Approximations for Distributions of Scan Statistics

JOSEPH I. NAUS*

Certain statistical applications deal with the extremal distributions of the number of points in a moving interval or window of fixed length. This article gives an approximation that is highly accurate for several of these distributions. Applications include the maximum cluster of points on a line or circle, multiple coverage by subintervals or subarcs of fixed size, the length of the longest success run in Bernoulli trials, and the generalized birthday problem.

KEY WORDS: Scan statistics; Window statistics; Maximum cluster; Longest run; Generalized birthday probability; Multiple coverage.

1. INTRODUCTION

A uranium prospector scans a region and focuses on clusters of high Geiger counts. A quality control expert develops zone tests based on too many of k successive observations falling within certain zones on the control chart. An operations analyst studies the multiple covering of a hardened target. The first two applications deal with the maximum and the third application with the minimum number of points in a moving window of fixed length.

Given a point process, let $n_{y,y+t}$ denote the number of points in the window $[y, y + t)$. Let

$$\tilde{n}_t(T) = \max_{0\le y\le T-t}\{n_{y,y+t}\};$$

$$\tilde{s}_t(T) = \min_{0\le y\le T-t}\{n_{y,y+t}\}.$$

Certain statistical applications require either the distribution of the maximum, $\tilde{n}_t(T)$, or the minimum, $\tilde{s}_t(T)$, as the window scans $(0, T)$. More generally, we will look at the probability of the event $E_T(t, n_1, n_2) = (\tilde{n}_t(T) < n_1 \cap \tilde{s}_t(T) > n_2)$, that the number of points in the sliding window never is as large as n_1 or as small as n_2. This probability is denoted by $Q_T(t, n_1, n_2)$ and when the arguments t, n_1, n_2 are fixed is abbreviated to Q_T.

Frequently the exact distribution Q_T either is not known or, if known, is not readily computable when t/T is small. We find that for certain important applications the following approximation is remarkably accurate:

$$Q_T \doteq Q_{2t}(Q_{3t}/Q_{2t})^{(T/t)-2}. \tag{1.1}$$

To see the reasoning underlying approximation (1.1), consider the case of a continuous time point process on $(0, Lt)$. Here T/t is an integer $L > 2$. Write the event $E_T(t, n_1, n_2)$ as $\cap_{i=1}^{L-1} E_i$, where E_i denotes the event:

$$(\max_{(i-1)t\le y\le it} n_{y,y+t} < n_1) \cap (\min_{(i-1)t\le y\le it} n_{y,y+t} > n_2).$$

$$Q_T = \Pr(\cap_{i=1}^{L-1} E_i) = \Pr(E_1)\prod_{k=2}^{L-1}\Pr(E_k \mid \cap_{j=1}^{k-1} E_j). \tag{1.2}$$

In certain cases it is reasonable to approximate $\Pr(E_k \mid \cap_{j=1}^{k-1} E_j)$ by $\Pr(E_k \mid E_{k-1})$ and there is also in some of these cases the symmetry $\Pr(E_k \mid E_{k-1}) = \Pr(E_2 \mid E_1) = \Pr(E_1 \cap E_2)/\Pr(E_1)$ for all k. We define Q_{2t} to be $\Pr(E_1)$ and Q_{3t} to be $\Pr(E_1 \cap E_2)$. Approximation (1.1) follows. When T/t is not an integer, approximation (1.1) still gives good interpolation accuracy.

Section 2 applies approximation (1.1) to a kth nearest-neighbor cluster problem on the line. Section 3 approximates the distribution of the longest success run in Bernoulli trials. Section 4 deals with the generalized birthday problem. Section 5 describes modifications of (1.1) to handle multiple clusters or coverage on the circle and the end effects for multiple coverage on the line.

Sections 2 through 5 deal with the "unconditional" problems in which the total number of points or covering arcs or intervals are Poisson and the total number of successes are binomially distributed random variables. For these cases, approximation (1.1) is remarkably accurate and yields, in addition to probabilities, approximations for the expected size of the longest success run and the expected waiting time till clusters. Exact expressions are given for Q_{2t} and Q_{3t} for various applications.

Section 6 deals with "conditional" problems in which the total number of points or successes is fixed. For these cases, approximation (1.1) is not as accurate as for the "unconditional" problems.

2. KTH NEAREST-NEIGHBOR WAITING TIME PROBABILITIES

Given a Poisson process on $(0, \infty)$ let $W_{n,t}$ denote the waiting time until the first occurrence within an interval of length t of n points. Given a fixed interval $(0, T)$, let $\tilde{p}_n$ denote the size of the smallest subinterval of $(0, T)$ that contains at least n points. $\tilde{n}_t$ denotes the largest number of points to be found in any subinterval of $(0, T)$ of

* Joseph I. Naus is Professor, Department of Statistics, Rutgers- The State University of New Jersey, New Brunswick, NJ 08903

© Journal of the American Statistical Association
March 1982, Volume 77, Number 377
Theory and Methods Section

length t. The distribution of $\tilde{n}_t$, $\tilde{p}_n$, and $W_{n,t}$ are related:

$$\Pr(W_{n,t} \le T) = \Pr(\tilde{p}_n \le t) = \Pr(\tilde{n}_t \ge n). \quad (2.1)$$

Denote this common probability by $P^*(n; \lambda T, t/T)$, where λ is the expected number of points per unit time in the Poisson process.

