Methodology and Computing in Applied Probability, 4, 409–419, 2002

System-Wide Probabilities for Systems with Runs and Scans Rules

GALIT SHMUELI gshmueli@rhsmith.umd.edu
University of Maryland, College Park, MD 20742

Received April 23, 2002; Accepted December 18, 2002

Abstract. Systems that include multiple decision rules are encountered in many fields. We focus on systems with several states or levels where the rules determining when and how to move between the levels are based on runs or scans. Our target is to evaluate the proportion of time spent at each level. This information is valuable since it is required for constructing various cost functions. In this paper we address two questions: How to incorporate the system of levels and decision rules into a probabilistic setting and which measures to use in order to evaluate the proportion of time spent at each level. We take a waiting-time approach, and use special features of runs and scans distributions in order to investigate long and short-term measures of the time spent at the different levels of a system. We introduce a short term per-cycle measure and compare it with the finite-time measure. We find that the two measures can differ significantly.

Keywords: runs and scans, short-term ratio, generating function, random number generation

1. Introduction

In this paper we deal with systems that have several levels or states, and which are governed by switching rules that are based on runs and scans. Two examples of such systems are continuous sampling plans (Dodge, 1943) and the Military Standard 105E acceptance sampling scheme (and its civilian counterpart ANSI/ASQC Z1.4). These systems include several levels of sampling, and switching between the different levels is governed by rules of the type ''switch following 10 consecutive accepted batches''.

When such systems of decisions are used in practice, the proportion of time spent at each level is an important feature which is used to construct various cost functions. We thus investigate the distribution of the proportion of time spent at each level.

We take a waiting-time approach meaning that we treat the time spent at a certain level of the system until switching to another level as a random variable. When the system is circular, i.e., each level leads only to the next level, we can use existing methods from waiting-time distributions that are based on runs and scans. A more complicated situation arises for systems that are not circular. Here each level may lead to more than one other level, thereby creating a compound decision rule. For example: ''switch to level 1 if event A_1 occurs or to level 2 if event A_2 occurs''. The resulting waiting time is thus of the ''sooner or later'' form. We show how a waiting-time approach can be taken for circular and non-circular systems using generating functions in Section 2.

Section 3 describes several measures for the time spent at each level. One option is to

use long-term proportions, such as the ratio of the expected time spent at a certain level divided by the sum of expected times under all levels. This measure is used in several applications and is relatively easy to compute. However, since waiting time distributions that are based on runs and scans are usually very skewed, such a measure might be uninformative. This means that although in the long run the ratio gives the probability of being at a certain level, the short term picture might differ greatly.

A second option is to look at short-term measures. We consider two such measures: The first is the per-cycle ratio, which is the proportion of time spent at a level within a single ''cycle''. The second short-term measure is the proportion of time spent at a level within a finite time-frame. Although the two measures seem similar, we find that the distributions of per-cycle ratios are highly skewed, while the finite-time proportions approach a normal distribution.

In all cases we take a generating function approach to compute the waiting times, and then use special features of runs and scans distributions to generate waiting time realizations efficiently.

We illustrate our method in Section 4 for the continuous sampling plan CSP-1. We compute the long-term and two short-term measures and compare the information contained in each. Concluding remarks are given in Section 5.

2. Distribution of the Waiting Time at Each Level

2.1 Circular Systems

We first deal with systems that are circular, in the sense that each level leads to a single other level. Systems with two levels that are connected are in definition circular. Two examples for circular systems are the CSP-1 and CSP-2 continuous sampling plans (Dodge, 1943). Each of these systems include two levels of inspection called screening and sampling. The switching rule from screening to sampling is the same in both systems, and it is based on a run of length i. Switching from sampling back to screening is governed by a scan for CSP-2. Figure 1 describes the two systems with their levels and switching rules. These systems are discussed further in Section 4. In both cases the systems are circular.

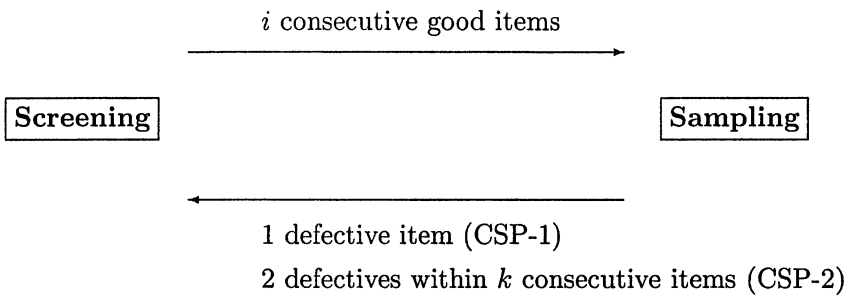

Figure 1. CSP-1 and CSP-2 circular systems for continuous sampling.

