Ann. Inst. Statist. Math.
Vol. 57, No. 1, 21–37 (2005)

A MARTINGALE APPROACH TO SCAN STATISTICS

VLADIMIR POZDNYAKOV[1], JOSEPH GLAZ[1], MARTIN KULLDORFF[2] AND J. MICHAEL STEELE[3]

[1] *Department of Statistics, University of Connecticut, 215 Glenbrook Road, U-4120, Storrs, CT 06269-4120, U.S.A.*

[2] *Department of Ambulatory Care and Prevention, Harvard Medical School and Harvard Pilgrim Health Care, 133 Brookline Avenue, Boston, MA 02215-3920, U.S.A.*

[3] *Wharton School, Department of Statistics, University of Pennsylvania, Huntsman Hall 447, Philadelphia, PA 19104, U.S.A.*

(Received October 23, 2003; revised March 12, 2004)

Abstract. Scan statistics are commonly used in biology, medicine, engineering and other fields where interest is in the probability of observing clusters of events in a window at an unknown location. Due to the dependent nature of the number of events in a large number of overlapping window locations, even approximate solutions for the simplest scan statistics may require elaborate calculations. We propose a new martingale method which allows one to approximate the distribution for a wide variety of scan statistics, including some for which analytical results are computationally infeasible.

Key words and phrases: Scan, run, pattern, martingale, stopping time.

1. Introduction

Scan statistics are used in a wide range of fields including brain imaging (Yoshida *et al.* (2003)), psychology (Margai and Henry (2003)), veterinary medicine (Enemark *et al.* (2002)), forestry (Coulston and Riitters (2003)), crime hot-spot analysis (Kaminski *et al.* (2000)), industrial quality control (Shmueli (2003*a*, 2003*b*)), and especially molecular biology (Durand and Sankoff (2003), Goldstein and Waterman (1992), Karlin and Brendel (1992), Naus and Sheng (1997), and Sheng and Naus (1994)). Four recent books summarize the current status of the field: Glaz and Balakrishnan (1999), Glaz *et al.* (2001), Balakrishnan and Koutras (2002) and Fu and Lou (2003).

Different applications use scan statistics of different kinds. In the simple form considered by Naus (1965), there is a temporal Poisson point process which is considered over a fixed time length T and there is a fixed size window of much shorter length. We then move (or scan) the window continuously from the start to end, counting at each location the number of events within the window. The *scan statistic* is then defined as the maximum number of events as the window moves over all possible locations. In most applications, the main question of interest is whether the cluster of events defined by the maximum is a likely chance occurrence or not, so the most common null-hypothesis is that the point process is a homogeneous Poisson process. That is, we are interested in the probability of observing at least the observed number of events as the maximum, given that the null-hypothesis is true. More generally, we are interested in the distribution of the test statistic.

The most commonly used variants of the scan statistic are (i) temporal and other

one-dimensional scan statistics versus spatial, spatio-temporal and higher dimensional scan statistics, (ii) continuous scan statistics where events can occur anywhere versus discrete scan statistics with a sequence of trials at which the event either occurs or does not occur, (iii) a homogeneous versus known inhomogeneous background intensity defining the null-hypotheses, (iv) a conditional or unconditional scan statistic where the conditioning is on the total number of events observed, (v) a fixed versus variable size scanning window, (vi) single scan statistics with only one type of events versus double scan statistics with two or more types of events, and (vii) univariate versus multivariate scan statistics, with the latter simultaneously scanning multiple data streams.

While simple to formulate, the probabilistic nature of the scan statistics is very complex due to the dependencies of the overlapping window locations considered. Exact derivations of the distribution function is only available for the simplest scenarios such as temporal scan statistics with fixed window size and a homogeneous null-hypothesis. Good approximations as well as lower and upper bounds are known for additional scan statistics, but for most practically important applications the scan statistic must be evaluated using simulations (Glaz *et al.* (2001)).

Martingales have been used successfully for many practical statistical and probability problems, and their introduction has major impacts on fields such as survival analysis (Aalen (1978), Andersen *et al.* (1993)). In this article we present a martingale approach to scan statistics with which it is possible to obtain good approximations for the distribution of several scan statistics for which analytical results are not readily available. Using martingales, Li (1980) derived the first moment and we derive the second moment of the waiting time until we observe a specified number of events within one or several windows of specified lengths. Using these two moments we obtain approximations for the distribution of this waiting time.

The martingale approach to derive the generating function and moments is an alternative approach to the Markov chain embedding method where the waiting time until reaching a pattern is represented as a hitting time at a state of a relevant Markov chain. Elaborations on the Markov chain embedding methods and its applications to the theory of runs and patterns are given, among others, in Fu (1986, 1996 and 2001), Chao and Fu (1991), Fu and Koutras (1994), Uchida (1998), Aki and Hirano (1999), Antzoulakos (2001), Robin and Daudin (2001), Balakrishnan and Koutras (2002), Fu and Chang (2002), Fu and Lou (2003) and Han and Hirano (2003). Related methods on the occurrence of patterns include the probabilistic methods based on recurrent event theory (Feller (1968), Breen *et al.* (1985), and Chrysaphinou and Papastavridis (1990)), the method of Markov renewal embedding (Blom and Thornburn (1982) and Biggins and Cannings (1987)) and Markov chain embedding which uses analysis of exponential Markov chains (Stefanov and Pakes (1997) and Stefanov (2000)). Recently, Stefanov (2003) introduced a new approach to evaluate the generating function of the waiting time for a pattern generated by both discrete and continuous processes.

The article is organized as follows. In Section 2, we present the martingale approach for deriving the first two moments and generating function of the distribution of the shortest waiting time until the occurrence of one of several predefined patterns in a sequence of iid discrete observations. In Section 3 we use the first two moments to approximate the waiting time distribution. In Section 4, we use these results to evaluate approximations for the distribution of fixed window scan statistics. The accuracy of these approximations is evaluated with the help of available lower and upper bounds. In Sections 5 through 7 new approximations are derived for the variable window scan

statistics, the double scan statistics and the multivariate scan statistics. Finally, some concluding remarks and open issues are reviewed in Section 8.

2. Moments and generating functions

Here we will derive the first and second moments and the generating function of the waiting time until we observe one element from a set of several predefined patterns. We will then show how these moments yields a computationally feasible approximation for the distribution of the waiting time.

2.1 *Expected time*

Let Z be an arbitrary discrete random variable which takes values in the set Σ, and let $\{Z, Z_k\}_{k\geq 1}$ be a sequence of independent, identically distributed random variables.

