2-2014

Numerical Procedures for Calculating the Probabilities of Recurrent Runs

Mark P. Silverman
Trinity College, mark.silverman@trincoll.edu

Open Journal of Statistics, **2014, 4, 122-131**
Published Online February 2014 (http://www.scirp.org/journal/ojs)
http://dx.doi.org/10.4236/ojs.2014.42014

Numerical Procedures for Calculating the Probabilities of Recurrent Runs

M. P. Silverman
Department of Physics, Trinity College, Hartford, USA
Email: mark.silverman@trincoll.edu

Received October 22, 2013; revised November 22, 2013; accepted November 30, 2013

ABSTRACT

Run count statistics serve a central role in tests of non-randomness of stochastic processes of interest to a wide range of disciplines within the physical sciences, social sciences, business and finance, and other endeavors involving intrinsic uncertainty. To carry out such tests, it is often necessary to calculate two kinds of run count probabilities: 1) the probability that a certain number of trials results in a specified multiple occurrence of an event, or 2) the probability that a specified number of occurrences of an event take place within a fixed number of trials. The use of appropriate generating functions provides a systematic procedure for obtaining the distribution functions of these probabilities. This paper examines relationships among the generating functions applicable to recurrent runs and discusses methods, employing symbolic mathematical software, for implementing numerical extraction of probabilities. In addition, the asymptotic form of the cumulative distribution function is derived, which allows accurate runs statistics to be obtained for sequences of trials so large that computation times for extraction of this information from the generating functions could be impractically long.

KEYWORDS

Recurrent Events; Theory of Runs; Time Series Analysis; Generating Functions; Probability Distributions

1. Introduction

1.1. Runs Tests for Non-Randomness

A stochastic process generates random outcomes in time or space. Such processes occur widely in the physical and social sciences, as well as in purely practical human activities such as finance, manufacturing, and commerce. Despite their random occurrence—indeed, precisely because of it—the outcomes of a stochastic process will display ordered patterns which a statistically naïve observer may mistakenly interpret as predictively useful information. In recent years, controversial issues over the information content of time series have arisen in a variety of disciplines such as nuclear physics [1] and econophysics (*i.e.* dynamics of the stock market) [2]. Although it is not possible to prove with certainty that a particular process is random, there are various statistical tests to demonstrate within specified confidence limits that it is not random. Among these, nonparametric runs tests are especially useful, in part because of their ease of implementation and statistical power [3].

1.2. Exclusive Runs

An exclusive run is an unbroken sequence of similar events, ordinarily of a binary nature. For example, a sequence of symbols *aabbbaa* comprises 2 runs of *a*'s of length 2 and 1 run of *b*'s of length 3. If the events *a* and *b* occur with fixed probabilities throughout the sequence, the stochastic process is of the Bernoulli kind, and the distribution theory of binary runs [4] can be used to test for non-randomness in permutational ordering of any such empirical sequence of outcomes.

It is not necessary for a stochastic process to generate binary events in order to be analyzed for runs. For example, a sequence of n different observations $x_1, x_2, \cdots, x_n$ of a continuous random variable will yield $n-1$ sequential differences that are either positive (+) or negative (–) and therefore again subject to binary runs analysis. The binary elements, however, are not Bernoulli variates since the probability of obtaining an element decreases with its position in the unbroken sequence. Nevertheless, one can test for non-randomness in permutational ordering with a different distribution theory [5].

Although developed initially for testing quality control in manufacturing, exclusive runs and up-down runs have been employed in analysis of a variety of experiments to test the fundamental prediction of quantum mechanics that transitions between quantum states occur randomly [6,7]. A problematic issue in the counting of exclusive or up-down runs is that the length of a run can be changed by future events. Thus, in the succession *aabbbaa*, the second run of 2 *a* could change to a run of 3 *a* or 4 *a* if the next two trials resulted in *ab* or *aa* respectively.

