Annals of the Institute of Statistical Mathematics (2020) 72:531–543
https://doi.org/10.1007/s10463-018-0696-3

Conditional waiting time distributions of runs and patterns and their applications

Tung-Lung Wu[1]

Received: 17 March 2018 / Revised: 16 October 2018 / Published online: 12 November 2018

Abstract

In this paper, a simple and general method based on the finite Markov chain imbedding technique is proposed to determine the exact conditional distributions of runs and patterns in a sequence of Bernoulli trials given the total number of successes. The idea is that given the total number of successes, the Bernoulli trials are viewed as random permutations. Then, we extend the result to multistate trials. The conditional distributions studied here lead to runs and patterns-type distribution-free tests whose applications are widespread. Two applications are considered. First, a distribution-free test for randomness is applied to rainfall data at Oxford from 1858 to 1952. The second application is to develop runs and patterns-type distribution-free control charts which can be used as Phase I and/or Phase II control charts. Numerical results for two commonly used runs-type statistics, the longest run and scan statistics, are also given.

Keywords Distribution-free tests · Conditional runs and patterns · Finite Markov chain imbedding · Control charts · Random permutation · Waiting time

1 Introduction

Many distribution-free tests are based on runs and patterns as they are easy to understand and easy to interpret. For example, Wald and Wolfowitz (1940) proposed a conditional run test for randomness using the number of runs given the total number of successes, and Lou (1996) considered the number of success runs and the length of the longest success run. To test for symmetry, many authors have considered runs tests; see, for example, Cohen and Menjoge (1988), McWilliams (1990), Gastwirth (1971) and Randles et al. (1980). As a special case of runs-type statistics, scan statistics are widely studied in many areas. For example, scan statistics can be seen in epidemiology (e.g., Kulldorff 1997), system reliability (e.g., Chang and Huang 2010), sensor

✉ Tung-Lung Wu
tw1475@msstate.edu

[1] Department of Mathematics and Statistics, Mississippi State University, Starkville, MS 39759, USA

network (e.g., Song et al. 2012) and DNA sequence analysis (e.g., Karwe and Naus 1997). Scan statistics are powerful in detecting local clusters, but their distributions are difficult to compute. Hence, it remains an open question to develop an efficient algorithm to compute the distributions of scan statistics.

Traditionally, distributions of runs and patterns are obtained using combinatorics. However, Fu and Koutras (1994) proposed the revolutionary finite Markov chain imbedding (FMCI) technique for distributions of runs and patterns. It is simple and computationally efficient, and it provides a unified framework to determine the exact *unconditional* distributions of runs and patterns. A forward and backward principle is also introduced to systematically construct desired Markov chains and state spaces (e.g., Fu and Lou 2003). It has been shown that the FMCI technique is a powerful numerical tool having many successfully applications in quality control (e.g., Fu et al. 2002), hypothesis testing (e.g., Lou and Fu 2007) and boundary crossing probabilities (e.g., Fu and Wu 2016). Although there are many existing results for unconditional distributions of runs and patterns, very little work has been done for conditional distributions of runs and patterns. Few exceptions include the conditional longest run by Lou (1996) and conditional scan statistics by Fu et al. (2012). It is the potential applications of conditional distributions of runs and patterns that motivate us to pursue this line of research.

In this paper, we develop a general framework for conditional distributions of runs and patterns. In Sect. 2, we study the conditional distributions of runs and patterns in a sequence of independent Bernoulli trials, including the longest run and scan statistics. The general result for conditional distributions of runs and patterns in a sequence of independent multistate trials is given in Sect. 3. Numerical results are given in Sect. 4. Two applications, distribution-free tests and distribution-free control charts, are given in Sect. 5. Summary is given in Sect. 6.

2 Conditional distributions of runs and patterns in a sequence of Bernoulli trials

In this section, we derive the exact distributions of runs and patterns given the total number of successes for a sequence of independent Bernoulli trials.

Consider a sequence of Bernoulli trials with two possible outcomes, success (1) and failure (0) and $S_2 = \{0, 1\}$. Let $X_1, \ldots, X_n$ be a sequence of Bernoulli random variables with $p = P(X_i = 1) = 1 - P(X_i = 0)$, $i = 1, 2, \ldots, n$. Let $N_1 = \sum_{i=1}^{n} X_i$ be the number of successes and $N_0 = n - N_1$ be the number of failures.

Let $\Lambda = \cup_{i=1}^{L} \Lambda_i$ be a compound pattern consisting of L simple patterns $\Lambda_1, \Lambda_2, \ldots, \Lambda_L$, where each simple pattern Λ_i is composed of a specified sequence of 2 symbols $\{0, 1\}$, and the length of Λ_i is fixed. Denote the waiting time of the first occurrence of the compound pattern Λ in a sequence by $W(\Lambda)$.