The distribution of $W_{n,t}$ is important in certain models of neuron discharge (Van de Grind et al. 1968), developability of film (Hamilton, Lawton, and Trabka 1972), and overflow in queueing (Newell 1963). $\tilde{p}_n$ denotes the size of the smallest nth nearest-neighbor distance. $\tilde{n}_t$ is the size of the largest cluster, and Conover, Bement, and Iman (1979) apply $\tilde{n}_t$ in a method for detecting uranium deposits. Neff and Naus (1980) describe various applications of $W_{n,t}$, $\tilde{p}_n$, and $\tilde{n}_t$.

Huntington and Naus (1975) give an exact formula for $P^*(n; \lambda T, t/T)$ that sums many products of determinants and for t/T small requires excessive computer time. Newell (1963) and Ikeda (1965) develop the asymptotic formula:

$$P^*(n; \lambda T, t/T) \doteq 1 - \exp(-\lambda^n t^{n-1} T/(n-1)!). \quad (2.2)$$

Conover, Bement, and Iman (1979) give the alternative approximation

$$P^*(n; \lambda L, 1/L) \doteq 1 - 1.5\, p_{n-1}^* \times \exp\{-\lambda(L-1)(1 - (p_{n-2}^*/p_{n-1}^*)\} + .5 \exp\{-\lambda L(1 - p_{n-2}^*)\}, \quad (2.3)$$

where

$$p_k^* = e^{-\lambda} \sum_{j=0}^{k} \lambda^j/j!.$$

Our approximation for this application is of the form (1.1):

$$Q^*(n; \lambda L, 1/L) \doteq Q^*(n; 2\lambda, \tfrac{1}{2}) [Q^*(n; 3\lambda, \tfrac{1}{3})/Q^*(n; 2\lambda, \tfrac{1}{2})]^{L-2}, \quad (2.4)$$

where $Q^*(n; \lambda L, 1/L) = 1 - P^*(n; \lambda L, 1/L)$. Approximation (2.4) is readily computable given the following exact results.

Theorem 1. For $n > 2$, $p_i = e^{-\lambda}\lambda^i/i!$, $F_n = \Sigma_{i=0}^{n} p_i$, $\lambda > 0$,

$$Q^*(n; 2\lambda, \tfrac{1}{2}) = F_{n-1}^2 - (n-1)p_n p_{n-2} - (n - 1 - \lambda)p_n F_{n-3}, \quad (2.5)$$

and

$$Q^*(n; 3\lambda, \tfrac{1}{3}) = F_{n-1}^3 - A_1 + A_2 + A_3 - A_4, \quad (2.6)$$

where

$$A_1 = 2p_n F_{n-1}((n-1)F_{n-2} - \lambda F_{n-3}), \quad (2.7)$$

$$A_2 = .5\, p_n^2 ((n-1)(n-2)F_{n-3} - 2(n-2)\lambda F_{n-4} + \lambda^2 F_{n-5}), \quad (2.8)$$

$$A_3 = \sum_{r=1}^{n-1} p_{2n-r} F_{r-1}^2, \quad (2.9)$$

$$A_4 = \sum_{r=2}^{n-1} p_{2n-r} p_r ((r-1)F_{r-2} - \lambda F_{r-3}), \quad (2.10)$$

where $F_i = 0$ for $i < 0$.

The proof of Theorem 1 is given in the Appendix. Approximation (2.4) can be applied even for noninteger $L = T/t$. Neff and Naus (1980) give tables of $P^*(n; 2\lambda, \frac{1}{2})$ and $P^*(n; 3\lambda, \frac{1}{3})$ for $n = 3(1)9$, $\lambda = .1(.1)10$. For other cases these probabilities can be computed using equations (2.5) through (2.10).

2.1 Comparing the Approximations

The Newell-Ikeda formula (2.2) is designed for small probabilities P^* and can be inaccurate in other cases. The Conover-Bement-Iman approximation (2.3) is designed for cases in which n is 7 or less. The new approximation (2.4) appears to be highly accurate over a wide range.

Example	Exact Value Neff & Naus	Approx. (2.4)	Conover-Bement-Iman	Newell-Ikeda
$P^*(4; 10, .1)$	.374	.374	.386	.811
$P^*(5; 12, \frac{1}{4})$	.765	.765	.816	1.000^-
$P^*(5; 18, \frac{1}{6})$	.896	.896	.945	1.000^-

2.2 The Expected Waiting Time Till a Cluster

Approximation (2.4) can be used to estimate the expected waiting time till $\tilde{n}_t \ge n$ for the Poisson process. Solov'ev (1966) and Glaz (1979) give approximations and bounds for this expectation. Without loss of generality define the units to make $t = 1$ and let λ denote the expected number of points per unit. Following the notation of Theorem 1 we have the approximation

$$E(W_{n,1}) \doteq .5 + F_{n-1} + Q_2^2/(Q_2 - Q_3), \quad (2.11)$$

where Q_2 and Q_3 denote respectively $Q^*(n; 2\lambda, \frac{1}{2})$ and $Q^*(n; 3\lambda, \frac{1}{3})$. An alternative approximation is

$$E(W_{n,1}) \doteq 2 + Q_2/\log_e(Q_2/Q_3). \quad (2.12)$$

For the reasoning behind approximation (2.11) suppose that $W_{n,1}$ could only assume integer values.