1.3. Runs of Recurrent Events

A third kind of runs analysis, based on Feller's theory of recurrent events [8], was recently employed to examine certain quantum optical processes for evidence of non-random behavior [9]. A recurrent run of length *t*, as defined by Feller, is a sequence of non-overlapping, uninterrupted successions of exactly *t* elements of the same kind. It is distinguished from the other two kinds of runs in that the concept of run length is so defined as to be independent of subsequent trials. For example, in the sequence *aaaabaaaaaa*, there are two runs of length 4 $[aaaa\,|\,b\,|\,aaaa\,|\,aa]$, three runs of length 3 $[aaa\,|\,ab\,|\,aaa\,|\,aaa]$, and five runs of length 2 $[aa\,|\,aa\,|\,b\,|\,aa\,|\,aa\,|\,aa]$. (Analyzed in terms of exclusive runs, there would have been 1 run of *a* of length 4 and 1 run of *a* of length 6, provided the sequence ended at the 11th trial). In a sequence of Bernoulli trials, a recurrent run of length *t* occurs at the n^{th} trial if the n^{th} trial adds a new run to the sequence. Thus, the recurrent runs of length 4 occur at positions 4, 9, and the recurrent runs of length 3 occur at positions 3, 8, 11.

The advantage of Feller's definition is that runs of a fixed length become recurrent events, and the statistical theory of recurrent events can then be applied to testing empirical data sequences for permutational invariance over a much wider variety of patterns than just those of unbroken sequences of identical binary elements. For example, one may be interested in testing the recurrence of a pattern *abab*, which, in a quantum optics experiment, might correspond to a sequence of alternate detections of left and right circularly polarized photons, or, in a series of stock price variations, might correspond to a sequence of alternative observations of rising and falling closing prices. Besides the application to runs, the same theoretical foundation may be applied to recurrent events in other forms such as return-to-origin problems, ladder-point problems (instances where a sum of random variables exceeds all preceding sums), and waiting-time problems.

The theory of recurrent runs, the relevant parts of which are examined in the following section, leads to generating functions from which the probability of a run of defined events of specified length can in principle be calculated exactly. As a practical matter, the extraction of these probabilities requires geometrically longer computation times with increasing sequence length. The availability of fast lap-top computers with large random access memory and of symbolic mathematical software of hitherto unparalleled ability to execute series expansions and perform differentiation and integration provides the analyst with computational power unimaginable to the creators of the statistical theory of runs. I report here mathematical strategies for reducing significantly the computation time for the probability of the widely applicable case of *k* occurrences of runs of length *t* in a Bernoulli sequence of length *n*.

2. Theory and Implementation of Recurrent Runs

2.1. Probability Generating Functions

Following Feller, I define the recurrent event E to be a run of successes of length *t* in a sequence of Bermoulli trials with *p* the probability of a single successful outcome and $q = 1 - p$ the probability of failure. Consider the following random variables:

$$T_k = \left[\text{number of trials ("waiting time") between } (k-1)^{th} \text{ and } k^{th} \text{ occurrence of E}\right] + 1 \tag{1}$$

$$S_k = \sum_{j=1}^{k} T_j = \text{number of trials up to and including } k^{th} \text{ occurrence of E} \tag{2}$$

$$N_n = \text{number of occurrences of E in } n \text{ trials}. \tag{3}$$

The distribution of the variable T is defined by

$$\Pr(T=n) \equiv f_n \text{ with } f_0 = 0 \tag{4}$$

where f_n is the probability that E occurs for the first time at the n^{th} trial. The generating function of the probabilities of first occurrence is expressed by

$$F(s) = \sum_{n=0}^{\infty} f_n s^n . \tag{5}$$

The number of trials to the k^{th} occurrence of E is then characterized by the random variable S_k in (2), which is a sum of the waiting times of k independent trials, from which it is follows that the associated generating function takes the form

$$F^{(k)}(s) = \sum_{n=0}^{\infty} f_n^{(k)} s^n = \left[F(s)\right]^k = \left(\sum_{n=0}^{\infty} f_n s^n\right)^k \tag{6}$$

where

$$\Pr(S_k = n) \equiv f_n^{(k)} \tag{7}$$

is the probability that the k^{th} occurrence of E first takes place at the n^{th} trial.

I leave to the cited literature the proof that the generating function (5) for runs of length t with individual probability of success p is given by

$$F(s,p,t) \equiv F(s) = \frac{p^t s^t (1-ps)}{1-s+qp^t s^{t+1}} \tag{8}$$

from which the mean and standard deviation of the recurrence times follow by differentiation

$$\mu(p,t) = \left.\frac{\mathrm{d}F(s)}{\mathrm{d}s}\right|_{s=1} = \frac{1-p^t}{qp^t} \quad (q = 1-p) \tag{9}$$

$$\sigma(p,t) = \sqrt{\left[\frac{\mathrm{d}^2F(s)}{\mathrm{d}s^2} - \left(\frac{\mathrm{d}F(s)}{\mathrm{d}s}\right)^2 + \frac{\mathrm{d}F(s)}{\mathrm{d}s}\right]_{s=1}} = \left(\left(qp^t\right)^{-2} - (2t+1)\left(qp^t\right)^{-1} - pq^{-2}\right)^{1/2} . \tag{10}$$