Given $N_1 = n_1$ and $N_0 = n - n_1 \equiv n_0$, a sequence of Bernoulli trials can be viewed as an $[n_1, n_0]$-specified random permutation, i.e., a random permutation of n_1 1's and n_0 0's. Let

$$\mathcal{P}_2 = \left\{ \boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n) : \pi_i = 0, 1 \text{ and } \sum_{i=1}^{n} \pi_i = n_1 \right\}$$

be the family of random permutations of n_1 1's and n_0 0's. Then, conditional distributions of runs and patterns given the total number n_1 of successes in a sequence of n Bernoulli trials are the same as distributions of runs and patterns in an $[n_1, n_0]$-specified random permutation $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$. It is worthwhile to mention that in an $[n_1, n_0]$-specified random permutation, the distributions of runs and patterns are independent of p. Specifically, the waiting time $W(\Lambda)$ in a sequence of Bernoulli trials can be viewed as the waiting time $W(\Lambda)$ in a random permutation of n_1 1's and n_0 0's. Next, we show how to construct a Markov chain for $P(W(\Lambda) > n)$ in an $[n_1, n_0]$-specified random permutation.

To explain the Markov chain imbedding procedure in an $[n_1, n_0]$-specified random permutation, we consider an urn consisting of n_1 balls labeled 1 and n_0 balls labeled 0. The balls are drawn one by one without replacement until we see any of the L patterns $\Lambda_1, \Lambda_2, \ldots, \Lambda_L$ or the urn is emptied. To form a Markov chain, a serial of so-called ending blocks is recorded toward the formation of a simple pattern (see, e.g., Fu and Lou 2003). Now define E_Λ as a set of ending blocks or the collection of subpatterns of the L simple patterns. For example, given a compound pattern $\Lambda = 1011 \cup 111$, the set of ending blocks is $\{0,1,10,11,101\}$, excluding the patterns 1011 and 111. Then, we can define a nonhomogeneous Markov chain $\{Y_t\}_{t=0}^{n}$ on the state space

$$\Omega = \{(\ell, \omega) : \ell = 0, 1, \ldots, n_1 \text{ and } \omega \in E_\Lambda \cup S_2\} \cup \{\emptyset, \alpha\}, \tag{1}$$

where E_Λ is the collection of all ending blocks, $\emptyset$ represents the initial state and α the absorbing state. At any time t (or t-th draw), a state $Y_t = (\ell_t, \omega_t)$ represents that during the sequential sampling process from the urn consisting of n_1 1's and n_0 0's, the total number of 1's observed during the first t draws is ℓ_t, and the longest subpattern observed up to time t is ω_t. The transition probabilities from state $u = (\ell_{t-1}, \omega_{t-1})$ to state $v = (\ell_t, \omega_t)$ are given by

$$\begin{aligned} p_{uv}(t) &= P(Y_t = (\ell_t, \omega_t) | Y_{t-1} = (\ell_{t-1}, \omega_{t-1})) \\ &= \begin{cases} \frac{N_1 - \ell_{t-1}}{n-t+1} & \text{if } \pi_t = 1, \ell_t = \ell_{t-1} + 1 \text{ and } \omega_t =< \omega_{t-1}, 1 >_\Omega, \\ \frac{N_0 - t + \ell_{t-1} + 1}{n-t+1} & \text{if } \pi_t = 0, \ell_t = \ell_{t-1} \text{ and } \omega_t =< \omega_{t-1}, 0 >_\Omega, \\ 1 & \text{if } \omega_t = \omega_{t-1} = \alpha, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{2}$$

where $< \omega_{t-1}, \pi_t >_\Omega$ denotes the longest subpattern after π_t is observed. The transition matrices are of the form

$$\mathbf{M}_t(n_1) = \left[\begin{array}{c|c} \mathbf{N}_t(n_1) & \mathbf{C}_t(n_1) \\ \hline \mathbf{0} & 1 \end{array} \right],$$

$t = 1, 2, \ldots, n$. We suppress n_1 in the parenthesis for notational simplicity.

Theorem 1 *Let $X_1, \ldots, X_n$ be a sequence of independent Bernoulli trials and Λ be a compound pattern. Then, given the total number of successes $N_1 = n_1$, the conditional waiting time distribution is given by*

$$P\left(W(\Lambda) > n \middle| N_1 = n_1\right) = \boldsymbol{\xi}_0 \prod_{t=1}^{n} \mathbf{N}_t \mathbf{1}^\top, \tag{3}$$

where $\boldsymbol{\xi}_0$ is a vector of initial probabilities, $\mathbf{N}_t, t = 1, 2, \ldots, n$ are the essential matrices whose entries are given in (2) and $\mathbf{1}$ is a row vector of ones.

Proof Based on the above construction, the waiting time random variable $W(\Lambda)$ is finite Markov chain imbeddable and it follows from Theorem 2.1 of Fu and Lou (2003) that the exact distribution is of the form in (3). The proof is completed. □

One can see that the conditional distribution in (3) is independent of p. Note that there are a few papers that have studied the conditional distributions of certain runs in the literature. Lou (1996) derived the exact conditional distributions of runs and the longest run. Lou treated the conditional distribution as the ratio of a joint distribution and a marginal distribution, and the two distributions are subsequently obtained by the FMCI technique. We consider Lou's method as an indirect method. Another paper by Fu et al. (2012) studied the conditional distribution of the discrete scan statistic for a sequence of Bernoulli trials given the total number of successes. There is a dual relationship that makes the scan statistic problem a special case of waiting time distribution of a compound pattern (e.g., Fu 2001). Considering the wide applications of the longest run and scan statistics (e.g., Woodall 2006), it is worth giving some details about these two special cases of Theorem 1 in the next two subsections.

