(k; N, 1/L)$ (where L is an integer), in terms of sums of $L \times L$ determinants, is derived in [43.6]; This is generalized in [43.7] for $P(k; N, r/L)$ in terms of sums of products of several determinants, and simplified further by [43.8] in terms of sums of products of two determinants. These general formulae are computationally intense for small w small as they involve summing determinants of large matrices; however, these formulae can be used to generate the piecewise polynomials that can be used to compute the probabilities for any w. A procedure to do this is given and implemented by [43.9], for $N/3 < k, N/2$. A systematic approach to generating the polynomials is described in [43.10], Table 3, which lists the piecewise polynomials for $N \le 20$. The polynomials are then used to generate (in their Table 1a), $P(k; N, w)$ for $w \le 0.5, k \le N/2$, for $N \le 25$, with w to three decimal places. (Table 1 in [43.10] gives values for all k, for $N \le 25$, with w to two places.)

A powerful general *spacings* approach is derived in [43.11], and is used to find the distribution of W_k, the minimum of the sum of $k-1$ adjacent spacings between times of events. Huffer and Lin also find the maximum sum of $k-1$ adjacent spacings, which is related to the minimum number of events in a scanning window. They use their method to increase the range of values of N and k for which polynomials can be computed, with N as large as 61 for k close to $N/2$.

Approximations

A variety of approximations for $P(k; N, w)$ have been developed based on various combinations of approaches: methods with moments based on spacings, Poisson-type approximations with and without declumping, averaging bounds, using product limit approximations ([43.3], Chapt. 10). To emphasize the connection between higher order spacings and the scan statistic, we describe approximations based on using the method of moments applied to k-th order spacings or gaps. If $X_1 X_2 \ldots X_N$ are the ordered values of the N points in $(0, T)$, then $X_2 - X_1, X_3 - X_2, \ldots$ are the first-order spacings; $X_3 - X_1, X_4 - X_2, \ldots$ are the second-order spacings; and $X_k - X_1, X_{k+1} - X_2, \ldots, X_N - X_{N-k+1}$ are the $(k-1)$-order spacings. (Instead of spacings they are sometimes referred to as gaps, or quasi-ranges). Let $Z_k(w)$ denote the number of $(k-1)$-order spacings that are $\le w$.

$$P(S_w \ge k) = P(Z_k(w) \ge 1) = 1 - P(Z_k(w) = 0) . \quad (43.6)$$

The distribution of $Z_k(w)$ is complex, but it is straightforward to compute the expectation of $Z_k(w)$, and with more effort its variance.

$$\begin{aligned} E[Z_k(w)] &= (N-k+1)P(X_k - X_1 \le w) \\ &= (N-k+1)P(Y_t \ge k-1) . \end{aligned} \quad (43.7)$$

Here Y_t has the binomial distribution described in (43.5). In the method of moments we approximate the distribution of $Z_k(w)$ by a simpler distribution with some of the same moments. For example, choosing the approximating distribution to be Poissonian, with the same first moment, gives the approximation

$$P(S_w \ge k) \approx 1 - \exp[-(N-k+1)P(Y_t \ge k-1)] . \quad (43.8)$$

Note that the same Poisson model could be used to find $P(Z_k(w) \ge n)$, which could be used to approximate the distribution of the number of k-within-w clusters. Approximation (43.8) is not very good in general, because the $(N-k+1)$ overlapping $(k-1)^{st}$-order spacings are not independent. If $X_k - X_1$ is very small, this implies that $X_k - X_2$ is even smaller, which makes for

a greater chance that $X_{k+1} - X_2$ will also be small. A local declumping technique is used [43.12] to adjust for this type of association, and find an approximation of the form $1 - e^{-\mu}$. Approximations of this form, but with different μ values, have been used by [43.13] and others, and [43.14] proves limiting results of this form which suggest their use as approximations; however, care must be taken because the limiting results converge very slowly. *Glaz* in [43.15, 16] and other papers develops better approximations and a variety of bounds.

Moments for Continuous Retrospective Case

To compute the expectation, variance and other moments of S_w or W_k, one could average over the distribution of the statistic, where the cumulative distribution function of W_k is given by $P(k; N, w)$, and of S_w by $1 - P*(k+1; g, w)$. Using this method, [43.6] proves for $(N+1)/2 < k \le N$ that

$$\begin{aligned} E(W_k) &= [k - 2(N-k+1)b]/(N+1)\,, \\ \text{var}\,(W_k) &= (N-k+1)\big[(N+k+1) \\ &\quad + 2(2k-N-1)b \\ &\quad - 4(N+2)(N-k+1)b^2\big] \\ &\quad /(N+1)^2(N+2)\,, \end{aligned} \tag{43.9}$$

where b denotes the binomial term $b[N-k+1; 2(N-k+1), .5]$. Tables for the expectation and variance of W_k, for $\{k = 3, 4, 5; N = k(1)19\}$, $\{k = 6; N = 6(1)17\}$, $\{k = 7; N = 7(1)20\}$, $\{k = 8; N = 8(1)23\}$, and $\{k = 9; N = 9(1)25\}$ are generated in [43.10].

Averaging over exact and simulated values, [43.6] gives means and variances of S_w for $N \le 10$, $w = .1(.1).9$, and [43.17] tabulate means and variances of S_w for $N = 2(1)40, 40(5)70, 85, 100, 125, 150, 200(100)500, 1000$; and $w = 1/T$, $T = 3(1)6, 8, 12$.

43.2.2 Prospective Continuous Case

In certain applications, the researcher is interested in the distribution of the scan statistic given that the total number of events in $(0, T)$ is a random variable. Events are viewed as occurring at random times according to some process. The Poisson process is one completely-at-random chance model. In this process, the number of events $Y_w(t)$ in any interval $[t-w, t)$ is Poisson-distributed with mean E_w. $P[Y_w(t) = k] = p(k; E_w)$ for $k = 0, 1, 2, \ldots$, where $p(k; \lambda)$ denotes the Poisson probability $\exp(-\lambda)\lambda^k/k!$. For the Poisson process, $E_w = wE_1$ where E_1 is sometimes denoted λ; the numbers of events in any disjoint (not overlapping) intervals are independently distributed. There are various other ways to characterize the Poisson process. For the Poisson process, the arrival times between points are independent exponential random variables. Conditional on there being a total of N points from the Poisson process in $[0, T)$, these N points are uniformly distributed over $[0, T)$.

Given that events occur at random over time, let $T_{k,w}$ denote the waiting time until we first observe at least k events in an interval of length w. Formally, $T_{k,w}$ equals $X_{(i+k-1)}$ for the smallest i such that $X_{(i+k-1)} - X_{(i)} \le w$. The three scan statistics S_w, W_k, and $T_{k,w}$ are related by $P(S_w \ge k) = P(W_k \le w) = P(T_{k,w} \le T)$. These probabilities only depend on k, E_1 (the expected number of points in a window of length 1), and the ratio w/T. Denote the common probabilities for the Poisson model case by $P*(k; E_T, w/T)$, where $E_T = TE_1$. In computing $P(S_w \ge k)$ or $P(W_k \le w)$, the formula is sometimes simplified by choosing the scale of measurement to make $T = 1$ and by denoting E_1 by λ; when applying the simplified formula or using tabled values, care must be taken to interpret a λ consistent with the scale of measurement. To avoid confusion in what follows, we use the notation $P^*(k; E_T, w/T)$.

Exact and Approximate Formulae for Cluster Probabilities

In [43.18], asymptotic formulae are derived for $P^*(k; E_T, w/T)$, but these converge very slowly. The exact formulae in [43.8] for $P^*(k; E_T, w/T)$ are computationally intensive.