For economy of expression, the parameters p, t will be suppressed in the arguments of $F(s,p,t)$, $\mu(p,t)$, and $\sigma(p,t)$ unless needed to avoid ambiguity. In general, these parameters will be chosen and fixed at the outset of any illustrative applications. Note, too, that to obtain a statistical moment from a probability generating function (pgf), the derivatives are evaluated at $s=1$, which leads to a sum of terms, whereas to obtain a probability the derivatives are evaluated at $s=0$, which leads to a single term.

For many applications the analyst's interest is not necessarily in the recurrence time (*i.e.* number of trials) to the k^{th} occurrence of E, but in the probability that E occurs k times in a fixed number n of trials. The relation connecting the two variates is

$$\Pr(N_n \geq k) = \Pr(S_k \leq n) . \tag{11}$$

The probability $p_{n,k}$ that k events E occur in n trials is then expressible as

$$p_{n,k} = \Pr(N_n = k) = \Pr(S_k \leq n) - \Pr(S_{k+1} \leq n) \tag{12}$$

and serves in the construction of two pgf's

$$G(z,n) = \sum_{k=0}^{\infty} p_{n,k} z^k \tag{13}$$

and

$$Q(s,k) = \sum_{n=1}^{\infty} p_{n,k} s^n = \frac{F(s)^k \left[1-F(s)\right]}{1-s} . \tag{14}$$

Note that the summation in (13) is over the number of occurrences, whereas the summation in (14) is over the number of trials. The second equality in (14) follows directly from Equation (11). Multiplying both sides of (13) by s^n and summing over *n* leads to the bivariate generating function

$$H(s,z) \equiv \sum_{n=1}^{\infty}\left[\sum_{k=0}^{\infty} p_{n,k} z^k\right] s^n = \frac{1-F(s)}{(1-s)(1-zF(s))} \tag{15}$$

from which the probabilities $p_{n,k}$ are calculable by series expansion of both sides of the equality.

A sense of the structure of the formalism can be obtained by considering the case of recurrent runs of length $t=3$ for a stochastic process with $p=\frac{1}{2}$. Substitution of these conditions into Equation (8) for $F(s)$ yields the following rational expression for the right side of Equation (15) and its corresponding Taylor-series expansion

$$\begin{aligned} G(s,z) &= \frac{-2\left(s^2+2s+4\right)}{\left(s^3+zs^3+2s^2+4s-8\right)} \\ &= 1+s+s^2+\left(\frac{1}{8}z+\frac{7}{8}\right)s^3+\left(\frac{3}{16}z+\frac{13}{16}\right)s^4+\left(\frac{1}{4}z+\frac{3}{4}\right)s^5+O\left(s^6\right) \end{aligned} \tag{16}$$

to order s^6. Recall that the powers of *s* designate the number of trials, and the powers of *z* designate the number of recurrences of runs of length 3. For a fixed power of *s*, the sum of the coefficients of the powers of *z* within each bracketed expression equal unity, as they must by the completeness relation for the probability of mutually exclusive outcomes. Note that the first three terms $\left(s^0+s^1+s^2\right)$ are independent of *z*—*i.e.* contain only powers z^0—since there cannot be runs of length 3 in a sequence of no more than 2 trials. For 3 trials, the probability of 0 runs of length 3 is 7/8 and the probability of 1 run of length 3 is 1/8. For 5 trials, however, the probability of 0 runs is 1/4 and the probability of 1 run is 3/4. This pattern persists: (a) to obtain a run of length *t*, the sequence of trials must be of length $s \geq t$, and (b) the greater the number of trials, the higher is the probability of obtaining longer runs.