2.1 The longest run

Let L_n denote the length of the longest run of 1's in a sequence of Bernoulli trials. Consider the pattern

$$\Lambda_d = \{\underbrace{11\cdots 1}_{d}\}.$$

The event $\{L_n < d\}$ occurs if and only if the pattern Λ_d does not occur in the sequence of Bernoulli trials. Thus, we have

$$P(L_n < d | N_1 = n_1) = P(W(\Lambda_d) > n | N_1 = n_1).$$

Example 1 Suppose that $n = 5$, $N_1 = 3$ and $d = 3$. The pattern corresponding to $\{L_5 < 3\}$ is $\Lambda_3 = 111$. The ending block is $E_{\Lambda_3} = \{1, 11\}$ and $E_{111} \cup S_2 = \{0, 1, 11\}$. The state space can be constructed according to (1) and is given by $\Omega = \{(0, 0), (1, 0), (1, 1), (2, 0), (2, 1), (2, 11), (3, 0), (3, 1), (3, 11)\} \cup \{\emptyset, \alpha\}$. Some redundant states are removed from the state space. For example, $(1, 11)$ would never occur. According to Theorem 1, the probability that the length of the longest

1's run is less than 3 is 0.7. The same probability can be obtained by enumeration. Given 3 successes (1's) in 5 Bernoulli trials, there are 10 possible outcomes $\{11100, 11010, 10110, 01110, 11001, 10101, 01101, 10011, 01011, 00111\}$, among which seven outcomes $\{11010, 10110, 11001, 10101, 01101, 10011, 01011\}$ contain the longest run of length less than three, and hence, the probability is 7/10.

2.2 Scan statistics

The scan statistic is defined as

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

where $S_n(r,t) = \sum_{i=t}^{t+r-1} X_i$ and r is the window size. The distribution of the scan statistic can be cast as the waiting time distribution of a compound pattern. The event $\{S_n(r) < s\}$ occurs if and only if an associated compound pattern $\Lambda_{r,s}$ does not appear in the sequence. For example, consider $r = 5$ and $s = 2$, the compound pattern associated with the event $\{S_n(5) < 2\}$ is $\Lambda_{5,2} = \{11, 101, 1001, 10001\}$. Thus, the probability $P(S_n(5) < 2)$ can be obtained through $P(W(\Lambda_{5,2}) > n)$. The relationship still holds true in the conditional case, i.e.,

$$P\left(S_n(5) < 2|N_1 = n_1\right) = P\left(W(\Lambda_{5,2}) > n|N_1 = n_1\right).$$

In general, for any window size r and s, a compound pattern $\Lambda_{r,s}$ can be defined accordingly, and the total number of simple patterns comprising the compound pattern $\Lambda_{r,s}$ can also be obtained. See Fu (2001) for details.

3 Conditional distributions of runs and patterns in a sequence of multistate trials

Let $\{X_i\}_{i=1}^n$ be a sequence of multistate trials. Each trial has m ($m \ge 2$) possible outcomes labeled $1, 2, \ldots, m$ with $P(X_1 = j) = p_j, j = 1, \ldots, m$ and $S_m = \{1, 2, \ldots, m\}$. Note that when $m = 2$, we let $S_2 = \{0, 1\}$. Let M_n be a finite multiset generated from S_m. If the multiplicities of the symbols (in increasing order) are $N_1, N_2, \ldots, N_m$ and $\sum N_i = n$, then a random permutation of M_n is called an $[N_1, N_2, \ldots, N_m]$-specified random permutation. Let $\mathcal{P}_m$ be the collection of all such random permutations of M_n.

Given a sequence of m-state trials such that $N_i = n_i, i = 1, \ldots, m$, the sequence is then an $[n_1, n_2, \ldots, n_m]$-specified random permutation, and conditional distributions of runs and patterns can again be viewed as distributions of runs and patterns in an $[n_1, n_2, \ldots, n_m]$-specified random permutation.

Let $\Lambda = \cup_{i=1}^L \Lambda_i$ be a compound pattern consisting of L simple patterns Λ_i, where each simple pattern Λ_i is composed of a specified sequence of m symbols and the length of Λ_i is fixed. To study the conditional waiting time distribution of Λ, a nonhomogeneous Markov chain is constructed in an $[n_1, n_2, \ldots, n_m]$-specified random permutation. Again, an urn model can be used to help explain how we form the nonhomogeneous Markov chain. We consider an urn consisting of n_1 balls labeled 1, n_2 balls labeled 2, …, and n_m balls labeled m and $\sum_{i=1}^{m} n_i = n$. The balls are drawn one by one without replacement toward forming the sequence until we see any of the L simple patterns or the urn is emptied. Let E_Λ be the set of ending blocks of the compound pattern Λ. Then, a nonhomogeneous Markov chain $\{Y_t\}_{t=0}^{n}$ can be defined on the state space

$$\Omega = \{(\ell_1, \ell_2, \ldots, \ell_m, \omega) : \ell_i = 0, 1, \ldots, n_i \text{ and } \omega \in E_\Lambda \cup S_m\} \cup \{\emptyset, \alpha\}.$$