In the case of circular systems the waiting time distribution in each level is an ordinary runs or scans related distribution. The waiting time for a run of length k is known as an order k geometric variable, and that for at least k events within d consecutive trials is known as a consecutive k-within-d variable. The probability functions of such variables can be obtained by using methods such as Markov Chain imbedding (Fu and Koutras, 1994) or the generating function method (Aki, 1992; Chryssaphinou *et al.*, 1994; Feller, 1968; Shmueli and Cohen, 2000). Both methods will yield a recursive formula for the probability function, and the generating function method can also yield a non-recursive formula by using partial fraction expansion.

2.2. Non-Circular Systems

In non-circular systems the levels are not inter-connected in a one-way circle by the decision rules. This means that there is at least one level that can shift to two or more other levels with a positive probability. An example for such a system is the set of switching rules between three types of sampling according to the Military Standard 105E scheme. As illustrated in Figure 2, there are three inspection levels: normal, reduced, and tightened inspection. The rules that determine when to shift from one level to the other are based on runs and scans. This system is not circular since it is possible to move from normal inspection to either reduced or tightened inspection. In comparison, the continuous sampling schemes that were described in the previous section were circular, since each state could only be followed by one other state.

We take the same approach as in circular systems, using the generating functions of the waiting time distributions. The only modification that is necessary is to treat the waiting time from a level that leads to more than one other level (e.g., the normal inspection level in Military Standard 105E) as a competition between two variables, also known as a ''sooner or later'' problem. To illustrate this, consider the rules in the Military Standard 105E system:

1. The rule for switching from reduced to normal inspection involves a single rejected batch, and thus the waiting time is a geometric variable.

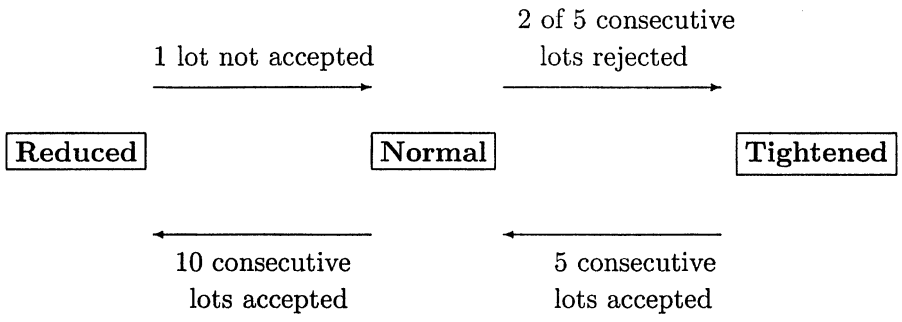

Figure 2. The Military Standard 105E system of rules for acceptance sampling.

2. The rule for switching from tightened to normal inspection involves a run of five accepted batches. The waiting time is thus an order five geometric variable.
3. The rules for switching from normal to reduced and tightened inspection involve a run and a scan. The waiting time in normal inspection until switching to either reduced or tightened inspection is thus a competition between an order 10 geometric variable and a consecutive-2-within-5 variable.

For the last rule, the generating function for the waiting time within normal inspection is given by

$$G_{\text{normal}}(s) = \frac{[1 - G_T(s)][1 - G_R(s)]}{1 - G_T(s)G_R(s)}, \tag{1}$$

where $G_T(s), G_R(s)$ are the generating functions corresponding to the switching rules from normal to tightened and from normal to reduced inspection, respectively. The first is the pgf of a 2-within-5 variable with parameter equal to the probability of rejecting a lot (under normal inspection), and the second is the pgf of a 10-order geometric with parameter equal to the probability of accepting a lot (under normal inspection). A recursive formula for the probability function that is obtained from (1) is given in Shmueli and Cohen (2000).

