On the analysis and design of CUSUM-Shewhart control schemes

by Emmanuel Yashchin

In recent years cumulative sum (CUSUM) control charts have become increasingly popular as an alternative to Shewhart's control charts. These charts use sequentially accumulated information in order to detect out-of-control conditions. They are philosophically related to procedures of sequential hypothesis testing (the relation being similar to that existing between Shewhart's charts and classical procedures for hypothesis testing). In the present paper we present a new approach to design of CUSUM-Shewhart control schemes and analysis of the associated run length distributions (under the assumption that the observations correspond to a sequence of independent and identically distributed random variables). This approach is based on the theory of Markov chains and it enables one to analyze the ARL (Average Run Length), the distribution function of the run length, and other quantities associated with a CUSUM-Shewhart scheme. In addition, it enables one to analyze situations in which out-of-target conditions are not present initially, but rather appear after a substantial period of time during which the process has operated in on-target mode (steady state analysis). The paper also introduces an APL package, DARCS, for design, analysis, and running of both one- and two-sided CUSUM-Shewhart control schemes and gives several examples of its application.

©Copyright 1985 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

1. Introduction

Since the early thirties, control charts have been widely used in industrial quality control as a means of monitoring the quality of manufactured products. Traditionally, the most commonly used are Shewhart's charts ($\bar{X}$-chart, R-chart, p-chart, etc.), where the process is said to be out of control once the last plotted observation falls outside the prescribed control limits. The main advantage of these "classical" charts is their simplicity. The main disadvantage is that typically these charts are not very sensitive with respect to moderate changes in the process level. In order to overcome this difficulty, several modifications of the basic procedure were proposed by various authors. These modifications call for additional signal criteria based on warning limits, runs, etc. However, the price for improved sensitivity is substantial, and loss of simplicity is only a minor part of it. Indeed, the classical chart enables one to analyze, in a relatively simple way, the distribution of the run length of the control chart for any values of the parameters—one can see that in this case the run length is a geometric random variable. With additional signal criteria, the run length distribution becomes much more complicated and in most cases can be analyzed only by means of simulation study. If such a study indicates that the scheme is not satisfactory (say, the probability of the false signal within a short period of time is too high), it is not always clear how the procedure can be changed in order to

meet the requirements. In the final account, the computational effort needed to find a scheme with desirable properties might be enormous. Another difficulty is the interpretation of the control chart in the presence of several signal criteria.

Analysis of the run length distribution of control schemes is especially important in situations where measurements are taken and processed automatically (and the operator is notified only if an out-of-control signal is triggered), as well as in situations where several sequences of observations are monitored simultaneously. In such situations frequent out-of-control signals associated with *economically* nonsignificant changes in process parameters may cause frequent unjustified corrective actions and/or eventually ruin the discipline of the operator. On the other hand, failure to detect a truly out-of-control situation rapidly may result in a substantial amount of poor-quality product. Thus, *any control scheme associated with automatic data processing and/or simultaneous control of several parameters should be thoroughly analyzed before it can be recommended for use.* The analysis should involve identification of various possible joint distributions of observations and investigation of the corresponding run length distributions. Its ultimate aim is to ensure that the run length of the scheme under consideration is sufficiently long if the changes in process parameters are not *economically* significant and sufficiently short if they are.

In situations associated with automatic data collection and processing, the following question becomes critical: Should we continue to use traditional control schemes just because they are most convenient for manual plotting and visual evaluation (which is based on intuition and is, therefore, highly subjective)? Indeed, some of the classical schemes were often preferred to schemes having much better statistical properties merely because of computational simplicity. For example, for samples of size $n > 2$ from a normal (or approximately normal) population, the sample range carries much less information about σ than the sample standard deviation, $\hat{\sigma}$ [1]; yet R-charts are often preferred to $\hat{\sigma}$-charts (possibly because fifteen years ago it was not very practical, because of computational difficulties, to run $\hat{\sigma}$-charts in an industrial environment). The answer to the above question is, probably, as follows: If the process is capable of meeting specifications and if relatively large variations of process parameters are not associated with significant economic losses, traditional control schemes do very well (as do many other "reasonable" schemes); otherwise, we must look for schemes with better statistical properties.

The desired properties of such an improved scheme might be as follows:

- It must be as sensitive as a comparable modified Shewhart's control chart with respect to substantial changes in the level of the controlled parameter and more sensitive than a comparable modified Shewhart's control chart with respect to moderate changes.
- Each scheme (chart) should be based on very few (one or two) signal criteria.
- It must be easily adjustable in order to meet reasonable requirements related to the behavior of the run length distribution for relevant values of controlled parameters.
- It must enable effective analytic evaluation of performance for a wide range of stochastic patterns of incoming data; in other words, its behavior must be easily predictable once the distribution of input data is given.
- It must be robust with respect to slight departures from the desired model that have no effect on the controlled parameter (for example, if the controlled parameter is the median of the distribution, we would not like the relevant scheme to overreact because of slight departures from assumed normality).
- It must enable relatively easy estimation of the current values of the controlled parameter, especially after an out-of-control signal has been triggered.
- It must enable easy visual interpretation by adequately trained personnel.
- The scheme must be convenient for graphic display on the screen of a computer terminal; in particular, graphic representations in which the chart does not systematically "run away" from the screen should be available.
- The scheme should enable easy implementation of the FIR (Fast Initial Response) feature; i.e., it must provide an instrument for initial setup which detects the *initial* out-of-control conditions earlier than similar conditions occurring later.

Are there control schemes possessing the mentioned properties? The answer is yes; in particular, some types of cumulative sum (CUSUM) control charts (first introduced by Page [2]) can serve as an adequate example. These charts use sequentially accumulated information in order to detect out-of-control conditions. They are philosophically related to procedures of *sequential* hypothesis testing [3], the relation being similar to that existing between Shewhart's charts and classical procedures of hypothesis testing. Several other schemes proposed as an alternative to the classical Shewhart's procedure (e.g., [4]) also meet the stated requirements to various extents. However, most of them either do not enable efficient study of the run length distribution or are associated with such unpleasant features as the necessity of specifying weights (in the weighted moving average charts), excessive algebraic manipulations, problems with visual interpretation, etc. Other reasons for the relative popularity of the CUSUM approach are due to its connection to the theory of Sequential Probability Ratio Tests [3, 5] and to the Central Limit Theory [6] as well as to the availability of certain optimality results [7, 8].

Though CUSUM charts are very useful for other purposes (e.g., retrospective data analysis, graphical data representation, diagnostics, forecasting, sequential hypothesis testing, etc.), our interests in the present work are primarily related to CUSUM as a means for detecting and analyzing out-of-control conditions. Extensive discussion on other applications of CUSUM schemes is contained, for example, in Refs. [9–12].

2. CUSUM control schemes

In this section we give a short description of typical CUSUM procedures. Let $x_1, x_2, \cdots$ be a sequence of observations related to a certain process. The observation x_i may represent, for example,

- the sample percentage of defective chips in the ith produced lot,
- the total number of defects found in the ith produced wafer,
- the sample mean of four diameters of ball bearings chosen at random during the ith production period,
- the sample standard deviation of ten simultaneous measurements (corresponding to various locations) of polyethylene film thickness taken during the ith sampling period,

and so on. Let us suppose that we would like our observations to fall as close as possible to some target value t_0. Without loss of generality, we can assume that $t_0 = 0$. If we are concerned about the possibility that the process might shift up from its target level, it would be natural to adopt the following three-parametric control scheme:

a. choose $h^+ > 0$ (signal level), k^+ (reference value), and $0 \le s_0^+ < h^+$ (headstart);
b. compute the sequence of cumulative sums:

$$s_i^+ = \max\{s_{i-1}^+ + (x_i - k^+), 0\}, \qquad i = 1, 2, \cdots; \qquad (1)$$

c. let N^+ be the first index i for which $s_i^+ \ge h^+$. Then trigger the out-of-control signal at time N^+.

The described procedure is called an upper Page's scheme with parameters (h^+, k^+, s_0^+). N^+ represents the run length of the scheme. If an additional signal criterion is introduced, namely

d. if a single observation x_i satisfies $x_i \ge c^+$, trigger an out-of-control signal at the moment i,

the procedure is called an upper Page's scheme with parameters (h^+, k^+, s_0^+) supplemented by Shewhart's limit c^+.

At this point, some comments about the meaning of the reference value, the headstart, and the Shewhart's limit are appropriate. The reference value k^+ acts as an "anchor," keeping the CUSUM from drifting in on-target situations. The headstart s_0^+ implements the FIR feature mentioned earlier. The rationale for using a headstart is as follows: If the process is on target, the Page's scheme is (most likely) brought to zero by the reference value, so that in this case the expected effect of the headstart is minimal; otherwise, however, the out-of-control signal is triggered much sooner (for example, see [13]). Finally, supplementing the scheme with a Shewhart's limit improves the sensitivity of the scheme with respect to substantial changes in the process level (for example, see [14]). There are also cases in which Shewhart's limits are introduced because of some special features of the associated production process or other considerations. Here and in what follows we refer to such (supplemented) Page's schemes as CUSUM-Shewhart control schemes.