At time t (or tth draw), a state $Y_t = (\ell_{1,t}, \ell_{2,t}, \ldots, \ell_{m,t}, \omega_t)$ represents that the number of balls labeled i observed during the first t draws is $\ell_{i,t}$, $i = 1, \ldots, m$ and the longest subpattern observed up to time t is ω_t. The transition probabilities from state $u = (\ell_{1,t-1}, \ell_{2,t-1}, \ldots, \ell_{m,t-1}, \omega_{t-1})$ to state $v = (\ell_{1,t}, \ell_{2,t}, \ldots, \ell_{m,t}, \omega_t)$ are given by

$$\begin{aligned} p_{uv}(t) &= P(Y_t = v | Y_{t-1} = u) \\ &= \begin{cases} \frac{n_j - \ell_{j,t-1}}{n-t+1} & \text{if} \quad \pi_t = j, \ell_{j,t} = \ell_{j,t-1} + 1 \text{ and } \omega_t =< \omega_{t-1}, j >_\Omega, \\ & j = 1, \ldots, m, \\ 1 & \text{if} \quad \omega_t = \omega_{t-1} = \alpha, \\ 0 & \text{otherwise}, \end{cases} \end{aligned} \tag{4}$$

where $< \omega_{t-1}, \pi_t >_\Omega$ denotes the longest subpattern after π_t is observed. Again, the transition matrices are of the form

$$\mathbf{M}_t = \left[\begin{array}{c|c} \mathbf{N}_t & \mathbf{C}_t \\ \hline \mathbf{0} & 1 \end{array}\right],$$

$t = 1, 2, \ldots, n$.

Theorem 2 *Let $X_1, \ldots, X_n$ be a sequence of independent m-state trials and let Λ be a compound pattern. Then, given the number of occurrences of each symbol, the conditional waiting time distribution is given by*

$$P\left(W(\Lambda) > n \middle| N_1 = n_1, N_2 = n_2, \ldots, N_m = n_m\right) = \boldsymbol{\xi}_0 \prod_{t=1}^{n} \mathbf{N}_t \mathbf{1}^\top,$$

where $\mathbf{N}_t$, $t = 1, \ldots, n$, are the essential matrices whose entries are given in (4).

Proof Based on the above construction, the waiting time random variable $W(\Lambda)$ is finite Markov chain imbeddable and this theorem is a straightforward result of Theorem 2.1 of Fu and Lou (2003). □

Note that the transition probabilities given in (4) do not depend on the probabilities p_j, $j = 1, \ldots, m$.

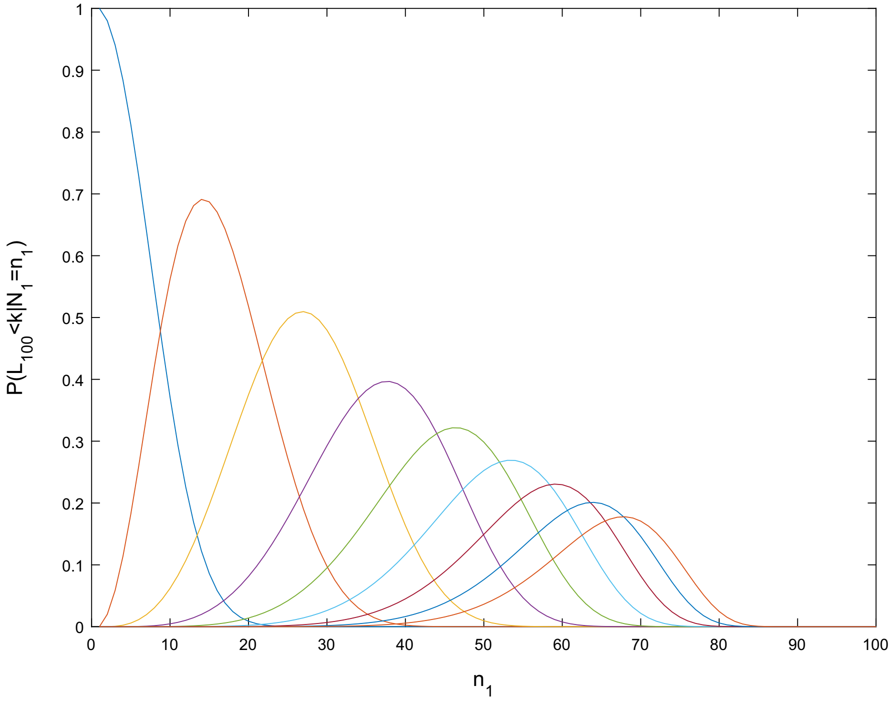

Fig. 1 $P(L_{100} = k|N_1 = n_1)$. From left to right: $k = 1, 2, \ldots, 9$

4 Numerical results

Plots of conditional probabilities of L_n and $S_n(r)$ are given in this section. The plot of conditional probabilities of the longest run is given in Fig. 1. The plots of conditional probabilities of three scan statistics of window sizes 8, 5 and 15 are given in Figs. 2, 3 and 4, respectively.