3. The Proportion of Time Spent at Each Level

We now tackle the problem of describing a system with runs- and scans-based switching rules by the proportion of time spent at each of the different levels. Denote the time spent at level i by X_i. This waiting time can be computed using any of the methods described in Section 2. In many applications it is required to know about the proportion of time spent at each of the system's levels. The most commonly used and published measures are ratios of waiting-time expectations of the form

$$\frac{E(X_i)}{\sum_j E(X_j)}, \tag{2}$$

which give the long-term proportion of time spent at level i. This is equivalent to computing the steady-state probabilities of a Markov chain. However, since the waiting time distributions of runs and scans variables tend to be very skewed, the ratio of expectations might differ significantly from the proportion of time spent at the system levels in the short run. We therefore investigate two short-term measures of the proportion of time spent at the different levels:

1. Per-cycle ratios of the form

$$p_i^{cyc} = \frac{X_i}{\sum_i X_i},$$

where X_i are independent runs/scans-related waiting times.

2. Short-run proportions of the form

$$p_i^N = \frac{\text{time spent at level } i \text{ until time } N}{N}.$$

The decision which short-term measure to use depends on factors such as the cost-function and the implementation of the system. From a cost-function perspective, if the major cost is switching between levels of the system then the per-cycle measure may be more suitable. If, on the other hand, costs of staying at the different levels are different, then the finite-time measure might be preferred. Another factor might be how long the system is left to run. Whether there is a time limit or a limit on the number of cycles would determine the adequacy of the short-term measure to be used.

In the previous section we derived the probability functions for the separate X_i variables and their sum. However, these probability functions are given either recursively or expressed as a function of roots of a polynomial, thus not enabling one to find the distribution of the per-cycle or short-run ratios directly. We suggest two solutions, one analytical and one computational for obtaining the probability functions of the above ratios.

3.1. A Theoretical Solution for Per-Cycle Proportions

From a theoretical point of view, the distribution of the ratios p_i^{cyc} can be computed by using conditional probabilities. Denoting $S = \sum_i X_i$,

$$P(p_i^{cyc} = t) = \sum_s P(p_i^{cyc} = t \,|\, S = s)P(S = s) \tag{3}$$

$$= \sum_s P(X_i = ts)P(S = s). \tag{4}$$

Although the distribution of X_i and S can be computed directly, this formula is not useful for practical purposes since S usually obtains an infinite number of values.

For computing the short-run distribution, a Markov chain can be used. McShane and Turnbull (1991) used a Markov chain setting to compute the short run expectation and variance for the CSP-1 inspection scheme (see Section 4), and then used a normal approximation. However, the required transition matrix might become very large, and thus prove to be computationally impractical (Yang, 1983).

3.2. A Computational Solution: Generating Random Data

A practical alternative to using the conditional probability formula or the Markov chain approach is to generate random values from the waiting time distributions X_i, and to use them for creating realizations from the per-cycle or short-run ratios' distributions.

Generating values from any run-related or scan-related distribution $P_T(t)$ can be done using the general inversion algorithm:

1. Generate a uniform $(0, 1)$ variate u.
2. Add the probabilities $P(T = 1) + P(T = 2) + \cdots$ until the sum exceeds u for the first time.
3. The largest value summed before exceeding u is the required generated value.

The efficiency of the inversion method depends on the parameters of the runs or scans distribution. For example, to generate data from an order k geometric distribution with probability p of success, the general inversion algorithm is very efficient when p is large because the distribution will be concentrated around k with a thin right tail. However, for small values of p or very large values of k the distribution will have a heavy right tail thereby causing the computation to take longer.

Next, we propose an alternative method for cases where the inversion method is likely to be inefficient. The method takes advantage of special probabilistic features of runs and scans distribution and thus leads to a more efficient generation of random data.

A key feature of order k (i.e., run-related) distributions, which implies an efficient generation algorithm, is that any discrete order k variable can be expressed as a random sum of i.i.d. k-truncated variables (Charalambides, 1986). For example, an order k geometric variable X can be expressed as the compound sum

$$X = \sum_{j=1}^{N} \Gamma_j, \tag{5}$$

where Γ_j are k-truncated i.i.d. geometric variables with parameter $q = 1 - p$, and N is a geometric variable (independent of Γ_j) with parameter p^k that counts successes until the first failure:

$$P(\Gamma_j = \gamma) = \frac{qp^{\gamma - 1}}{1 - p^k} \quad \gamma = 1, 2, \ldots \tag{6}$$

$$P(N = n) = p^k(1 - p^k)^n \quad n = 0, 1, \ldots \tag{7}$$

To generate a value from X we thus generate a value for N and then generate N values from Γ. In both cases the inverse-CDF can be found easily and used for the random number generation. To generate N:

1. Generate a uniform $(0, 1)$ variate u.
2. The required value is given by $\lceil \log u / \log(1 - p^k) \rceil$.

To generate a value for Γ:

1. Generate a uniform $(0, 1)$ variate u.
2. The required value is given by $\lceil \log(1 - up(1 - p^k)) / \log p \rceil$.

