J. Appl. Prob. **50,** 1089–1101 (2013)
Printed in England

DISCRETE, CONTINUOUS AND CONDITIONAL MULTIPLE WINDOW SCAN STATISTICS

TUNG-LUNG WU * *** AND
JOSEPH GLAZ,* **** *University of Connecticut*
JAMES C. FU,** *University of Manitoba*

Abstract

The distributions of discrete, continuous and conditional multiple window scan statistics are studied. The finite Markov chain imbedding technique has been applied to obtain the distributions of fixed window scan statistics defined from a sequence of Bernoulli trials. In this manuscript the technique is extended to compute the distributions of multiple window scan statistics and the exact powers for multiple pulse and Markov dependent alternatives. An application in blood component quality monitoring is provided. Numerical results are also given to illustrate our theoretical results.

Keywords: Scan statistic; multiple window; finite Markov chain imbedding; random permutation; Poisson process; power

2010 Mathematics Subject Classification: Primary 60E05
Secondary 60J10

1. Introduction

The literature on scan statistics is substantial and growing rapidly due to the widespread applications in many areas (see, e.g. [2], [13], and [19]). In particular, scan statistics have been widely applied in quality control (see, e.g. [18]) to increase the sensitivity of detecting an out-of-control signal. The development of the scan statistic method on blood component quality monitoring can be found in [11].

Let $N(t)$ be a Poisson process with intensity λ on (0,1]. For $0 < \omega \leq 1$, let $S(\omega, t) = N(t + \omega) - N(t)$ denote the number of events that have occurred in the interval $(t, t + \omega]$, where ω is the window size. An unconditional continuous scan statistic is defined as

$$S(\omega) = \sup_{0<t\leq 1-\omega} S(\omega, t).$$

The exact distribution of $S(\omega)$ has been derived (see [20]) over a limited range of parameters. Many other authors have derived approximations and tight bounds for distribution of $S(\omega)$ (see, e.g. [6], [10], and [12]).

Let $X_1, \ldots, X_n$ be a sequence of independent, identically distributed Bernoulli trials with $\mathbb{P}(X_1 = 1) = p$ and $\mathbb{P}(X_1 = 0) = q = 1 - p$. For $1 \leq r \leq n$, let $S_n(r, i) = \sum_{v=i}^{i+r-1} X_v$.

Received 23 November 2012; revision received 13 February 2013.
* Postal address: Department of Statistics, University of Connecticut, Storrs, CT 06269-4120, USA.
** Postal address: Department of Statistics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada.
Email address: fu@umanitoba.ca
*** Email address: tung-lung.wu@uconn.edu
**** Email address: glaz@uconnvm.uconn.edu

An unconditional discrete scan statistic is defined as

$$S_n(r) = \max_{1 \le i \le n-r+1} S_n(r, i).$$

The unconditional probability $\mathbb{P}(S_n(r) > a)$ can be obtained by the conditional probability weighted by the binomial distribution, i.e.

$$\mathbb{P}(S_n(r) > a) = \sum_{N=0}^{n} \binom{n}{N} p^N q^{n-N} \mathbb{P}\left(S_n(r) > a \,\middle|\, \sum_{i=1}^{n} X_i = N \right).$$

The exact unconditional and conditional distributions have been obtained by many authors; see, e.g. [16] for the conditional case and [4] for the unconditional case. Approximations and bounds have also been extensively studied; see, e.g. [7] and [9].

For testing the null hypothesis H_0 of uniformity against the alternative hypothesis H_a of presence of clusters, the general likelihood ratio test rejects the null hypothesis if the observed scan statistic is large. In the discrete case, under H_a, we usually specify a block of r observations $\{X_\tau, \ldots, X_{\tau+r-1}\}$, of which the probability $p_1 = \mathbb{P}(X_i = 1) > p$ for $i = \tau, \ldots, \tau + r - 1$, where r is a known constant. A variable window scan statistic proposed by Nagarwalla [15] does not require the length of the scanning window to be fixed, and this motivates us to study the following general theory: given an integer k and $1 \le r_1 < r_2 < \cdots < r_k \le n$, the distribution of a discrete multiple window scan statistic is given by

$$\mathbb{P}(S_n(r_j) < a_j, \text{ for all } j = 1, \ldots, k), \tag{1.1}$$

where $a_1 < \cdots < a_k$ are constants. For convenience, the probability given in (1.1) is sometimes denoted by $\mathbb{P}(S_n(r_1, \ldots, r_k) < (a_1, \ldots, a_k))$. A continuous version of the multiple window scan statistic can be defined analogously. Glaz and Zhang [8] derived simple approximations for the above probability for one and two dimensional cases.

Although many accurate approximations (see, e.g. [17]) have been proposed for the distribution of a multiple window scan statistic, there is no theoretical result for their exact α level and power. In this manuscript the finite Markov chain imbedding (FMCI) technique (see, e.g. [5]) is extended to study the distributions of multiple window scan statistics for both discrete and continuous, as well as, conditional and unconditional cases. In Section 2 we obtain the exact distributions of conditional and unconditional discrete multiple window scan statistics. The approximation for conditional and unconditional continuous cases is given in Section 3. In Section 4 the exact powers of the unconditional discrete case are obtained for (i) the multiple pulse alternative and (ii) the Markov dependent alternative. An application in blood component quality monitoring is given in Section 5. Numerical results are given in Section 6. Section 7 provides a summary and discussion.

2. Discrete multiple window scan statistics

In this section the exact distributions of unconditional and conditional discrete multiple window scan statistics are obtained through the direct extension of the approaches in [4] and [6].