Table 2 in [43.10] gives $P^*(k; E_T, w/T)$ for $k = 3(1)9$, and a range of values for E_T and w/T. In Table 2, λ denotes E_T. Table 2a in [43.10] gives $P^*(k; 2E_w, 1/2)$, $P^*(k; 3E_w, 1/3)$, $P^*(k; 4E_w, 1/4)$ for $k = 3(1)9$, and a range of values for λ which denote $2E_w, 3E_w and 4E_w$ respectively in that table. Application (d) in [43.10, p. 4] illustrates how to use these values to accurately approximate $P^*(k; 2LE_w, 1/2L)$.

Reference [43.19] derives readily computable formulae for $P^*(k; 2E_w, 1/2)$ and $P^*(k; 3E_w, 1/3)$ and uses them to give the following highly accurate approximation for $P^*(k; E_T, w/T)$. Denote $1 - P^*(k; E_T, w/T)$, by $Q^*(k; E_T, w/T)$; $\exp(-\psi)\psi^j/j!$ by $p(j; \psi)$, and $\Sigma_{i \le k} p(j; \psi)$ by $F_p(k; \psi)$.

$$\begin{aligned} Q^*(k; E_T, w/T) \approx\; & Q^*(k; 2E_w, 1/2) \\ & \times [Q^*(k; 3E_w, 1/3) \\ & / Q^*(k; 2E_w, 1/2)]^{(T/w)-2}\,. \end{aligned}$$

$$
\begin{aligned}
Q^*(k;2\psi,1/2) &= [F_p(k-1;\psi)]^2 - (k-1)\\
&\quad \times p(k;\psi)p(k-2;\psi)\\
&\quad -(k-1-\psi)p(k;\psi)\\
&\quad \times F_p(k-3;\psi)\,,\\
Q^*(k;3\psi,1/3) &= (F_p(k-1;\psi))^3\\
&\quad -A_1+A_2+A_3-A_4\,, \qquad (43.10)
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= 2p(k;\psi)F_p(k-1;\psi)[(k-1)F_p(k-2;\psi)\\
&\quad -\psi F_p(k-3;\psi)]\,;\\
A_2 &= 0.5[p(k;\psi)]^2\left[(k-1)(k-2)F_p(k-3;\psi)\right.\\
&\quad \left.-2(k-2)\psi F_p(k-4;\psi)+\psi^2 F_p(k-5;\psi)\right]\,,\\
A_3 &= \sum_{r=1}^{k-1} p(2k-r;\psi)[F_p(r-1;\psi)]^2\,,\\
A_4 &= \sum_{r=2}^{k-1} p(2k-r;\psi)p(r;\psi)[(r-1)\\
&\quad \times F_p(r-2;\psi)-\psi F_p(r-3;\psi)]\,.
\end{aligned}
$$

Subsequent researchers [43.20] note the remarkable accuracy of this approximation. Tight bounds for $Q^*(k;E_T,w/T)$ are derived by [43.21], who proves that approximation (43.10) falls within the bounds. Our experience is that it gives great accuracy over the entire range of the distribution. For example, $P^*(4,10;0.1) = 0.374$, $P^*(5;12,0.25) = 0.765$; $P^*(5;8,1/6) = 0.896$ by both the approximation and the exact tabled values in [43.10]. One can readily compute $Q^*(k;2E_w,1/2)$ and $Q^*(k;3E_w,1/3)$ for any k or E_w, or use tabled values. An even better approximation can be obtained by taking

$$
\begin{aligned}
Q^*(k;\lambda T,w/T) &\approx Q^*(k;\lambda,1/3)\\
&\quad \times [Q^*(k;4\lambda,1/4)\\
&\quad /Q^*(k;\lambda,1/3)]^{(T/w)-3}\,. \qquad (43.11)
\end{aligned}
$$

One can use values from Table 2a in *Neff* and *Naus* for $Q^*(k;\lambda,1/3)$ and $Q^*(k;4\lambda,1/4)$; or alternatively compute $Q^*(k;4\lambda,1/4)$ using the results of [43.8]. This generalizes naturally to even more accurate approximations. Recently [43.22] other highly accurate approximations for $Q^*(k;L\lambda,1/L) = Q_L$ have been developed, together with error bounds. Using terms of the form $Q^*(k;\lambda,1/r) = Q_r$, for $r = 2,3,4,\ldots$; for example, approximation (1.18) from that work uses Q_2 and Q_3, and has a relative error $< 3.3(L-1)(1-Q_2)^2$, for $L > 4$, if the error bound is small relative to 1. Approximation (1.17) uses Q_r for $r = 2,3,4,5$, and has a smaller error bound, under certain conditions.

A simpler approximation is derived by [43.23], which is computable on a calculator, and is reasonably accurate for small to moderate values of $P^*(k;\lambda T,w/T)$ that might be used when testing hypotheses for unusual clusters.

$$
\begin{aligned}
P^*(k;E_T,w/T) &\approx 1-F_p(k-1;E_w)\\
&\quad \times\exp[-[(k-E_w)/k)]\\
&\quad \times\lambda(T-w)p(k-1;E_w)]\,. \qquad (43.12)
\end{aligned}
$$

For larger values of $P^*(k;E_T,w/T)$, (43.12) may not be accurate. For example, (43.12) gives $P^*(5,12,25) \approx 0.555$, as compared to the exact value of 0.765. In certain applications one seeks the distribution or moments of distribution of S_w, or the related statistics W_k, or $T_{k,w}$. If formulae for the moments are not available, one could average over the approximate distribution of the statistic, but in this case one would want to use an approximation that is accurate over the range of the distribution.

Example 2: A telecommunications engineer seeks to develop a system with the capacity to handle the possibility of multiple calls being dialed simultaneously. Dialing times start at random according to a Poisson process, with a 10 s dialing time. During an average 8 h busy period, 57 600 calls are dialed. The engineer asks how likely it is that at some point in the 8 h busy period there will be 50 or more phone calls being dialed simultaneously. There are an infinite number of overlapping intervals, each of 10 s duration, in an 8 h period. The maximum number of calls in any of the infinite number of overlapping windows is the scan statistic S_w. Here we are scanning a $T = 28\,800$ s period that has an expected number of calls $E_{28800} = 57600$, with a scanning window of $w = 10$ s, and asking how likely it is that $S_{10} \geq 50$. The answer needs to take into account the multiple comparisons involved in scanning the infinite number of overlapping 10 s periods within an 8 h period, and is given by $P^*(50;57600,10/28800)$, computed by (43.10) or (43.12).

Moments of Scan Statistic Distributions: Continuous Prospective Case

To compute the expectation, variance and other moments of S_w, W_k or $T_{k,w}$, one could average over the distribution of the statistic, where the cumulative distribution functions of W_k or $T_{k,w}$ are given by $P^*(k;E_T,w/T)$, and of S_w by $1-P^*(k+1;E_T,w/T)$. To derive formula or compute the moments, one could use either the

exact formula or approximation (43.10), which is highly accurate over the range of the distribution.

Example 3: A window information flow control scheme is described in [43.24] where a sender of information packets stops sending when there is evidence of overload. An open-loop control mechanism avoids feedback delays by basing the control mechanism on the maximum number of information packets in a sliding time window of fixed prespecified length, the scan statistic. Strong approximations are used in [43.24] to derive asymptotic (for $T \gg w > \log T$) results for $P(S_w \geq k)$. For $w = 20$, $T = 1\,000\,000$ and a Poisson process with an average of one observation per unit of time, their asymptotic approximation gives $\mathrm{AVE}(S_w) = 42$, compared to their simulated value of 44.5. Reference [43.3] (pp. 33–34) uses approximation (43.10) to compute $Q^*(k; 1\,000\,000, 20/1\,000\,000)$ for $k = 41(1)52$), which gives all of the distribution needed to compute $AVE(S_w) = \Sigma_{k>1}[1 - Q^*(k)] \approx 44.84$. This is because for $k < 41$, $Q^*(k) < Q^*(41) \approx 6.6E-7$; and for $k > 52$, $Q^*(k) > Q^*(52) \approx 0.9999$.