5 Applications

Two applications are given in this section. The first application is to construct a distribution-free test for randomness for annual rainfall data at Oxford from 1858 to 1952. The second application is to develop distribution-free control charts for monitoring location shifts.

5.1 A distribution-free test for randomness

A distribution-free test for randomness based on runs-type statistics is proposed for meteorological data (Foster and Stuart 1954). The plot of total annual rainfalls at Oxford from 1858 to 1952 is given in Fig. 5.

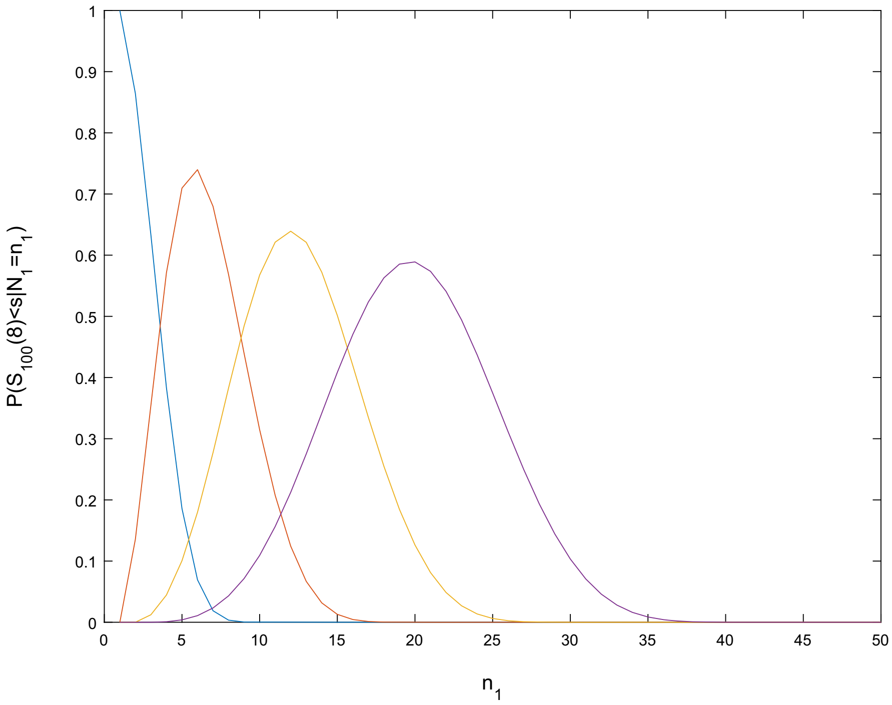

Fig. 2 $P(S_{100}(8) = s|N_1 = n_1)$. From left to right: $s = 1, 2, 3, 4$

To test whether there is an upward trend in mean, we choose to use the longest run as the test statistic. The data are standardized so that the mean is zero and variance is one. Next, we define Bernoulli trials $\{X_t\}_{t=1}^{95}$ as

$$X_t = \begin{cases} 0 & \text{if} \quad Y_t < c, \\ 1 & \text{if} \quad Y_t \geq c, \end{cases} \tag{5}$$

where Y_t is the standardized total annual rainfall of year t. The threshold c may be determined according to the strength of the trend in mean. The choice of c is critical to the performance of the test. A simple choice of c is zero, but there would be too much noise (too many 1's). To reduce the noise and limit the number of 1's, we use $c = 0.1$ instead of 0. This gives 48 years where the standardized total annual rainfalls are above 0.1, and the length of the longest run, located in years 1975–1983, is 9 with p-value 0.0678 (see Table 1). As in Foster and Stuart (1954), we obtain a similar conclusion that the proposed test is not significant at $\alpha = 0.05$. However, it would be significant if we choose $\alpha = 0.1$, and this indicates the evidence is fairly strong that we should not ignore this signal. The histogram in Fig. 5 shows that the distribution of total annual rainfall is not normal, and this makes our distribution-free test a desired method for studying such non-normal data.