Consider a collection of finite sequences $\{A_j\}_{1\leq j\leq K}$ over Σ, and without loss of generality assume that no sequence contains another as a subsequence. Next, we denote by τ_{A_j} the waiting time until A_j occurs as a run in the series $Z_1, Z_2, \ldots$. We are interested in both expected time of

$$\tau = \min\{\tau_{A_1}, \ldots, \tau_{A_K}\} \tag{2.1}$$

and probabilities $\pi_j = \boldsymbol{P}(\tau = \tau_{A_j})$.

The martingale approach to this problem was introduced in an elegant paper of Li (1980), and it has been further developed by Gerber and Li (1981), Williams (1991), Blom *et al.* (1994), and Pozdnyakov and Kulldorff (2003). For clarity of presentation, we will briefly review some of these results.

Following Li (1980), we introduce a measure of the amount of overlap between two sequences. Let $A = (a_1, \ldots, a_m)$ and $B = (b_1, \ldots, b_k)$ be two sequences over the alphabet Σ, and for each pair (i, j) we write

$$\delta_{ij} = \begin{cases} 1/\boldsymbol{P}(Z = b_j) & \text{if } \ 1 \leq i \leq m, \ 1 \leq j \leq k, \ \text{and } \ a_i = b_j \\ 0 & \text{otherwise.} \end{cases}$$

Next, we define $A * B$ by setting

$$A * B = \delta_{11}\delta_{22}\cdots\delta_{mm} + \delta_{21}\delta_{32}\cdots\delta_{mm-1} + \cdots + \delta_{m1}, \tag{2.2}$$

and we set $\Pi = (\pi_1, \ldots, \pi_K)^{\perp}$, $Y = (y_1, \ldots, y_K)^{\perp}$. Finally, we consider the matrix

$$M = \begin{bmatrix} A_1 * A_1 & A_1 * A_2 & \cdots & A_1 * A_K \\ A_2 * A_1 & A_2 * A_2 & \cdots & A_2 * A_K \\ \vdots & \vdots & & \vdots \\ A_K * A_1 & A_K * A_2 & \cdots & A_K * A_K \end{bmatrix}, \tag{2.3}$$

which Gerber and Li (1981) proved to be nonsingular. One has two notable results of Li (1980):

THEOREM 2.1. (*Li*, 1980). *The expected value of* τ *is given by*

$$\boldsymbol{E}(\tau) = \frac{1}{y_1^* + \cdots + y_K^*},$$

where $Y^* = (y_1^*, \ldots, y_K^*)^\perp$ *is the unique solution to the linear system* $MY = \mathbf{1}$, *and* $\mathbf{1} = (1, \ldots, 1)^\perp$.

THEOREM 2.2. (*Li*, 1980). *The vector of probabilities* $\Pi = (\pi_1, \ldots, \pi_K)^\perp$ *satisfy equation* $M^\perp \Pi = \boldsymbol{E}(\tau)\mathbf{1}$.

2.2 *Generating function*

Martingale arguments for finding the generating function of the waiting time in the case of one pattern were originally developed by Gerber and Li (1981). In their method, the transition from one pattern to many is based on some results on hitting times in a Markov chain, but our approach is based on matching expressions of the stopped martingale for different terminal patterns. This alternate method is intuitive and simple; moreover, it can be employed to get higher order moments.

To see how this works, we first consider a simple example first, and then we will show how it can be generalized.

Example 2.1. We flip a fair coin and we wait for one of two sequences: $A_1 = HH$ and $A_2 = HTH$. We are interested in the generating function of $\tau = \min\{\tau_{A_1}, \tau_{A_2}\}$.

Assume that we have two teams of gamblers. Before n-th round a new gambler from the first team joins the game and starts betting $y_1\alpha^n$ dollars on the sequence A_1; here $0 < \alpha < 1$ and y_1 is a number that we will choose later. If $Z_n \neq H$, then he leaves the game with nothing. If $Z_n = H$, he doubles his money, and bets the whole fortune on the event that $Z_{n+1} = H$. If he win, he leaves the game with $4y_1\alpha^n$ dollars. If he loses, then again he leaves with nothing. The second team bets in the similar fashion on the sequence A_2 but the initial bet of the gambler who joins game at n-th round is $y_2\alpha^n$. Let X_n be the net casino gain at moment n. Since the amount of each bet at n-th round is always determined by the history up to the moment $n-1$, and in each case the odds are fair, therefore, the net casino gain is a martingale. It is easy to see that

$$X_\tau = \begin{cases} (y_1+y_2)\alpha\frac{\alpha^\tau-1}{\alpha-1} - [y_1 \times (4\alpha^{\tau-1} + 2\alpha^\tau) + y_2 \times 2\alpha^\tau], & \text{if } \tau = \tau_{A_1}, \\ (y_1+y_2)\alpha\frac{\alpha^\tau-1}{\alpha-1} - [y_1 \times 2\alpha^\tau + y_2 \times (8\alpha^{\tau-2} + 2\alpha^\tau)], & \text{if } \tau = \tau_{A_2}, \end{cases}$$

which simplifies to

$$X_\tau = \begin{cases} (y_1+y_2)\alpha\frac{\alpha^\tau-1}{\alpha-1} - [(4/\alpha+2)y_1 + 2y_2]\alpha^\tau, & \text{if } \tau = \tau_{A_1}, \\ (y_1+y_2)\alpha\frac{\alpha^\tau-1}{\alpha-1} - [2y_1 + (8/\alpha^2+2)y_2]\alpha^\tau, & \text{if } \tau = \tau_{A_2}. \end{cases}$$

Now let us assume that we can choose the initial bets (y_1^*, y_2^*) in such way that

$$\begin{aligned} (4/\alpha+2)y_1^* + 2y_2^* &= 1 \\ 2y_1^* + (8/\alpha^2+2)y_2^* &= 1. \end{aligned}$$

Then regardless which sequence occurs first the stopped martingale is given by

$$X_\tau = (y_1^* + y_2^*)\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \alpha^\tau.$$

Since the expected value of τ is finite, and the increments of the martingale X_n is almost sure bounded, we find by the Optional-Stopping Theorem that

$$0 = \boldsymbol{E}X_\tau = (y_1^* + y_2^*)\frac{\alpha}{\alpha-1}\boldsymbol{E}\alpha^\tau - (y_1^* + y_2^*)\frac{\alpha}{\alpha-1} - \boldsymbol{E}\alpha^\tau,$$

and we may solve for $\boldsymbol{E}\alpha^{\tau}$ to obtain

$$\boldsymbol{E}\alpha^{\tau} = 1 - \frac{1}{\frac{\alpha}{1-\alpha}(y_1^* + y_2^*) + 1}.$$