Various approximations are given by [43.25] and [43.26] for moments of $T_{k,w}$. For the Poisson process, [43.26] gives approximations and bounds for the expectation and variance of $T_{k,w}$, and bounds for the expectation for general point processes with i.i.d. interarrival times between the points. Details are given for the Poisson, Bernoulli, and compound Poisson processes. We now discuss Samuel–Cahn's results for the Poisson case. Let $\delta_{k,w}$ denote the total number of points observed until the first cluster of k points within an interval of length w occurs. Note that $\delta_{k,w}$, $T_{k,w}$, S_w, and W_k are different but interrelated statistics associated with the scanning process. For a Poisson process with mean λ per unit time, the expected waiting times between points is $1/\lambda$. She applies Wald's lemma, to find for the Poisson case,

$$E(T_{k,w}) = E(\delta_{k,w})/\lambda \,, \tag{43.13}$$

and derives a series of approximations for $E(\delta_{k,w})$. The simplest of these is

$$E(\delta_{k,w}) \approx k + \left\{ [F_p(k-2; \lambda w)]^2 / P(\delta_{k,w} = k+1) \right\} , \tag{43.14}$$

where

$$P(\delta_{k,w} = k+1) = \sum_{i=0}^{k-2} (-1)^{k-2-i} p(i; \lambda w) + (-1)^{k-1} \exp(-2\lambda w) \,, \tag{43.15}$$

and where $p(i; \lambda w)$ and $F_p(k-2; \lambda w)$ are Poisson terms defined before (43.10).

43.2.3 Discrete Binary Trials: The Prospective Case

In start-up tests for a piece of equipment, the equipment might perform successfully on the first test trial, then fail on the second. Consecutive points in a QC chart may be in or out of a warning zone. In a stream of items sampled from an assembly line, some are defective while some are acceptable. Here the data is viewed as a sequence of T binary outcome trials. Each trial t results in a "success" or "failure." For $t = w, w+1, \ldots, T$, $Y_t(w)$ and $E_t(w)$ are the observed and expected number of "successes" within the w consecutive trials, $t-w+1, t-w+2, \ldots, t$.

The scan statistic S_w is the maximum number of successes within any w contiguous trials within the T trials. For the special case where $S_w = w$, a *success-run* of length w has occurred within the T trials. When $S_w = k$, a *quota* of k successes within m consecutive trials has occurred. Related statistics include W_k, the smallest number of consecutive trials that contain k ones; $T_{k,w}$, the number of trials until we first observe at least k ones in an interval of length w; and V_r, the length of the longest number of consecutive trials that have at most r failures. V_0 is the length of the longest success run. The statistics are related by $P(S_w \geq k) = P(W_k \leq w) = P(T_{k,w} \leq T)$, and $P(V_r \geq k+r) = P(S_{k+r} \geq k)$. We illustrate these statistics in the following example.

Example 4: The DNA molecule most often consists of two complementary strands of nucleotides each consisting of a deoxyribose residue, a phosphate group, and a nucleotide base. The four nucleotide bases are denoted A, C, G, T, where an A on one strand links with a T on the other strand, and similarly C with G. Molecular biologists sometimes compare DNA from two different sources by taking one strand from each, viewing each as a linear sequence of the letters A, C, G, T, aligning the two sequences by a global criterion, and then looking for long perfectly or almost perfectly matching "words" (subsequences). For illustration, consider the following two aligned sequences from two different plant proteins. If letters in the same position in the two sequences match, we put an "s" at that position; if not an "f"

Source 1: A A A C C G G G C A C T A C G G T G A G A C G T G A

Source 2: A A T C C C C C G T G C C C T T A G A G G C G T G G
Match: s s f s s f f f f f f f f s s s f s s s f s s s s f
The longest perfectly *matching word* is CGTG, corresponding to a success run of length 4. For the s/f sequence, $V_o = 4$. We note that the longest word with at most one mismatch is of length eight letters (underlined). Here $V_1 = 8$. If we had scanned the sequence of $T = 26$ aligned pairs of letters, looking for the largest number of matches within $w = 8$ letters, we would find $S_8 = 7$. If we had looked for the smallest number of consecutive letters containing seven matches, we would find $W_7 = 8$. The waiting time until we first observe seven matches within eight consecutive letters is $T_{7,8} = 25$.

Exact Results

There are a variety of algorithms to compute the distribution of the prospective discrete scan statistic exactly. Recursion relations and generating functions for the special cases of $k = w - 1$ and $k = w$ are given in [43.27]. The Markov chain imbedding approach was applied in [43.28] and [43.29] to derive an exact formula for the expected waiting time until a k-in-w quota, and is refined and unified in [43.30] to be efficient for computing the distributions of runs and scans. Useful recurrence relations and other formulae are given in [43.31–33]. Recently, a martingale approach has been used to find generating functions and moments of scan statistics [43.34].

Recent studies [43.35] and [43.36] into the computational complexity of the Markov chain approach to find exact results for the discrete scan statistic shows that it is computationally feasible. Many of the results are motivated by quality control and acceptance sampling applications [43.27, 33, 37]. Other recent results are motivated by the reliability of linear systems, where the system fails if any k within w consecutive components fail [43.1, 30, 38–40].

Approximate Results

There are asymptotic results for $P(S_w \geq k)$ for a variety of probability models that are called *Erdös–Rényi laws*, or Erdös–Rényi–Shepp theorems. DNA and protein sequence matching has stimulated further generalizations (see [43.41–44]). A simple random model is the Bernoulli trials model, where the T trials are independent binary trials, with a probability of "success" on trial t equal to a constant value p. Reference [43.45] reviews and proves some important general limit law results (as T tends to infinity), and shows how they apply in the special case of a Bernoulli process. The asymptotic results converge quite slowly, and for certain applications give only rough approximations ([43.3], pp. 233-235). Various approximations to $P(S_w k)$ are derived using the method of moments, a Poisson approximation using declumping, and other methods [43.46].

For the Bernoulli process, denote $P(S_w \geq k)$ by $P'(k|w; T; p) = 1 - Q'(k|w; T; p)$.

In [43.19], the following highly accurate approximation is given for $Q'(k|w; T; p)$. Let $b(k; w, p)$ be the binomial probability defined in (43.5), and let

$$
\begin{aligned}
F_b(r; w, p) &= \Sigma_{i<r} b(i; w, p)\,, \\
Q'(k|w; T; p) &\approx Q'(k|w; 2w; p)\left[Q'(k|w; 3w; p)\right. \\
&\quad \left./\,Q'(k|w; 2w; p)\right]^{(T/w)-2}\,, \\
Q'(k|w; 2w; p) &= [F_b(k-1; w, p)]^2 \\
&\quad -(k-1)b(k; w, p) \\
&\quad \times F_b(k-2; w, p) \\
&\quad + wpb(k; w, p) \\
&\quad \times F_b(k-3; w-1, p)\,, \\
Q'(k|w; 3w; p) &= (F_b(k-1; w, p))^3 \\
&\quad - A_1 + A_2 + A_3 - A_4\,, \qquad (43.16)
\end{aligned}
$$

where

$$
\begin{aligned}
A_1 &= 2b(k; w, p)F_b(k-1; w, p)\,[(k-1) \\
&\quad \times F_b(k-2; w, p) - wpF_b(k-3; w-1, p)]\,; \\
A_2 &= 0.5[b(k; w, p)]^2\left[(k-1)(k-2)\right. \\
&\quad \times F_b(k-3; w, p) \\
&\quad -2(k-2)wpF_b(k-4; w-1, p) \\
&\quad \left.+(w-1)p^2F_b(k-5; w-2, p)\right]\,; \\
A_3 &= \sum_{r=1}^{k-1} b(2k-r; w, p)[F_b(r-1; w, p)]^2\,; \\
A_4 &= \sum_{r=2}^{k-1} b(2k-r; w, p)b(r; w, p)\,[(r-1) \\
&\quad \times F_b(r-2; w, p) - wpF_b(r-3; w-1, p)]\,.
\end{aligned}
$$