$$E(W_{n,1}) = \sum_{T=0}^{\infty} \Pr(W_{n,1} > T) \doteq 1 + F_{n-1} + \sum_{T=2}^{\infty} Q_2(Q_3/Q_2)^{T-2}. \quad (2.13)$$

Subtract .5 as a correction for continuity to find approximation (2.11). We could also interpret approximation

(2.4) to be good even for noninteger T and write

$$E(W_{n,1}) = \int_0^\infty \Pr(W_{n,1} > T)\, dT$$

$$\doteq 2 + \int_2^\infty Q_2(Q_3/Q_2)^{T-2} dT = \text{rhs of (2.12)}.$$

2.3 Example: $n = 3, \lambda = 1$

Here $Q_2 = .77817792$, $Q_3 = .66251375$ and approximation (2.11) gives $E(W_{3,1}) \doteq 6.7$ and (2.12) gives 6.8. Glaz (1979) gives the bounds $5.8 < E(W_{3,1}) < 8.6$. A simulation based on 10,000 replications gave $E(W_{3,1}) \doteq 6.63$ with standard error $\doteq .062$.

The bounds in Glaz (1979) can be quite wide. For example, for $\lambda = 1$, $56.6 < E(W_{5,1}) < 211.7$. Here $n = 5$, $Q_2 = .983483383$, $Q_3 = .971277358$, and approximation (2.12) gives 80.8. The simulation gave $E(W_{5,1}) \doteq 81.065$ with standard error $\doteq .8074$.

3. LENGTH OF THE LONGEST SUCCESS RUN IN BERNOULLI TRIALS

Given a sequence of n Bernoulli trials with probability of success on a given trial p, and $q = 1 - p$, let X denote the length of the longest success run. Various derivations and expressions are given for the distribution of X by Uspensky (1937, p. 79), Bradley (1968, p. 267), Bateman (1948, p. 112) and others. Bateman's formula is computationally the simplest and is

$$\Pr(X \geq m) = \sum_{j=1}^{[n/m]} (-1)^{j+1} \times \left(p + \frac{(n - jm + 1)q}{j}\right) \binom{n-jm}{j-1} p^{jm} q^{j-1}, \quad (3.1)$$

where $[x]$ denotes the largest integer in x. However, for n large relative to m, (3.1) is tedious to use. Feller (1957, p. 310) gives the approximation

$$\Pr(X \geq m) \doteq 1 - \exp(-nq\,p^m). \quad (3.2)$$

For this application, let $Q_L = 1 - \Pr(X \geq m \mid n = m)$. Approximation (1.1) becomes

$$\Pr(X \geq m) \doteq 1 - Q_2(Q_3/Q_2)^{(n/m)-2}, \quad (3.3)$$

where

$$Q_2 = 1 - \Pr(X \geq m \mid n = 2m) = 1 - p^m(1 + mq), \quad (3.4)$$

and

$$Q_3 = 1 - \Pr(X \geq m \mid n = 3m) = 1 - p^m(1 + 2mq) + \tfrac{1}{2}p^{2m}(2mq + m(m-1)q^2). \quad (3.5)$$

Equations (3.4) and (3.5) are special cases of (3.1).

Grant (1947) gives tables of $\Pr(X \geq m)$, the probability of at least one success run of length at least m in n Bernoulli trials each with probability of success p.

Example: $m/n;p$	Grant's tables Exact probability	Our approx. (3.3)	Approx. (3.2)
$\Pr(X \geq 7 \mid 50; \frac{1}{2})$	.1653	.1653	.1774
$\Pr(X \geq 10 \mid 50; \frac{1}{2})$	.0204	.0204	.0241
$\Pr(X \geq 5 \mid 50; \frac{1}{3})$	.1214	.1214	.1282
$\Pr(X \geq 4 \mid 40; \frac{1}{3})$	.2739	.2739	.2805
$\Pr(X \geq 2 \mid 16; \frac{1}{5})$	.4107	.4106	.4007

The Expected Size of the Longest Run

For n very large it is convenient to find the expectation of X by

$$E(X) = \sum_{m=1}^{\infty} \Pr(X \geq m), \quad (3.5)$$

using approximation (3.3) for $\Pr(X \geq m)$. In practice only a few probability terms need to be computed. For example, for $n = 20{,}000$, $p = \frac{1}{2}$, $\Pr(X \geq 10) \doteq .99994522$ and $\Pr(X \geq 22) \doteq .00238$. We can approximate $E(X) \doteq 9 + \sum_{m=10}^{22} \Pr(X \geq m) \doteq 13.62$ using either approximation (3.2) or (3.3).