This method also works in the general situation of K stopping sequences, provided that one makes the natural alterations. First we introduce a slightly modified measure of the amount of overlap between two sequences. If $A = (a_1, \ldots, a_m)$ and $B = (b_1, \ldots, b_k)$ are two sequences over Σ then we define

$$A * B(\alpha) = \delta_{11}\delta_{22}\cdots\delta_{mm}/\alpha^{m-1} + \delta_{21}\delta_{32}\cdots\delta_{mm-1}/\alpha^{m-2} + \cdots + \delta_{m1}/1. \tag{2.4}$$

Assume that we have K teams that bet on the K sequences in the correspondence with the rules of fair odds as they are described in the above example and the n-th player from j-th team start his betting on the sequence A_j with an initial bet of $y_j\alpha^n$ dollars. The net casino gain at time τ is given by

$$X_\tau = \begin{cases} (y_1 + \cdots + y_K)\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \sum_{i=1}^K A_1 * A_i(\alpha) y_i \alpha^\tau, & \text{if } \tau = \tau_{A_1}, \\ (y_1 + \cdots + y_K)\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \sum_{i=1}^K A_2 * A_i(\alpha) y_i \alpha^\tau, & \text{if } \tau = \tau_{A_2}, \\ \vdots & \vdots \\ (y_1 + \cdots + y_K)\alpha\frac{\alpha^\tau - 1}{\alpha - 1} - \sum_{i=1}^K A_K * A_i(\alpha) y_i \alpha^\tau, & \text{if } \tau = \tau_{A_K}. \end{cases}$$

Let

$$M(\alpha) = \begin{bmatrix} A_1 * A_1(\alpha) & A_1 * A_2(\alpha) & \cdots & A_1 * A_K(\alpha) \\ A_2 * A_1(\alpha) & A_2 * A_2(\alpha) & \cdots & A_2 * A_K(\alpha) \\ \vdots & \vdots & & \vdots \\ A_K * A_1(\alpha) & A_K * A_2(\alpha) & \cdots & A_K * A_K(\alpha) \end{bmatrix}. \tag{2.5}$$

Note that $M(1) = M$ and as it was shown in Gerber and Li (1981) the matrices $M(\alpha)$ are non-singular for all $0 < \alpha \leq 1$.

The method of Example 2.1 then yields a general result.

THEOREM 2.3. *The generating function of τ is given by*

$$\boldsymbol{E}\alpha^{\tau} = 1 - \frac{1}{\frac{\alpha}{1-\alpha}(y_1^* + \cdots + y_K^*) + 1},$$

where $Y^* = (y_1^*, \ldots, y_K^*)^{\perp}$ is the unique solution to the linear system $M(\alpha)Y = \mathbf{1}$.

2.3 *Second moment*

It is perhaps surprising that a more elaborate scheme is needed to apply this general idea of matching the stopped martingale to compute of the second moment of τ. The crucial idea is to introduce *two* teams for each sequence (i.e. in total we have $2K$ teams), and to illustrate the idea, we again consider a sequence of Bernoulli trials.

Example 2.2. We flip a fair coin and we wait until we observe one of two sequences: $A_1 = HH$ and $A_2 = HTH$. Our goal is to find the second moment of waiting time

$\tau = \min\{\tau_{A_1}, \tau_{A_2}\}$. The gambling is organized now in the following way. When a gambler from the first team of those two that bet on A_j joins the game at the n-th round he starts his betting with $y_j n$ dollars, a gambler from the second team bets z_j dollars.

The net casino gain at the moment τ is given by

$$X_\tau = \begin{cases} (y_1+y_2)\frac{\tau(\tau+1)}{2} + (z_1+z_2)\tau \\ \qquad - [y_1(4(\tau-1)+2\tau) + y_2\tau + 6z_1 + 2z_2], & \text{if} \quad \tau = \tau_{A_1}, \\ (y_1+y_2)\frac{\tau(\tau+1)}{2} + (z_1+z_2)\tau \\ \qquad - [y_1 2\tau + y_2(8(\tau-2)+2\tau) + 2z_1 + 10z_2], & \text{if} \quad \tau = \tau_{A_2}. \end{cases}$$

Rearranging terms we get

$$X_\tau = \begin{cases} (y_1+y_2)\frac{\tau(\tau+1)}{2} + (z_1+z_2)\tau \\ \qquad - [(6y_1+2y_2)\tau + 4(-1)y_1 + 6z_1 + 2z_2], & \text{if} \quad \tau = \tau_{A_1}, \\ (y_1+y_2)\frac{\tau(\tau+1)}{2} + (z_1+z_2)\tau \\ \qquad - [(2y_1+10y_2 2)\tau + 8(-2)y_2 + 2z_1 + 10z_2], & \text{if} \quad \tau = \tau_{A_2}. \end{cases}$$

Now, let us assume that we can choose the initial bets (y_1^*, y_2^*) and (z_1^*, z_2^*) in such way that we have the relation

$$\begin{aligned} 6y_1^* + 2y_2^* &= 1 \\ 2y_1^* + 10y_2^* &= 1 \end{aligned}$$

and the relation

$$\begin{aligned} 4(-1)y_1^* + 6z_1^* + 2z_2^* &= 1 \\ 8(-2)y_2^* + 2z_1^* + 10z_2^* &= 1. \end{aligned}$$

For such a choice of initial bets the stopped martingale is given by

$$X_\tau = (y_1^* + y_2^*)\frac{\tau(\tau+1)}{2} + (z_1^* + z_2^*)\tau - \tau - 1.$$

After taking the expected value of both sides of the last equation and solving it with respect to $\boldsymbol{E}\tau^2$ we get a formula for the second moment. Naturally one needs to employ the Optional-Stopping Theorem here, and, a bit later, we will show that this is indeed justified.