Definition 2.1. A compound pattern generated by a set of simple patterns is said to be an effective compound pattern, denoted by Λ^E, if any simple pattern in Λ^E is not a segment of another pattern in Λ^E, i.e. the simple patterns in Λ^E are all distinct.

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

It follows from the above definition that a compound pattern Λ generated by a set of simple patterns can always be reduced to an effective compound pattern Λ^E by removing those simple patterns which contain another simple pattern as a segment, and, for $n \geq 0$,

$$\mathbb{P}(W(\Lambda) > n) = \mathbb{P}(W(\Lambda^E) > n), \tag{2.1}$$

where $W(\Lambda)$ is the waiting time until the first occurrence of the compound pattern Λ.

Given an integer k and $1 \leq r_1 < \cdots < r_k \leq n$, we are interested in the events $\{S_n(r_j) < a_j\},\ j = 1, \ldots, k$, occurring simultaneously. Following Fu [4], the event $\bigcap_{j=1}^{k}\{S_n(r_j) < a_j\}$ occurs if and only if none of the compound patterns Λ_{r_j,a_j}, associated with each fixed window scan statistic, occur in the sequence of n Bernoulli trials. We give a simple example for illustration. Given $k = 2$, $r_1 = 4$, $r_2 = 6$, $a_1 = 3$, and $a_2 = 4$, we scan two windows of sizes 4 and 6 over a sequence. The event $\{S_n(4) < 3\}$ corresponds to the compound pattern $\Lambda_{4,3} = \{111, 1011, 1101\}$, and the event $\{S_n(6) < 4\}$ corresponds to another compound pattern $\Lambda_{6,4}$ = {1111, 10111,11011,11101,100111,101011,101101,110011,110101,111001}. Then, $\{S_n(4) < 3\}$ and $\{S_n(6) < 4\}$ both occur if and only if the compound pattern generated by $\Lambda_{4,3}$ and $\Lambda_{6,4}$ does not occur in the sequence of n trials. Note that the above compound pattern is not an effective one and the corresponding effective compound pattern is given by {111,1011,1101,110011}.

Lemma 2.1. *Given an integer k, $1 \leq r_1 < \cdots < r_k \leq n$, and $1 \leq a_1 < \cdots < a_k \leq r_k + 1$, we have*

$$\mathbb{P}(S_n(r_j) < a_j,\ j = 1, \ldots, k) = \mathbb{P}(W(\Lambda_k^E) > n) = \boldsymbol{\xi}_0(\boldsymbol{N}^E(k; p))^n \mathbf{1}^\top, \tag{2.2}$$

where Λ_k^E is the effective compound pattern associated with the multiple window scan statistic, $\boldsymbol{N}^E(k; p)$ is the essential transition probability matrix of the imbedded Markov chain of $W(\Lambda_k^E)$, $\boldsymbol{\xi}_0$ is an appropriate initial distribution, and $\mathbf{1}^\top$ is the transpose of the vector (1,...,1).

Proof. Let Λ_k be the compound pattern associated with the multiple window scan statistic. From Fu [4] together with (2.1) follows

$$\mathbb{P}(S_n(r_j) < a_j,\ j = 1, \ldots, k) = \mathbb{P}(W(\Lambda_k) > n) = \mathbb{P}(W(\Lambda_k^E) > n) = \boldsymbol{\xi}_0(\boldsymbol{N}^E(k; p))^n \mathbf{1}^\top$$

and this completes the proof.

The state space and the essential transition probability matrix $\boldsymbol{N}^E(k; p)$ of the imbedded Markov chain can be obtained from [4]. Thus, the exact distribution of the discrete multiple window scan statistic can be calculated via Lemma 2.1. Note that it is easy to see that the compound pattern Λ_k is generated by a number of such simple patterns and how many is given by

$$\sum_{j=1}^{k} \sum_{\nu=0}^{r_j - a_j} \binom{a_j - 2 + \nu}{\nu}. \tag{2.3}$$

In the above example, for $k = 2$, the resulting effective compound pattern is Λ^E = {111,1011, 1101,110011} consisting of only 4 simple patterns instead of 13 according to (2.3). Thus, the total number of simple patterns in the effective compound pattern Λ_k^E is substantially less than $\sum_{j=1}^{k} \sum_{\nu=0}^{r_j - a_j} \binom{a_j - 2 + \nu}{\nu}$. Some details on this matter are given in Section 7. In what follows, when we mention a compound pattern, it always refers to the corresponding effective compound pattern and is denoted by Λ, for simplicity, without the superscript E.

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

Remark 2.1. Note that $\{a_j\}$ must be in strictly increasing order. If there is some i and j such that $a_i \geq a_j$ and $r_i < r_j$, then the occurrence of $\{S_n(r_j) < a_j\}$ implies the occurrence of $\{S_n(r_i) < a_i\}$. Hence, in this case, the statement $S_n(r_i) < a_i$ is redundant. On the other hand, $S_n(r_j) \geq a_j$ is redundant when $S_n(r_i) \geq a_i$ is also considered.