It is clear that in order to affect the performance of the control scheme, the Shewhart's limit must satisfy $c^+ < h^+ + k^+$. On the other hand, if $c^+ \le k^+$, an out-of-control signal can be triggered only if the Shewhart's limit has been violated. Therefore, classical (Shewhart's) control schemes can be viewed as special cases of CUSUM-Shewhart schemes. An alternate way to represent a Shewhart's scheme as a special case of a CUSUM-Shewhart scheme is to set h^+ and k^+ to zero and the desired Shewhart's limit, respectively.

In a similar way, if we are concerned about the possibility that the process might shift down from its target level, the appropriate *lower* Page's scheme with parameters ($h^- > 0$, k^-, $0 \le s_0^- < h^-$) calls for computing the sequence of lower cumulative sums

$$s_i^- = \max\{s_{i-1}^- + (-x_i - k^-), 0\}, \qquad i = 1, 2, \cdots \qquad (2)$$

and triggering an out-of-control signal at the first moment N^- for which $s_{N^-} \ge h^-$. If an additional signal criterion (calling for a signal at the moment i if $x_i \le c^-$) is introduced, we say that the lower scheme is supplemented by Shewhart's limit c^-.

When we would like to detect rapidly both types of shift of process from its target level, it makes sense to run both schemes simultaneously. This procedure is called a two-sided Page's scheme with parameters $(h^+, k^+, s_0^+, h^-, k^-, s_0^-)$, possibly supplemented by Shewhart's limits (c^-, c^+). The run length of the two-sided scheme is denoted by N. Clearly, $N = \min(N^+, N^-)$.

Now we give some comments related to selection of the scheme parameters. First, the domain of possible values of a single controlled parameter (say ξ) usually consists of a "target" region and one or two "taboo" regions (depending on what types of deviations of ξ from its target level ξ_0 are considered undesirable). (See, for example, [15(a)].) The control procedure is usually applied to a sequence $\hat{\xi}_1 - \xi_0, \hat{\xi}_2 - \xi_0, \cdots$ of estimates of $\xi - \xi_0$ corresponding to sequential moments of time.

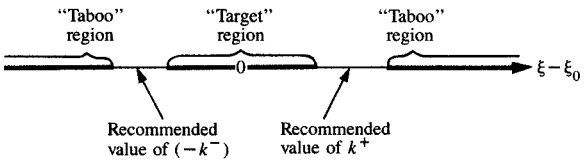

Figure 1

Typical structure of the domain corresponding to a single parameter ξ.

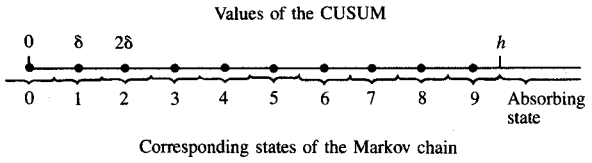

Figure 2

Discretization of the values of a one-sided CUSUM scheme.

The reference values k^+, k^- are usually recommended to be chosen about midway between the appropriate bounds of the "target" and "taboo" regions, as shown in **Figure 1**. Such a choice of k^+, k^- has certain (asymptotic) optimality properties (for example, see [8]) and, moreover, it is known to work very well for most stochastic patterns of observations encountered in practical applications. When the target is chosen to be the origin, the above recommendations always call for positive values of k^+, k^-, which is very convenient for a user, especially in the case of two-sided procedures. However, in some cases (often occurring in one-sided control situations) it is hard to tie the observed quantities to any particular "target" value. In these cases, instead of looking for some artificial target value, we can simply pick the reference values in accordance with the above recommendations and apply them directly to the observations. Though the resulting reference values may now be negative, the logic of cusum plotting is unaffected and confusing situations are unlikely to occur.

Once the reference values are fixed, the performance of the scheme depends primarily on the values of signal levels and Shewhart's limits. Typically, one is interested in schemes satisfying certain requirements related to on-target performance. Thus, values of (h^+, c^+, h^-, c^-) appropriate in a given situation must not only satisfy these requirements but also ensure the best possible off-target performance of the scheme. In Sections 6 and 8 we outline the design procedures for achieving this goal.

The values of the headstarts are usually selected in the final stage of the design procedure. As we know, setting the headstart to a nonzero value improves the sensitivity of the scheme with respect to initial out-of-control conditions. On the other hand, it increases the probability of a false signal within any given period of time. Thus, there usually exists a "trade-off" value of the headstart for which improvement in sensitivity is not associated with substantial increase in probability of triggering a false signal; the situation here is somewhat similar to that occurring in problems related to testing of a hypothesis when one considers "trade-off" between Type 1 and Type 2 errors. In most practical situations the value of the headstart does not exceed half of the corresponding signal level.

3. Analysis of one-sided Page's schemes

As we mentioned in the introduction, the run length distribution is the primary criterion of performance of a control scheme. In this section we introduce the methodology for deriving the basic quantities associated with this distribution. Our basic assumption is that the observations $x_1, x_2, \cdots$ are realizations of a sequence $X_1, X_2, \cdots$ of independent and identically distributed (iid) random variables. The distribution function of X_i is denoted by $F(x)$.

At present, the most popular method for analysis of run length distributions (first introduced in 1972 by Brook and Evans [16]) is based on discretization of the values of CUSUM and then treating it as a Markov chain (on other methods of analysis see, for example, [17]). Let $\{h, k, s_0\}$ be a Page's scheme applied to the sequence of observations $x_1, x_2, \cdots$. Then it is clear that the values $s_0, s_1, \cdots$ form a Markov chain which is discrete in time but may be continuous in space. The levels 0 and h are reflecting and absorbing barriers of the chain, respectively.

For computational purposes we discretize the values of $s_0, s_1, \cdots$ as shown in **Figure 2**. In other words, the values of $s_0, s_1, \cdots$ are rounded to the center of a corresponding group. The number of states of the Markov chain (excluding the absorbing state) is termed the *level of discretization* of the scheme and denoted by d. For example, in the case represented by Fig. 2, the level of discretization is $d = 10$.

Note that the length of an interval corresponding to a single state, δ, is always related to the level of discretization by means of the formula

$$\delta = h/(d - 0.5). \tag{3}$$

Thus, the centers of the groups are at points $0, \delta, 2\delta, \cdots, (d-1)\delta$ and $h = (\delta/2) +$ (center of the last group). Such a method of discretization usually gives approximations of good quality and is recommended in many sources (for example, [16]). The transition matrix of the Markov chain can be easily expressed in terms of $F(x)$. Analysis of this matrix enables one to find the average run length (ARL) and the standard deviation of the run length (SDRL), as well as

the higher moments of the run length N. One can also compute $P\{N > r\}$, $r = 0, 1, \cdots$. Some of the relevant formulas can be found in the mentioned paper by Brook and Evans.

Next we give some comments about the effect of discretization. Extensive case studies indicate that levels of discretization of about $d \approx 30$ give results which are satisfactory for most practical purposes. The reason for that is related to the fact that we discretize the *states* of the CUSUM chart, but *not* the observations themselves. Thus, relatively low sensitivity with respect to level of discretization is explained by compensation of roundoff errors when subsequent values of the scheme are computed. As an example, let us apply the scheme ($h = 3$, $k = 1$, $s_0 = 0$) to three sequences of normal observations corresponding to $\mu = 0$, 0.5, and 1 and $\sigma = 1$. **Table 1** contains the values of ARL as well as the lower and upper 5% quantiles of the run length distribution (in parentheses) corresponding to levels of discretization ranging from 10 to 100. It indicates that levels of discretization as low as 10 enable one to roughly assess the properties of the run length distribution.

Finally, we discuss two special topics related to the one-sided Page's scheme.

One-sided Page's scheme supplemented by Shewhart's limit c If the one-sided scheme is supplemented by Shewhart's limit c, the cusum $s_0, s_1, \cdots$ is still a Markov chain, with the transition matrix being a modified version of one considered earlier. Analysis of this transition matrix results in the basic quantities associated with the CUSUM-Shewhart scheme under consideration. Another way of looking at the scheme supplemented by Shewhart's limit c is as follows: Replace $F(x)$, the distribution function (d.f.) of X, with a d.f. of an improper random variable X^* defined by $X^* = X$ if $X \le c$ and $X^* = \infty$ otherwise. The (improper) d.f. $F^*(x)$ of X^* is, of course,

$$F^*(x) = \begin{cases} F(x), & x < c, \\ F(c), & x \ge c, \end{cases} \tag{4}$$

and analysis can easily be performed by using F^* as the d.f. of the observations instead of F.