4. THE GENERALIZED BIRTHDAY PROBLEM

Given a sequence of n Bernoulli trials with probability of success on a given trial p and $q = 1 - p$, let $n_{y,y+m}$ denote the number of successes in trials y, $y + 1$, $y + 2, \ldots, y + m - 1$. Let $k_m(n) = \max_{1 \leq y \leq n-m+1} n_{y,y+m}$. When $k_m(n) \geq k$ we say that a *quota* of at least k successes within some m consecutive trials has occurred. A quota is a generalization of a success run; the special case $k = m$ is considered in Section 3.

Saperstein (1972) and Naus (1974) consider the conditional problem in which the total number of successes in n trials is a constant **a** and all $\binom{n}{\mathbf{a}}$ orders of **a** successes and $n - \mathbf{a}$ failures are equally likely. The unconditional probability $\Pr(k_m(n) \geq k)$ can be gotten by averaging the conditional probability $\Pr(k_m(n) \geq k \mid \mathbf{a})$ over the binomial distribution of **a**. In general the computation of the conditional probability involves the sum of many determinants and is particularly complex for m small relative to n. In this section we develop an approximation for the unconditional probability $\Pr(k_m(n) \geq k)$ that is denoted by $1 - Q'_L$, where $L = n/m$.

$$Q'_L \doteq Q'_2\,(Q'_3/Q'_2)^{L-2}, \quad (4.1)$$

where

$$Q'_2 = \Pr(k_m(2m) < k) \quad \text{and} \quad Q'_3 = \Pr(k_m(3m) < k).$$

From Naus (1974) we can get explicit expressions for Q'_2 and Q'_3 in a way similar to that in which Q_2 and Q_3 were derived in Theorem 1 of Section 2. To find Q'_2 we average over the conditional probability in corollary 1 in Naus (1974). To find Q'_3 we apply Theorem I in Naus (1974) and expand terms and average over the joint distribution of (n_1, n_2, n_3) where n_i is the number of successes in trials $(i - 1)m + 1$ through im. (In Naus 1974, equation

(2.1) $\sum_{r=1}^{j-1}$ should be $\sum_{r=i}^{j-1}$.) The separate evaluations of the various sums follow the approach leading to equations (2.7) through (2.10).

Theorem 2. Let $b_k = b(k; m, p) = \binom{m}{k}p^k q^{m-k}$ and $F_b(r; s, p) = \sum_{i=0}^{r} b(i; s, p)$. For $n > 2$,

$$Q'_2 = F_b^2(k-1; m, p) - (k-1)b_k F_b(k-2; m, p) + mpb_k F_b(k-3; m-1, p). \tag{4.2}$$

$$Q'_3 = F_b^3(k-1; m, p) - A'_1 + A'_2 + A'_3 - A'_4, \tag{4.3}$$

where

$$A'_1 = 2b_k F_b(k-1; m, p)((k-1)F_b(k-2; m, p) - mpF_b(k-3; m-1, p)); \tag{4.4}$$

$$A'_2 = \tfrac{1}{2}b_k^2((k-1)(k-2)F_b(k-3; m, p) - 2(k-2)mpF_b(k-4; m-1, p) + m(m-1)p^2F_b(k-5; m-2, p)); \tag{4.5}$$

$$A'_3 = \sum_{r=1}^{k-1} b_{2k-r}F_b^2(r-1; m, p); \tag{4.6}$$

$$A'_4 = \sum_{r=2}^{k-1} b_{2k-r}b_r((r-1)F_b(r-2; m, p) - mpF_b(r-3; m-1, p)). \tag{4.7}$$

Huntington (1976) derives for the (unconditional) case of a sequence of Bernoulli trials the expected waiting time till the first k in m quota. Following the reasoning leading to approximation (2.11), we can use approximation (4.1) to develop an approximation for the expected waiting time till the first quota:

$$\begin{aligned} &E(\text{w.t. till } k \text{ in } m \text{ quota}) \\ &= \sum_{n=0}^{\infty} \Pr(\text{w.t.} > n) \\ &= \sum_{n=0}^{\infty} \Pr(k_m(n) < k) \\ &\doteq 2m + \sum_{n=2m}^{\infty} Q'_2(Q'_3/Q'_2)^{(n/m)-2} \\ &= 2m + Q'_2/(1 - (Q'_3/Q'_2)^{1/m}). \end{aligned} \tag{4.8}$$

Example: Expected waiting time till a $k = 5$ in $m = 6$ quota

p	Q'_2	Q'_3	Exact Expected w.t. Huntington (1974)	Approximate Exp. w.t. Approx. (4.8)
.1	.99970639	.99946882	25,250	25,250
.2	.9924992	.98668186	1,025.50	1,025.50

5. MULTIPLE COVERAGE AND LARGE CLUSTERS ON LINE AND CIRCLE

5.1 Multiple Coverage on the Line

Let N subintervals, each of length t, be dropped so that their midpoints are uniformly distributed on $(0, T)$. We say that a point x is covered by a subinterval with midpoint y if $y - t/2 \le x \le y + t/2$. We say that a point is multiply n-covered if it is covered by at least n subintervals. All points in $(t/2, T - t/2)$ will be n-covered iff the scanning window of length t always contains at least n points, that is, if $\tilde{s}_t(T) \ge n$. Huntington (1978) derives the distribution of $\tilde{s}_t(T)$ as the sum of determinants with the greatest complexity occurring when T/t is large. Approximation (1.1) can be applied here in a manner similar to that in Section 2.