The following simpler, and fairly accurate, approximation is suggested in [43.47].

$$
P'(k|w; T; p) \approx 1 - \left[C(D/C)^{(T/w)-2}\right]\,, \qquad (43.17)
$$

where

$$
\begin{aligned}
C &= 2\Sigma_{i<k} b(i; w, p) - 1 - (k-1-wp)b(k; w, p)\,, \\
D &= 2\Sigma_{i<k} b(i; w, p) - 1 - (2k-1-2wp) \\
&\quad \times b(k; w, p)\,.
\end{aligned}
$$

Example: $k = 8$, $w = 10$, $p = 0.3$, $T = 50$. Applying (43.16), $Q'(8|10; 20; 0.3) = 0.991032$; $Q'(8|10; 30; 0.3) = 0.983827$; $P'(8|10; 50; 0.3) \approx 0.03$.

Applying (43.17) gives $C = 0.991032$; $D = 0.9837985$; $P'(8|10; 50; 0.3) \approx 0.03$. Both of these results agree with the exact result from the Markov Chain imbedding approach.

In some applications the researcher is interested in scanning for unusual clusters of successes or clusters of failures. One can use the above results to bound the probability of either type, or can use the Markov Chain imbedding or another approach to compute the probability directly [43.39]. In other applications, individual trials can result in an integer number of points and [43.48] derives accurate approximations and tight bounds for this case.

Moments of Discrete Trial Scan Statistics

An accurate approximation for the expected waiting time until a k in w cluster in a Bernoulli trials, with a probability p of *success* on an individual trial, is given in [43.19]

$$E(T_{k,w}) \approx 2w + Q'(k|w; 2w; p) / \left\{1 - [Q'(k|w; 3w; p) / Q'(k|w; 2w; p)]^{1/w}\right\} . \quad (43.18)$$

More complicated formulae for the exact expectation, and methods to find it, are given in [43.28, 29], and other approximations in [43.26] for the expectation and variance of the waiting time for the Bernoulli and more general processes.

An important generalization is to the case of independent binary trials with unequal probabilities. For many results see [43.1]. Reliability engineers have been studying the reliability of a linearly ordered set of N independent components with different probabilities of being defective. In the *k-within-consecutive-m-out-of-N systems*, the system fails if there is a quota of k defectives within any m consecutive components in the system (see [43.40, 49]).

43.2.4 Discrete Binary Trials: The Retrospective Case

In this discrete scenario, the data is viewed as a sequence of T binary outcome trials. Each trial t results in a "success" or "failure." S_w is the largest number of successes in any w consecutive trials. A simple random model is where there are N successes distributed at random over the T trials. Denote $P(S_w \geq k)$ by $P(k|w; N, T)$.

The exact distribution of $P(k|w; N, T)$ for the case $k > N/2$ is given in [43.31]. This formula can also be used to approximate small values of $P(k|w; N, T)$ for $k < N/2$, and the approximation is given below in (43.19). Reference [43.50] derives the exact distribution for all N and k, for $m/N = 1/L$, L an integer, and more generally for $m = cR$, $N = cL$, where c, R, and L are integers, and $c > 1$. These general formulae are computationally complex, but can be used to derive simpler formulae for special cases. These formulae can also be used to approximate other cases.

Approximations and bounds for $P(k|w; N, T)$ are given in [43.3] (pp. 56–58, 212–216) and in [43.1] (pp. 319–323). These can then be used to approximate the moments of the distributions of S_w and W_k. A simple approximation is given by

$$P(k|w; N, T) = [(k - E_w)(1/w) + 1]P(Y_t = k) + 2P(Y_t > k) , \quad (43.19)$$

where

$$E_w = N(w/T) ;$$
$$P(Y_t = k) = H(k, N, w, T) = \binom{w}{k}\binom{T-w}{N-k} / \binom{T}{N} .$$

For the case where w and T are large, the discrete retrospective scan probability $P(k|w; N, T)$, can be approximated by the continuous retrospective scan probability $P(k; N, w/T)$.

43.2.5 Ratchet-Scan: The Retrospective Case

In the ratchet scenario, time is divided into T disjoint intervals (hours, days, or weeks) and the reported data consists of the number of events in each interval. For $t = w, w+1, \ldots, T$, $Y_t(w)$ and $E_t(w)$ are the observed and expected number of events within the w consecutive intervals, $t - w + 1, t - w + 2, \ldots, t$. For the constant background retrospective case there are N events spread at random over the T intervals. (The model is multinomial where each of the N balls is equally likely to fall in any of the T cells, independently of the other balls.) Here $\lambda = wN/T$. Approximations and bounds for $P(k; \lambda, w, T)$ are described in [43.3] (pp. 327–328), [43.2] (p. 81–91), and [43.51]. The simple Bonferroni upper bound is

$$P(k; \lambda, w, T) < (T - w + 1)\Sigma_{i \geq k} b(i; N, p) , \quad (43.20)$$

where $N = T\lambda/w$; $p = 1/T$. Note also that $P(k; \lambda, w, T)$ for the ratchet scan must be less than that for the continuous scan. Thus, $P(k; \lambda, w, T)$ for the retrospective ratchet scan must be less than the right-hand side of (43.5)

43.2.6 Ratchet-Scan: The Prospective Case

In the ratchet scenario, time is divided into T disjoint intervals (hours, days, or weeks) and the reported data consists of the number of events in each interval. For $t = w, w+1, \ldots, T$, $Y_t(w)$ and $E_t(w)$ are the observed and expected number of events within the w consecutive intervals, $t-w+1, t-w+2, \ldots, t$. For the constant background prospective case the number of events (counts) in the T intervals are T independently and identically distributed Poisson random variables, each with expectation λ/w. For the prospective ratchet scan constant background case, the approach in [43.5] gives the following approximation:

$$P(k; \lambda, w, T) \approx (T-w+1)G(k, \lambda) - (T-w) \times \Sigma_{j=0} p(j, \lambda)[G(k-j, \lambda/w)]^2 , \quad (43.21)$$

Here $p(j, \lambda)$ is the Poisson probability defined in (43.20), and

$$G(k, \lambda) = \Sigma_{j \geq k} p(j, \lambda) .$$

For the case where the number of events in the T intervals are independently and identically distributed random variables, [43.48] derives accurate approximations and tight bounds for $P(k; \lambda, w, T)$.

43.2.7 Events Distributed on the Circle

When studying seasonality effects of disease patterns over time, one might be interested in unusual clustering during certain periods of the year. The researcher may view the time period as a circle, with January following December. In studying the directions of flights of birds or insects, the directions may be viewed as points on a circle. Many of the distributional results described above for the line have also been derived for scan statistics on the circle. The circular ratchet scan was introduced in [43.52] and further developed in [43.51]; accurate approximations for the continuous scan on the circle and line are given in [43.19]; and for the discrete scan on the circle in [43.53].