Analysis of steady state situations Our previous discussion was related to the situations in which deviations of the process from target conditions occur at time $i = 0$. Under this assumption, one of the primary questions of interest was this: How fast will the relevant control schemes detect the presence of various types of out-of-control conditions? However, in many cases, we would like to analyze the run length distributions corresponding to deviations of the process from the target conditions, when these deviations occur after a substantial period of time during which the process operated in on-target mode characterized by some distribution function of the observations F.

Table 1 Effect of the level of discretization on ARL and 5% quantiles (in parentheses) corresponding to the scheme ($h = 3$, $k = 1$, $s_0 = 0$). The observations are iid normal with $\sigma = 1$. The entries are rounded to the nearest integer.

d	$\mu =$ 0	0.5	1
10	1918 (100, 5741)	117 (8, 343)	17 (3, 45)
20	1952 (102, 5842)	117 (8, 345)	17 (3, 45)
30	1958 (102, 5860)	117 (9, 345)	17 (3, 45)
50	1961 (102, 5869)	118 (9, 345)	17 (3, 45)
100	1962 (102, 5873)	118 (9, 345)	17 (3, 45)

First of all, let us ask the following question: What are the probabilities of various states of the Markov chain (associated with our control scheme) after a long period of time *given* that the out-of-control signal was *not* triggered during this period of time? It is well known (for example, see [18]) that the relevant probabilities are given by the normalized left eigenvector corresponding to the maximal real eigenvalue λ_0 of R, the $(d \times d)$ principal minor of the transition matrix. Thus, we denote

$$q(j) \stackrel{\text{def}}{=} \lim_{k \to \infty} P\{(j - 0.5)\delta < s_k < (j + 0.5)\delta \mid N > k\},$$
$$j = 0, 1, \cdots, d - 1, \tag{5}$$

provided the limit exists. Note that in this case the distribution $\{q(j)\}$ does not depend on s_0.

Now let us assume that after a long period of time the d.f. of the observations switched from F (on-target d.f.) to $\tilde{F}$. Let N be the remaining run length until the signal is triggered. Then

$$E\{N\} = \sum_{j=0}^{d-1} q(j) E_{\tilde{F}}\{N \mid s_0 = j\delta\},$$
$$P\{N > r\} = \sum_{j=0}^{d-1} q(j) P_{\tilde{F}}\{N > r \mid s_0 = j\delta\}, \tag{6}$$

where the subscript $\tilde{F}$ emphasizes that the corresponding expectation and probability are computed under the assumption that the observations are generated by the d.f. $\tilde{F}$. Other quantities related to N can also be found by using formulas of type (6).

4. Some basic information related to the analysis of two-sided Page's schemes

Analysis of two-sided schemes is known to be technically much more complicated than that of one-sided schemes. Indeed, after discretization of both schemes, one can see that the two-sided scheme represents a pair of dependent Markov chains operating simultaneously. This pair may be treated as a single Markov chain with the set of possible states represented by pairs (i, j), where i and j are the states corresponding to the upper and lower schemes, respectively.

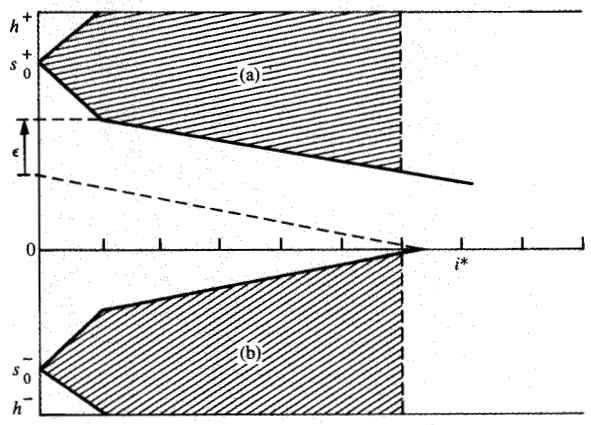

Figure 3

Nonsignal regions corresponding to the event $\{N \geq i^*\}$.

Unfortunately, the corresponding transition matrix is usually so large that any detailed analysis of the run length distribution becomes practically impossible. For example, if each scheme is discretized into $d = 30$ states, the transition matrix of the two-sided scheme is about (900×900). This example provides a partial explanation for the skepticism of many potential users of cumulative sum techniques.

Clearly, an alternative, much more efficient method of analysis was needed. Indications that such a method might exist are related to the fact that upper and lower schemes are strongly correlated; thus, it is reasonable to expect that the transition matrix of the two-sided scheme is "too big" for the information it contains. Indeed, Woodall [15(b)] showed that some reduction of the number of states of a two-sided scheme is possible. However, even with this reduction, the size of the transition matrix remains substantial. One can also approach the problem by trying to find a relation between the basic quantities associated with a two-sided scheme and those associated with corresponding one-sided schemes. Some important results in that direction were obtained by Khan [19] and Lucas and Crosier [13]. The comprehensive theory (and associated method) for analysis of general two-sided CUSUM-Shewhart schemes on the basis of this approach can be found in [20].

In this section we give some basic information related to this method. This information is helpful for understanding the principles upon which our approach to analysis and the associated software are based. The reader is referred to [20] for a more detailed discussion and proofs of the presented theoretical results. To simplify the notation we shall assume, without loss of generality, that $h^+ \geq h^-$.

First let us introduce two constants which play a crucial role in the analysis of two-sided CUSUM schemes. Let $(h^+, k^+, s_0^+, h^-, k^-, s_0^-)$ be a two-sided Page's scheme. The first constant, i^*, represents the smallest non-negative integer for which

$$s_0^+ + s_0^- - h^+ - i^*(k^+ + k^-) \leq 0. \tag{7}$$

One can show [20] that this constant has the following property: Whatever the realization of the observations and index i may be, $s_i^+ + s_i^- \leq h^+$ if and only if $i \geq i^*$.

Analysis of the two-sided Page's scheme is heavily dependent on the question of whether the associated upper and lower schemes do or do not interact. By definition, the upper and lower schemes do not interact if no realization x_1, x_2, $\cdots$ of observations exists for which one of the schemes signals while, at the same moment, the value of the opposite scheme is not zero. Absence of interaction means that the successive points of signal of the two-sided scheme form a *renewal process* [21] which makes the analysis much easier. The second constant of interest, ε, defined by

$$\varepsilon \stackrel{\text{def}}{=} (h^+ - h^-) - (k^+ + k^-), \tag{8}$$

represents a measure of the "amount of interaction" present in a given two-sided scheme. Indeed, let $x_1, x_2, \cdots$ be any realization for which the run length N is greater than i^*. As shown in [20, Theorem 4.1], for such a realization

$$s_N^+ \geq h^+ \quad \text{implies that} \quad s_N^- = 0, \tag{9a}$$

$$s_N^- \geq h^- \quad \text{implies that} \quad 0 \leq s_N^+ \leq \max(0, \varepsilon). \tag{9b}$$

Thus, if $\varepsilon \leq 0$ and it is known that no signal was triggered within the first i^* steps of the scheme, the subsequent behavior of the two-sided scheme is as if no interaction were present. On the other hand, if $\varepsilon > 0$ and it is known that no signal was triggered within the first i^* steps, it may happen that one of the schemes signals while at the same time the value of the opposite scheme is not zero—but this value never exceeds ε. So, if ε is small compared to h^+, bounds and approximations of very high quality for quantities associated with two-sided schemes are usually available. [Most practical situations (in particular, when $h^+ = h^-$) are related to the case $\varepsilon \leq 0$, so that no need for such bounds and approximations typically arises.] If $\varepsilon > 0$, we say that the two-sided Page's scheme satisfies the *intrinsic interaction* condition. In this case ε is called the *power* of intrinsic interaction and ε/h^+ is called the *relative power* of intrinsic interaction.

So far, we have considered the situation in which *it is known* that no signal was triggered within the first i^* steps. But analysis of the run length distribution must take into account the possibility that a signal may be triggered within this period of time. One can show that such analysis can be based on only one of the schemes. Indeed, assume that $i^* > 0$ and consider the regions a and b given by

$$0 \leq s_0^+ + s_0^- - h^- - i(k^+ + k^-) < s_i^+ < h^+, \tag{10a}$$

$$0 \leq s_0^+ + s_0^- - h^+ - i(k^+ + k^-) < s_i^- < h^-, \tag{10b}$$

$0 \le i < i^*$ (see **Figure 3**). As shown in [20, Theorem 4.2]), cusum paths corresponding to any realization of observations for which $N \ge i^*$ must lie within these regions. Moreover, if for some realization $x_1, x_2, \cdots$ one of the relations (10) holds for every $0 \le i < i^*$, then (for this realization) the other relation also holds for every $0 \le i < i^*$ and $N \ge i^*$.