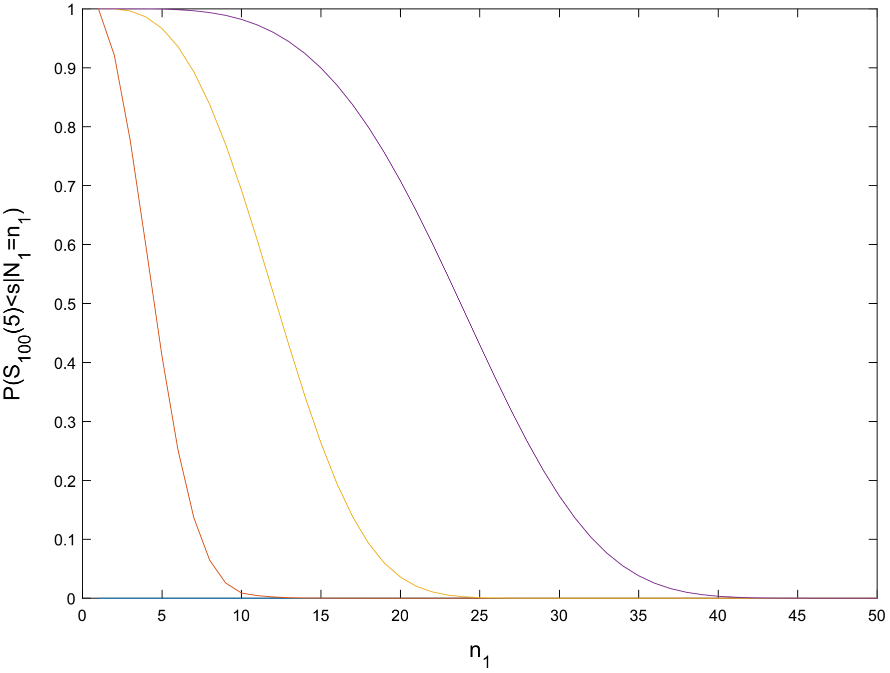

Fig. 3 $P(S_{100}(5) < s|N_1 = n_1)$. From left to right: $s = 2, 3, 4$

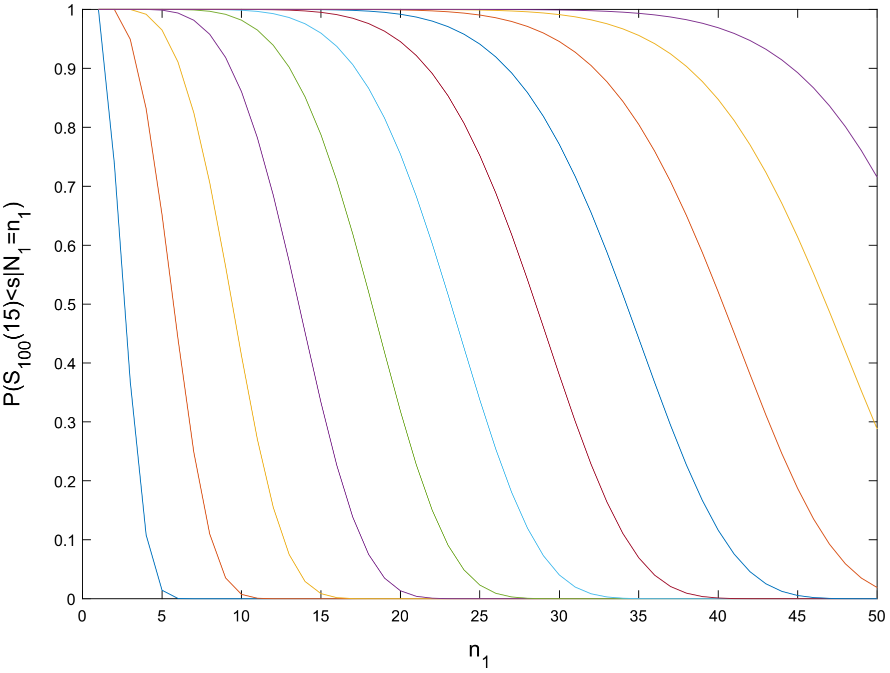

Fig. 4 $P(S_{100}(15) < s|N_1 = n_1)$. From left to right: $s = 2, 3, \ldots, 12$

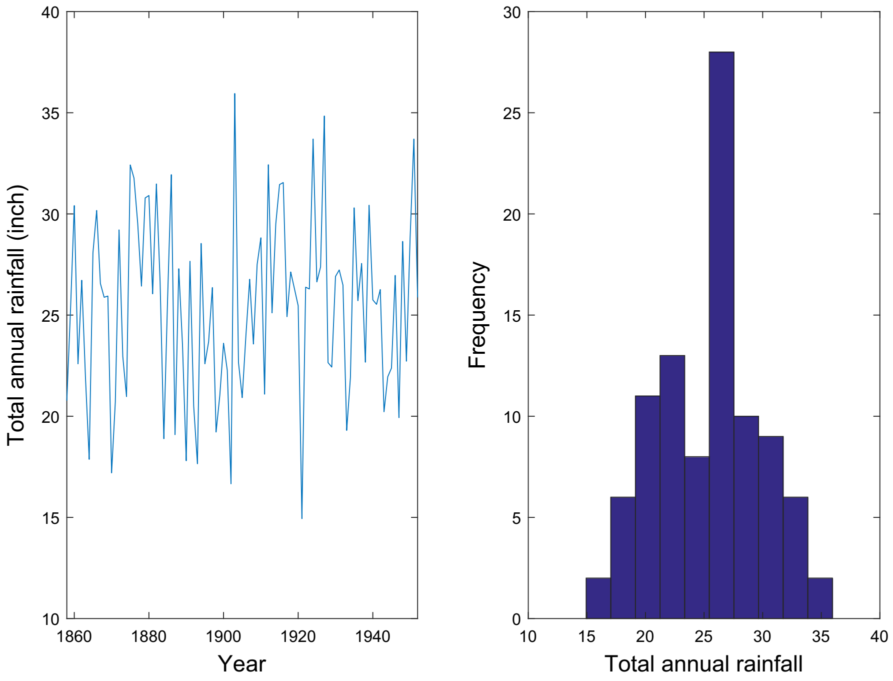

Fig. 5 The total annual rainfalls at Oxford from 1858 to 1952

Table 1 $P(L_{95} < k| \sum_{t=1}^{95} X_t = 48)$

$k = 5$	$k = 6$	$k = 7$	$k = 8$	$k = 9$	$k = 10$
0.1584	0.4550	0.7062	0.8556	0.9322	0.9690