2.2. Moment Generating Functions

It is not necessary to know the individual $p_{n,k}$ to determine the mean number of recurrent runs

$$\langle N_n \rangle = \sum_{k-0}^{\infty} k p_{n,k} \tag{17}$$

which, for many applications in the physical sciences and elsewhere, is the experimentally observed quantity of interest. Multiplying both sides of Equation (17) by s^n and summing *n* over the range $(1,\infty)$ leads to the generating function for the distribution of $\langle N_n \rangle$

$$M_1(s) = \sum_{n=1}^{\infty} \langle N_n \rangle s^n = \frac{F(s)}{(1-s)(1-F(s))}. \tag{18}$$

Starting from the relation

$$\left\langle N_n^2 \right\rangle = \sum_{k-0}^{\infty} k^2 p_{n,k} \tag{19}$$

and following the same procedure that led to (18) yields the generating function for the distribution $\left\langle N_n^2 \right\rangle$

$$M_2(s) = \sum_{n=1}^{\infty} \left\langle N_n^2 \right\rangle s^n = \frac{F(s)+F(s)^2}{(1-s)\left(1-F(s)\right)^2}. \tag{20}$$

From the generating functions (18) and (20) one can deduce the asymptotic relations for mean and variance

$$\langle N_n \rangle \approx n\mu^{-1} \tag{21}$$

and

$$\mathrm{var}(N_n) \approx n\sigma^2\mu^{-3}. \tag{22}$$

2.3. Numerical Procedures

The statistics (probabilities and expectation values) for any physically meaningful choice of probability of success p, run length t, and number of trials n are deducible exactly from expressions (15) and (18) in the manner previously illustrated. For many applications, however, particularly where it is possible to accumulate long sequences of data as is often the case in atomic, nuclear and elementary particle physics experiments or investigations of stock market time series, the tests for evidence of non-random behavior are best made by examining long runs. Suppose, for example, one wanted the probability of obtaining the number of occurrences of runs of length 50 in a sequence of 100 trials. One approach, leading directly to all non-vanishing probabilities, would be to extract the 100th term

$$p_{100,50} = \frac{1}{562949953421312}\left[\left(\frac{1125899906842623}{2251799813685248} + \frac{1}{2251799813685248}z\right)(z+1)\right]$$
$$\approx 1.0000 + 2.3093\times 10^{-14} z + 7.8886\times 10^{-31} z^2$$

from the series expansion of the bivariate generating function (15). Powerful symbolic mathematical software such as *Maple* or *Mathematica* permits one to do this up to a certain order limited by the speed and memory of one's computer, but these calculational tools may become insufficient when one is seeking exact probabilities of runs in data sequences of thousands to millions of bits.

Using complex variable theory, one can extract expressions for $p_{n,k}$ and $\langle N_n \rangle$ by evaluation of contour integrals

$$\begin{aligned} p_{n,k} &= \frac{1}{2\pi \mathrm{i}} \oint_C P_n(\xi)\xi^{-(k+1)}\mathrm{d}\xi \\ &= \left(\frac{1}{2\pi \mathrm{i}}\right)^2 \oint_C \mathrm{d}\zeta \frac{1-F(\zeta)}{(1-\zeta)\zeta^{n+1}} \oint_C \frac{\mathrm{d}\xi}{(1-\xi F(\zeta))\xi^{k+1}} \\ &= \frac{1}{n!}\left\{\frac{\mathrm{d}^n}{\mathrm{d}s^n}\left[(1-s)^{-1}(1-F(s))F(s)^k\right]\right\}\Bigg|_{s=0} \end{aligned} \tag{23}$$

and similarly

$$\langle N_n \rangle = \frac{1}{2\pi \mathrm{i}} \oint_C M_1(\xi)\xi^{-(k+1)}\mathrm{d}\xi = \frac{1}{n!}\left\{\frac{\mathrm{d}^n}{\mathrm{d}s^n}\left[(1-s)^{-1}(1-F(s))^{-1}F(s)\right]\right\}\Bigg|_{s=0} \tag{24}$$

where C is the unit circle and the generating function $F(s)$, given by Equation (8), specifies the single-event probability p and run length t. Contrary to first impression, however, the execution of expressions (23) or (24) for $p_{n,k}$ or $\langle N_n \rangle$ directly by differentiation, instead of by a series expansion of the corresponding generators up to the order that yields the desired $p_{n,k}$ or $\langle N_n \rangle$, is not computationally economic. The computer, in fact, executes the series expansion of the generator much more rapidly than it performs symbolic differentiation.

Because the generator $M_1(s)$ [Equation (18)] is univariate, one can take advantage of the rapidity with which symbolic mathematical software executes a series expansion to obtain exact values $\langle N_n \rangle$ for very long sequence lengths by the following simple procedure:

1) For given p and t, express $M_1(s)$ as a rational function $g(s)/h(s)$ of s.