The last result implies that the probability of any event related to the behavior of the run length of a two-sided scheme within the first i^* steps can be found by analyzing the upper (or lower) scheme only. For example, $N \ge i^*$ if and only if the path of the upper scheme (see Fig. 3) lies within the shaded region corresponding to (10a) [or, alternatively, if and only if the path of the lower scheme lies within the shaded region corresponding to (10b)].

In light of the previous discussion, we are able to suggest a very simple criterion for noninteraction of the upper and lower schemes, namely: The upper and lower schemes do not interact if and only if $\varepsilon \le 0$ and $s_0^+ + s_0^- - h^- - (k^+ + k^-) \le 0$ or, equivalently, if and only if

$$\varepsilon \le \min\{0,\ h^+ - (s_0^+ + s_0^-)\}. \tag{11}$$

Moreover, the results given above enable one to develop a unified approach to analysis of two-sided control schemes. This approach is based on analysis of the probabilistic behavior of the scheme prior to i^* and after i^* separately. The first part of the analysis (as well as the conditional distribution of values of the upper and lower schemes at the moment i^*, given that no out-of-control signal was triggered within the first i^* steps) can be performed by considering one of the schemes only, and therefore is not associated with a substantial computational effort. Thus, we assume, without loss of generality, that $i^* = 0$ and proceed to the second part.

Let $L^+(p \mid s_0^+)$, $L^-(p \mid s_0^-)$, and $L(p \mid s_0^+, s_0^-)$ be Laplace transforms of the run lengths of the upper scheme with headstart s_0^+, the lower scheme with headstart s_0^-, and the two-sided scheme, respectively. Also, define

$$A_\varepsilon(p) = \frac{L^+(p \mid s_0^+)[1 - L^-(p \mid 0)] + L^-(p \mid s_0^-)[1 - L^+(p \mid \varepsilon)]}{1 - L^-(p \mid 0)L^+(p \mid \varepsilon)}, \qquad p \ne 0,$$

$$A_\varepsilon(0) = 1. \tag{12}$$

We start the analysis of a given two-sided Page's scheme by finding ε. If $\varepsilon \le 0$, one can show (see [20, Theorem 5.1]) that $L(p \mid s_0^+, s_0^-) = A_0(p)$. Thus, expansion of (12) into power series enables one to obtain expressions for ARL, SDRL, and $P(UP) = P\{N^+ < N\}$ (i.e., the probability that the signal is triggered by the upper scheme) in terms of the ARL's and SDRL's of the associated one-sided schemes.

The formula (12) also enables one to determine the run length distribution of the two-sided scheme. Indeed, once discretization is performed, $L^+(\log p \mid s_0^+)$, $L^-(\log p \mid s_0^-)$ and, consequently, $L(\log p \mid s_0^+, s_0^-)$ become ratios of polynomials with real coefficients. The run length distribution can therefore be analyzed by finding the roots of the denominator in (12), expanding (12) into the sum of partial fractions, and subsequent termwise inversion of the Laplace transform.

If $\varepsilon > 0$, one can prove [20, Section 8] that for $p > 0$,

$$A_0(p) \le L(p \mid s_0^+, s_0^-) \le A_\varepsilon(p). \tag{13}$$

This inequality leads immediately to bounds for ARL and, after some additional analysis, to bounds for higher-order moments as well as for the run length distribution itself. In practical terms, however, the Laplace transform of the run length is much closer to the left bound in (13) than to its right bound. In fact, one has

$$L(p \mid s_0^+, s_0^-) = A_0(p) + P \cdot C(p), \tag{14}$$

where P is the probability that

1. the signal is triggered by the lower scheme;
2. the value of the upper scheme at the moment of signal is not equal to 0 [note that by (9), it must be less than ε];
3. the upper scheme does not reach 0 (before it signals) during its subsequent path;
4. the value of the upper scheme at the moment of its signal does not exceed h^+ by more than the minimal value of the upper scheme achieved during its subsequent path;

and $C(p)$ is the "correction term" associated with the above event. One can show that as $p \to 0$, $C(p) = -\mu_P p + o(p)$, where μ_P is approximately equal to half of the ARL of the upper scheme with headstart $(h^+ - \varepsilon)$. It is intuitively clear that whatever the stochastic pattern of observations may be, P must be very small (especially for moderate powers of intrinsic interaction), and if, in addition, the trend of the CUSUM path is upwards, then also the impact of $C(p)$ is small. Since in the vast majority of the problems encountered in practice the relative power of intrinsic interaction is 0.5 or less, we performed an extensive study of the second term in (14) in this domain in order to decide whether its impact was significant enough to justify additional computational effort. This study was based on the exact transition matrix of the bivariate Markov chain as well as on simulated runs, and it led to the conclusion that approximations based on the first term of (14) produced results sufficient for most practical purposes.

The results given so far are relevant with respect to any stochastic pattern of incoming iid observations. It is clear that any restrictions on the nature of observations can lead only to a *decrease* in actual power of intrinsic interaction. For example, consider the case in which the two-sided scheme is supplemented by Shewhart's limits c^-, c^+. Clearly, if $-c^- \le k^-$ (i.e., the lower scheme is of Shewhart's type), the

```
ANALYSIS OF ONE-SIDED CUSUM SCHEME WITH PARAMETERS  H,K,C = 3 1 3.5
THE LEVEL OF DISCRETIZATION IS 30
THE VARYING PARAMETER IS THE HEADSTART, (HEADST).
  HEADST      ARL     SDRL       10       20       30       50      100
       0   1507.3   1505.4  .99430   .98772   .98118   .96823   .93660
 .163E01   1478.9   1505.1  .97559   .96913   .96272   .95001   .91897
 .173E01   1472.3   1504.9  .97124   .96481   .95842   .94577   .91487
 .183E01   1464.4   1504.7  .96602   .95963   .95327   .94069   .90996
```

Figure 4

```
ANALYSIS OF ONE-SIDED CUSUM SCHEME WITH PARAMETERS  H,K,C = 3 1 3.5
THE LEVEL OF DISCRETIZATION IS 30
THE HEADSTART IS OUT OF RANGE; STEADY STATE ANALYSIS ASSUMED
THE CHANGING PARAMETER NAME IS SIGMA
    SIGMA       ARL      SDRL        5       10       15       20       50      100
 .800E00  47185.9  47194.2  .99971   .99960   .99949   .99939   .99875   .99769
 .900E00   6279.8   6280.8  .99896   .99816   .99736   .99657   .99182   .98396
 .100E01   1505.9   1505.3  .99668   .99338   .99009   .98680   .96733   .93573
 .110E01    530.1    529.1  .99146   .98215   .97291   .96376   .91064   .82853
 .120E01    241.8    240.7  .98190   .96174   .94197   .92260   .81449   .66171
```

Figure 5

question whether interaction is present or not can be ignored. Otherwise, all the results given earlier remain relevant, but instead of ε one should use $\varepsilon_1 = \varepsilon - \beta(k^+ + k^-)$ where $\beta = -1 - h^-/(k^- + c^-)$. One can see that in this case, whenever introducing Shewhart's limits affects the run length distribution, β is positive.

5. Software support

Procedures for analysis of CUSUM-Shewhart control schemes described in the previous sections are used primarily for designing a control scheme appropriate in a given situation. In cases where it is really important to have a "good" scheme, the design procedure is likely to require a substantial amount of work. First of all, the designer should study the available data (and perform experimental work, if necessary) in order to identify the relevant on-target and off-target stochastic patterns of incoming observations. Subsequently, he must choose the parameters of the scheme in such a way that its performance is satisfactory with respect to these patterns. Clearly, in most nontrivial situations the design procedure requires appropriate software. As a minimum, such software should enable one a) to obtain any quantity associated with a given CUSUM-Shewhart procedure for an arbitrary distribution of iid observations, and b) to apply CUSUM-Shewhart procedures to actual sets of data and, for more complicated stochastic behavior of the observations, study the properties of a scheme by simulation.

In this section we give a short description of an APL software package, DARCS, for design, analysis, and running of CUSUM-Shewhart control schemes developed recently in the Department of Mathematical Sciences of the IBM Thomas J. Watson Research Center, with some examples of its application. This material is helpful for a better understanding of the design procedures given in the following sections. In the present work we discuss the functions appropriate for situations in which the observations form an iid sequence only. A more detailed description of the package as well as additional examples can be found in [22].