Fu *et al.* [6] extended Fu's [4] result and showed that the conditional fixed window scan statistic ($k = 1$) is finite Markov chain imbeddable. Given $\sum_{i=1}^{n} X_i = N$, window size r, and a, the distribution of the conditional fixed window scan statistic $\mathbb{P}(S_n(r) < a \mid \sum_{i=1}^{n} X_i = N)$ is treated as the waiting time distribution of the first occurrence of a corresponding compound pattern $\Lambda_{r,a}$ in a $[n - N, N]$-specified random permutation $\boldsymbol{\pi}$, i.e.

$$\mathbb{P}\left(S_n(r) < a \,\middle|\, \sum_{i=1}^{n} X_i = N\right) = \mathbb{P}(W(\Lambda_{r,a}) > n \mid \boldsymbol{\pi}) = \boldsymbol{\xi}_0\left(\prod_{t=1}^{n} \boldsymbol{N}_t(r, a)\right)\mathbf{1}^{\top},$$

where $\boldsymbol{N}_t(r, a)$ is the essential transition probability matrix of the imbedded Markov chain of $W(\Lambda_{r,a})$. Although they did not point it out, they, in fact, not only obtained the distribution of the conditional fixed window scan statistic, but also extended the solution of the waiting time problem of runs and patterns to random permutations. Hence, we can readily adopt their approach to derive the exact distribution of the conditional discrete multiple window scan statistic. Given $1 \leq r_1 < \cdots < r_k \leq n$ and $a_1 < \cdots < a_k$, following Fu *et al.* [6], we have

$$\mathbb{P}\left(S_n(r_j) < a_j,\, j = 1, \ldots, k \,\middle|\, \sum_{i=1}^{n} X_i = N\right) = \mathbb{P}(W(\Lambda_k) > n \mid \boldsymbol{\pi}) = \boldsymbol{\xi}_0\left(\prod_{t=1}^{n} \boldsymbol{N}_t(k)\right)\mathbf{1}^{\top}, \tag{2.4}$$

where Λ_k is the corresponding effective compound pattern, and $\boldsymbol{N}_t(k)$ can be constructed using (2.3) from [6].

In some applications, scan statistics are used for two purposes. In a retrospective study where the total number of successes $\sum_{i=1}^{n} X_i = N$ is known, a conditional multiple scan statistic can be deployed and the test can be implemented exactly based on (2.4). Once a cluster is detected, its location, where the number of successes is above the threshold, is also revealed. There are various algorithms to identify the locations of clusters. For example, Cucala [3] proposed a procedure to identify multiple clustering by removing the existing cluster and rescaling the spacings between observations so that the procedure can be repeated until no significant cluster can be found.

In a prospective study, the scan statistic detects the clusters that currently exist. To implement a prospective scan statistic, we may conduct a pilot study with size n_0 or use historical data to accurately estimate the probability of success p. Suppose that the pilot data is independent of the current data and they are from the same population. For a given $n > 0$, it follows from Slutsky's theorem that, as $n_0 \to \infty$,

$$\mathbb{P}\left(S_n(r_j) < a_j,\, j = 1, \ldots, k \,\middle|\, \hat{p} = \frac{\sum_{i=1}^{n_0} X_i}{n_0}\right) \to \mathbb{P}(S_n(r_j) < a_j,\ j = 1, \ldots, k \mid p).$$

The result is based on the fact that the probability of a scan statistic in (2.2) is a polynomial function of p which is continuous. Using such an estimate, for a given type-I error, a threshold is determined for detecting clusters according to (2.2). The distribution of a scan statistic is sensitive to the probability of success p which is closely related to the performance (power) of the test. An overestimation of p would lead to a loss of power in detecting clusters. In practice,

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

TABLE 1: 10 simulations for $\mathbb{P}(S_{100}(10, 15, 20) < (3, 5, 7) \mid \hat{p} = 0.05)$.

Simulations	1	2	3	4	5	6	7	8	9	10
Results	0.552	0.773	0.935	0.773	0.935	0.935	0.441	0.935	0.665	0.441

we may expect $p \leq 0.1$ which leads to an upper bound, for the standard deviation of the estimator, of $\sqrt{0.09/100} = 0.03$ if $n_0 = 100$. The approximate 95% confidence interval is $(0.36, 0.64)$ if $n_0 = 1000$ and $\bar{x} = 0.05$. We ran 10 simulations for the probability of a multiple window scan statistic of window size $(10, 15, 20)$ when $\hat{p} = 0.05$ and $n_0 = 100$. The numerical results are given in Table 1 and show that the median is an accurate estimate for $\mathbb{P}(S_{100}(10, 15, 20) < (3, 5, 7) \mid \hat{p} = 0.05) = 0.773$. Further theoretical investigations on the efficiency of estimates may be presented in a subsequent paper.

3. Continuous multiple window scan statistics

For a Poisson process $N(t)$, given $N(1) = N$, the N points are distributed according to a uniform distribution on (0,1]. Given an integer k and window sizes $0 < \omega_1 < \cdots < \omega_k \leq 1$, we first consider the distribution of conditional continuous multiple window scan statistic

$$\mathbb{P}(S(\omega_j) < a_j,\ j = 1, \ldots, k \mid N).$$

Given a large integer n, the interval (0,1] is divided into n subintervals $(0 = t_0, t_1], \ldots, (t_{n-1}, t_n = 1]$, each of equal length $t_i - t_{i-1} = \Delta t = 1/n$, $i = 1, \ldots, n$. Then, all the induced $[n - N, N]$-specified permutations are of equal probability. The following lemmas from [6] will be used to prove our main result, Theorem 3.1.

Lemma 3.1. *For all window sizes $0 < \omega_j \leq 1,\ j = 1, \ldots, k$, the following holds*

$$\max_{1 \leq i \leq n - [n\omega_j] + 1} S_n([n\omega_j], i) \leq \sup_{0 < t \leq 1 - \omega_j} S(\omega_j, t) \leq \max_{1 \leq i \leq n - [n\omega_j] - 1} S_n([n\omega_j] + 2, i),$$

where $[n\omega_j]$ is the integer part of $n\omega_j$.