Glaz and Naus (1979) find the probability that all points in $(0, T)$ are n-covered. Requiring $(0, t/2)$ and $(T - t/2, T)$ to be n-covered increases the complexity of the problem. Glaz and Naus deal with the conditional problem in which the total number of dropped subintervals, N, is fixed. We will deal here with the unconditional problem in which N is Poisson.

To see how approximation (1.1) can be modified in this case to handle the ends of the interval, consider the case in which T/t is an integer L and define E_i as in Section 1 but with $n_1 = \infty$, $n_2 = n$. Let E_1^* be $E_1 \cap$ (at least n points in $(0, t/2)$). Let E_{L-1}^* be $E_{L-1} \cap$ (at least n points in $(T - t/2, T)$).

$$\begin{aligned} Q_L &= \Pr(E_1^* \cap E_{L-1}^* \cap \textstyle\bigcap_{i=2}^{L-2} E_i) \\ &\doteq \Pr(E_1^*)\Pr(E_2 \mid E_1^*) \\ &\quad \times [(\Pr(E_3 \mid E_2)]^{L-4}\Pr(E_{L-1}^* \mid E_{L-2}) \\ &= (\Pr(E_1^* \cap E_2))^2 \\ &\quad \times (\Pr(E_2 \cap E_3)/\Pr(E_2))^{L-4}/\Pr(E_2). \end{aligned} \tag{5.1}$$

$\Pr(E_2)$ and $\Pr(E_2 \cap E_3)$ can be found as before; $\Pr(E_1^* \cap E_2)$ can be found following the approach in Glaz and Naus (1979).

5.2 Clusters and Coverage on the Circle

Glaz and Naus (1979) observe that the problem of multiply covering the circle is directly related to the problem of multiple clusters on the circle. This observation applies both for the conditional and for the unconditional (on total number of points or arcs) problems. For the unconditional problem in which the total number of points is Poisson with mean $L\lambda$, approximation (2.4) can be modified to handle the case of clusters on the circle. Let $Q_c^*(n; \lambda L, 1/L)$ denote the probability that there does not exist a subarc of length $1/L$ of the circle with unit circumference that contains at least n points. For $L \ge 5$,

$$Q_c^*(n; \lambda L, 1/L) \doteq Q^*(n; 4\lambda, \tfrac{1}{4})(Q^*(n; 3\lambda, \tfrac{1}{3}))^{L-2}/(Q^*(n; 2\lambda, \tfrac{1}{2}))^{L-1}, \tag{5.2}$$

where Theorem 1 of Section 2 gives explicit forms for $Q^*(n; 2\lambda, \frac{1}{2})$ and $Q^*(n; 3\lambda, \frac{1}{3})$, and Neff and Naus (1980) give tables for these and $Q^*(n; 4\lambda, \frac{1}{4})$. The reasoning behind approximation (5.2) is similar to that of approximation (1.1). Let E_i be defined as in Section 1 with $t = 1$, $T = L$. For the circle,

$$Q_c^*(n; L\lambda, 1/L) = \Pr(E_1) \prod_{k=2}^{L} \Pr(E_k \mid \cap_{i=1}^{k-1} E_i). \quad (5.3)$$

For $k = 2, \ldots, L-1$ approximate $\Pr(E_k \mid \cap_{i=1}^{k-1} E_i)$ by $\Pr(E_k \mid E_{k-1})$. Approximate

$$\Pr(E_L \mid \cap_{i=1}^{L-1} E_i) \quad \text{by} \quad \Pr(E_L \mid E_1 \cap E_{L-1})$$
$$= \Pr(E_L \cap E_1 \cap E_{L-1})/\Pr(E_1)\Pr(E_{L-1})$$
$$= Q^*(n; 4\lambda, \tfrac{1}{4})/(Q^*(n; 2\lambda, \tfrac{1}{2}))^2.$$

Example: $n = 5$, $L = 6$, $\lambda = 1$. From Neff and Naus (1980), $Q^*(5; 2, \frac{1}{2}) = .983483383$, $Q^*(5; 3, \frac{1}{3}) = .971277358$, $Q^*(5; 4, \frac{1}{4}) = .959220621$. Approximation (5.2) gives $Q_c^*(5; 6, \frac{1}{6}) \doteq .9278$. This is within .0001 of the exact result derived by averaging exact conditional probabilities from Wallenstein (1971) over the Poisson distribution.

6. CONDITIONAL PROBLEMS

Sections 2 through 5 deal with "unconditional" cases in which the total number of points, arcs, and successes are random variables with Poisson or binomial distributions (depending on the application). For these cases approximations of the form (1.1) give great accuracy. We will see in this section that for problems conditional on the total number of points, covering arcs, or successes approximations of the form (1.1) provide only rough estimates. The greater dependence of the E_i's in the conditional cases causes the approximation $\Pr(E_k \mid \cap_{j=1}^{k-1} E_j) \doteq \Pr(E_k \mid E_{k-1})$ to be less accurate than in the unconditional case.