5.2 Statistical process control

Control chart is the main tool in statistical process control (SPC). In general, SPC consists of two steps: Phase I and Phase II. In Phase I, a set of historical data is analyzed to determine if they can be considered from the in-control process and used to estimate the values of parameters and control limits for the Phase II control chart. Here, we consider the problem in Phase I to determine whether the data are from the in-control process or not. For Phase II control charts, see Wu (2018) for distribution-free control charts with data-dependent control limits. For review of recent advances on nonparametric/distribution-free control charts with supplementary runs rules, see Koutras et al. (2007) and Chakraborti et al. (2011).

We propose distribution-free control charts based on some runs and patterns-type statistics. The frequently used runs rules are

Rule 1 One or more points fall outside the three-sigma control limits;
Rule 2 Two of three consecutive points fall outside the two-sigma warning limits;
Rule 3 Four of five consecutive points fall outside the one-sigma limits;
Rule 4 Eight points in a row fall on one side of the center line.

Table 2 Simple patterns with respect to rule (i) and rule (ii)

Rule (i)	Rule (ii)
33, 313, 323	2222, 21222, 22122, 22212
	3222, 31222, 32122, 32212
	2322, 21322, 23122, 23212
	2232, 21232, 22132, 22312
	2223, 21223, 22123, 22213
	3223, 31232, 32132, 32213
	31223, 32123, 23213
	23123

The above rules are often referred to as Western Electric rules; see Western Electric Company (1956). Rules 2 and 3 can be considered as scan rules and Rule 4 as a run rule. A general runs and patterns rule, denoted by $R(k, r, Z)$, is proposed by Shmueli and Cohen (2003). The control chart signals an out-of-control alert if k of the last r tested points fall in the region Z. To utilize these rules, the real line is partitioned into three zones as follows:

Zone 1 = the interval $(-\infty, c_1)$
Zone 2 = the interval (c_1, c_2)
Zone 3 = the interval (c_2, ∞).

Without loss of generality, suppose the target value is 0. To detect an upward shift, we consider two runs rules so that the process is said to be out-of-control if (i) two of three consecutive points fall in Zone 3 and (ii) four of five consecutive points fall in Zone 2 or Zone 3. We want to test H_0 : the process is in-control against H_1 : the process is out-of-control. Note that here we consider the problem in Phase I to determine in-control samples that will be used to estimate the values of parameters of Phase II control charts. Therefore, it is a retrospective study on a given data set. The above rules (i) and (ii) can be cast as waiting time problem of a compound pattern. The out-of-control signal can be viewed as whether the compound pattern occurs or not. According to the rules (i) and (ii), the compound pattern consists of 3 simple patterns with respect to rule (i) and 28 simple patterns with respect to rule (ii). Those simple patterns are given in Table 2.

Finally, we apply the proposed method to the piston ring data in Table 6.3 of Montgomery (2009) to determine whether the process is in-control. Since we are conducting a retrospective analysis, we do not monitor each data point sequentially. We simply examine the entire 25 samples using our proposed rules. For illustrative purpose, the thresholds $c_1 = 74.014$ $(\bar{x} + 1.3\hat{\sigma})$ and $c_2 = 74.026$ $(\bar{x} + 2.5\hat{\sigma})$ are chosen so that the size of the distribution-free test is controlled at 0.0546. It can be seen from Fig. 6 that there are 14 points in Zone 2 and 1 point in Zone 3, but none of the patterns in Table 2 occurs. So we conclude that the process is in-control. The conclusion is consistent with the $\bar{x}$ and s control charts used in Montgomery (2009).

Remark 1 The distribution-free test for reference data considered here is a retrospective analysis. We only test the data once after the data have been collected. However, we

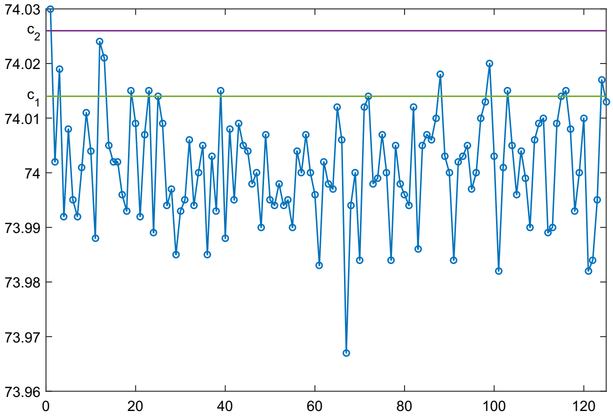

Fig. 6 The proposed patterns-type distribution-free control chart with $c_1 = 74.014$ and $c_2 = 74.026$

can also construct a Phase II control chart to monitor the process sequentially. The details can be found in Wu (2018).