Proof. For each $1 \leq j \leq k$, the continuous scan statistics $S(\omega_j)$ can be expressed as

$$\sup_{0 < t \leq 1 - \omega_j} S(\omega_j, t) = \max_{1 \leq i \leq n - [n\omega_j]} \sup_{t_{i-1} < t \leq t_i} S(\omega_j, t).$$

For simplicity of notation, with understanding, the last scanning window stops at time $t = 1$. It follows from the definition, for $i = 1, \ldots, n - [n\omega_j]$,

$$\max(S_n([n\omega_j], i), S_n([n\omega_j], i + 1)) \leq \sup_{t_{i-1} < t \leq t_i} S(\omega_j, t).$$

Taking the maximum yields

$$\max_{1 \leq i \leq n - [n\omega_j] + 1} S_n([n\omega_j], i) \leq \sup_{0 < t \leq 1 - \omega_j} S(\omega_j, t).$$

Similarly, we also have

$$\sup_{0 < t \leq 1 - \omega_j} S(\omega_j, t) \leq \max_{1 \leq i \leq n - [n\omega_j] - 1} S_n([n\omega_j] + 2, i).$$

This completes the proof.

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

Lemma 3.2. *For a given integer k,* $0 < \omega_1 < \cdots < \omega_k \le 1$, *and* $a_1 < \cdots < a_k < N$,

$$|\mathbb{P}(S_n([n\omega_j]+2) > a_j \mid N) - \mathbb{P}(S_n([n\omega_j]) > a_j \mid N)| \le \frac{2N^2}{n}, \qquad j = 1, \ldots, k.$$

Proof. See proof of Theorem 3.1 in [6].

In the sequel, we generalize the result of Fu *et al.* [6] to the conditional continuous multiple window scan statistic.

Theorem 3.1. *For a given integer k,* $0 < \omega_1 < \cdots < \omega_k \le 1$, *and* $a_1 < \cdots < a_k < N$,

$$\begin{aligned}&\mathbb{P}(S(\omega_j) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &\quad = \lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k \mid N).\end{aligned}$$

Proof of Theorem 3.1. It follows, from the definition, that $\{S_n([n\omega_j]) > a_j\}$ implies that $\{S(\omega_j) > a_j\}$, and $\{S(\omega_j) > a_j\}$ implies that $\{S_n([n\omega_j]+2) > a_j\}$ for $j = 1, \ldots, k$. Given $N(1) = N$, from Lemma 3.1 follows

$$\begin{aligned}&\mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &\quad \le \mathbb{P}(S(\omega_j) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &\quad \le \mathbb{P}(S_n([n\omega_j]+2) > a_j, \text{ for some } j = 1, \ldots, k \mid N).\end{aligned}$$

Thus, from Lemma 3.2 and the fact that $\{S_n([n\omega_j]) > a_j\} \subseteq \{S_n([n\omega_j]+2) > a_j\}$, we have

$$\begin{aligned}&|\mathbb{P}(S_n([n\omega_j]+2) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &- \mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k \mid N)|\\ &\quad \le \sum_{j=1}^{k} [\mathbb{P}(S_n([n\omega_j]+2) > a_j \mid N) - \mathbb{P}(S_n([n\omega_j]) > a_j \mid N)]\\ &\quad \le \frac{2kN^2}{n}.\end{aligned}$$

This completes the proof.

Remark 3.1. Note that, as remarked in [6], the theorem still holds for fixed integers $\ell \ge 0$ and $h \ge 2$, i.e.

$$\begin{aligned}&\mathbb{P}(S(\omega_j) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &\quad = \lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]+h) > a_j, \text{ for some } j = 1, \ldots, k \mid N)\\ &\quad = \lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]-\ell) > a_j, \text{ for some } j = 1, \ldots, k \mid N).\end{aligned}$$

For the unconditional case, from the viewpoint of the Poisson process with intensity λ, let $\{X_i\}_{i=1}^n$ be a sequence of Bernoulli trials with probability $p_n = \lambda/n$. For $1 \le r_1 < \cdots < r_k \le n$, it follows that

$$\begin{aligned}&\mathbb{P}(S_n(r_j) < a_j, \ j = 1, \ldots, k)\\ &\quad = \sum_{N=0}^{n} \binom{n}{N} p_n^N (1-p_n)^{n-N} \mathbb{P}\left(S_n(r_j) < a_j, \ j = 1, \ldots, k \,\middle|\, \sum_{i=1}^{n} X_i = N\right).\end{aligned}$$

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

Taking $r_j = [n\omega_j]$, and since $\sum_{i=1}^{n} X_i$ converges in the limit of large n to a Poisson random variable with parameter λ, the above equation yields the following result: for sufficiently large n

$$\lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]) < a_j,\ j = 1, \ldots, k) = \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda} \mathbb{P}(S(\omega_j) < a_j,\ j = 1, \ldots, k \mid N).$$

This is equivalent to saying

$$\lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]) < a_j,\ j = 1, \ldots, k) = \mathbb{P}(S(\omega_j) < a_j,\ j = 1, \ldots, k),$$

or

$$\begin{aligned} &\lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]) > a_j - 1, \text{ for some } j = 1, \ldots, k) \\ &\quad = \mathbb{P}(S(\omega_j) > a_j - 1, \text{ for some } j = 1, \ldots, k). \end{aligned}$$

Theorem 3.2. *For a given integer k, $0 < \omega_1 < \cdots < \omega_k \le 1$, and $a_1 < \cdots < a_k$,*

(i) *we have*

$$\begin{aligned} &|\mathbb{P}(S(\omega_j) > a_j, \textit{ for some } j = 1, \ldots, k) \\ &\quad - \mathbb{P}(S_n([n\omega_j]) > a_j, \textit{ for some } j = 1, \ldots, k)| \\ &\qquad \le \frac{2k(\lambda^2 + \lambda)}{n}, \end{aligned}$$