The package contains two basic functions for analysis of the run length distribution. The first one determines, for any given pattern of iid observations, the ARL, SDRL, and run length distribution corresponding to a given one-sided CUSUM-Shewhart scheme (including steady state analysis). The second function performs a similar analysis for a given two-sided scheme. In addition, it enables one to determine the probability $P(UP)$ that the out-of-control signal is triggered by the upper scheme.

Each of the mentioned functions can operate in one of the following three modes:

I-mode is used for interactive analysis of the performance of a given scheme with respect to a fixed d.f. of incoming observations;

V-mode is used for a noninteractive analysis of a sequence of schemes (depending on a single varying parameter) with respect to a given fixed d.f. of the observations;

E-mode is used for a noninteractive analysis of performance of a fixed scheme with respect to a set of several d.f.'s of the observations corresponding to different values of a specified parameter.

To illustrate application of these functions, let us consider several examples.

Suppose that we would like to analyze the run length of a one-sided scheme $h = 3$, $k = 1$ supplemented by Shewhart's limit $c = 3.5$, for all (discretized) values of s_0 between 1.6 and 1.8 and $s_0 = 0$ when $\{x_i\}$ are distributed normally with $\mu = 0$ and $\sigma = 1$. Suppose also that this distribution corresponds to an on-target situation and that we would like (for the purpose of future steady state analysis) to store the steady state probabilities $\{q(j)\}$. In addition, we would like to compute, for each headstart, the probabilities $P\{run\ length > r\}$ for $r = 10$, 20, 30, 50, and 100. Application of an appropriate function (in V-mode) results in the printout shown in **Figure 4**. The computed values of the steady state distribution are $\{q(j), j = 0, 1, \cdots, 30\} = \{0.8155, 0.0241, \cdots, 0.0001\}$; the intermediate values are omitted.

Figure 4 serves as an illustration of how unreliable ARL can be as a performance criterion for a control scheme. Note that although some of these schemes have ARL's over 1400, the probability of a signal before 100 observations are taken is close to 0.1.

To illustrate the application of the E-mode of analysis, let us ask the following question: After the scheme considered previously runs for a long time in on-target mode (i.e., with observations coming from the standard normal distribution), what is the effect of change in σ (in the steady state situation)

from its on-target value (1) to 0.8, 0.9, 1.1, or 1.2? Application of the analysis function results in the printout shown in **Figure 5**, which illustrates the high sensitivity of performance of a CUSUM with respect to slight changes in σ. It also explains why there is a reason to be suspicious about the performance of any scheme that has been derived "under the assumption that σ is known."

In our next example, let us analyze the run length of a symmetric two-sided scheme $h^+ = 3$, $k^+ = 1$, $h^- = 3$, $k^- = 1$, supplemented by Shewhart's limits $c^- = -3.5$, $c^+ = 3.5$, the observations $\{x_i\}$ being distributed normally with $\mu = 0$ and $\sigma = 1$. Let us also compute the probabilities $P\{run\ length > r\}$ for $r = 10$, 20, 30, 50, and 100. In order to provide the reader with the chance to compare the results with those obtained earlier (note that our scheme is combined from two one-sided schemes considered in the first example), we chose (0, 0), (1.627, 1.627), and (1.627, 1.831) as pairs of headstarts of interest. Now application of the function for analysis of two-sided schemes (in V-mode, the varying parameters being s_0^+ and s_0^-) produces the results shown in **Figure 6**.

On the basis of Fig. 6, one can verify that the ARL of a two-sided scheme can be roughly approximated by a harmonic mean of ARL's corresponding to one-sided schemes [23]. In addition, one can see that $P\{run\ length > r\}$ can be approximated by a product of analogous probabilities corresponding to one-sided schemes. For example, in accordance with the first printout of this section, $P\{R.L. > 100 \mid s_0^+ = 1.627, s_0^- = 1.831\} = 0.83557 \approx 0.91897 \times 0.90996 = 0.83622$. This property is related to one mentioned above and also to the fact that the run length distribution of a CUSUM-Shewhart scheme can usually be roughly approximated (especially in the tail area) by an appropriate geometric distribution. We use it in Section 8 for the purpose of designing two-sided schemes.

Finally, let us examine the effect of a shift of μ from 0 to 0.1, 0.25, 0.5, 1.0, 1.5, or 2.0 on the run length of the scheme. After applying the appropriate function in E-mode we obtain the results shown in **Figure 7**, indicating that for $\mu \geq 0.5$ performance of the two-sided scheme is roughly equivalent to that of the upper scheme.

6. Design of a one-sided CUSUM-Shewhart scheme

Until the early seventies, design and analysis of CUSUM control schemes centered about the notion of ARL. Though the use of ARL as a primary criterion for evaluation of the performance of the scheme was criticized by some authors (for example, [11, 24]), the choice of alternative measures of efficacy was very limited because of theoretical as well as computational difficulties. Typically, the approach to the problem of choice of an appropriate scheme was based on the assumption that the observations come from a normal population with a known standard deviation σ. Under this

```
ANALYSIS OF THE TWO-SIDED CUSUM SCHEME WITH PARAMETERS:
H+,K+,C+ = 3 1 3.5 AND H-,K-,C- = 3 1 -3.5
THE LEVELS OF DISCRETIZATION ARE D+,D-= 30 30
 HEADSTARTS       P(UP)      ARL     SDRL       10       20       30       50      100
      0     0     .500    753.6    751.7  .98863  .97556  .96267  .93740  .87708
.163E01 .163E01   .500    725.3    751.2  .95137  .93880  .92639  .90207  .84402
.163E01 .183E01   .495    718.1    750.9  .94185  .92940  .91712  .89304  .83557
```

Figure 6

```
ANALYSIS OF THE TWO-SIDED CUSUM SCHEME WITH PARAMETERS:
H+,K+,C+ = 3 1 3.5 AND H-,K-,C- = 3 1 -3.5
THE LEVELS OF DISCRETIZATION ARE D+,D-= 30 30
THE VALUES OF THE HEADSTARTS ARE 0 0
THE CHANGING PARAMETER NAME IS MEU
    MEU  P(UP)      ARL     SDRL        5       10       20       50      100
.100E00   .749    653.9    651.9  .99462  .98707  .97204  .92832  .85978
.250E00   .937    365.7    363.3  .99129  .97791  .95136  .87595  .76333
.500E00   .995    110.5    107.6  .97494  .93183  .84916  .64251  .40367
.100E01  1.000     17.1     14.1  .84008  .59530  .29071  .03376  .00093
.150E01  1.000      6.3      3.9  .48629  .12590  .00780  .00000  .00000
.200E01  1.000      3.7      1.8  .13647  .00455  .00000  .00000  .00000
```

Figure 7

assumption, ARL was studied as a function of the process mean μ, similarly to the way in which the operating characteristic (OC) curve is studied in standard hypothesis testing (or acceptance sampling) problems (for example, see [25, Fig. 22.7]). The decision whether to adopt a given scheme was then based on the properties of this curve, the major points of interest being μ_0 (good quality) and μ_1 (bad quality). In particular, in the case of cumulative $\bar{X}$-charts σ is a function of the sample size n, and then it is always possible to find h, k, and n (we assume that $s_0 = 0$) such that ARL (μ_0) is approximately equal to some fixed large number L_a and, at the same time, ARL (μ_1) is approximately equal to some fixed small number L_r. Nomograms and examples related to the solution of this two-point problem can be found, for example, in [25, pp. 471–479] and [26].

One of the major drawbacks of such an approach is that it is solely based on the notion of ARL. Though ARL is probably meaningful in the *off-target* situation, it can be highly misleading when the on-target case is under study (primarily because the set of possible CUSUM paths includes "too many" extremely "short" members). Indeed, as we have learned from the first example in Section 5, schemes with ARL $\approx$ 1500 can signal within the first 100 observations with probability approximately 0.1. This feature of the run length distribution might represent a serious source of problems, especially in the cases in which one uses headstarts. Indeed, an increase in the probability of a false signal resulting from the use of a headstart is not as visible in terms of ARL, because the right tail behavior of the run length distribution is (practically) independent of the headstart. An additional drawback of ARL as a criterion for choice of an appropriate control scheme is that it is extremely sensitive with respect to slight departures from normality as well as with respect to slight variations in σ.