Now, to write the value of the net casino gain at the moment τ we first need to introduce the following notation. If $A = (a_1, \ldots, a_m)$ and $B = (b_1, \ldots, b_k)$ are two sequences over Σ, then we define

$$A \star B = -\delta_{11}\delta_{22}\cdots\delta_{mm}(m-1) - \delta_{21}\delta_{32}\cdots\delta_{mm-1}(m-2) - \cdots - \delta_{m1}0. \tag{2.6}$$

The stopped martingale X_τ is given by

$$X_\tau = \begin{cases} \sum_{i=1}^K y_i \frac{\tau(\tau+1)}{2} + \sum_{i=1}^K z_i \tau \\ \quad - \sum_{i=1}^K A_1 * A_i y_i \tau - \sum_{i=1}^K A_1 \star A_i y_i - \sum_{i=1}^K A_1 * A_i z_i, & \text{if} \quad \tau = \tau_{A_1}, \\ \sum_{i=1}^K y_i \frac{\tau(\tau+1)}{2} + \sum_{i=1}^K z_i \tau \\ \quad - \sum_{i=1}^K A_2 * A_i y_i \tau - \sum_{i=1}^K A_2 \star A_i y_i - \sum_{i=1}^K A_2 * A_i z_i, & \text{if} \quad \tau = \tau_{A_2}, \\ \qquad\qquad \vdots & \qquad\qquad \vdots \\ \sum_{i=1}^K y_i \frac{\tau(\tau+1)}{2} + \sum_{i=1}^K z_i \tau \\ \quad - \sum_{i=1}^K A_K * A_i y_i \tau - \sum_{i=1}^K A_K \star A_i y_i - \sum_{i=1}^K A_K * A_i z_i, & \text{if} \quad \tau = \tau_{A_K}. \end{cases}$$

Let us define

$$N = \begin{bmatrix} A_1 \star A_1 & A_1 \star A_2 & \cdots & A_1 \star A_K \\ A_2 \star A_1 & A_2 \star A_2 & \cdots & A_2 \star A_K \\ \vdots & \vdots & & \vdots \\ A_K \star A_1 & A_K \star A_2 & \cdots & A_K \star A_K \end{bmatrix}. \tag{2.7}$$

Suppose that we can find such $Y^* = (y_1^*, \ldots, y_K^*)^\perp$ and $Z^* = (z_1^*, \ldots, z_K^*)^\perp$ that

$$MY^* = \mathbf{1}$$
$$NY^* + MZ^* = \mathbf{1}$$

then the stopped martingale X_τ is given by

$$X_\tau = \sum_{i=1}^K y_i^* \frac{\tau(\tau+1)}{2} + \sum_{i=1}^K z_i^* \tau - \tau - 1.$$

Now it is time to apply the Optional Stopping Theorem. However, the increments of the net casino gain X_n are no longer bounded almost sure, so we need a stronger version. The classical Doob's Optional-Stopping Theorem (e.g., Shiryaev (1995) p. 485) will do the trick; one just needs to note that X_n is at most $O(n^2)$, but $\boldsymbol{P}(\tau > n)$ goes to zero at exponential rate. After some algebra we get a general formula for $\boldsymbol{E}\tau^2$.

THEOREM 2.4. *Let* $Y^* = (y_1^*, \ldots, y_K^*)^\perp$ *and* $Z^* = (z_1^*, \ldots, z_K^*)^\perp$ *be the unique solution to the linear system*

$$MY^* = \mathbf{1}$$
$$NY^* + MZ^* = \mathbf{1}$$

then

$$\boldsymbol{E}\tau^2 = \frac{1 + (1 - \sum_{i=1}^K z_i^* - \sum_{i=1}^K y_i^*/2)\boldsymbol{E}\tau}{\sum_{i=1}^K y_i^*/2}.$$

3. Approximating the distribution of the waiting time

With the first two moments in hand, we can approximate the distribution of the waiting time τ with the help from several possible benchmark distributions. This choice is critical, and the most natural choices may not be the best. In some circumstances one can do better than to use exponential, gamma or Weibull.

When selecting the best approximation, it is important to realize that for our purposes the accuracy in the tail of the distribution is important because we are interested in the probability of the waiting time being larger than T, where T is relatively far away from 0. Moreover, as time goes on without observing the desired event, the process is more and more independent of the starting conditions, and hence, $P(\tau = T \mid \tau > T - s)$ is approximately equal to $P(\tau = T - 1 \mid \tau > T - s - 1)$ for large T. This is the property of a homogeneous Poisson process, and hence we would expect that the tail of the waiting time distribution is approximately exponential. This leads us to suggest using the distribution of random variable $c + X$ to approximate the distribution of τ, where $c = \mu - \sigma$ is a constant, X is exponentially distributed with parameter σ, $\mu = \boldsymbol{E}(\tau)$, and $\sigma^2 = \mathbf{Var}(\tau)$. This ensures that the approximate distribution has the same first two moments as the true distribution. We call this the *shifted exponential* distribution, and it suggests that

$$\boldsymbol{P}(\tau \leq n) \approx 1 - \exp(-(n + 0.5 + \sigma - \mu)/\sigma),$$

where the 0.5 term is a continuity correction.

To show that this is indeed a good approximation of the distribution, we will compare it with two other candidates:

1) *exponential*

$$\boldsymbol{P}(\tau \leq n) \approx 1 - \exp(-(n - l)/\mu),$$

where l is the length of the shortest sequence

2) *gamma*

$$\boldsymbol{P}(\tau \leq n) \approx \frac{1}{\Gamma(a)} \int_0^{(n-l)/b} x^a e^{-x} dx,$$

where l is again the length of the shortest sequence, $b = \sigma^2/\mu$, and $a = \mu/b$.

Here the factor l has been introduced to improve the performance of these two approximations, but we will see that even with the best choice of l the shifted exponential distribution does better than each of these. We also investigated the Weibull distribution based approximation. But the Weibull approximations are significantly worse than those of the exponential and the gamma, so we omit them.

4. Fixed window scan statistics

Example 4.1. Assume that we observe a sequence of Bernoulli trials where the probability of failure is known and relatively small -5%. We have an *alert* if we observe too many failures during a short period of time. Specifically, we stop the process if we observe three or more failures in any 5 sequential trials.

The first question is how long we have to wait for an alert which is caused purely by randomness, and this problem can be easily addressed by Theorem 2.1. Indeed, we have an alert when the following runs occur first time: (1) 3-out-of-3—FFF, (2) 3-out-of-4—$FFSF, FSFF$, (3) 3-out-of-5—$FFSSF, FSFSF, FSSFF$.