(ii) *and*

$$\begin{aligned} &\mathbb{P}(S(\omega_j) > a_j, \textit{ for some } j = 1, \ldots, k) \\ &\quad = \lim_{n\to\infty} \mathbb{P}(S_n([n\omega_j]) > a_j, \textit{ for some } j = 1, \ldots, k). \end{aligned}$$

Proof. For given n, it follows that

$$\begin{aligned} &|\mathbb{P}(S(\omega_j) > a_j, \text{ for some } j = 1, \ldots, k) - \mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k)| \\ &\quad \le \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda} |\mathbb{P}(S(\omega_j) > a_j, \text{ for some } j = 1, \ldots, k \mid N) \\ &\qquad\qquad - \mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k \mid N)| \\ &\quad \le \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda} |\mathbb{P}(S_n([n\omega_j] + 2) > a_j, \text{ for some } j = 1, \ldots, k \mid N) \\ &\qquad\qquad - \mathbb{P}(S_n([n\omega_j]) > a_j, \text{ for some } j = 1, \ldots, k \mid N)| \\ &\quad \le \sum_{N=0}^{\infty} \frac{\lambda^N}{N!} \mathrm{e}^{-\lambda} \frac{2kN^2}{n} \\ &\quad = \frac{2k(\lambda^2 + \lambda)}{n}. \end{aligned}$$

This completes the proof of (i). Result (ii) follows directly from result (i).

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

4. Power

In this section we compute the exact powers of unconditional discrete multiple window scan statistics. The powers of unconditional continuous multiple window scan statistics can then be approximated by the powers of discrete ones. The approximations for null distributions of scan statistics have been derived by many authors (see, e.g. [1]), while there are not many theoretical results for the power in the literature. A pulse alternative was proposed in [21], where, under H_a, the probability of success p_1 in a given window of length d is greater than p_0, the probability of success under H_0. We consider the following two generalizations: (i) the multiple pulse alternative and (ii) the Markov dependent alternative. The details are given as follows. Given $r_1, \ldots, r_k$, $a_1, \ldots, a_r$, under H_0, the p-value of a multiple window scan statistic is defined as

$$\mathbb{P}(S_n(r_j) \geq a_j, \text{ for some } j = 1, \ldots, k \mid H_0).$$

4.1. Multiple pulse alternative

Let p_i denote the probability of success of the ith trial, $i = 1, \ldots, n$. The pulse alternative is given below. Under H_0, the probability of success is the same for all Bernoulli trials and the null hypothesis is given by

$$H_0 : p_i = p_0, \quad i = 1, \ldots, n.$$

The alternative hypothesis assumes the presence of a cluster in a certain interval and is given by

$$\begin{aligned} H_a\colon\ & p_i = p_0, \quad i = 1, \ldots, \tau - 1, \tau + d, \ldots, n, \quad \text{and} \\ & p_i = p_1 > p_0, \quad i = \tau, \ldots, \tau + d - 1, \end{aligned}$$

where τ is the starting location of changes and is unknown, and d is the pulse size. For computing the power of a fixed window scan statistic of window size r, under H_a, we usually specify a block

$$B(t, r) = \{X_t, X_{t+1}, \ldots, X_{t+r-1}\},$$

where $\mathbb{P}(X_i = 1) = p_1 > p_0$, $i = t, \ldots, t + r - 1$.

We generalize the pulse alternative to a multiple pulse alternative by allowing multiple blocks (clusters) of different sizes. Specifically, for a given m, a multiple pulse alternative is given by

$$\begin{aligned} H_a\colon\ & p_i = p_1 > p_0, \quad i = \tau_\nu, \ldots, \tau_\nu + d_\nu - 1, \ \nu = 1, \ldots, m, \quad \text{and} \\ & p_i = p_0, \quad \text{otherwise.} \end{aligned}$$

Note that each block $B_\nu(\tau, d) = \{X_{\tau_\nu}, X_{\tau_\nu+1}, \ldots, X_{\tau_\nu+d_\nu-1}\}$ of size d_ν does not overlap, or it will reduce to fewer blocks of larger sizes.

The FMCI technique can be applied to calculate the exact power for a multiple pulse alternative with a minor modification of the formula in Lemma 2.1. Given $\tau_\nu = t_\nu$ and d_ν, $\nu = 1, \ldots, m$, the power is given by

$$\begin{aligned} &\mathbb{P}(S_n(r_j) \geq a_j, \text{ for some } j = 1, \ldots, k \mid H_a) \\ &\quad = 1 - \mathbb{P}(W(\Lambda_k) > n \mid H_a) \\ &\quad = 1 - \boldsymbol{\xi}_0 \prod_{\nu=1}^{m} [\boldsymbol{N}(k; p_0)^{\tau_\nu - \tau_{\nu-1} - d_{\nu-1}} \boldsymbol{N}(k; p_1)^{d_\nu}] \boldsymbol{N}(k; p_0)^{n - \tau_m - d_m + 1} \mathbf{1}^\top, \end{aligned}$$

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

where $\tau_0 = 0, d_0 = 1$ and $\boldsymbol{N}(k;\, p_1)$ is the essential probability transition matrix of the imbedded Markov chain associated with the probability of success p_1.