43.3 Higher Dimensional Scans

Scan statistic have been applied extensively to study the clustering of diseases over space. In two dimensions, circular, rectangular, elliptical and other shaped scanning windows are used. Regions scanned can have arbitrary shapes (typical in epidemiology), or the surface of a sphere (astronomy or ships at sea), or approximately rectangular (blood cells on a slide). In three dimensions, two of which may be spatial and one temporal, cylindrical scanning windows may be used. In the case where all three dimensions are spatial, a scanning sphere or cube may be used. Scan distributions have been approximated or simulated for events from homogeneous or heterogeneous Poisson processes, uniform, or more generally distributed points in continuous space, and for various models for two-dimensional grids. Simulation is the most widely applied approach, but can be computationally intensive for a large number of points. Not only must the points be generated, but one must check all possible positions of the scanning window, and in variable-size window applications [43.54–57] each of the window sizes must be checked too. A variety of algorithms have been developed to simulate scan probabilities in two dimensions. Efficient Monte Carlo algorithms for one and two dimensions are developed in [43.55, 58, 59]; (see [43.3], Chapt. 16). Importance sampling can be used to reduce the computational effort of the simulation [43.60]. A web-based program, SatScan [43.56], is available that scans generally distributed points over an arbitrarily shaped region with circular scanning windows (with a variety of diameters).

In two dimensions, the points are randomly distributed over a two-dimensional region. A circular (or rectangular or other shaped) window scans the region. The scan statistic S_w is the largest number of points in any window of diameter w; the scan statistic W_k generalizes to the diameter of the smallest scanning window that contains k points. The distributions of the statistics S_w and W_k are still related, $P(S_w \geq k) = P(W_k \leq w)$. However, in more than one dimension, W_{k+1} is not equivalent to the smallest k-th-nearest neighbor distance among the N points. Section 43.3.1 discusses the retrospective case of a fixed number of points in the unit square; Sect. 43.3.2 discusses the prospective case where the number of points is Poisson distributed.

43.3.1 Retrospective Continuous Two-Dimensional Scan

We first focus on the problem of scanning a rectangular region with a rectangular window with sides of length u and v that are oriented parallel to the sides of the square. There are a fixed number, N, of points distributed over the rectangular region, The units of measurement are chosen to make the rectangular region the unit square. The results for the unit square give the results for an $a \times b$ scanning window within an $S \times T$ rectangular region, by choosing the units of measurement for the $x-$ and $y-$axes to make $S = 1$ unit on the x-axis, and $T = 1$ unit on the y-axis, and by setting $u = a/S$ and $v = b/T$.

Given N points distributed at random over the unit square, let $S_{u,v}$ denote the maximum number of points in any subrectangle with sides of length u and height v parallel to the sides of the unit square. Let $P(k; N, u, v)$ denote $P(S_{u,v} \geq k)$. Bounds for $P(k; N, u, v)$, that converge for small u, v are given in [43.61], and the approximation for this case is

$$P(k; N, u, v) \cong k^2 \binom{N}{k} (uv)^{k-1} . \tag{43.22}$$

An exact formula is available for $P(N-1; N; u, v)$, (see [43.3], Chapt. 16). The following approximation by [43.62] is based on large deviation theory

$$\begin{aligned} P(k; N, u, v) \cong & \left\{ \left[N^2 w(1-u)(1-v)E^3 \right. \right. \\ & \left. \left. /(1-w)^3(1+E) \right] + C \right\} \\ & \times b(k; N, w) , \end{aligned} \tag{43.23}$$

where

$$\begin{aligned} w = uv,\ E &= (k/Nw) - 1 ; \\ b(k; N, w) &= \binom{N}{k} w^k (1-w)^{N-k} , \\ C &= [Nv(1-u)E/(1-w)] \\ &\quad + \left[Nu(1-v)E^2/(1+E)(1-w)^2 \right] \\ &\quad + [(1+E)(1-w)/E] . \end{aligned}$$

The approximation is refined further in [43.62], but (43.23) appears to give good accuracy for small $P(k; N, u, v)$. Simulation is used to evaluate larger values of $P(k; N, u, v)$. Order the points by the X coordinates, with $X_1 < X_2 \leq ... \leq X_N$. For each $i = 1, ..., N - k + 1$, check whether $(X_{k+i-1} - X_i \leq u)$, and if so, whether the corresponding Y's fall within a distance v.

For the case of a circular scanning window of radius r, the algorithm looks at pairs of points within a distance $2r$, finds the two circles of radius r on which the two points fall on the circumference, and counts the number of points in each of the circles.

The above generalizes to higher dimensions. Scan the r-dimensional unit cube with a rectangular block with sides of length $u_1, ..., u_r$, oriented parallel to the sides of the unit cube. Let $w = \prod u_i$ denote the volume of the scanning rectangle. Let $P(k; N, u_1, u_2, ..., u_r)$ denote the probability that at least one scanning rectangular block with sides $(u_1, u_2, ..., u_r)$ parallel to those of the unit cube contains at least k points. In corollary 2.3, [43.63] gives the approximation for $r = 1, 2, 3$

$$\begin{aligned} P(k; N, u_1, u_2, ..., u_r) \cong & \{1 - [Nw/k(1-w)]\}^{2r-1} \\ & \times (k^r/w) b(k : N, w) , \end{aligned} \tag{43.24}$$

where $w = \prod u_i$, and $b(k; N, w)$ is the binomial probability in (43.23). For the case $r = 2$, (43.24) does not reduce exactly to (43.23), but gives similar values for small probabilities.

43.3.2 Prospective Continuous Two-Dimensional Scan

Let the number of points in the unit square be a Poisson-distributed random variable with mean λ. Scan the unit square with a subrectangle with sides of length u and v that are parallel to the sides of the square. Let $P^*(k; \lambda, u, v)$ denote the probability that at least one uxv-scanning subrectangle contains at least k points. Several approximations from [43.64] and [43.65] are given for $P^*(k; \lambda, u, v)$, with the best approximation being (from [43.65]):

$$\begin{aligned} P^*(k; \lambda, u, v) \cong & 1 - F_p(k-1; \lambda uv) \\ & \exp\{ -\zeta(1 - (\lambda uv/k))\lambda v(1-u) \\ & \times p(k-1; \lambda uv) \} , \\ \zeta = & [1 - (\lambda uv/k)]\lambda u(1-v) \\ & \times [P^*(k-1; \lambda v, u) \\ & - P^*(k; \lambda v, u)] ; \\ F_p(k-1; \lambda uv) = & \sum_{i=0}^{k-1} p(i; \lambda uv) , \end{aligned} \tag{43.25}$$

where $P^*(k; \lambda v, u)$ is the one-dimensional scan statistic. A simpler but rougher approximation is

$$\begin{aligned} P^*(k; \lambda, u, v) \cong & 1 - [1 - P^*(k-1; \lambda v, u)] \\ & \exp(-\zeta) . \end{aligned} \tag{43.26}$$

For the case of scanning an $S \times T$ rectangular region with a circular window of radius r, let $r\pi^{0.5}/S = u$, $r\pi^{0.5}/T = v$, and choose the X- and Y-scales so that $S = 1$ and $T = 1$. Let $P_c^*(k; \lambda, u, v)$ denote the probability that the maximum number of points in the scanning window is at least k, where λuv denotes the expected number of points in a circle of radius r. The following approximation is given in [43.64]:

$$P_c^* \cong 1 - \exp[-k\lambda(1-2u)(1-2v)p(k-1; \lambda uv)] . \quad (43.27)$$

The simulations in [43.65, 66] show that for small u, v, $P_c^* \cong P^*(k; \lambda, u, v)$. For small u, v, we approximate the circular window by a square window of the same area.

Researchers continue to develop more accurate approximations for the distribution of the two-dimensional scan statistic for the Poisson process; see [43.67] and [43.25]. Other researchers seek to generalize the asymptotic results and some approximations to higher dimensions or more general scanning regions. The approach of [43.65] is generalized in [43.66] to give approximations for more than two dimensions. Recent research [43.68] derives approximate results with error bounds for one and higher dimensions for general distributions, and general bounded scanning sets, and uses these results to prove various asymptotic results. Poisson approximation and large deviation theory has been used in [43.69] to derive very general results for scanning spatial regions of two and more dimensions.