6.1 The Conditional Cluster Problem

Let N denote the total number of points in $(0, T)$. Given a Poisson process, N has a Poisson distribution and, conditional on N fixed, the points are uniformly distributed on $(0, T)$. Let $Q(n; N, t/T)$ denote $\Pr(\bar{n}_t(T) < n \mid N$ points in $(0, T))$. $Q(n; N, t/T)$ is the conditional analog of $Q^*(n; \lambda T; t/T)$ approximated in Section 2. The corresponding approximation, again letting $t = 1$, $L = T$ for simplicity, is

$$Q(n; N, 1/L) \doteq Q_2 (Q_3/Q_2)^{L-2}, \quad (6.1)$$

where

$$Q_2 = \sum_{R=0}^{2n-2} Q(n; R, \tfrac{1}{2})\, b(R; N, 2/L) \quad (6.2)$$

and

$$Q_3 = \sum_{R=0}^{3n-3} Q(n; R, \tfrac{1}{3})\, b(R; N, 3/L), \quad (6.3)$$

where

$$b(R; N, p) = \binom{N}{R} p^R (1-p)^{N-R}.$$

Example: $n = 3$, $N = 19$, $L = 100$. From Neff and Naus (1980), $Q(3; 19, .01) = .783$. Approximation (6.1) gives

$$Q(3; 19, .01) \doteq .9967643\ (.9944237/.9967643)^{98} = .792.$$

Example: $n = 5$, $N = 18$, $L = 6$. From Neff and Naus, $Q(5; 18, \frac{1}{6}) = .007$. Approximation (6.1) gives .059.

6.2 Length of Longest Success Run Given s Successes in n Trials

Given s successes in n trials, such that all $\binom{n}{s}$ arrangements of s successes and $n - s$ failures are equally likely, let X_s denote the length of the longest success run. The distribution of X_s is (see Bradley 1968, p. 257)

$$\Pr(X_s \geq m) = \sum_{i=1}^{[s/m]} (-1)^{i+1} \binom{n-s+1}{i} \binom{n-im}{n-s}/\binom{n}{s}. \quad (6.4)$$

Burr and Cane (1961) develop and survey various approximations for $\Pr(X_s \geq m)$. The approach in the present article (see Sec. 3) suggests the approximation

$$\Pr(X_s \geq m) \doteq 1 - Q^{**}_2 (Q^{**}_3/Q^{**}_2)^{(n/m)-2}, \quad (6.5)$$

where to find Q^{**}_2 (or Q^{**}_3) we condition on the number of successes in $(1, 2m)$ (or in $(1, 3m)$), apply (6.4) for the special cases $n = 2m, 3m$, and average over the number of successes to find

$$1 - Q^{**}_2 = (m\binom{n-m-1}{s-m} + \binom{n-m}{s-m})/\binom{n}{s}. \quad (6.6)$$

$$1 - Q^{**}_3 = (2m\binom{n-m-1}{s-m} + \binom{n-m}{s-m} - \binom{m}{2}\binom{n-2m-2}{s-2m} - m\binom{n-2m-1}{s-2m})/\binom{n}{s}. \quad (6.7)$$

Example: $s = 20$, $n = 40$, $m = 5$. Approximation (6.5) gives $\Pr(X_{20} \geq 5) \doteq .420$. Mosteller (1941) gives the exact value .450.

Example: $s = 60$, $n = 205$, $m = 3$. Approximation (6.5) gives $\Pr(X_{60} \geq 3) = .9772$. Burr and Cane (1961) give the exact value of .99024. They mention that a previous approximation gives .9755, and they give a superior approximation for this case of .99043.

APPENDIX: PROOF OF THEOREM 1

Divide $(0, 1)$ into L equal parts and condition on $(n_1, n_2, \ldots, n_L)$ the numbers of points in the L parts. Naus (1966) proved that conditional on the n_i

$$\Pr(\text{no } n \text{ in } 1/L \text{ cluster} \mid n_1, \ldots, n_L) = \det \mid 1/c_{ij}! \mid n_1! \cdots n_L! \quad \text{for all } n_i \leq n-1$$
$$= 0 \quad \text{otherwise}, \quad (A.1)$$

where

$$c_{ij} = (j - i)n - \sum_{r=i}^{j-1} n_r + n_i \quad \text{for } i < j$$

$$= (j - i)n + \sum_{r=j}^{i} n_r \qquad \text{for } i \geq j.$$

Average the conditional probability (A.1) over the joint Poisson distribution of the n_i to find

$$Q^*(n; \lambda L, 1/L) = \sum_{n_1=0}^{n-1} \cdots \sum_{n_L=0}^{n-1} \det \mid e^{-\lambda}\lambda^{c_{ij}}/c_{ij}! \mid . \tag{A.2}$$

To find (2.5), let $L = 2$, and note that the conditional probability (A.1) is

$$\begin{aligned} &\Pr(\text{no } n \text{ in } \tfrac{1}{2} \text{ cluster} \mid n_1 = i, n_2 = j) \\ &\quad = 1 - \binom{i+j}{n}/\binom{i+j}{i} \quad \text{for} \quad i + j \geq n, i, j < n, \\ &\quad = 1 \quad \text{for } i + j < n \\ &\quad = 0 \quad \text{if either} \quad i \quad \text{or} \quad j \geq n. \end{aligned} \tag{A.3}$$