This algorithm is likely to be more efficient than the inversion method for small values of p or large values of k.

Another example is the order k negative binomial variable, which is the waiting time for the r-th run of length k to occur. Following Charalambides (1986), it can be expressed as in (5) with N being a truncated negative binomial variable with parameters r and p^k.

4. An Example: Continuous Sampling Plan 1

To illustrate our method we apply it to the continuous sampling scheme CSP-1 which was described in Section 2.1.

The two waiting times that arise in this system are the number of items inspected under screening and the number of items inspected under sampling. The former is an order i geometric variable while the last is an ordinary geometric variable (assuming probability sampling).

An important performance measure of continuous sampling plans is the operating characteristic (OC) function, which is the proportion of items produced during sampling phases. Denote by X the number of items produced during screening, and by Y the number of items produced during sampling. The traditional definition of the OC function is given by

$$OC(p) = \frac{E(Y)}{E(X+Y)}, \tag{8}$$

where p is the probability of producing a non-conforming item.

Here X is an order i geometric variable with parameter $q = 1 - p$, since the number of produced items is the same as the number of inspected items under screening. Y, the number of items produced during sampling, is a geometric variable with parameter pf, where f is the probability of an item being inspected. The two probability generating functions are given by

$$G_X(s) = \frac{(qs)^i(1-qs)}{1-s+pq^i s^{i+1}}, \tag{9}$$

$$G_Y(s) = \frac{pfs}{1-(1-pf)s}. \tag{10}$$

The expectations can be derived from these generating functions by differentiation to obtain the OC function

$$OC_{CSP1}(p) = \frac{1}{1-f+f/q^i}. \tag{11}$$

Since there are only two levels of inspection the long-run proportion of items produced under sampling is $OC(p)$, and under screening is $1 - OC(p)$. This is equivalent to constructing a Markov chain and computing the steady state probabilities. Previous work on continuous sampling plans obtained other measures that are based on ratios of

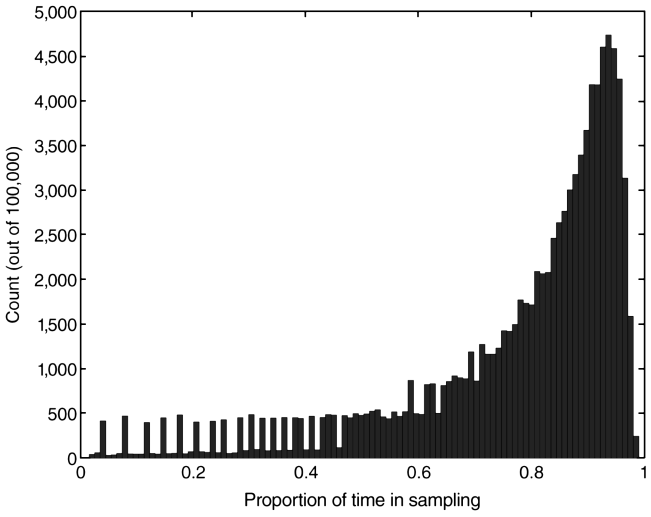

Figure 3. Distribution of $Y/(X+Y)$ for 100,000 simulated values for a CSP-1 scheme with $i = 23, f = 0.5$ and proportion non-conforming $p = 0.01$.

expectations (e.g., the average fraction of inspected items) through a Markov chain approach by computing the steady-state probabilities (McShane and Turnbull, 1991). From a computational point of view, a Markov chain for the CSP-1 scheme requires $i + 5$ states (McShane and Turnbull, 1991), where i ranges between 3 and 17,420 according to the Military Standard 1235B. For more complicated schemes, such as CSP-2, an even larger number of states might be required.

The *OC* function is based on expectations which usually come from highly skewed distributions. This means that even though the *OC* tells us about the expected proportion of time that will be spent in sampling, this proportion might vary widely from the *OC* value in the short term.

In order to find the distributions of short-term measures such as the per-cycle *OC* or the finite-time *OC*, we generate random values from values from $P_X(x)$ and $P_Y(y)$. Generating values from the latter is straightforward (see Devroye, 1986, pp. 498–501). For the i-order geometric variable the general inversion algorithm is very efficient when q is large (which is very probable in practice). For small values of q or huge values of i we prefer the more efficient generating algorithm described in Section 3.2.