In general, the user of a CUSUM scheme probably feels uneasy about specifying a particular ARL for the on-target

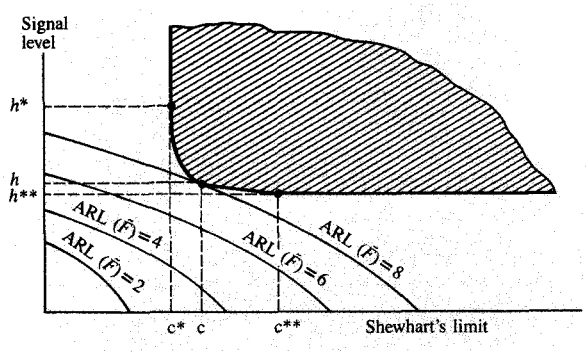

Figure 8

The set values (h, c) corresponding to (15) (shaded) and level curves of ARL corresponding to some off-target distribution $\bar{F}$ (k is fixed and $s_0 = 0$).

situation; what he typically wants is that the scheme will not generate a *false* alarm within a certain period of time (say, a shift) with probability of at least, say, 0.99. Indeed, the most fruitful applications of CUSUM techniques in industry are related to situations associated with automated collection and processing of data generated in a systematic way by a production line. In such situations costs associated with unjustified troubleshooting (possibly caused by a false out-of-control signal) are usually of primary concern for the designer of a control procedure. Since this procedure typically monitors several (sometimes thousands of) sequences of observations at the same time, the necessity of keeping the overall probability of a false signal small leads to corresponding requirements related to each individual (univariate) scheme. This argument serves as a basis for our approach to the problem of design. It also emphasizes the need for having the ability to analyze the run length *distribution* and not just its first moment, the ARL, in order to choose a scheme with "good" properties.

In this section we give some recommendations that might be helpful for designing a one-sided (say, upper) CUSUM-Shewhart scheme. Let $x_1, x_2, \cdots$ be the sequence of observations the scheme is intended to monitor. In all practical cases we would like this sequence to be iid or at least stationary in time. So, let us suppose that some process history is available, and it indicates that such behavior of the observations is in principle achievable. Moreover, in the initial stage we assume that the observations form an iid sequence with common d.f. $F(x)$. Let ξ be some measure of central tendency (mean, median, etc.) of F. Then our scheme is primarily supposed to control the level of this parameter (though it is possible that F depends on other parameters as well).

To choose an appropriate one-sided (say, upper) Page's scheme one must in general specify four parameters and possibly some other "shadow" parameters which affect the distribution of the observations (e.g., sample size n when the observations form a sequence of sample means) but are related to the particular sampling routine rather than to the nature of the process of interest. In the present work we assume that the values of all parameters of this type are fixed. The main reason for excluding the "shadow" parameters from consideration is related to our desire to keep the discussion as simple as possible; our experience shows that the design procedures we are going to present together with DARCS software enable one to determine, relatively quickly, the values of such parameters by trial and error. Another reason is related to the fact that in many practical situations the sampling intensity is determined by the production process itself and therefore cannot be easily changed (for example, in cases where *every* produced item is subject to automated inspection, the sampling intensity is determined by the production speed).

Under the stated assumptions the following sequence of steps in most cases leads to a scheme with "good" properties.

Step 1 Choose the "target" and "taboo" regions for ξ (see Fig. 1). This may require evaluation (or estimation) of various parameters of F (or F itself) on the basis of the previous history. It may also be associated with cost analysis.

Step 2 Define the *extreme on-target* family, i.e., the family of d.f.'s that are likely to represent the on-target observations and for which ξ is equal to the *upper* bound of the "target" region. (For example, in the case when the "target" region is $\xi \le 0.5$, one may include in the extreme on-target family all the normal d.f.'s with $\mu = 0.5$, $\sigma \le 0.1$ as well as all gamma d.f.'s with $\beta\alpha = 0.5$, $\beta\sqrt{\alpha} \le 0.1$, $\alpha \ge 25$.) In a similar way, define the *extreme off-target* family, i.e., the family of d.f.'s that are likely to represent the off-target observations and for which ξ is equal to the *lower* bound of the "taboo" region.

Step 3 Choose k in accordance with the recommendations of Section 2.

Step 4 Choose K (in accordance with sampling intensity) and $\alpha \approx 0$, for which it is desired that

$$P\{\textit{run length} > K \mid \text{Process is on target}\} \ge 1 - \alpha. \qquad (15)$$

Step 5 It is not difficult to see that the set of all pairs (h, c) for which (15) holds form a convex region of the type shown in **Figure 8**.

One is primarily interested in schemes corresponding to the curved boundary of this region. Moreover, in most practical cases it turns out that the scheme of interest corresponds to one of the extreme points $[(h^*, c^*)$ or $(h^{**}, c^{**})]$ of this region. These points can be found by means of the following procedure.

From the extreme on-target family, choose the "worst" representative F_0 for which the out-of-control signal is still undesirable. [For example, in the case considered in Step 2, such a representative is, probably, the gamma d.f. with $\beta = 0.02$, $\alpha = 25$. (The "worst" representatives of the on-target family of distributions are those having the "shortest" run lengths; similarly, the "worst" representatives of the off-target family are those having the "longest" run lengths.)] Find a *minimal* c^* for which a *pure* Shewhart's scheme with upper control limit c^* satisfies

$$F_0(c^*) \geq (1 - \alpha)^{1/K}. \tag{16}$$

[Note that in the case where the observations are coming from the population F_0, the resulting "pure" Shewhart's control scheme satisfies (15).] Continue the analysis for F_0 and find a *minimal* value of the signal level (h^*) for which the scheme "practically never" signals within K observations [i.e., for which (15) is still valid]. (Typically, this results in a scheme for which a false out-of-control signal is most likely related to violation of the Shewhart's limit; thus, the scheme corresponding to this extreme point can be viewed as a Shewhart's scheme "supplemented" by the CUSUM control criterion. This also explains why it is usually possible to introduce a "moderate" headstart without affecting substantially the on-target performance of the scheme.) Choose a headstart value s_0^* for which the resulting scheme still "practically never" signals within K observations provided that they are generated by F_0.

Next find a scheme corresponding to the second extreme point of the boundary region (see Fig. 8). Start with the scheme (h^*, k) without either headstart or Shewhart's limit. Decrease h until the *minimal* value is found (h^{**}) for which the scheme "practically never" signals within K observations, provided that they are generated by F_0. Increase c (starting from c^*) until the *minimal* value is found (c^{**}) for which (15) is still valid. Supplement the derived scheme with an appropriate headstart s_0^{**}.

Step 6 Examine the behavior of *both* derived schemes with respect to "worst" representatives of the extreme off-target family. Make slight adjustments to the parameters of the schemes, if necessary.

Step 7 Usually both schemes perform well for the values of ξ corresponding to the "target" and "taboo" regions. The scheme based on (h^{**}, k, c^{**}) is generally more sensitive with respect to the values of ξ corresponding to the "intermediate" region. On the other hand, the scheme based on (h^*, k, c^*) is more likely to "catch" extremely large deviations from the target level within a very short period of time. Therefore, the final choice of a scheme depends on the relative importance of these properties in a given situation. In some cases the user may prefer a scheme based on some (h, k, c) corresponding to a "tradeoff" between these properties. For example, if one is interested in a scheme having the smallest possible ARL for some given off-target d.f. $\tilde{F}$, the appropriate choice could be based on (h, c), derived as shown in Fig. 8. Thus, in this step we choose an appropriate scheme and examine its sensitivity with respect to the "worst" representatives of the extreme off-target family. If it is found to be satisfactory, the scheme is ready for steady state analysis. Otherwise, it is worthwhile to examine the effect of slight variations in the parameters of the scheme (including k) on the overall performance of the scheme. If this examination does not lead to a satisfactory scheme, it is most likely that, with present sampling intensity, a scheme with desired properties is, in principle, unachievable. The possibilities corresponding to this situation are as follows:

- Increase the sampling intensity, redetermine K accordingly, and then repeat the search procedure. One of the ways to increase the sampling intensity is to pick n observations (instead of a single observation) at a time, if possible; the control scheme is then applied to a corresponding sequence of estimators for ξ.
- Decrease the value of h until some kind of "trade-off" between the probability of a false alarm and sensitivity is achieved.

Step 8 Let $\hat{F}_0$ be a d.f. which seems most likely to represent the data corresponding to an on-target situation. Perform a steady state analysis with respect to the "worst" representatives of the extreme off-target family. Adjust the parameters of the scheme slightly if necessary.

Step 9 Examine the performance of the scheme with respect to selected on-target and off-target distributions F which are likely to appear in practice.

Though the above sequence of steps seems to be somewhat lengthy, one finds out, after some practice, that intuition and common sense lead quickly to a scheme with the desired properties, provided the latter exists.

Next we discuss some additional aspects which (in certain cases) should be taken into consideration when performing the final analysis of a control scheme.

If it is known that the observations may correspond to a sequence of *dependent* random variables (e.g., a stationary Gaussian sequence), it is very desirable to examine the behavior of the derived scheme with respect to simulated sequences corresponding to "on-target" and "off-target" conditions. Some theoretical results related to the performance of CUSUM schemes in situations where observations are serially correlated can be found, for example, in the two-part paper by Bagshaw and Johnson [27, 28]. These results are based on the fact that under certain rather general conditions, the (properly normalized) CUSUM path converges weakly to the Wiener process; thus,

the run length distribution of a Page's scheme can be approximated by that of the time to absorption of a Wiener process with reflecting and absorbing barriers. The theoretical basis for such an approximation can be found in [29, Theorems 20.1 and 21.1].