6 Summary

Given the configuration of a sequence of multistate trials, conditional distributions of runs and patterns of the sequence can be viewed as distributions of runs and patterns in a random permutation. Thus, the conditional distributions are independent of the parameters. We derive exact conditional waiting time distributions of runs and patterns using the FMCI technique. The fact that the conditional distribution is independent of the parameters leads to applications like distribution-free tests and distribution-free control charts. Distribution-free tests for randomness and symmetry can be constructed with proper choice of runs and patterns. Some typical choices include number of runs, the longest run, scan statistics, number of rises and successions and order-preserving patterns. Distribution-free control charts have recently received considerable attention in SPC. Limited work has been done with runs and patterns-type control charts due to the lack of a unified approach to handle conditional distributions of runs and patterns. Our work fills this gap and provides a tool to design new distribution-free control charts.

References

Chakraborti, S., Human, S. W., Graham, M. A. (2011). Nonparametric (distribution free) quality control charts. In N. Balakrishnan (Ed.), *Handbook of methods and applications of statistics*: *Engineering, quality control and physical sciences* (pp. 298–329). New York: Wiley.

Chang, Y.-M., Huang, T.-H. (2010). Reliability of a 2-dimensional k-within-consecutive-rX s-out-of-m X n : F system using finite Markov chains. *IEEE Transactions on Reliability*, *59*, 725–733.

Cohen, J. P., Menjoge, S. S. (1988). One-sample run tests of symmetry. *Journal of Statistical Planning and Inference*, *18*, 93–100.

Foster, F. G., Stuart, A. (1954). Distribution-free tests in time-series based on the breaking of records. *Journal of the Royal Statistical Society-Series B Methodological*, *16*, 1–13.

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

Fu, J. C., Koutras, M. V. (1994). Distribution theory of runs: A Markov chain approach. *Journal of the American Statistical Association*, *89*, 1050–1058.

Fu, J. C., Lou, W. Y. W. (2003). *Distribution theory of runs and patterns and its applications—a finite Markov chain imbedding approach*. River Edge: World Scientific Publishing Co., Inc.

Fu, J. C., Wu, T.-L. (2016). Boundary crossing probabilities for high-dimensional Brownian motion. *Journal of Applied Probability*, *53*, 543–553.

Fu, J. C., Spiring, F. A., Xie, H. (2002). On the average run lengths of quality control schemes using a Markov chain approach. *Statistics & Probability Letters*, *56*, 369–380.

Fu, J. C., Wu, T.-L., Lou, W. Y. W. (2012). Continuous, discrete, and conditional scan statistics. *Journal of Applied Probability*, *49*, 199–209.

Gastwirth, J. L. (1971). On the sign test for symmetry. *Journal of the American Statistical Association*, *66*, 821–823.

Karwe, V. V., Naus, J. I. (1997). New recursive methods for scan statistic probabilities. *Computational Statistics & Data Analysis*, *23*, 389–402.

Koutras, M. V., Bersimis, S., Maravelakis, P. E. (2007). Statistical process control using shewhart control charts with supplementary runs rules. *Methodology and Computing in Applied Probability*, *9*, 207–224.

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

Lou, W. Y. W. (1996). On runs and longest run tests: A method of finite Markov chain imbedding. *Journal of the American Statistical Association*, *91*, 1595–1601.

Lou, W. Y. W., Fu, J. C. (2007). On exact type I and type II errors of Cochran's test. *Statistics & Probability Letters*, *77*, 1282–1287.

McWilliams, T. P. (1990). A distribution-free test for symmetry based on a runs statistic. *Journal of the American Statistical Association*, *85*, 1130–1133.

Montgomery, D. C. (2009). *Introduction to statistical quality control*. New York: Wiley.

Randles, R. H., Fligner, M. A., Policello, G. E, I. I., Wolfe, D. A. (1980). An asymptotically distribution-free test for symmetry versus asymmetry. *Journal of the American Statistical Association*, *75*, 168–172.

Shmueli, G., Cohen, A. (2003). Run-length distribution for control charts with runs and scans rules. *Communications in Statistics-Theory and Methods*, *79*, 122–128.

Song, X., Willett, P., Glaz, J., Zhou, S. (2012). Distributed detection with a scan statistic: Global to local inference. In *2012 IEEE 7th sensor array and multichannel signal processing workshop (SAM)* (pp. 485–488).

Wald, A., Wolfowitz, J. (1940). On a test whether two samples are from the same population. *Annals of Mathematical Statistics*, *11*, 147–162.

Western Electric Company. (1956). *Statistical quality control handbook*. Indianapolis: Western Electric Corporation.

Woodall, W. H. (2006). The use of control charts in health-care and public-health surveillance. *Journal of Quality Technology*, *38*, 89–104.

Wu, T.-L. (2018). Distribution-free runs-based control charts. arXiv e-prints arXiv:1801.06532.