To illustrate the computational procedure we select a CSP-1 scheme with $i = 23$ and $f = 0.5$, which yields an AOQL of 1.22% (i.e., the worst average outgoing quality after applying this scheme will not exceed 1.22% non-conforming items). The screening waiting time is a 23-order geometric variable with parameter q and the sampling stage waiting time is a geometric variable with parameter $0.5p$. $P_X(x)$ can be calculated through the website http://iew3.technion.ac.il/sqconline/k_geo.html, which also gives a plot of the cumulative probability (useful for finding quantiles).

Using the formula in (8) we calculate the long-term measure $OC(0.01) = 0.885$. To learn about the variability around the OC values, we look at the distribution of the two short-term OC measures. For the per-cycle OC, we simulate the distribution of $Y/X+Y$ by generating 100,000 values for X and Y as described above. Figure 4 displays the distribution of the simulated values $Y/(X+Y)$ for $p = 0.01$. The distribution is very skewed to the left, thus making the value of $OC(0.01) = 0.885$ not very informative.

For the finite time OC, 1,000 values were generated for X and Y until time $N = 100, 500, 1{,}000$, and 5,000. Plots of the simulated distributions are given in Figure 4. It is clear that as N increases the distribution of the short-term measure OC_N becomes more normal. The proportion of time spent in sampling is the sum of the proportions within each cycle (including the last partial cycle). The per-cycle proportions are i.i.d. variables and thus when N is large the finite-time OC is a sum of many i.i.d. random variables, which reaches a normal distribution.

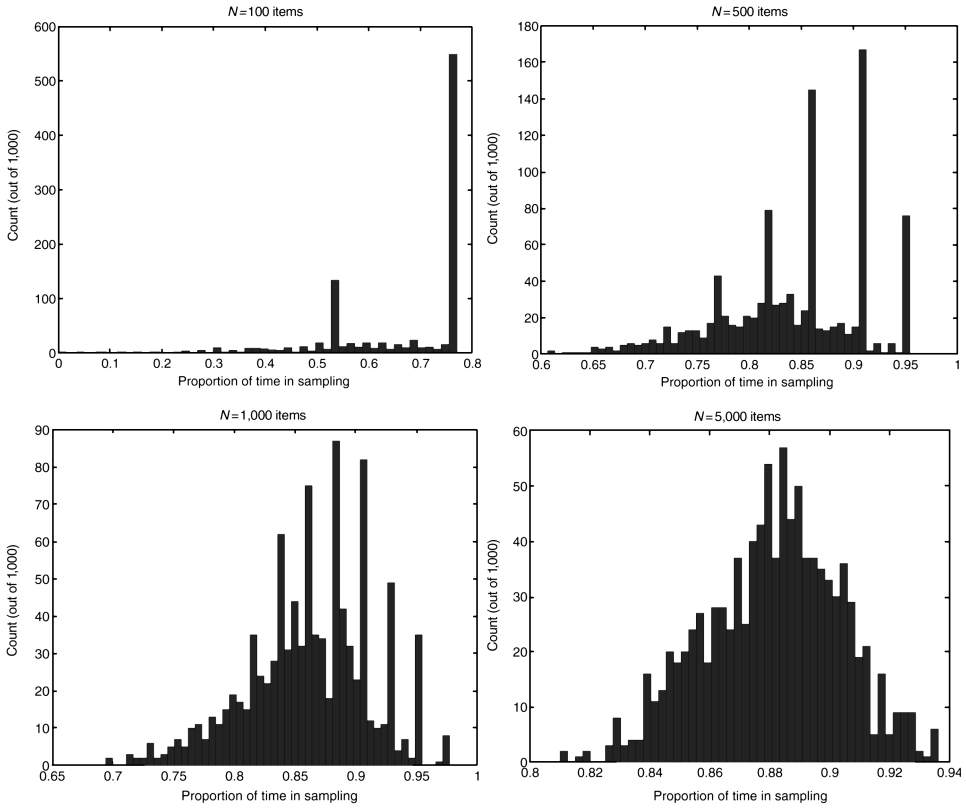

Figure 4. Distribution of the proportion of time spent in sampling within $N = 100, 500, 1{,}000$, and 5,000 produced items, for a CSP-1 scheme with $i = 23$, $f = 0.5$ and proportion non-conforming $p = 0.01$.