2) Convert $g(s)/h(s)$ to a power series $\sum_{j=0}^{\infty} \langle N_j \rangle s^j$. In *Maple* this can be done by the command

$convert(\cdots, FormalPowerSeries, s)$ where the ellipsis $(\cdots)$ represents either the expression to be converted or the *Maple* equation number of that expression.

3) Extract the single desired term $\langle N_n \rangle$. In *Maple* this can be done by filtering the sum: $\langle N_n \rangle = \sum_{j=n}^{n} \langle N_j \rangle s^j$.

Alternatively, one can convert the series generated to order $O(n+1)$ to a polynomial of degree n, and use the command *lcoeff* to extract the leading coefficient of the polynomial.

As an illustration, consider the calculation (by means of *Maple*) of the exact mean number of runs of length $t = 4$ in a sequence of 1 million trials with probability of success $p = 0.5$. Following the foregoing steps, we have

1) $M_1(s)=\dfrac{F(s)}{(1-s)(1-F(s))}=\dfrac{1}{2}\dfrac{s^4}{(s^4+s^3+2s^2+4s-8)(s-1)}$

2) $convert(M_1(s),FormalPowerSeries,s)\rightarrow\sum_{j=0}^{\infty}(\cdots)_j\,s^j$

3) $evalf\left(expand\left(\sum_{j=n}^{n}(\cdots)_j\,s^j\right)\right)\rightarrow N_n s^n$

where $(\cdots)_j$ is to be replaced by the actual j^{th} numerical element in the sum obtained in step (2). The arrow, inserted by the author, symbolically points to the form of the output. Extraction of this element for the specified conditions led to $\langle N_{1,000,000}\rangle=33{,}333.258$ in a fraction of a second. In *Maple*, the command *evalf* calls for numerical evaluation of expressions; omitting this command results in an exact fraction, which for a sequence length of 1 million is too unwieldy to be useful. Use of the command *lcoeff* yields the same result, but executes more slowly for large *n*.

The procedure described above for converting the rational expression of *s* into a formal power series in *s* did not work with the bivariate generator $H(s,z)$, which involved a product of *z* with a power of *s* (depending on run length *t*) in the denominator, and required, for the conversion, solution of the roots of a high-order (>2) algebraic equation. *Maple* did return, however, the recursion relation for the coefficients of the formal power series. An alternative procedure to isolate the values $p_{n,k}$ for fixed *n*, which still relied on the computational speed of series expansion and worked well for sequence lengths in the thousands, is the following:

1) For given *p* and *t*, express $H(s,z)$ as a rational function of *s* and *z*.
2) Generate a series expansion of $H(s,z)$ to order *n* and $n+1$.
3) Convert the series expansions into polynomials $P_{(n+1)}$ and $P_{(n)}$.
4) Subtract one polynomial from the other to obtain an expression of the form

$$P_{(n+1)}-P_{(n)}\rightarrow\left[p_{n,0}+p_{n,1}z+p_{n,2}z^2+\cdots+p_{n,\left\lfloor\frac{n}{t}\right\rfloor}z^{\left\lfloor\frac{n}{t}\right\rfloor}\right]s^n$$

where $\left\lfloor\dfrac{n}{t}\right\rfloor$ is the largest integer *k* such that $kt\le n$. The coefficients $p_{n,k}$ are given as exact fractions.

5) Evaluate the set $\{p_{n,k}\}$ as floating-point numbers, if desired.

As an example, the procedure led in under 10 seconds to the full set $p_{1000,k}\{k=0,\cdots,200\}$ for the probability of *k* occurrences of runs of length 4 in a sequence of 1000 trials. The calculations described in this section were performed with a laptop computer (Intel-based Mac Powerbook) running *Maple* 16.

One final procedure, particularly suitable when only selected probabilities of the full set $\{p_{n,k}\}$ are desired, is to obtain these probabilities directly from the generating function $Q(s,k)$ in (14). As in the previous examples, this can be accomplished in either of two ways: (1) by evaluating the leading coefficient of the polynomial in the desired degree *n*, or (2) by filtering the power series for the n^{th} term. As an example, consider the probability $p_{100,4}$ for obtaining 4 runs of length $t=4$ in a series of 100 trials with probability of success $p=0.5$. The generating function then takes the form

$$Q(s,k,p,t)=Q(s,4,0.5,4)=-\frac{2s^{16}(s^3+2s^2+4s+8)}{(s^4+2s^3+4s^2+8s-16)^5}$$

In *Maple*, method (1) proceeds as follows:

1) $convert\left(-\dfrac{2s^{16}(s^3+2s^2+4s+8)}{(s^4+2s^3+4s^2+8s-16)^5},FormalPowerSeries,s\right)$

2) $evalf\left(expand\left(\sum_{k=100}^{100}(\cdots)\right)\right)\rightarrow 0.2007906348s^{100}$

where the ellipsis is to be replaced by the k^{th} term produced in step (1). In using *Maple* to execute method (2), one proceeds in a single step once the rational expression $Q(s,4,0.5,4)$ has been obtained:

$$evalf\left(lcoeff\left(convert\left(series\left(Q(s,4,0.5,4),s=0,101\right),polynom\right)\right)\right)\rightarrow 0.2007906348\,.$$

Note that the series must be expanded to $n+1$ terms in order to obtain the coefficient of s^n, since the summation index begins at 0.

3. Generating Function of Cumulative Probability

For many applications in the physical sciences and elsewhere, the full set of probabilities $\{p_{n,k}\}$ provides more information than is desirable or usable. Moreover, because $p_{n,k}$ for large *k* may be very small and the variance relatively large, the more observationally stable statistic is the probability of obtaining *k or more* occurrences of the specified event, or in other words, the complementary cumulative probability (ccp)

$$\Pr(N_n \geq k) \equiv \tilde{P}_{nk} = \sum_{j=k}^{\infty} p_{nj} \tag{25}$$

introduced in Equation (11). Experimental situations calling for preferential usage of a cumulative probability distribution over a probability function abound in the physical sciences, as, for example, in the analysis of fragmentation [10] and other stochastic processes leading to a power-law distribution.

The generating function for the ccp is derivable from $Q(s,k)$ by summing over the recurrence index *k*

$$C(s,k) \equiv \sum_{n=1}^{\infty} \tilde{P}_{nk} s^n = \sum_{n=1}^{\infty} \left[\sum_{j=k}^{\infty} Q(s,j) \right] s^n = \frac{F(s)^k}{1-s}. \tag{26}$$

(It is understood that $F(s)$, and therefore $C(s,k)$, are also functions of *p* and *t*). Thus, one can calculate $\Pr(N_n \geq k)$ directly from the generator (26) by use of the methods previously described. A modern laptop running *Maple* can return results for $\Pr(N_n \geq k)$ within minutes for sequence lengths *n* on the order of many 1000's.

4. Asymptotic Distributions

One can show by application of the Central Limit Theorem (CLT) to relation (11) that for sufficiently large number of trials *n* and number of occurrences *k*, the number N_n of runs of length *t* produced in *n* trials is approximately normally distributed with mean and variance given by relations (21) and (22). The approximation, whose relative accuracy improves in the limit of increasing *n*, is actually quite good even for modest values of *n*, as shown in **Table 1** for $n = 100$. Expansion of the generating function $M_1(s)$ yielded the exact mean value as an integer or fraction, which was then expressed as a floating-point number to three significant figures for comparison with the Gaussian approximation. It is to be noted from the Table that the Gaussian approximation overstates the mean values, and that the absolute error $\left[p_{n,k}^{\text{Gauss}} - p_{n,k}^{\text{Exact}} \right]$, in contrast to the relative error $\left[\left(p_{n,k}^{\text{Gauss}} - p_{n,k}^{\text{Exact}} \right) \Big/ p_{n,k}^{\text{Exact}} \right]$ increases with run length and number of trials.

The Gaussian approximation, however, does not work well in estimating the cumulative probability $\tilde{P}_{nk}$ for large *n* and long runs. However, one can obtain substantial improvement by taking account of the asymptotic behavior of $\mu(p,t)$ and $\sigma(p,t)$, and the equivalence in Equation (11). Expansion in probability *p* for fixed run length *t* of expressions (9) and (10) yields the series

$$\left.\begin{matrix} \mu(p,t) \\ \sigma(p,t) \end{matrix}\right\} \rightarrow p^{-t} + p^{-t+1} + \cdots + p^{-1} + \begin{cases} O(p^0) \\ -c(t) + O(p) \end{cases} \tag{27}$$

in which the constant $c(t)$ is orders of magnitude smaller than the leading terms. Thus, in practical terms, $\mu(p,t)$ and $\sigma(p,t)$ are virtually equal for long runs, as illustrated in **Table 2** for $p = 0.5$.