Table 1. Fixed window scans: at least 3 out of 10, $\boldsymbol{P}(F) = .01$, $\mu = 30822$, $\sigma = 30815$.

n	exponential	shifted exponential	gamma	upper bound	lower bound
500	0.01600	0.01589	0.01597	0.01588	0.01589
1000	0.03183	0.03173	0.03179	0.03171	0.03174
1500	0.04741	0.04731	0.04736	0.04729	0.04733
2000	0.06274	0.06265	0.06267	0.06262	0.06267
2500	0.07782	0.07773	0.07775	0.07770	0.07776
3000	0.09266	0.09258	0.09258	0.09254	0.09261
4000	0.12162	0.12155	0.12154	0.12150	0.12169
5000	0.14966	0.14960	0.14957	0.14954	0.14965

Table 2. Fixed window scans: at least 4 out of 20, $\boldsymbol{P}(F) = .05$, $\mu = 481.59$, $\sigma = 469.35$.

n	exponential	shifted exponential	gamma	upper bound	lower bound
50	0.09110	0.07827	0.08268	0.07713	0.07940
60	0.10977	0.09770	0.10059	0.09543	0.09989
70	0.12807	0.11672	0.11828	0.11337	0.11991
80	0.14599	0.13534	0.13573	0.13095	0.13949
90	0.16354	0.15357	0.15292	0.14819	0.15864
100	0.18073	0.17141	0.16985	0.16508	0.17736

By Theorem 2.1 and easy numerical calculations one finds the expected time is 1608.4. Moreover, Theorem 2.4 tells us that the standard deviation of the waiting time is 1604.8, a value that is notably close to the mean. Still, this is not surprising; an alert is a rare event and dependence between two consecutive alerts is weak, so one expects the distribution of the waiting time to be approximately exponential.

For the fixed window scan statistic, Glaz and Naus (1991) developed tight lower and upper bounds which are presented in Tables 1 and 2 along with the approximations based on the exponential, shifted exponential, and gamma distributions. As can be seen, the shifted exponential approximation performs consistently well, and it has the reassuring feature of staying between the lower and upper bounds. When μ is large and σ is close to μ, the differences between the various approximations are marginal and all of the estimates are close to the true probability, but one should note if μ is relatively small and σ differs from μ, the approximations based on the exponential and gamma distributions do not perform as well as the shifted exponential approximations.

In conclusion, we see that the first two moments are sufficient to obtain a very good approximation for the fixed window scan statistic. We will see shortly that the martingale approach can be successfully used for other scan statistics, even those for which no good bounds or approximations were known earlier.

5. Scan statistics with a variable window size

When searching for clusters, the cluster size is often unknown. That means that we do not know the proper window size to use. For example, if we use a window size of 3 days we may be unable to detect a 3 week cluster, or vice versa. To solve this problem, Loader (1991) and Kulldorff (1997) used the likelihood function instead of the event count to rank the potential clusters. This means, for example, that a cluster with 5 events during 10 days may be ranked higher than both a cluster with 6 events during 20 days and a cluster with 2 events during 4 days.

Example 5.1. Suppose that in a sequence of 30 Bernoulli trials with probability of failure $p = .25$ we observe a window of size 7 with 5 events, and we want to know the probability of observing a cluster of this or higher likelihood during 30 random trails. The first step is then to find other cluster with higher likelihood, which turn out to be a window of 5 with 4 events and a window of 3 with 3 events. This is, we should monitor for the following three types of alerts: (1) when we observe an F run of length 3, (2) at least 4 F out of 5 consecutive trials, (3) at least 5 F out of 7 consecutive trials.

It is easy to see then that the alerts of all kinds are produced by only three sequences: FFF, $FFSFF$, $FFSFSFF$. Therefore, by Theorems 2.1 and 2.4 we find that the expected time for an alert is 72.345, and standard deviation is 69.828.

By Theorem 2.3 and the help of `Mathematica`, one can show also that for an arbitrary p one has

$$\boldsymbol{E}(\alpha^{\tau}) = \frac{P(\alpha)}{Q(\alpha)},$$

where

$$P(\alpha) = p^3\alpha^3 + p^4q\alpha^5 + p^5q^2\alpha^7$$

and where

$$\begin{aligned} Q(\alpha) = {} & 1 + (-1+p)\alpha + (-p+p^2)\alpha^2 \\ & + (-p^2+p^3+p^2q)\alpha^3 + (-p^2q+p^3q)\alpha^4 + (-p^3q+p^4q+p^3q^2)\alpha^5 \\ & + (-p^3q^2+p^4q^2)\alpha^6 + (-p^4q^2+p^5q^2)\alpha^7. \end{aligned}$$

Since

$$\boldsymbol{E}(\tau) = \frac{\partial \boldsymbol{E}(\alpha^{\tau})}{\partial \alpha}\bigg|_{\alpha=1},$$

one can get the expected time via differentiation.

Now, going back to the original problem, we can see that observing a cluster 5-out-of-7 failures in the sequence of 30 trials is not a rare event since the expected time till having this cluster (or a more extreme one) is about 70. The shifted exponential approximation gives a p-value which is approximately equal to .33. The simulated (10000 simulations) p-value is also $\approx$.33.

Example 5.2. Assume that we observe iid Bernoulli trials with $p = .01$ and we scan for (1) at least 2 failures in 10 consecutive trials, (2) or at least 3 in 50 consecutive trials.

We are interested in the approximation for the distribution of the waiting time till one of these two situations occur. The total number of stopping patterns that trigger

Table 3. Variable window: at least 2 out of 10 or at least 3 out of 50, $\boldsymbol{P}(F) = .01$, $\mu = 795.33$, $\sigma = 785.85$.

n	exponential	shifted exponential	gamma	simulated $N = 100000$
50	0.05857	0.05085	0.05542	0.05029
60	0.07033	0.06285	0.06685	0.06187
70	0.08195	0.07470	0.07817	0.07404
80	0.09342	0.08640	0.08939	0.08623
90	0.10474	0.09796	0.10050	0.09718
100	0.11593	0.10936	0.11150	0.11058

these two alerts is 224. In this case, the exponential and gamma approximations are especially interesting, because it is difficult to get the exact distribution of τ, to the best knowledge the most efficient method is the computationally heavy Markov chain embedding method given by Antzoulakos (2001). The introduced approximations could be useful provided they are accurate, and as we will see they are.

The numerical results are given in Table 3, and compared with estimated probabilities based on 100000 replications. We see that the two moment approximation based on the shifted exponential distribution performs quite well, and these approximations are the first approximations that anyone has given for this variable window scan statistic.

6. Double scan statistics

Naus and Wartenberg (1997) and Naus and Stefanov (2002) considered double scan statistic where one is interested in the probability of observing a cluster where the window contains at least k_1 events of type 1 and at least k_2 of events of type 2. The martingale approach works for these types of scan statistics as well.