Another feature which is clearly seen in Figure 4 is the spikes at certain values. These spikes occur due to the specific setting of the selected CSP-1 scheme. Since we selected $i = 23$ and $1 - p = 0.99$, and the scheme starts with the screening phase, there is a very large chance that the screening stage will end right after the first 23 screened items. Therefore, when $N = 100$ the most likely scenarios are one or two screening phases, each consisting of $X = 23$ which lead to the two spikes at $1 - 23/100$ and $1 - 46/100$. The same argument explains the spikes for $N = 500$ at $1 - 23/500$, $1 - 46/500$, $1 - 69/500, \ldots$, etc. The figure also gives a sense of the number of cycles that one should expect to see when applying the CSP-1 scheme with this setting. It remains to investigate the exact distribution of the number of cycles within a finite number of produced items (i.e., finite time).

5. Concluding Remarks

In multi-level systems a major interest is to evaluate the proportion of time spent at each of the levels. In this paper we investigate and compare long and short term measures for the ratios of time spent at each level of the system. In particular, we deal with systems where switching between the levels is governed by runs and scans rules. In such cases the distributions of the waiting times in the different levels tend to be highly skewed. The long term measure for the proportion of time spent at a certain level is the ratio of expectations of waiting times, and can thus be insufficient in describing the variability in the short term. The two short term measures discussed here, namely, the per-cycle ratio and the finite-time ratio, give more information about the distribution of time spent at each of the system's levels. Although both are short-term measures and are related, their distributions can differ significantly, as illustrated for the bi-level CSP-1 scheme. In this case the distribution of the proportion of time spent at the sampling phase is shown to be highly skewed for the per-cycle ratio, and much more symmetrical for the finite-time ratio.

Distributions of runs and scans related waiting times have been investigated and several methods for deriving them exist in the literature. However, when it comes to functions of such variable, and in particular to ratios of waiting times, obtaining their distribution is not straightforward. We suggest to use simulated data for this purpose, taking advantage of features of runs and scans distributions in order to generate random numbers efficiently.

Although this paper deals with i.i.d. waiting times, the results can be carried over to the Markov-dependent case. The waiting time distributions that arise in the context of the circular and non-circular systems can be calculated using the generating function method (e.g., Aki, 1992; Shmueli, 2002) or Markov-Chain embedding (Balakrishnan and Koutras, 2002, pp. 46–48; Fu and Koutras, 1994). For purposes of simulation the inversion method can always be used once the waiting time distribution is specified. For k-order i.i.d. variables we showed how the random sum representation can be used for more efficient generation in some cases. It remains open to find similar representations for scans-related i.i.d. waiting times and for the Markov-dependent case.

References

S. Aki, ''Waiting time problems for a sequence of discrete random variables,'' *Annals of the Institute of Statistical Mathematics* vol. 44 pp. 363–378, 1992.

N. Balakrishnan and M. V. Koutras, *Runs and Scans with Applications*, Wiley & Sons: New York, 2002.

Ch. A. Charalambides, ''On discrete distributions of order *k*,'' *Annals of the Institute of Statistical Mathematics* vol. 38A pp. 557–568, 1986.

O. Chryssaphinou, S. Papastavridis, and T. Tsapelas, ''On the waiting time of appearance of given patterns,'' In *Runs and Patterns in Probability*, Kluwer Academic Publishers, pp. 231–241, 1994.

L. Devroye, *Non-Uniform Random Variate Generation*, New York: Springer-Verlag, 1986.

H. F. Dodge, ''A sampling inspection plan for continuous production,'' *Annals of Mathematical Statistics* vol. 14 pp. 264–279, 1943.

W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1, Wiley, 1968.

J. C. Fu and M. V. Koutras, ''Distribution theory of runs: a Markov chain approach,'' *J. Amer. Statist. Assoc.* vol. 89(427) pp. 1050–1058, 1994.

L. M. McShane and B. W. Turnbull, ''Probability limits on outgoing quality for continuous sampling plans,'' *Technometrics* vol. 33(4), pp. 393–404, 1991.

G. Shmueli and A. Cohen, ''Run-related probability functions applied to sampling inspection,'' *Technometrics* vol. 42(2), pp. 188–202, 2000.

G. Shmueli, ''Computing consecutive-type reliabilities non-recursively,'' *IEEE Tranactions on Reliability*, to appear, 2002.

G. L. Yang, ''A renewal-process approach to continuous sampling plans,'' *Technometrics* vol. 25(1) pp. 59–67, 1983.