43.3.3 Clustering on the Lattice

In some applications, events can only occur on a grid of points; a recent example [43.70] deals with the reliability of a two-dimensional system. Another example involves clustering of diseased plants in a field of evenly spaced plants. In other applications, the method of measurement limits the observed events to occurring on a rectangular R by T lattice of points. The researcher scans the grid looking for unusual clusters. Results have been developed for several models; The researcher scans the lattice with a rectangular m_1 by m_2 sublattice with sides parallel to those of the lattice; events are independent and equally likely to occur at any point of the lattice. This case is the discrete analog of the continuous prospective case with an oriented rectangular scanning window, as discussed in Sect. 43.3.2.

Let X_{ij} for $i = 1, .., R$; $j = 1, ..., T$ denote a rectangular lattice of independent and identically distributed Bernoulli random variables, where $P(X_{ij} = 1) = p = 1 - P(X_{ij} = 0)$. View the lattice with position (1,1) in the lower left corner. Let $Y_{r,s}(m_1, m_2)$ denote the number of events (ones) in an m_1 by m_2 subrectangle whose lower left corner is at (r, s) in the lattice. Denote the *two-dimensional discrete scan statistic* $S'_{m1,m2}$, the maximum number of events in the scanning subrectangle when we scan the R by T lattice with an m_1 by m_2 rectangle of points (with sides oriented with the sides of the lattice).

Algorithms are given in [43.71] to find the largest rectangle with all 1's in the lattice, and [43.71] generalizes the limit law to higher dimensions, and to allow for some 0's in the subrectangle. Approximations for $P\left(S'_{m1,m2} \geq k\right)$ are given in [43.72]. For the special case of a square lattice, $R = T$, and a square scanning subrectangle $m_1 = m_2 = m$, let B_s denote the event $[Y_{1,s}(m, m) < k]$. Then,

$$P\left(S'_{m,m} \geq k\right) \cong 1 - q(2m-1) \times [q(2m) / q(2m-1)]^{(T-2m+1)(T-m+1)} , \quad (43.28)$$

where

$$q(2m-1) = P(B_1 B_2 ... B_m) ,$$
$$q(2m) = P(B_1 B_2 ... B_{m+1}) .$$

The terms $q(2m-1)$ and $q(2m)$ can be evaluated using an algorithm in [43.73].

For m_1, m_2, R, and T all large, the discrete scan probability can be approximated by the continuous case probability in Sect. 43.3.2.

$$P^*(k; \lambda, u = m_1/R; v = m_2/T) \cong P\left(S'_{m,m} \geq k\right) . \quad (43.29)$$

43.4 Other Scan Statistics

43.4.1 Unusually Small Scans

Scan statistics have been developed to test for unusually sparse intervals of time, or regions of space. Given N points independently drawn from the uniform distribution on [0, 1), let D_w denote the smallest number of points in an interval of length w; let V_k denote the size of the largest subinterval of [0, 1) that contains k points.

The statistics are related:

$$P(D_w \le k) = P(V_k \ge w)\,. \tag{43.30}$$

The scan statistic V_{k+1} can be viewed as the maximum of the sum of k consecutive spacings. The interval V_{r+1} is called the "maximum r-th-order gap".

A variety of results have been derived for the related statistics D_w and V_k. In [43.14] asymptotic results are derived for the distributions of order statistics of r-th order gaps, for i.i.d. distributed spacings (gaps between consecutive points) and some more general uniform mixing stationary processes. For the special case of a minimum r-scans of $n(= N+1)$ i.i.d. uniform [0, 1) spacings, they give

$$\lim_{n->\infty} P\{V_{r+1} \le [\log_e n + (r-1) \times \log_e \log_e n + y]/n\} = \exp[\,e^{-y}/(r-1)!]\,. \tag{43.31}$$

The asymptotic convergence of (43.31) is very slow, and care must be taken in using it as an approximation. For the case of N uniformly distributed points on the line, the exact distribution of D_w and V_k are derived in [43.74]. The formulae are computationally intensive, but can be applied to derive special cases that can be used to find highly accurate approximations similar in form to (43.16). A general approach to finding the distribution of V_k is described in [43.11].

For a sequence of Bernoulli trials, the distribution of the minimum scan can be computed using the formula for the maximum scan. Let D_w^* denote the minimum number of successes in any w consecutive trials. Then $D_w^* \ge k$ if the maximum number of failures in any w trials is less than or equal to $w-k$.

The distribution of the minimum scan for the line and circle is related to a multiple coverage problem. The multiple coverage problem is as follows: Given N subarcs each of length w dropped at random on the circumference of the unit circle, what is the probability that the arcs completely cover the circumference of the circle at least m times? To relate the coverage to the scan problem, let the N points in the scan problem correspond to the midpoints of the subarcs in the coverage problem. If the N subarcs do not cover the circle m times, there must be some point on the circumference not covered enough. This implies that the subinterval of length w centered at that point contains fewer than m of the N midpoints of the N subarcs; in this case, the minimum number of the N (mid)points in a scanning window of length w must be less than m.

43.4.2 The Number of Scan Clusters

There are several ways to count clusters depending on the overlap allowed between two clusters. For example, suppose we scan (0,1) with a window of width $w = 0.20$, looking for clusters of at least $k = 3$ points. Assume that we observe the six points: 0.10, 0.15, 0.28, 0.34, 0.41, 0.62. One approach to counting the number of 3-within-0.20 clusters is to see how many of the 2-spacings are $< w$. Here, $0.28 - 0.10 = 0.18 < w$, $0.34 - 0.15 = 0.19 < w$, $0.41 - 0.28 = 0.13 \le w$, and $0.62 - 0.34 > w$. We would count three clusters. Note that two clusters share some (but not all) points, and this case is sometimes referred to as a nonoverlapping case. For the conditional case, expressions are available for the expectation and variance of the number of clusters. See [43.75] and [43.11] for this method of counting. Results are derived in [43.14] for the distribution of the number of clusters that do not overlap with a previously counted cluster A Markovian declumping is used in [43.76] and [43.77], where an r-spacing$< w$ is counted so long as the immediately previous r-spacing is not counted. A compound Poisson approach to counting the number of clumps is applied in [43.11, 76, 78] and [43.12]. This is generalized in [43.79] to two and higher dimensions. See ([43.3], Chapt. 17) for a summary of many of the results for the continuous scan.

For a sequence of trials, ([43.1], Chapt. 10) describes in detail the Markov chain approach to finding the distribution of the waiting time until the k-th discrete scan cluster, for different ways of counting clusters. In Chapt. 4, [43.1] discusses the discrete counting approach for the case of runs, and there is a great deal of literature on the number of overlapping and nonoverlapping runs; see also [43.80]. In [43.81], a compound Poisson approximation is used for the multiple cluster problem, and the motivating example is the clustering of individual claims exceeding threshold risks.

43.4.3 The Double-Scan Statistic

In [43.82], a scan-type statistic called the *double-scan statistic* is defined based on the number of "declumped" (a type of nonoverlapping) clusters that contain at least one of each of two types of event. The expectation and approximate distribution of the number of declumped clusters is derived for this test statistic for two chance models. Define the event $E(i)$ to have occurred if there are at least one of each of the two types of events anywhere within the w consecutive days $i, i+1, \ldots, i+w-1$. The event $E(i)$ indicates the occurrence

of a two-type w-day cluster. Let $Z(i) = 1$ if $E(i)$ occurs and none of $E(i-1)$, $E(i-2)$, ..., $E(i-w+1)$ occur; $Z(i) = 0$ otherwise. Let $S_{w(2)} = \Sigma_{1<i<N-w+1} Z_i . S_{w(2)}$ is the *double scan statistic*. This method counts the number of times that an $E(i)$ occurs with no previously overlapping $E(i)$'s. When the events are relatively rare and distributed according to certain chance models, the number of these declumped clusters is approximately Poisson-distributed [43.64]. This model fits well when the two types of events do not occur too frequently. As the density of the number of events increases, there comes a point where more events lead to fewer clusters. In [43.83], the distributions of a family of double-scan statistics are derived using a different approach to count clusters. This method treats clusters as recurrent events, and counts the number of times that an $E(i)$ occurs with no previously overlapping $E(i)$'s that were counted in a cluster.