Example 6.1. Assume that we have two types of failures F_1 and F_2 and suppose that we stop if we have three failures of the first type in a row or at least two F_2 out of three consecutive trials. The waiting time for an alert caused by randomness is determined by the first occurrence of any of the following four runs: (1) $F_1F_1F_1$, (2) F_2F_2, (3) $F_2F_1F_2$, and (4) F_2SF_2.

If we let $\boldsymbol{P}(F_1) = p_1$, $\boldsymbol{P}(F_2) = p_2$, and $\boldsymbol{P}(S) = q = 1 - p_1 - p_2$, then the matrix $M(\alpha)$ is given by

$$M(\alpha) = \begin{bmatrix} \frac{1}{p_1^3\alpha^2} + \frac{1}{p_1^2\alpha} + \frac{1}{p_1} & 0 & 0 & 0 \\ 0 & \frac{1}{p_2^2\alpha} + \frac{1}{p_2} & \frac{1}{p_2} & \frac{1}{p_2} \\ 0 & \frac{1}{p_2} & \frac{1}{p_2^2q\alpha^2} + \frac{1}{p_2} & \frac{1}{p_2} \\ 0 & \frac{1}{p_2} & \frac{1}{p_2} & \frac{1}{p_2^2q\alpha^2} + \frac{1}{p_2} \end{bmatrix},$$

and by solving the system $M(\alpha)Y = \mathbf{1}$ we get generating function for τ

$$\boldsymbol{E}(\alpha^\tau) = 1 - \left(1 + \frac{\alpha}{1-\alpha}\left(\frac{1}{\frac{1}{p_1} + \frac{1}{p_1^2} + \frac{1}{\alpha^2 p_1^3}} + \frac{\frac{1}{p_1} + \frac{1}{q} + \frac{1}{p_1 q}}{\frac{1}{p_2}\left(\frac{1}{q} + \frac{1}{p_1}\left(1 + \frac{1}{q} + \frac{1}{\alpha p_2 q}\right)\right)}\right)\right)^{-1}.$$

Table 4. Double scans: three F_1 in a row or at least two F_2 out of 3, $\boldsymbol{P}(F_1) = .04$, $\boldsymbol{P}(F_2) = .01$, $\mu = 324.09$, $\sigma = 318.34$.

n	exponential	shifted exponential	gamma	simulated $N = 100000$
10	0.02438	0.01480	0.02175	0.01401
15	0.03932	0.03015	0.03568	0.03084
20	0.05403	0.04527	0.04959	0.04508
25	0.06851	0.06015	0.06342	0.06169
30	0.08277	0.07479	0.07714	0.07590
35	0.09681	0.08921	0.09074	0.09134
40	0.11064	0.10340	0.10419	0.10529
45	0.12425	0.11738	0.11749	0.11878
50	0.13766	0.13113	0.13063	0.13342

Table 5. Double scans: at least two F_2 out of 10 or at least three of any kind out of 10, $\boldsymbol{P}(F_1) = .01$, $\boldsymbol{P}(F_2) = .005$, $\mu = 3571.8$, $\sigma = 3566.2$.

n	exponential	shifted exponential	gamma	simulated $N = 100000$
100	0.02706	0.02625	0.02681	0.02713
200	0.05393	0.05318	0.05352	0.05489
300	0.08004	0.07936	0.07955	0.08052
400	0.10544	0.10481	0.10488	0.10639
500	0.13014	0.12957	0.12953	0.13299

Here for a natural numerical example, we note that if $p_1 = .04$, $p_2 = .01$, and $q = .95$, then we get

$$\boldsymbol{E}(\tau) = \frac{\partial \boldsymbol{E}(\alpha^\tau)}{\partial \alpha}\bigg|_{\alpha=1} = 3897.7.$$

To find the standard deviation of the waiting time, we now only need to take the second derivative of the generating function, the standard deviation can also be calculated via Theorem 2.4. In particular, when $p_1 = .04$, $p_2 = .01$, and $q = .95$, the standard deviation is equal to 3895.6. The closeness of μ and σ suggests that again the exponential approximation to the distribution of τ may be appropriate.

Example 6.2. Assume we have a scanning window of length 10 and we stop the scanning process if we have one of the following two situations: (1) at least two failures of type two, F_2, (2) at least three failures of any kind.

The total number of stopping sequences is 153. We have

1) 9 sequences with exactly two F_2

$$F_2F_2, F_2SF_2, \ldots, F_2SSSSSSSF_2,$$

2) 108 sequences with exactly two F_1 and one F_2

$$F_2F_1F_1, F_2SF_1F_1, F_2F_1SF_1, \ldots, F_2SSSSSSSF_1F_1, \ldots, F_2F_1SSSSSSSF_1,$$

$$F_1F_2F_1, F_1SF_2F_1, F_1F_2SF_1, \ldots, F_1SSSSSSSF_2F_1, \ldots, F_1F_2SSSSSSSF_1,$$
$$F_1F_1F_2, F_1SF_1F_2, F_1F_1SF_2, \ldots, F_1SSSSSSSF_1F_2, \ldots, F_1F_1SSSSSSSF_2,$$

3) and 36 sequences with exactly three F_1

$$F_1F_1F_1, F_1SF_1F_1, F_1F_1SF_1, \ldots, F_1SSSSSSSF_1F_1, \ldots, F_1F_1SSSSSSSF_1.$$

As we can see from Tables 4 and 5 all the approximations do well if μ is large, and the shifted exponential does better if μ is relatively small.

7. Multivariate scan statistics

For a multivariate scan statistic, we have multiple data streams and we have common scanning window. We are interested in the probability of simultaneously observing a specified number of events in each data stream. For example, we may be interested in the probability of seeing at least 3 events in data stream A and 5 events in data stream B during any 10 day period. The probability may be different for the events in the different data streams.

Example 7.1. Let $\{Z_i\}_{i\geq 1}$ will be iid sequence of bivariate random variables, i.e. $Z_i = [Z_i^{(1)}, Z_i^{(2)}]^{\perp}$. Assume that

$$Z_i^{(j)} \in \{1,2,3\}, \qquad j = 1,2$$

and

$$p_{km} = \boldsymbol{P}(Z_i^{(1)} = k, Z_i^{(2)} = m), \qquad k,m = 1,2,3.$$

We stop at time τ if (1) $Z_{\tau-1}^{(1)} + Z_{\tau}^{(1)} \geq 5$ or (2) $Z_{\tau-1}^{(2)} + Z_{\tau}^{(2)} = 6$.