Remark 4.1. We can even compute the power for a multiple pulse alternative, where the probabilities p_ν are different for blocks $B_\nu(\tau, d),\ \nu = 1, \ldots, m$. The power is then given by

$$1 - \boldsymbol{\xi}_0 \prod_{\nu=1}^{m} [\boldsymbol{N}(k;\, p_0)^{\tau_\nu - \tau_{\nu-1} - d_{\nu-1}} \boldsymbol{N}(k;\, p_\nu)^{d_\nu}] \boldsymbol{N}(k;\, p_0)^{n - \tau_m - d_m + 1} \mathbf{1}^\top.$$

4.2. Markov dependent alternative

The Markov dependent alternative tends to form a clump if the Bernoulli trials are positively correlated. We can model this by Markov dependent trials with high probability from state 1 to state 1. Hence, a Markov dependent alternative is given by H_a: $\{X_n\}$ is a sequence of Markov dependent trials with transition matrix

$$\Pi = \begin{matrix} 0 \\ 1 \end{matrix} \begin{bmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{bmatrix},$$

where $p_{11} > p_0$, and p_0 is the probability of success under H_0. Let

$$\boldsymbol{p^0} = (\mathbb{P}(X_1 = 0), \mathbb{P}(X_1 = 1)).$$

Then it follows from the FMCI technique that the power can be computed by

$$\begin{aligned} \mathbb{P}(S_n(r_j) \geq a_j,\ \text{for some } j = 1, \ldots, k \mid H_a) &= 1 - \mathbb{P}(W(\Lambda_k) > n \mid H_a) \\ &= 1 - \boldsymbol{\xi}_0 \boldsymbol{N}(k;\, \Pi)^n \mathbf{1}^\top, \end{aligned}$$

where the essential matrix $\boldsymbol{N}(k;\, \Pi)$ can be similarly constructed by replacing p_0 and $1 - p_0$ by p_{11}, p_{01} and p_{10}, p_{00}, respectively.

5. An application

The scan statistic has been recognized and applied in various areas, such as genetics [11] and quality control [14]. We consider the problem for the processing of blood and blood components for transfusion. As noted in [14], the blood component quality monitoring presents several challenges in practice due to the link between blood component and individual facilities. The first challenge is the low volume of blood product in some small facilities. The second challenge is the low expected frequency of nonconforming blood components. The third challenge is to be able to detect the occurrence of a nonconforming process as soon as possible. Therefore, a test which can signal the occurrence of a nonconforming process at an early stage needs to be developed. Lachenbruch *et al.* [14] compared three models: (i) binomial model, (ii) negative binomial model, and (iii) scan statistic method. In the binomial model, to assure no more than 5% nonconforming lots, a process is said to be conforming if no failures are observed in 59 consecutive observations, and no failures in 299 consecutive observations assure a conforming process with no more than 1% nonconforming lots. The binomial and negative binomial models are not able to detect clusters of failures occurring near the boundaries of two sets of observations, while scan statistics can overcome this shortcoming. We generalized the scan statistic method by adopting the multiple window scan statistic to enhance the performance of detecting nonconforming processes.

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

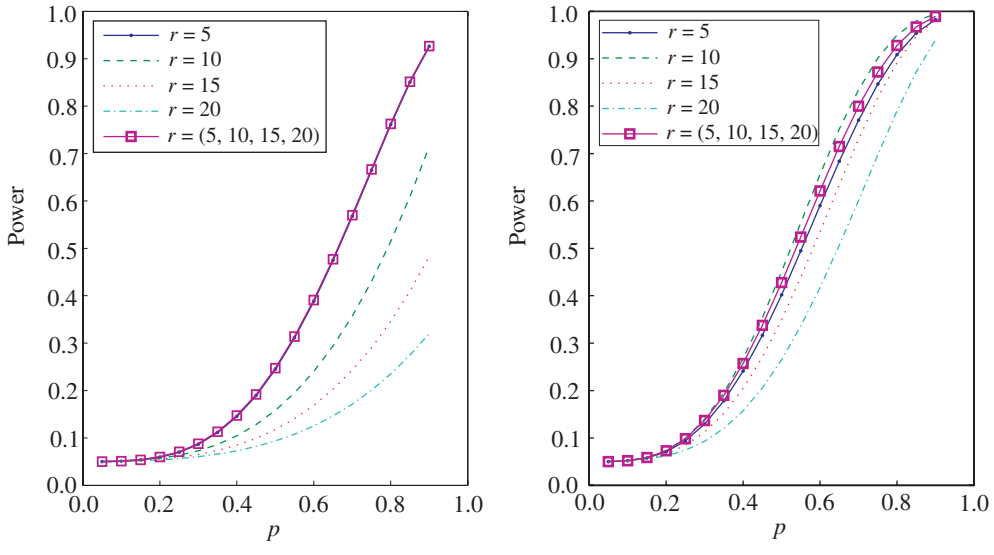

FIGURE 1: OC curves for pulse sizes equal to 5 (*left*) and 8 (*right*).

According to Lachenbruch *et al.* [14], for blood product, the minimal suggested standard is to test at least 60 units per month for quality assurance and for the blood centers to replace their equipment about once every three years. This suggests that we should use $n = 2000$ and let $p_0 = 0.05$. The powers are computed, for the pulse sizes 5 and 8, based on a randomized test with type-I error equal to 0.05. Figure 1 is the plot of operation characteristic curves (OC) of four fixed window scan statistics of window sizes 5, 10, 15, and 20, and a multiple window scan statistic of window size (5, 10, 15, 20), denoted by $S_n(5, 10, 15, 20)$. It is known that a fixed window scan statistic of size r is most powerful for the pulse alternative of size equal to r. For the pulse size equal to 5, the fixed window scan statistic $S_{2000}(5)$ and the multiple window scan statistic $S_{2000}(5, 10, 15, 20)$ perform equally well and better than other fixed window scan statistics. In fact, from the numerical results (not provided), $S_{2000}(5, 10, 15, 20)$ performs slightly better than $S_{2000}(5)$, and we suspect that this is due to the use of a randomized test. While in the case of pulse size equal to 8, the powers of $S_{2000}(5, 10, 15, 20)$ are slightly lower than those of $S_{2000}(10)$ but higher than other fixed window scan statistics. Thus, we may conclude that if the pulse size is unknown, a multiple window scan statistic is preferable to a fixed window scan statistic.