In the cases where observations may be contaminated by outliers, one may be willing to consider a modification of the basic control procedure that calls for special treatment of cases in which an out-of-control signal is suspected of being triggered by an outlier. Since such modification typically leads to some loss of sensitivity with respect to off-target situations, the modified scheme should be re-examined with respect to relevant stochastic patterns of the input observations. The effect of some of the possible modifications (for example, ignore the outlier, winsorize the outlier, etc.) is discussed in [30].

7. Examples

In this section we give several examples illustrating the use of some functions in relation to the design and analysis of CUSUM-Shewhart schemes. We limit ourselves to outlining only the basic steps of the design and analysis, so that the interested reader can verify each step by using the software.

Example 7.1 (Cumulative $\hat{\sigma}$-chart) In the oxidation process of silicon wafers, we are interested in keeping the "within-the-lot" variability of the thickness of the grown SiO_2 layer as small as possible. In order to achieve that, we take n measurements of film thickness per lot and monitor the resulting sequence of sample standard deviations,

$$\hat{\sigma}_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_i)^2}, \qquad i = 1, 2, \cdots,$$

where $y_{i1}, y_{i2}, \cdots y_{in}$ are the observations corresponding to the ith lot and $\bar{y}_i$ is their sample mean.

We do not have a particular "target" value for the sequence $\hat{\sigma}_1, \hat{\sigma}_2, \cdots$; instead, we have a "target" region: namely, we want the underlying "true" standard deviation σ to lie within the interval $0 \le \sigma \le 0.02$ nm. On the other hand, we would like to detect as quickly as possible the situation in which $\sigma \ge 0.04$ nm.

Suppose that for every i, $\{y_{i1}, y_{i2}, \cdots y_{in}\}$ are realizations of a normal random variable with certain mean and standard deviation σ. Then $\hat{\sigma}_1, \hat{\sigma}_2, \cdots$ can be considered as realizations of a random variable $\Sigma(\sigma, n) = \sigma\sqrt{\chi^2(n-1)/(n-1)}$, where $\chi^2(n-1)$ is a chi-square random variable with $(n-1)$ degrees of freedom. Under the additional assumptions that $\hat{\sigma}_1, \hat{\sigma}_2, \cdots$ are realizations of *independent* random variables, and that

a. the sample size is fixed ($n = 4$),
b. the planning "horizon" does not extend beyond $K = 200$ samples (because of the regular equipment maintenance operations), and
c. if the process is on-target, the probability of a false signal within 200 samples should be at most $\alpha = 0.0001$,

we would like to find a CUSUM-Shewhart scheme with the best possible out-of-target performance.

We start by noting that the "target" and "taboo" regions are $\sigma \le 2$ and $\sigma \ge 4$, respectively. Moreover, since we are in fact dealing with a one-parametric family corresponding to $\Sigma(\sigma, 4)$, the only member of the extreme on-target family is the d.f. of $\Sigma(2, 4)$; analogously, the only member of the extreme off-target family corresponds to $\Sigma(4, 4)$. In accordance with the recommendations of Section 2, we choose the starting reference value $k = 3$. Next, one can see, by analyzing a "pure" Shewhart's scheme with respect to $\Sigma(2, 4) \sim F_0$, that the solution of (16) is $c^* \approx 6.6$. Further, we find, by using the V-mode of the function for analysis of one-sided schemes, that $h^* = 5$. Finally, for the headstart $s_0^* = 1.02$, the probability of a false signal (under F_0) is still approximately equal to 0.0001.

In order to obtain the parameters of the scheme corresponding to the second extreme point of Fig. 4, we start with a "pure" CUSUM scheme ($h^* = 5$, $k = 3$) and find, by using the V-mode of the software, that $h^{**} = 4.5$. Further, we can supplement the scheme with $c^{**} = 7.2$ and $s_0^{**} = 0.3$ without affecting its on-target performance (with respect to F_0).

In the off-target situation corresponding to $\Sigma(4, 4)$, the ARL of both derived schemes is 6.7; the SDRL's are 4.8 and 4.5, respectively. To illustrate the point we discussed in Step 7, consider the off-target situation corresponding to $\Sigma(8, 4)$. The probability that the run length will exceed 1 is 0.44 (for the first scheme) and 0.51 for the second one. On the other hand, the first scheme is less sensitive with respect to moderate deviations of σ from its target region; for example, if the observations are generated by $\Sigma(3, 4)$, the ARL's of the first and second schemes are 60.1 and 50.7, respectively. In **Table 2** we give the values of ARL corresponding to both derived schemes and a "pure" Shewhart's scheme having

Table 2 Values of the ARL corresponding to schemes considered in Example 7.1 as functions of the standard deviation, σ ($n = 4$): a) Shewhart's scheme with $c = 6.55$; b) $h^* = 5$, $k = 3$, $s_0^* = 1.02$, $c^* = 6.6$; c) $h^{**} = 4.5$, $k = 3$, $s_0^{**} = 0.3$, $c^{**} = 7.2$.

Scheme	$\sigma =$	*2*	*2.5*	*3*	*3.5*	*4*	*5*	*6*	*7*	*8*
a		2.08×10^6	7820.3	396.4	68.0	22.2	6.2	3.2	2.2	1.8
b		2.03×10^6	2095.1	60.1	13.5	6.7	3.3	2.3	1.9	1.6
c		1.98×10^6	1368.7	50.7	13.0	6.7	3.5	2.5	2.0	1.7

comparable on-target performance (note that for each of these schemes $P\{R.L. > 200 \mid \sigma = 2\} = 0.9999$). This table can serve as a good illustration of the superiority of CUSUM techniques compared to the classical ones.

Our next example is related to the control of the parameter λ of the Poisson population. The classical (Shewhart's) method of control is usually referred to in the literature as a *c*-chart.

Example 7.2 (Cumulative c-chart) [31] Spin dryers are used as one of the steps in the production of integrated circuit chips from semiconductor wafers. Typically, the process steps are followed by rinses with deionized, filtered water. After the rinsing, the water is removed by placing the wafers in the spin dryer (centrifugal device), which spins the water off the wafers (and accelerates evaporation by using dry filtered gas).

Periodically, test wafers are run through the rinse and drying cycle and the particles on the wafer that are larger than a specified diameter are counted. The recorded counts $\{o_1, o_2, \cdots\}$ serve as a basis for the decision to clean and retest the spin dryer. On the basis of theoretical considerations, there is reason to believe that, during a certain initial period of time, the process $\{o_i\}$ corresponds to a sequence of iid Poisson random variables with parameter λ. Under normal conditions, the level of this process does not exceed 6.5. Levels of the process exceeding 11.5 are associated with a high rate of defective production—situations in which the process of counted contaminating particles reaches this level should be detected as soon as possible. On the other hand, since cleaning and retesting represent an expensive and tedious procedure, we are interested in a CUSUM control scheme for which the probability of a false signal within 100 tests is not more than 0.01, and, at the same time, sensitivity with respect to the levels of the process exceeding 11.5 is as high as possible.

Since the first four steps of the design are straightforward (clearly, in our case $\alpha = 0.01$, $K = 100$, $k = 9$), we proceed towards Step 5 and find that $c^* = 18.5$. Subsequently, we find that $h^* = 13.5$. Note that in this part of the analysis we use the nonstandard level of discretization, $d = 14$. [If one deals with counted data and chooses an integer value of k, the values of the CUSUM scheme become limited to the set of non-negative integers (provided the headstart is an integer). Thus, the choice of h and d in accordance with the rule $h = j - 0.5$, $d = j$, where j is an integer, ensures the absence of a roundoff error.] Finally, we choose $s_0^* = 5$ in accordance with Step 7 and proceed towards finding a scheme corresponding to the second extreme point of the boundary region. We start with a "pure" CUSUM scheme ($h^* = 13.5$, $k = 9$) and find that $h^{**} = 12.5$. Subsequently, we find that $c^{**} = 19.5$ and $s_0^{**} = 3$.