Recall that $\mu(p,t)$ and $\sigma(p,t)$, which were derived from the generating function $F(s,p,t)$, are respectively the mean number of trials to the *first* occurrence of event E, which is a run of length *t*. The equivalence of the mean and standard deviation suggests that the asymptotic distribution of the random variable T in Equation (1) is exponential [11] $E(\lambda)$ with parameter $\lambda = \mu^{-1}$. Then the random variable S_k in Equation (2), which is a sum of *k* independent exponential random variables $\sum_{j=1}^{k} T_j$, follows a gamma distribution $Gam(\lambda,k)$ with cumulative probability function

$$\Pr(S_k \leq n) \approx \frac{1}{\Gamma(k)} \int_0^{n\mu^{-1}} x^{k-1} \mathrm{e}^{-x} \mathrm{d}x \tag{28}$$

Table 1. Comparison of exact and Gaussian mean numbers of runs for $n = 100$ trials with $p = 0.5$.

Run Length t	Mean Number of Runs (Exact)	Mean Number of Runs (Gaussian Approximation)
1	50	50
2	16.556	16.667
3	7.041	7.143
4	3.258	3.333
5	1.562	1.613
6	7.611 (−1)	7.937 (−1)
7	3.738 (−1)	3.937 (−1)
8	1.842 (−1)	1.961 (−1)
9	9.098 (−2)	9.785 (−2)
10	4.496 (−2)	4.888 (−2)
11	2.222 (−2)	2.443 (−2)
12	1.099 (−2)	1.221 (−2)
13	5.433 (−3)	6.104 (−3)
14	2.686 (−3)	3.052 (−3)
15	1.328 (−3)	1.526 (−3)
16	6.561 (−4)	7.630 (−4)
17	3.243 (−4)	3.815 (−4)
18	1.602 (−4)	1.907 (−4)
19	7.916 (−5)	9.537 (−5)
20	3.910 (−5)	4.768 (−5)
40	2.819 (−11)	4.547 (−11)
60	1.820 (−17)	4.337 (−17)

Table 2. Asymptotic mean and SD of S_1.

Run Length t	$\mu(0.5,t)$	$\sigma(0.5,t)$
1	2.00	1.41
5	62.00	58.22
10	2046.00	2037.47
20	2.09715×10^{6}	2.09713×10^{6}
30	2.14748×10^{9}	2.14748×10^{9}

where $\Gamma(k)$ is the standard gamma function equal to $(k-1)!$ for integer k.

By virtue of equivalence (11), Equation (28) also yields a closed-form asymptotic relation for the sought-for cumulative probability distribution $\Pr(N_n \geq k)$.

Consider, for example, the probability of 1 or more runs of length 20 in 2000 trials with individual probability of success 0.5. A comparison of the results of (a) Equation (28), (b) the cumulative Gaussian distribution with mean and variance given by relations (21) and (22), and (c) the exact calculation obtained from the generating function (26)

$$\Pr\left(N_{2000}\geq 1\right)\Big|_{\text{Gamma Dist}}=9.532\times 10^{-4}$$
$$\Pr\left(N_{2000}\geq 1\right)\Big|_{\text{Gaussan Dist}}=1.588\times 10^{-1}$$
$$\Pr\left(N_{2000}\geq 1\right)\Big|_{\text{Exact Generator}}=9.451\times 10^{-4}$$

supports the distribution (28). If the number of trials is increased to 10^6, the Gamma and Gaussian asymptotic distributions lead to comparable results

$$\Pr\left(N_{10^6}\geq 1\right)\Big|_{\text{Gamma Dist}}=0.3793$$
$$\Pr\left(N_{10^6}\geq 1\right)\Big|_{\text{Gauss Dist}}=0.3004$$
$$\Pr\left(N_{10^6}\geq 1\right)\Big|_{\text{Exact Generator}}=0.3793$$

but the former is still superior to the latter when compared to the probability calculated from the exact generating function.

The relation (28), by which one can calculate the cumulative probability $\Pr\left(N_n\geq k\right)$ for large values of n, also allows one to calculate the individual probabilities p_{nk} as a function of number of occurrences k through the identity

$$p_{nk}=P\left(N_n\geq k\right)-P\left(N_n\geq k+1\right)\approx\frac{1}{\Gamma\left(k\right)}\int_0^{n\mu^{-1}}x^{k-1}\left(1-\frac{x}{k}\right)e^{-x}\mathrm{d}x \tag{29}$$

5. Conclusions

The theory of recurrent runs provides a statistical basis for rejecting the hypothesis that a series of observations (in time or space) are random. This is a matter that often arises in experimental investigations in atomic, optical, nuclear, and elementary particle physics, as well as in other sciences, finance, and commerce, which may entail a very large number—perhaps in the thousands to millions—of trials or observations.