This stopping rule is determined by 33 stopping sequences:

$$\begin{bmatrix}33\\11\end{bmatrix} \quad \begin{bmatrix}33\\21\end{bmatrix} \quad \begin{bmatrix}33\\12\end{bmatrix} \quad \begin{bmatrix}33\\31\end{bmatrix} \quad \begin{bmatrix}33\\13\end{bmatrix} \quad \ldots.$$

Now the question is how to compute $\boldsymbol{E}(\tau)$. At first glance this "two-dimensional" situation seems significantly different from the considered earlier examples, but it is not. To see how easy it is, we first introduce the following 9-letter alphabet of 2-tuples:

$$\begin{bmatrix}1\\1\end{bmatrix} \quad \begin{bmatrix}1\\2\end{bmatrix} \quad \begin{bmatrix}1\\3\end{bmatrix} \quad \begin{bmatrix}2\\1\end{bmatrix} \quad \begin{bmatrix}2\\2\end{bmatrix} \quad \begin{bmatrix}2\\3\end{bmatrix} \quad \begin{bmatrix}3\\1\end{bmatrix} \quad \begin{bmatrix}3\\2\end{bmatrix} \quad \begin{bmatrix}3\\3\end{bmatrix}.$$

In this alphabet each of the 33 sequences is identified with a 2-letter word, so we can again apply our earlier results without any changes. For example, if probabilities p_{km} are given by

$$\begin{matrix}.7 & .05 & .02\\ .1 & .04 & .01\\ .05 & .02 & .01\end{matrix}$$

then the expected waiting time is 37.007 and the standard deviation is 35.633.

Table 6. Bivariate multinomial scans: $\mu = 494.92$, $\sigma = 493.45$.

n	exponential	shifted exponential	gamma	simulated $N = 100000$
10	0.01603	0.01814	0.01570	0.01833
20	0.03572	0.03784	0.03513	0.03758
30	0.05500	0.05714	0.05425	0.05718
40	0.07391	0.07606	0.07301	0.07497
50	0.09243	0.09459	0.09143	0.09507
60	0.11058	0.11276	0.10950	0.11422
70	0.12838	0.13056	0.12723	0.13214
80	0.14581	0.14800	0.14461	0.14902
90	0.16290	0.16509	0.16166	0.16301
100	0.17964	0.18184	0.17838	0.17905

Table 7. Bivariate Bernoulli scans: $\mu = 786.31$, $\sigma = 783.49$.

n	exponential	shifted exponential	gamma	simulated $N = 100000$
25	0.02883	0.02853	0.02822	0.02828
50	0.05922	0.05904	0.05826	0.05857
75	0.08866	0.08859	0.08747	0.08842
100	0.11718	0.11721	0.11584	0.11776
125	0.14481	0.14494	0.14336	0.14627
150	0.17157	0.17179	0.17005	0.17118

Finally, let us provide numerical results in the case of more realistic multivariate iid sequences. Specifically, let us consider a sequence with a different distribution over the 9-letter alphabet:

$$\begin{matrix} .9 & .03 & .02 \\ .02 & .01 & .005 \\ .005 & .005 & .005 \end{matrix}.$$

Table 6 contains the numerical results for this example.

Example 7.2. Assume $\{Z_i\}_{i\geq 1}$ is an iid sequence of bivariate random variables, i.e. $Z_i = [Z_i^{(1)}, Z_i^{(2)}]^{\perp}$, where each component is a Bernoulli random variable with the following joint distribution:

$$\boldsymbol{P}(Z_i^{(1)} = 0, Z_i^{(2)} = 0) = .98, \quad \boldsymbol{P}(Z_i^{(1)} = 1, Z_i^{(2)} = 0) = .005,$$
$$\boldsymbol{P}(Z_i^{(1)} = 0, Z_i^{(2)} = 1) = .005, \quad \boldsymbol{P}(Z_i^{(1)} = 1, Z_i^{(2)} = 1) = .01.$$

In each row we have a scanning window of length 5, and we stop if in one of the two windows we have at least 2 ones. As before first let us introduce the following 4-letter alphabet:

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

In this new alphabet we have 40 stopping sequences that correspond to the stopping rule described above. Numerical results are presented in Table 7.

8. Discussion

The martingale approach yields a formula like that of Theorem 2.4 for *any* moment of τ, and, in theory, higher moments should provide better scan statistics approximations. Nevertheless, for the scan statistics of importance in practice, it is evident two moments are all one needs to get very good estimates.

We used the martingale approach for a number of different scan statistics, but we view it as a general tool of wide applicability. We believe that the martingale methods can also be applied for continuous, inhomogeneous, or spatial scan statistics—all of which are of practical importance. We are less optimistic about the utility of the martingale approach for conditional scan statistics, except to the extent that the unconditional scan statistic is sometimes a good approximation of the conditional scan statistic.

When one compares the martingale approach to the Markov chain embedding method recently developed by Antzoulakos (2001), Fu (2001) and Fu and Chang (2002), one finds that neither method dominates the other—each has its own advantages and disadvantages. The martingale approach always results in a smaller set of linear equations to be solved, sometimes significantly reducing computational complexity. Also the martingale method can be used to obtain higher moments. On the other hand, the Markov chain embedding method works for Markov dependent trials, but the martingale approach does not seem to be able to cover this case.

References

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics*, **6**, 701–726.

Aki, S. and Hirano, K. (1999). Sooner and later waiting time problems for runs in Markov dependent bivariate trials, *Annals of the Institute of Statistical Mathematics*, **51**, 17–29.

Andersen, P. K., Borgan, O., Gill, R. D. and Keiding, N. (1993). *Statistical Methods Based on Counting Processes*, Springer Series in Statistics, Springer-Verlag, New York.

Antzoulakos, D. (2001). Waiting times for patterns in a sequence of multistate trials, *Journal of Applied Probability*, **38**, 508–518.

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications*, Wiley Series in Probability and Statistics, John Wiley & Sons, New York.

Biggins, J. D. and Cannings, C. (1987). Markov renewal processes, counters and repeated sequences in Markov chains, *Advances in Applied Probability*, **19**, 521–545.

Blom, G. and Thorburn, D. (1982). How many random digits are required until given sequences are obtained?, *Journal of Applied Probability*, **19**, 518–531.

Blom, G., Holst, L. and Sandell, D. (1994). *Problem and Snapshots from the World of Probability*, Springer-Verlag, New York.