Define p_i and F_n as in Theorem 1. Average (A.3) over the joint distribution of n_1, n_2 to find

$$Q^*(n; 2\lambda, \tfrac{1}{2}) = {F_{n-1}}^2 - \sum\sum p_i p_j \binom{i+j}{n}/\binom{i+j}{i}, \tag{A.4}$$

where the sum is over $i \leq n - 1, j \leq n - 1, i + j \geq n$. Let $k = i + j$ and then $r = k - n$ to find

$$\begin{aligned} Q^*(n; 2\lambda, \tfrac{1}{2}) &= {F_{n-1}}^2 - \sum_{k=n}^{2n-2} \sum_{i=k-n+1}^{n-1} e^{-2\lambda}\lambda^k/n!(k - n)! \\ &= {F_{n-1}}^2 - p_n \sum_{r=0}^{n-2} (n - r - 1)\lambda^r e^{-\lambda}/r! \\ &= \text{rhs of equation (2.5).} \end{aligned} \tag{A.5}$$

Similarly, to find (2.6) we consider the case $L = 3$, where the conditional probability (A.1) becomes

$$\begin{aligned} &\Pr(\text{no } n \text{ cluster} \mid n_1, n_2, n_3) = 1 \\ &\quad - n_1!n_2!n_3!(a^{-1} + b^{-1} - c^{-1} - d^{-1} + e^{-1}), \end{aligned} \tag{A.6}$$

where

$$\begin{aligned} a &= n_1!\, n!\, (n_2 + n_3 - n)! \quad ; \\ b &= n_3!\, n!\, (n_1 + n_2 - n)! \\ c &= n!n!(n_1 + n_2 + n_3 - 2n)! \quad ; \\ d &= (2n - n_2)!(n_1 + n_2 - n)!(n_2 + n_3 - n)! \\ e &= (2n - n_2)!n_2!(n_1 + n_2 + n_3 - 2n)! \end{aligned}$$

To find the unconditional probability, sum (A.6) over the joint Poisson distribution of n_1, n_2, n_3 treating each term separately but noting the symmetry in a and b sums. In each triple sum, $n_i \leq n - 1$ for $i = 1, 2, 3$ but the range of summation has other conditions that differ for the different terms. For example, for term a we have the condition $n_2 + n_3 \geq n$. For term c we have the condition $(n_1 + n_2 + n_3) \geq 2n$. The details of the simplifications of the individual sums follow:

$$Q^*(n; 3\lambda, \tfrac{1}{3}) = {F_{n-1}}^3 - A_1 + A_2 + A_3 - A_4, \tag{A.7}$$

where

$$A_1 = 2 \sum_{n_1=0}^{n-1} \sum_{n_2} \sum_{n_3 \mid n_2 + n_3 \geq n, n_2 \leq n-1, n_3 \leq n-1} p_{n_1} p_n p_{n_2 + n_3 - n}. \tag{A.8}$$

Let $k = n_2 + n_3$.

$$\begin{aligned} A_1 &= 2\, p_n F_{n-1} \sum_{k=n}^{2n-2} \sum_{n_2=k-n+1}^{n-1} p_{k-n} \\ &= 2p_n F_{n-1} \sum_{k=n}^{2n-2} (n - 1 - (k - n))p_{k-n}. \end{aligned}$$

Let $r = k - n$ to get rhs of equation (2.7).

$$A_2 = {p_n}^2 \sum_{n_1} \sum_{n_2} \sum_{n_3 \mid n_1 + n_2 + n_3 \geq 2n, n_1 \leq n-1, n_2 \leq n-1, n_3 \leq n-1} \times p_{n_1 + n_2 + n_3 - 2n}. \tag{A.9}$$

Let $k = n_1 + n_2 + n_3$.

$$\begin{aligned} A_2 &= {p_n}^2 \sum_{k=2n}^{3n-3} (n - 1 - (k - 2n)) \\ &\qquad \times (n - 2 - (k - 2n))p_{k-2n}/2 \\ &= \text{rhs (2.8).} \end{aligned}$$

$$A_3 = \sum_{n_1} \sum_{n_2} \sum_{n_3 \mid n_1 + n_2 \geq n, n_2 + n_3 \geq n, n_i \leq n-1 \text{ for } i=1,2,3} \times p_{2n - n_2} p_{n_1 + n_2 - n} p_{n_2 + n_3 - n} \tag{A.10}$$

Let $r = n_2$, $i = n_1 + n_2$, $j = n_2 + n_3$.

$$\begin{aligned} A_3 &= \sum_{r=1}^{n-1} p_{2n-r} \sum_{i=n}^{r+n-1} p_{i-n} \sum_{j=n}^{r+n-1} p_{j-n} \\ &= \text{rhs (2.9).} \end{aligned}$$

$$A_4 = \sum_{n_1} \sum_{n_2} \sum_{n_3 \mid n_1 + n_2 + n_3 \geq 2n, n_i \leq n-1 \text{ for } i=1,2,3} \times p_{2n - n_2} p_{n_2} p_{n_1 + n_2 + n_3 - 2n} \tag{A.11}$$

Let $k = n_1 + n_2 + n_3$.

$$\begin{aligned} A_4 &= \sum_{n_2=2}^{n-1} \sum_{k=2n}^{2n+n_2-2} \sum_{n_1=k-n_2-(n-1)}^{n-1} p_{2n-n_2} p_{n_2} p_{k-2n} \\ &= \sum_{n_2=2}^{n-1} p_{2n-n_2} p_{n_2} \sum_{k=2n}^{2n+n_2-2} (n_2 - 1 - (k - 2n))p_{k-2n} \end{aligned}$$

Let $n_2 = r$ and $s = k - 2n$ and simplify to find the rhs of (2.10). This completes the proof of Theorem 1.