The double-scan statistic can be generalized, for the case of two types of events, to clusters where there are at least r type one and s type two events within a w-day period. For other applications, the statistics can be generalized to more than two types of events, and the distribution of the number of declumped clusters can be derived.

43.4.4 Scanning Trees and Upper Level Scan Statistics

The scanning approach has been extended to a variety of data structures. In two-dimensional geographic scanning, the SatScan [43.56] simulation approach uses a range of circular window sizes to detect unusual clusters. When studying patterns of disease, a researcher may be looking for connected regions (such as counties) with above-average concentrations of disease. These regions may not all fall within a compact circular, elliptical or other simple shape. One can still scan for these linked high-density regions, and [43.84] develops a "higher-level scan" approach for assessing the unusualness of clusters that arise.

In other situations, the researcher may not be looking for clusters of disease in space, but instead for some other variable such as occupation. One might try to scan a data set combining occupations that have above-average incidence. However, scanning all possible combinations involves so many multiple comparisons that the test statistic adjusted for this would not be powerful, and the results would be difficult to interpret. In the case of occupations (and many other variables) the data can be placed in the form of a tree structure (for example, a carpenter, electrician, plumber are all under building trades). In [43.85] an approach is given to scanning data in a tree structure, looking for unusual clusters, and adjusting for the multiple comparisons made.

In some applications one seeks to scan with a range of window sizes. Simulation is typically used to test unusual clusters [43.54,56]. Asymptotic results [43.62,63, 68,69] and approximations [43.57] are also available for a variety of models.

References

43.1 N. Balakrishnan, M.V. Koutras: *Runs and Scans with Applications*, Vol. 1 (Wiley, New York 2001)

43.2 J. Glaz, N. Balakrishnan: *Scan Statistics and Applications* (Birkhauser, Boston 1999)

43.3 J. Glaz, J. Naus, S. Wallenstein: *Scan Statistics* (Springer, Berlin Heidelberg 2001)

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

43.5 S. Wallenstein, N. Neff: An approximation for the distribution of the scan statistic, Stat. Med. **6**, 197–207 (1987)

43.6 J. I. Naus: Some probabilities, expectations, variances for the size of the largest clusters, smallest intervals, J. Am. Stat. Assoc. **61**, 1191–1199 (1966)

43.7 S. Wallenstein, J. Naus: Probabilities for a k–th nearest-neighbor problem on the line, Ann. Prob. **1**, 188–190 (1973)

43.8 R. J. Huntington, J. I. Naus: A simpler expression for K–th nearest neighbor coincidence probabilities, Ann. Prob. **3**, 894–896 (1975)

43.9 S. Wallenstein, J. Naus: Probabilities for the size of largest clusters, smallest intervals, J. Am. Stat. Assoc. **69**, 690–697 (1974)

43.10 N. Neff, J. Naus: *Selected Tables in Mathematical Statistics, Volume VI: The distribution of the Size of the Maximum Cluster of Points on a Line* (Am. Math. Soc., Providence 1980)

43.11 F. W. Huffer, C-T. Lin: Computing the exact distribution of the extremes of sums of consecutive spacings, Comput. Stat. Data Anal. **26**, 117–132 (1997)

43.12 J. Glaz, J. Naus, M. Roos, S. Wallenstein: Poisson approximations for the distribution, moments of ordered m-spacings, J. Appl. Prob. **31A**, 271–81 (1994)

43.13 N. Cressie: On the minimum of higher order gaps, Australian J. Stat. **19**, 132–143 (1977)

43.14 A. Dembo, S. Karlin: Poisson approximations for r-scan processes, Ann. Appl. Prob. **2**, 329–357 (1992)
43.15 J. Glaz: Approximations, bounds for the distribution of the scan statistic, J. Am. Stat. Assoc. **84**, 560–569 (1989)
43.16 J. Glaz: Approximations for tail probabilities, moments of the scan statistic, Stat. Med. **12**, 1845–1852 (1993)
43.17 S. Wallenstein, M. S. Gould, M. Kleinman: Use of the scan statistic to detect time-space clustering, Am. J. Epidemiology **130**, 1057–1064 (1989)
43.18 G. F. Newell: *Distribution for the Smallest Distance Between any Pair of K–th Nearest-Neighbor Random Points on a Line*, ed. by M. Rosenblatt (Wiley, New York 1963) pp. 89–103
43.19 J. I. Naus: Approximations for distributions of scan statistics, J. Am. Stat. Assoc. **77**, 177–183 (1982)
43.20 M. Berman, G. K. Eagleson: A useful upper bound for the tail probability of the scan statistic when the sample size is large, J. Am. Stat. Assoc. **84**, 560–566 (1985)
43.21 S. Janson: Bounds on the distributions of extremal values of a scanning process, Stochastic Proc. Appl **18**, 313–328 (1984)
43.22 G. Haiman: Estimating the distributions of scan statistics with high precision, Extremes **3**, 4349–361 (2000)
43.23 S. E. Alm: On the distribution of a scan statistic of a Poisson process. In: *Probability and Mathematical Statistics*, ed. by A. Gut, L. Helst (Univ. Press, Upsalla 1983) pp. 1–10
43.24 A. W. Berger, W. Whitt: Asymptotics for open-loop window flow control, J. App. Math. Stochastic Anal. , 70 (1993) Special issue in honor of Lajos Takac's 70^{th} birthday
43.25 M. Mansson: Poisson approximation in connection with clustering of random points, Ann. Appl. Prob. **9**, 465–492 (1999)
43.26 E. Samuel-Cahn: Simple approximations to the expected waiting time for a cluster of any given size, for point processes, Adv. Appl. Prob. **15**, 21–38 (1983)
43.27 S. W. Roberts: Properties of control chart zone tests, Bell System Tech. J. **37**, 83–114 (1958)
43.28 I. Greenberg: The first occurrence of N successes in M trials, Technometrics **12**, 627–634 (1970)
43.29 R. J. Huntington: Mean recurrence times for k successes within m trials, J. Appl. Prob. **13**, 604–607 (1976)
43.30 M. V. Koutras, V. A. Alexandrou: Runs, scans, urn model distributions: A unified Markov chain approach, Ann. Inst. Stat. Math. **47**, 743–766 (1995)
43.31 B. Saperstein: The generalized birthday problem, J. Am. Stat. Assoc. **67**, 425–428 (1972)
43.32 B. Saperstein: Note on a clustering problem, J. Applied Prob. **12**, 629–632 (1975)
43.33 B. Saperstein: The analysis of attribute moving averages: MIL-STD-105D reduced inspection plans, Sixth Conf. Stochastic Proc. Appl. Tel Aviv (1976)
43.34 V. J. Pozdnyakov, J. Glaz, M. Kulldorff, M. Steele: A martingale approach to scan statistics, Ann. Inst. Stat. Math. (2004) in press
43.35 F. K. Hwang, P. E. Wright: An O(n/log n) algorithm for the generalized birthday problem, Comp. Stat. Data Anal. **23**, 443–451 (1997)
43.36 H. P. Chan, T. L. Lai: Boundary crossing probabilities for scan statistics, their applications to change point detection, Method. Comp. Appl. Prob. **4**, 317–336 (2002)
43.37 G. Shmueli: System wide probabilities for systems with runs, scans rules, Method. Comp. Appl. Prob. **4**, 401–419 (2003)
43.38 J. C. Chang, R. J. Chen, F. K. Hwang: A minimal-automation-based algorithm for the reliability of Con(d,k,n) system, Method. Comp. Appl. Prob. **3**, 379–386 (2001)
43.39 J. C. Fu: Distribution of the scan statistic for a sequence of bistate trials, J. Appl. Prob. **38**, 908–916 (2001)
43.40 S. Kounias, M. Sfakianakis: The reliability of a linear system, its connection with the generalized birthday problem, Stat. Appl. **3**, 531–543 (1991)
43.41 R. A. Arratia, L. Gordon, M. S. Waterman: The Erdös Rényi Law in distribution for coin tossing, sequence matching, Ann. Stat. **18**, 539–570 (1990)
43.42 R. A. Arratia, M. S. Waterman: The Erdös Rényi strong law for pattern matching with a given proportion of mismatches, Ann. Prob. **17**, 1152–1169 (1989)
43.43 S. Karlin, F. Ost: Counts of long aligned word matches among random letter sequences, Adv. Appl. Prob. **19**, 293–251 (1987)
43.44 M. Mansson: On compound Poisson approximation for sequence matching, Comb. Prob. Comp. **9**, 529–548 (2000)
43.45 P. Deheuvels: On the Erdös Rényi theorem for random fields, sequences, its relationships with the theory of runs, spacings, Z. Wahrsch. **70**, 91–115 (1985)
43.46 R. A. Arratia, L. Goldstein, L. Gordon: Poisson approximation, the Chen–Stein method, Stat. Sci. **5**, 403–434 (1990)
43.47 S. Wallenstein, J. Naus, J. Glaz: Power of the scan statistic in detecting a changed segment in a Bernoulli sequence, Biometrika **81**, 595–601 (1995)
43.48 J. Glaz, J. Naus: Tight bounds, approximations for scan statistic probabilities for discrete data, Ann. Appl. Prob. **1**, 306–318 (1991)
43.49 M. V. Koutras: Consecutive k, r-out-of-n: DFM systems, Microelectronics Reliab. **37**, 597–603 (1997)
43.50 J. I. Naus: Probabilities for a generalized birthday problem, J. Am. Stat. Assoc. **69**, 810–815 (1974)