6. Numerical results

Under H_0, let the probability of success $p_0 = 0.1$ and $n = 90$. Throughout this section we compare the powers of three fixed window scan statistics of windows sizes 10, 15, and 20, and a multiple window scan statistic $S_n(10, 15, 20)$. To compare powers, a randomized test is used to set the type-I error to be 0.05. We consider the following two cases for multiple pulse alternatives.

Case 1: Let $m = 1$.

$$H_a: p_i = p_1 > p_0, \quad i = \tau_1, \ldots, \tau_1 + d_1 - 1, \quad \text{and}$$
$$p_i = p_0, \quad \text{otherwise.}$$

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

TABLE 2: Powers for multiple pulse alternative Case 1.

Scan statistics	$p_1 = 0.3$	Simulation	$p_1 = 0.5$	Simulation
$S_{90}(10)$	0.3687	0.3692	0.8553	0.8547
$S_{90}(15)$	0.3981	0.3992	0.8898	0.8895
$S_{90}(20)$	0.3705	0.3682	0.8572	0.8569
$S_{90}(10, 15, 20)$	0.3764	0.3754	0.8657	0.8655

TABLE 3: Powers for multiple pulse alternative Case 2.

Scan statistics	$p_1 = 0.3$	Simulation	$p_1 = 0.5$	Simulation
$S_{90}(10)$	0.4754	0.4750	0.9381	0.9384
$S_{90}(15)$	0.5006	0.5017	0.9502	0.9498
$S_{90}(20)$	0.4692	0.4693	0.9297	0.9293
$S_{90}(10, 15, 20)$	0.4848	0.4854	0.9440	0.9438

TABLE 4: Powers for Markov dependent alternatives with transition matrices Π_1 and Π_2, $n = 90$, and $\boldsymbol{p}^0 = [0.9,\ 0.1]$.

Scan statistics	Π_1	Simulation	Π_2	Simulation
$S_{90}(10)$	0.3063	0.3066	0.8244	0.8256
$S_{90}(15)$	0.2841	0.2832	0.7953	0.7942
$S_{90}(20)$	0.2667	0.2657	0.7718	0.7719
$S_{90}(10, 15, 20)$	0.3130	0.3123	0.8338	0.8356

Case 2: Let $m = 2$.

$$H_a\colon p_i = p_1 > p_0, \quad i = \tau_\nu, \ldots, \tau_\nu + d_\nu - 1,\ \nu = 1, 2, \quad \text{and}$$
$$p_i = p_0, \quad \text{otherwise.}$$

For Case 1, we choose $\tau_1 = 10$ and $d_1 = 15$, and the powers are given in Table 2. It is clear that the multiple window scan statistic outperforms the fixed window scan statistics of window sizes not equal to 15, as $S_{90}(15)$ is most powerful. Similar results can be seen in Table 3 for Case 2, where there are 2 clusters with $\tau_1 = 10$, $\tau_2 = 50$, $d_1 = 8$, and $d_2 = 16$. It can be expected that the fixed window scan statistic of size 15 would perform well and it turns out to be the case. While $S_{90}(15)$ is still the most powerful among those tests considered, the multiple window scan statistic $S_{90}(10, 15, 20)$ performs better than other fixed window scan statistics. The powers based on 10000 simulation runs are given for comparison. Also note that scan statistics are insensitive to the locations of clusters.

Table 4 gives the powers for two Markov dependent alternatives where the transition probability matrices are given by

$$\Pi_1 = \begin{matrix} 0 \\ 1 \end{matrix}\begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix}, \qquad \Pi_2 = \begin{matrix} 0 \\ 1 \end{matrix}\begin{bmatrix} 0.9 & 0.1 \\ 0.4 & 0.6 \end{bmatrix}.$$

It clearly shows that the multiple window scan statistic is more powerful than other fixed window

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

TABLE 5: The unconditional discrete case with $\mathbb{P}(X_1 = 1) = 0.1$.

r_1	r_2	r_3	a_1	a_2	a_3	n	FMCI	Simulation
5	10		2	3		100	0.0691	0.0692
5	10	15	4	6	10	100	0.9682	0.9683
5	10	20	3	4	8	100	0.6045	0.6049

TABLE 6: The unconditional continuous case.

ω_1	ω_2	ω_3	a_1	a_2	a_3	λ	n	FMCI	Simulation
0.01	0.02	0.03	2	3	4	3	500	0.9177	0.9176
0.01	0.02	0.03	2	3	4	5	500	0.7924	0.7915
0.02	0.04		3	4		3	500	0.9954	0.9949
0.02	0.04		3	4		5	500	0.9794	0.9773
0.03	0.06		3	6		3	300	0.9905	0.9893
0.03	0.06		3	6		5	300	0.9596	0.9558

scan statistics. Tables 5 and 6 provide probabilities for unconditional discrete and continuous multiple window scan statistics, respectively, for various combinations of parameters.

7. Summary and discussion

We have extended the FMCI technique to discrete and continuous multiple window scan statistics for both conditional and unconditional cases. The exact and approximate distributions of discrete and continuous multiple window scan statistics are obtained, respectively. The exact power of unconditional discrete multiple scan statistics is also derived. One of the new results established in this manuscript is the rate of convergence for unconditional continuous scan statistics associated with a Poisson process.