Breen, S., Waterman, M. and Zhang, N. (1985). Renewal theory for several patterns, *Journal of Applied Probability*, **22**, 228–234.

Chao, M. T. and Fu, J. C. (1991). The reliability of large series systems under Markov structure, *Advances in Applied Probability*, **23**, 894–908.

Chrysaphinou, O. and Papastavridis, S. (1990). The occurrence of a sequence of patterns in repeated dependent experiments, *Theory of Probability and Applications*, **35**, 145–152.

Coulston, J. and Riitters, K. (2003). Geographic analysis of forest health indicators using spatial scan statistics, *Environmental Management*, **31**, 764–773.

Durand, D. and Sankoff, D. (2003). Tests for gene clustering, *Journal of Computational Biology*, **10**, 453–482.

Enemark, L., Ahrens, P., Juel, D., Petersen, E., Petersen, R., Andersen, J., Lind, P. and Thamsborg, S. (2002). Molecular characterization of Danish Cryptosporidium parvum isolates, *Parasitology*, **125**, 331–341.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. 1, 3rd ed., Wiley, New York.

Fu, J. C. (1986). Reliability of consecutive-k-out-of-n: F systems with $(k-1)$-step Markov dependence, *IEEE Transactions on Reliability*, **R35**, 602–606.

Fu, J. C. (1996). Distribution theory of runs and patterns associated with a sequence of multi-state trials, *Statistics Sinica*, **6**, 957–974.

Fu, J. (2001). Distribution of the scan statistics for a sequence of bistate trials, *Journal of Applied Probability*, **38**, 908–916.

Fu, J. and Chang, Y. (2002). On probability generating functions for waiting time distribution of compound patterns in a sequence of multistate trials, *Journal of Applied Probability*, **39**, 70–80.

Fu, J. C. and Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach, *Journal of the American Statistical Association*, **78**, 168–175.

Fu, J. C. and Lou, W. Y. W. (2003). *Distribution Theory of Runs and Patterns*, World Scientific Publishing, Singapore.

Gerber, H. and Li, S. (1981). The occurrence of sequence patterns in repeated experiments and hitting times in a Markov chain, *Stochastic Processes and Their Applications*, **11**, 101–108.

Glaz, J. and Balakrishnan, N. (eds.) (1999). *Recent Advances on Scan Statistics*, Birkhauser Publishers, Boston.

Glaz, J. and Naus, J. (1991). Tight bounds for scan statistics probabilities for discrete data, *Annals of Applied Probability*, **1**, 306–318.

Glaz, J., Naus, J. and Wallenstein, S. (2001). *Scan Statistics*, Springer, New-York.

Goldstein, L. and Waterman, M. S. (1992). Poisson, compound Poisson and process approximations for testing statistical significance in sequence comparisons, *Bulletin of Mathematical Biology*, **54**, 785–812.

Han, Q. and Hirano, K. (2003). Sooner and later waiting time problems for patterns in Markov dependent trials, *Journal of Applied Probability*, **40**, 73–86.

Kaminski, R., Jefferis, E. and Chanhatasilpa, C. (2003). A spatial analysis of American police killed in the line of duty, *Atlas of Crime: Mapping the Criminal Landscape* (eds. L. S. Turnbull, E. H. Hendrix and B. D. Dent), Oryx Press, Phoenix, Arizona.

Karlin, S. and Brendel, V. (1992). Chance and statistical significance in protein and DNA sequence analysis, *Science*, **257**, 39–49.

Kulldorff, M. (1997). A spatial scan statistic, *Communications in Statistics: Theory and Methods*, **26**, 1481–1496.

Li, S. (1980). A martingale approach to the study of occurrence of sequence patterns in repeated experiments, *the Annals of Probability*, **8**, 1171–1176.

Loader, C. (1991). Large deviation approximations to distribution of scan statistics, *Advances in Applied Probability*, **23**, 751–771.

Margai, F. and Henry, N. (2003). A community-based assessment of learning disabilities using environmental and contextual risk factors, *Social Science and Medicine*, **56**, 1073–1085.

Naus, J. I. (1965). The distribution of the size of the maximum cluster of points on a line, *Journal of the American Statistical Association*, **60**, 532–538.

Naus, J. I. and Sheng, K. N. (1997). Matching among multiple random sequences, *Bulletin of Mathematical Biology*, **59**, 483–496.

Naus, J. I. and Stefanov, V. T. (2002). Double-scan statistics, *Methodology and Computing in Applied Probability*, **4**, 163–180.

Naus, J. I. and Wartenberg, D. A. (1997). A double-scan statistic for clusters of two types of events, *Journal of the American Statistical Association*, **92**, 1105–1113.

Pozdnyakov, V. and Kulldorff, M. (2003). On the occurrence of sequence patterns: An alternative proof and extended results (preprint).

Robin, S. and Daudin, J.-J. (2001). Exact distribution of the distances between any occurence of a set of words, *Annals of the Institute of Statistical Mathematics*, **53**, 895–905.

Sheng, K.-N. and Naus, J. (1994). Pattern matching between two non-aligned random sequences, *Bulletin of Mathematical Biology*, **56**, 1143–1162.

Shiryaev, A. N. (1995). *Probability*, 2nd ed., Springer, New York.

Shmueli, G. (2003*a*). Computing consecutive-type reliabilities non-recursively, *IEEE Transactions on Reliability*, **52**, 367–372.

Shmueli, G. (2003*b*). System-wide probabilities for systems with runs and scans rules, *Methodology and Computing in Applied Probability*, **4**, 401–419.

Stefanov, V. T. (2000). On some waiting time problems, *Journal of Applied Probability*, **37**, 756–764.

Stefanov, V. T. (2003). The intersite distances between pattern occurrences in strings generated by general discrete- and continuous-time models: An algorithmic approach, *Journal of Applied Probability*, **40**, 881–892.

Stefanov, V. T. and Pakes, A. G. (1997). Explicit distributional results in pattern formation, *Annals of Applied Probability*, **7**, 666–678.

Uchida, M. (1998). On generating functions of waiting time problems for sequence patterns of discrete random variables, *Annals of the Institute of Statistical Mathematics*, **50**, 655–671.

Williams, D. (1991). *Probability with Martingales*, Cambridge University Press, Cambridge.

Yoshida, M., Naya, Y. and Miyashita, Y. (2003). Anatomical organization of forward fiber projections from area TE to perirhinal neurons representing visual long-term memory in monkeys, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 4257–4262.