In this paper theoretical and numerical methods based on different generating functions were derived and investigated to determine (a) the probability p_{nk} for k recurrence runs of length t in n Bernoulli trials, (b) the complementary cumulative probability $\tilde{P}_{nk}=\Pr\left(N_n\geq k\right)$, and (c) the mean number of runs $\left\langle N_n\right\rangle$.

The methods reported here can be implemented on modern laptop computers running commercially available symbolic mathematical software, such as *Maple* (which was the application used by the author). Computation times for application of these methods to data sequences up to millions of trials could range from seconds to minutes.

To compute runs statistics for sequences of intermediate to very long trial numbers, the asymptotic distribution for the number of trials up to and including the k^{th} occurrence $\left(\lfloor n/t\rfloor\geq k\geq 1\right)$ of a specified run length t was derived and found to be a Gamma distribution $Gam\left(\mu\left(p,t\right)^{-1},k\right)$ to excellent approximation. In the limit of very large (technically, infinite) n and k, the Central Limit Theorem (CLT) predicts an asymptotic distribution $\Pr\left(N_n\geq k\right)$ of Gaussian form. Both the Gamma and Gaussian asymptotic distributions give comparable results for $\Pr\left(N_n\geq k\right)$ under these circumstances, but the Gaussian approximation is less accurate and fails entirely for values of n and k for which the CLT does not apply.

REFERENCES

[1] M. P. Silverman and W. Strange, "Search for Correlated Fluctuations in the β^+ Decay of Na-22," *Europhysics Letters*, Vol. 87, No. 3, 2009, pp. 32001-32006. http://dx.doi.org/10.1209/0295-5075/87/32001

[2] D. Sornette, "Critical Market Crashes," *Physics Reports*, Vol. 378, No. 1, 2003, pp. 1-98. http://dx.doi.org/10.1016/S0370-1573(02)00634-8

[3] J. Wolfowitz, "On the Theory of Runs with Some Applications to Quality Control," *The Annals of Mathematical Statistics*, Vol. 14, No. 3, 1943, pp. 380-288. http://dx.doi.org/10.1214/aoms/1177731421

[4] A. M. Mood, "The Distribution Theory of Runs," *The Annals of Mathematical Statistics*, Vol. 11, No. 4, 1940, pp. 367-392. http://dx.doi.org/10.1214/aoms/1177731825

[5] H. Levene and J. Wolfowitz, "The Covariance Matrix of Runs Up and Down," *The Annals of Mathematical Statistics*, Vol. 15,

No. 1, 1944, pp. 58-69. http://dx.doi.org/10.1214/aoms/1177731314

[6] M. P. Silverman, W. Strange, C. Silverman and T. C. Lipscombe, "Tests for Randomness of Spontaneous Quantum Decay," *Physical Review A*, Vol. 61, No. 4, 2000, Article ID: 042106. http://dx.doi.org/10.1103/PhysRevA.61.042106

[7] M. P. Silverman and W. Strange, "Experimental Tests for Randomness of Quantum Decay Examined as a Markov Process," *Physics Letters A*, Vol. 272, No. 1-2, 2000, pp. 1-9. http://dx.doi.org/10.1016/S0375-9601(00)00374-1

[8] W. Feller, "Fluctuation Theory of Recurrent Events," *Transactions of the American Mathematical Society*, Vol. 67, 1949, pp. 98-119. http://dx.doi.org/10.1090/S0002-9947-1949-0032114-7

[9] D. Branning, A. Katcher, W. Strange and M. P. Silverman, "Search for Patterns in Sequences of Single-Photon Polarization Measurements," *Journal of the Optical Society of America B*, Vol. 28, No. 6, 2011, pp. 1423-1430. http://dx.doi.org/10.1364/JOSAB.28.001423

[10] M. P. Silverman, W. Strange, J. Bower and L. Ikejimba, "Fragmentation of Explosively Metastable Glass," *Physica Scripta*, Vol. 85, 2012, Article ID: 065403. http://dx.doi.org/10.1088/0031-8949/85/06/065403

[11] A. M. Mood, F. A. Graybill and D. C. Boes, "Introduction to the Theory of Statistics," 3rd Edition, McGraw-Hill, New York, 1974, pp. 540-541.