[*Received June 1980. Revised February 1981.*]

REFERENCES

BATEMAN, G.I. (1948), "On The Power Function of the Longest Run as a Test for Randomness in a Sequence of Alternatives," *Biometrika*, 35, 97–112.

BRADLEY, J.V. (1968), *Distribution-Free Statistical Tests*, Englewood Cliffs: Prentice-Hall Inc.

BURR, E.J., and CANE, G. (1961), "Longest Run of Consecutive Observations Having a Specified Attribute," *Biometrika*, 48, 461–465.

CONOVER, W.J., BEMENT, T.R., and IMAN, R.L. (1979), "On a Method For Detecting Clusters of Possible Uranium Deposits," *Technometrics*, 21, 277–282.

FELLER, W. (1957), *An Introduction to Probability Theory and Its Applications*, *I*, (2nd ed.), New York: John Wiley.

GLAZ, J. (1979), "Expected Waiting Time for the Visual Response," *Biological Cybernetics*, 35, 39–41.

GLAZ, J., and NAUS, J. (1979), "Multiple Coverage of the Line," *Annals of Probability*, 7, 900–906.

GRANT, D.A. (1947), "Additional Tables of the Probability of 'Runs' of Correct Responses in Learning and Problem-Solving," *Psychological Bulletin*, 44, 276–279.

GRIND, W.A. VAN DE, SCHALM, T. VAN, and BOUMAN, M.A. (1968), "A Coincidence Model of the Processing of Quantum Signals by the Human Retina," *Kybernetik*, 4, 141–146.

HAMILTON, J.F., LAWTON, W.H., and TRABKA, E.A. (1972), "Some Spatial and Temporal Point Processes in Photographic Science," in *Stochastic Point Processes: Statistical Analysis, Theory and Applications*, ed. P.A.W. Lewis, New York: John Wiley.

HUNTINGTON, R.J. (1974), "Distributions and Expectations for Clusters in Continuous and Discrete Cases, With Applications," Ph.D. Dissertation, Rutgers University, New Brunswick, New Jersey.

——— (1976), "Mean Recurrence Times for K Successes Within M Trials," *Journal of Applied Probability*, 13, 604–607.

——— (1978), "Distributions of the Minimum Number of Points in a Scanning Interval on the Line," *Stochastic Processes and Their Applications*, 7, 73–77.

HUNTINGTON, R.J., and NAUS, J.I. (1975), "A Simpler Expression for Kth Nearest-Neighbor Coincidence Probabilities," *Annals of Probability*, 3, 894–896.

IKEDA, S. (1965), "On Bouman-Velden-Yamamoto's Asymptotic Evaluation Formula for the Probability of Visual Response in a Certain Experimental Research in Quantum Biophysics of Vision," *Annals of Institute of Statistical Mathematics*, 17, 295–310.

MOSTELLER, F. (1941), "Note on an Application of Runs to Quality Control Charts," *Annals of Mathematical Statistics*, 12, 228–232.

NAUS, J.I. (1966), "Some Probabilities, Expectations and Variances for the Size of Largest Clusters and Smallest Intervals," *Journal of the American Statistical Association*, 61, 1191–1199.

——— (1974), "Probabilities for a Generalized Birthday Problem," *Journal of the American Statistical Association*, 69, 810–815.

NEFF, N.D., and NAUS, J.I. (1980), *The Distribution of the Size of the Maximum Cluster of Points on a Line*, Vol. VI in IMS Series of Selected Tables in Mathematical Statistics, Providence: American Mathematical Society.

NEWELL, G.F. (1963), "Distribution for the Smallest Distance Between Any Pair of Kth Nearest-Neighbor Random Points on a Line," in *Time Series Analysis, Proceedings of a Conference Held at Brown University*, ed. M. Rosenblatt, New York: Academic Press.

SAPERSTEIN, B. (1972), "The Generalized Birthday Problem," *Journal of the American Statistical Association*, 67, 425–428.

SOLOV'EV, A.D. (1966), "A Combinatorial Identity and Its Application to the Problem Concerning the First Occurrence of a Rare Event," *Theory of Probability and Its Applications*, 11, 276–282.

USPENSKY, J.V. (1937), *Introduction to Mathematical Probability*, New York: McGraw-Hill.

WALLENSTEIN, S.R. (1971), "Coincidence Probabilities Used in Nearest-Neighbor Problems on the Line and Circle," Ph.D. Dissertation, Rutgers University, New Brunswick, New Jersey.