The distribution of a multiple window scan statistic is connected to the waiting time distribution of the corresponding compound pattern. The corresponding compound pattern is generated by compound patterns associated with each fixed window scan statistic of window sizes included in the multiple window scan statistic. The computational load is not proportional to the number of window sizes k. When k increases, the number of simple patterns in the effective compound pattern is significantly less than $\sum_{j=1}^{k}\sum_{v=0}^{r_j-a_j}\binom{a_j-2+v}{v}$ as many of them are redundant. However, there is no simple rule to calculate the total number of simple patterns comprising the effective compound pattern. The method based on compound patterns for exact results may become computationally infeasible when the window size is too large. An accurate approximation based on a smaller essential transition probability matrix remains an open question.

A multiple window scan statistic should not be blindly used for an arbitrarily large k and a wide range of consecutive integers for $\{r_j\}$. It turns out that an arbitrarily large value of k leads to a decrease in power. For example, in the discrete case if we select $k = 15$ and $(r_1, \ldots, r_{15}) = (5, \ldots, 20)$, the power for multiple pulse alternative Case 1 is decreased to 0.3687, while the power of the multiple window scan statistic of window size $(10, 15, 20)$ is 0.3764.

https://doi.org/10.1239/jap/1389370101 Published online by Cambridge University Press

It is known that a fixed window scan statistic is most powerful under the pulse alternative if the cluster size is equal to the window size. An application in the quality control of blood components in Section 5 shows that a multiple window scan statistic is a better choice if the cluster size d is unknown. In contrast to the binomial and negative binomial models, scan statistics offer constant monitoring during the entire period of a monitoring process. If there is a plausible cluster size d, then we suggest using slightly larger window sizes; as seen from the OC curves in Figure 1, $S_{2000}(10)$ performs better than $S_{2000}(5)$ when $d = 8$. In addition, we proposed two alternatives: multiple pulse and Markov dependent alternatives. The numerical results show that the multiple window scan statistic performs very well under both cases, especially for the Markov dependent alternative.

Acknowledgement

The authors thank the referee for the helpful suggestions and comments which led to the improvement of the presentation of the results in this article.

References

[1] Chen, J. and Glaz, J. (2009). Approximations for two-dimensional variable window scan statistics. In *Scan Statistics*, eds J. Glaz, V. Pozdnyakov and S. Wallenstein, Birkhäuser, Boston, MA, pp. 109–128.
[2] Cressie, N. A. C. (1991). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley, New York.
[3] Cucala, L. (2008). A hypothesis-free multiple scan statistic with variable window. *Biometrical J.* **50,** 299–310.
[4] Fu, J. C. (2001). Distribution of the scan statistic for a sequence of bistate trials. *J. Appl. Prob.* **38,** 908–916.
[5] Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns and Its Applications*. World Scientific Publishing Co. Inc., River Edge, NJ.
[6] Fu, J. C., Wu, T.-L. and Lou, W. Y. W. (2012). Continuous, discrete, and conditional scan statistics. *J. Appl. Prob.* **49,** 199–209.
[7] Glaz, J. (1989). Approximations and bounds for the distribution of the scan statistic. *J. Amer. Statist. Assoc.* **84,** 560–566.
[8] Glaz, J. and Zhang, Z. (2004). Multiple window discrete scan statistics. *J. Appl. Statist.* **31,** 967–980.
[9] Glaz, J., Pozdnyakov, V. and Wallenstein, S. (eds) (2009). *Scan Statistics*. Birkhäuser, Boston, MA.
[10] Haiman, G. (2000). Estimating the distributions of scan statistics with high precision. *Extremes* **3,** 349–361.
[11] Hoh, J. and Ott, J. (2000). Scan statistics to scan markers for susceptibility genes. *Proc. Nat. Acad. Sci. USA* **97**, 9615–9617.
[12] Janson, S. (1984). Bounds on the distributions of extremal values of a scanning process. *Stoch. Process. Appl.* **18,** 313–328.
[13] Koutras, M., Papadopoulos, G. and Papastavridis, S. (1993). Reliability of 2-dimensional consecutive-k-out-of-n: F systems. *IEEE Trans. Reliab.* **42,** 658–661.
[14] Lachenbruch, P. A., Foulkes, M. A., Williams, A. E. and Epstein, J. S. (2005). Potential use of the scan statistic for quality control in blood product manufacturing. *J. Biopharmaceutical Statist.* **15,** 353–366.
[15] Nagarwalla, N. (1996). A scan statistic with a variable window. *Statist. Med.* **15,** 845–850.
[16] Naus, J. (1974). Probabilities for a generalized birthday problem. *J. Amer. Statist. Assoc.* **69,** 810–815.
[17] Naus, J. I. and Wallenstein, S. (2004). Multiple window and cluster size scan procedures. *Methodology Comput. Appl. Prob.* **6,** 389–400.
[18] Rakitzis, A. C. and Antzoulakos, D. L. (2011). Chi-square control charts with runs rules. *Methodology Comput. Appl. Prob.* **13,** 657–669.
[19] Rosenfeld, A. (1978). Clusters in digital pictures. *Inform. and Control* **39,** 19–34.
[20] Wallenstein, S. R. and Naus, J. I. (1974). Probabilities for the size of largest clusters and smallest intervals. *J. Amer. Statist. Assoc.* **69,** 690–697.
[21] Wallenstein, S., Naus, J. and Glaz, J. (1994). Power of the scan statistic in detecting a changed segment in a Bernoulli sequence. *Biometrika* **81,** 595–601.