43.51 J. Krauth: Bounds for the upper tail probabilities of the circular ratchet scan statistic, Biometrics **48**, 1177–1185 (1992)

43.52 S. Wallenstein, C. R. Weinberg, M. Gould: Testing for a pulse in seasonal event data, Biometrics **45**, 817–830

43.53 J. Chen, J. Glaz: Approximations for discrete scan statistics on the circle, Stat. Prob. Lett. **44**, 167–176 (1999)

43.54 N. Nagarwalla: A scan statistic with a variable window, Stat. Med. **15**, 845–50 (1996)

43.55 J. Glaz, N. Balakrishnan: Spatial scan stat: Models, calculations and applications. In: *Scan Statistics and Applications*, ed. by M. Kulldorff (Birkhauser, Boston 1999) pp. 303–322

43.56 M. Kulldorff, G. Williams: *SatScan v. 1.0. Software for the Space and Space-Time Scan Statistics* (National Cancer Inst, Bethesda 1997)

43.57 J. I. Naus, S. Wallenstein: Multiple window, cluster size scan procedures, Method. Comp. Appl. Prob. **6**, 389–400 (2004)

43.58 N. H. Anderson, D. M. Titterington: Some methods for investigating spatial clustering with epidemiologic applications, J. R. Stat. Soc. **A160**, 87–105 (1997)

43.59 D. Q. Naiman, C. Priebe: Computing scan statistic p-values using importance sampling, with applications to genetics, medical image analysis, J. Comp. Graph. Stat. **10**, 296–328 (2001)

43.60 C. E. Priebe, D. Q. Naiman, L. M. Cope: Importance sampling for spatial scan analysis: computing scan statistic p-values for marked point processes, Comp. Stat. Data Anal. **35**, 475–485 (2001)

43.61 J. I. Naus: Clustering of random points in two dimensions, Biometrika **52**, 263–267 (1965)

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

43.63 I. P. Tu: Theory and Applications of Scan Statistics. Ph.D. Thesis (Stanford University, Palo-Alto 1997)

43.64 D. Aldous: *Probability Approximations via the Poisson Clumping Heuristic* (Springer, Berlin Heidelberg 1989)

43.65 S. E. Alm: On the distribution of scan statistics of a two-dimensional Poisson processes, Adv. Appl. Prob. **29**, 1–18 (1997)

43.66 S. E. Alm: Approximation, simulation of the distribution of scan statistics for Poisson processes in higher dimensions, Extremes **1**, 111–126 (1998)

43.67 G. Haiman, C. A. Preda: A new method for estimating the distribution of scan statistics for a two-dimensional Poisson Process, Method. Comp. Appl. Prob. **4**, 393–408 (2002)

43.68 M. D. Penrose: Focusing of the scan statistic and geometric clique number, Adv. Appl. Prob. **34**, 739–753 (2002)

43.69 D. Siegmund, B. Yakir: Tail probabilities for the null distribution of scanning stattistics, Bernoulli **6**, 191–213 (2000)

43.70 T. Akiba, T. Yamamoto, H. Yamamoto: Reliability of a 2-dimensional k-within-consecutive rxs out of mxn: F system.e=48, pages=625-637, year=2001,, Naval Res Log

43.71 R. W. R. Darling, M. S. Waterman: Extreme value distribution for the largest cube in a random lattice, SIAM J. Appl. Math. **46**, 118–132 (1986)

43.72 J. Chen, J. Glaz: Two dimensional discrete scan statistics, Stat. Prob. Lett. **31**, 59–68 (1996)

43.73 V. Karwe, J. Naus: New recursive methods for scan statistic probabilities, Comp. Stat. Data Anal. **23**, 389–402 (1997)

43.74 R. J. Huntington: Distribution of the minimum number of points in a scanning interval on the line, Stochastic Proc. Appl. **7**, 73–77 (1978)

43.75 J. Glaz, J. Naus: Multiple clusters on the line, Commun. Stat. Theory Meth. **12**, 1961–1986 (1983)

43.76 X. Su, S. Wallenstein: New approximations for the distribution of the *r*-scan statistic, Stat. Prob. Lett. **46**, 411–419 (2000)

43.77 X. Su, S. Wallenstein, D. Bishop: Non-overlapping clusters: Approximate distribution, application to molecular biology, Biometrics **57**, 420–426 (2001)

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

43.79 J. Glaz, N. Balakrishnan: On Poisson Approximation for Continuous Multiple Scan Statistics in Two Dimensions. In: *Scan Statistics and Applications*, ed. by M. Mansson (Birkhauser, Boston 1999)

43.80 A. Godbole, S. G. Papastavrides: *Runs and Patterns in Probability* (Kluwer, Dordrecht 1994)

43.81 M. V. Boutsikas, M. V. Koutras: Modeling claim exceedances over thresholds, Insur. Math. Economics **30**, 67–83 (2002)

43.82 J. I. Naus, D. Wartenberg: A double scan statistic for clusters of two types of events, J. Am. Stat. Assoc. **92**, 1105–1113 (1997)

43.83 J. I. Naus, V. Stefanov: Double scan statistics, Method. Comp. Appl. Prob. **4**, 163–180 (2002)

43.84 G. P. Patil, C. Taille: Geographic, network surveillance via scan statistics for critical area detection, Stat. Sci. **18**, 457–465 (2003)

43.85 M. Kulldorff, Z. Fang, S. J. Walsh: A tree based scan statistic for database disease surveillance, Biometrics **59**, 323–331 (2003)