To analyze the performance of the first derived scheme for $6.5 \le \lambda \le 11.5$, we apply the E-mode of the function for analysis. This gives the results shown in **Figure 9**.

```
ANALYSIS OF ONE-SIDED CUSUM SCHEME WITH PARAMETERS  H,K,C = 13.5 9 18.5
THE LEVEL OF DISCRETIZATION IS 14
THE VALUE OF THE HEADSTART IS 5
THE CHANGING PARAMETER NAME IS LAMBDA
  LAMBDA      ARL      SDRL        5        10        20        40       100       200
.650E01 11367.9 11381.4  .99848  .99790  .99701  .99526  .99003  .98137
.750E01   653.7   660.1  .98563  .97507  .95989  .93123  .85033  .73080
.850E01    62.1    62.3  .92371  .84147  .71429  .51852  .19846  .04004
.950E01    15.0    13.3  .76041  .51587  .24243  .05396  .00060  .00000
.105E02     7.0     5.1  .50851  .18627  .02479  .00044  .00000  .00000
.115E02     4.5     2.7  .26435  .03713  .00070  .00000  .00000  .00000
```

Figure 9

By using the DARCS software, one can easily verify that in this example the properties of the second scheme are roughly the same and also that additional variation of the parameters of the derived schemes leads to only very minor improvements in their performance.

8. Design of a two-sided CUSUM-Shewhart scheme

Design of a two-sided CUSUM-Shewhart scheme represents, in general, a more difficult problem than design of a one-sided scheme. Though the "first draft" of a scheme with desired properties can be obtained in a relatively straightforward way, the final adjustments and analysis can be time-consuming, as they are associated with manipulations of eight parameters.

Under the assumptions of Section 6, the following sequence of steps usually leads to a scheme with desirable properties:

Step 1 and Step 2 The same as in Section 6.

Step 3 Choose k^+ and k^- in accordance with the recommendations of Section 2.

Step 4 If the "target" region consists of a single point or its width is very small compared to widths of both "intermediate" regions, design (by using the procedure described in Section 6) an upper Page's scheme (h^+, k^+, s_0^+, c^+) and a lower scheme (h^-, k^-, s_0^-, c^-) such that relation (15) holds *for each* scheme. Otherwise, design both schemes so that

$$P\{run\ length > K \mid \text{Process is on target}\} \ge \sqrt{1 - \alpha} \qquad (17)$$

holds for each scheme.

Step 5 Combine the derived one-sided schemes into a single two-sided scheme. This scheme represents a first approximation which in most practical situations is fairly close to what is desired. Indeed, the behavior of the combined scheme with respect to off-target patterns of incoming data is primarily determined by one of "well-designed" one-sided schemes. Further, if the width of the "target" region is very small compared to widths of both "intermediate" regions, then, by the property discussed in

```
ANALYSIS OF THE TWO-SIDED CUSUM SCHEME WITH PARAMETERS:
H+,K+,C+ = 2.1 3 1E60 AND H-,K-,C- = 3.5 2 -5.1
THE LEVELS OF DISCRETIZATION ARE D+,D-= 18 30
THE VALUES OF THE HEADSTARTS ARE 0.93 0.7
THE CHANGING PARAMETER NAME IS MEU
     MEU  P(UP)      ARL     SDRL        5       10       50      100      200
 -.40E01   .000      2.0       .7   .00046   .00000   .00000   .00000   .00000
 -.10E01   .000   4976.2   4977.4   .99868   .99764   .98966   .97976   .96028
 -.50E00   .002  .162E06  .162E06   .99994   .99991   .99966   .99935   .99874
       0   .420  .241E07  .241E07   .99998   .99998   .99996   .99994   .99990
  .500E00   .995  .427E06  .427E06   .99985   .99984   .99974   .99963   .99939
  .100E01  1.000 36076.0 36108.4   .99895   .99881   .99771   .99632   .99357
  .400E01  1.000      2.0       1.3   .02206   .00043   .00000   .00000   .00000
  .600E01  1.000      1.0       .2   .00000   .00000   .00000   .00000   .00000
```

Figure 10

Section 5, the combined scheme (approximately) satisfies the relation (15). Otherwise, the "on-target" behavior of the scheme is mostly determined by one of the one-sided schemes and, by construction, (15) is once more approximately satisfied.

Step 6 Adjust the parameters of the combined scheme so that (15) is satisfied.

Step 7 Examine the behavior of the resulting scheme with respect to the "worst" representatives of the extreme off-target family. If sensitivity is found satisfactory, proceed to the next step. Otherwise, examine the effect of slight variations in the parameters (including k^+, k^-). If it does not improve the performance of the scheme, it is most likely that the scheme with desired properties cannot, in principle, be achieved. The possibilities corresponding to such a situation are analogous to those described in Section 6, Step 7.

Step 8 Examine the performance of the scheme with respect to selected on-target and off-target distributions which are likely to appear in practice.

The effects of some other factors that one may be willing to consider when designing a two-sided CUSUM-Shewhart scheme (for example, see Section 6) can be studied by simulation or by means of approximations based on the properties of the related Wiener process.

For an application of the above procedure, consider the following example.

Example 8.1 (Cumulative $\bar{X}$-chart) Suppose that in the oxidation process of silicon wafers considered in Example 7.1 we are also interested in keeping the difference between the actual mean thickness of the grown SiO_2 layer and the target value t_0 (we denote this difference by Δ) as close to zero as possible. The consequences of systematic deviations between the actual mean thickness and t_0 depend not only on the magnitude but also on the sign of the deviation, so we would like to guard ourselves against situations in which Δ is more than 0.06 nm or less than -0.04 nm.

With observations taken in the same way as in Example 7.1, we shall try to design a CUSUM-Shewhart procedure for controlling Δ on the basis of the sequence $x_i = \bar{y}_i - t_0$, $i = 1, 2, \cdots$.

Suppose that for every i, $\{y_{i1}, y_{i2}, \cdots y_{in}\}$ are realizations of a normal random variable with certain mean and standard deviation σ. We also suppose, as in Example 7.1, that $n = 4$, and, if the process is on target, the probability of a false signal within $K = 200$ samples should not exceed 0.0001. Under these assumptions, we would like to find a scheme with the best performance with respect to the off-target region consisting of two parts, $\{\Delta \le -0.04$ nm$\}$ and $\{\Delta \ge 0.06$ nm$\}$.

Of course, before proceeding further, we need to specify the on-target family of distributions of interest. In order to illustrate the point of Step 4, we consider two possibilities:

a. The on-target family is a set of all normal d.f.'s $N(\Delta, \sigma)$ for which $-1 \le \Delta \le 2$ and $\sigma \le 2$.
b. The on-target family is a set of all normal d.f.'s for which $\Delta = 0$ and $\sigma \le 2$.

In Case a the "worst" representatives of the extreme on-target family are d.f.'s $N(\Delta = -1, \sigma = 2)$ and $N(\Delta = 2, \sigma = 2)$. Further, the width of the "target" region (3) is comparable to the widths of both "intermediate" regions, which are 3 and 4. Therefore, we design, by using the method of Section 5, an upper scheme so that $P\{R.L. > 200 \mid \Delta = 2, \sigma = 2\} \approx 0.9999$. One of the appropriate schemes is $h^+ = 3.2$, $k^+ = 4$, $s_0^+ = 0.33$, $c^+ = 7$). Next, we design a lower scheme satisfying $P\{R.L. > 200 \mid \Delta = -1, \sigma = 2\} \approx 0.9999$ ($h^- = 4.6$, $k^- = 2.5$, $s_0^- = 1.25$, $c^- = -6$). Finally, combining these schemes leads to a two-sided scheme with the desired properties.

In Case b the only "worst" representative of the extreme on-target family is $N(\Delta = 0, \sigma = 2)$. Since the "target" region consists of a single point, 0, we design an upper scheme for which $P\{R.L. > 200 \mid \Delta = 0, \sigma = 2\} \approx \sqrt{0.9999} = 0.99995$ ($h^+ = 2.1$, $k^+ = 3$, $s_0^+ = 0.93$, no Shewhart's limit), and a lower scheme for which the same property holds ($h^- = 3.5$, $k^- = 2$, $s_0^- = 0.7$, $c^- = -5.1$). Finally, we combine the derived schemes and examine the behavior of the resulting two-sided scheme with respect to $\Delta = 0, \pm 0.5, \pm 1, \pm 4$, and 6 by applying the E-mode of the function for analysis of two-sided schemes. This gives the results shown in **Figure 10**. One can see that the derived scheme satisfies (15). Moreover, application of the software shows that further variations of parameters will not lead to dramatic improvements of its performance.

Concluding remarks

In this paper we have presented an approach for complete analysis of one-sided CUSUM-Shewhart control schemes. The methods which we believe to be original include

a. An approach to the problem of design of CUSUM-Shewhart control schemes (Sections 6 and 8).
b. "Steady state analysis" of the performance of a CUSUM-Shewhart scheme (Section 3).

c. Complete (not asymptotic) analysis of the run length distribution (implemented in the software).
d. Software for design, analysis, and running of CUSUM-Shewhart schemes (Section 5).

Acknowledgments

I wish to thank B. Flehinger (IBM Research) for very constructive criticisms and consultations. I also thank the referees for suggestions that improved the paper substantially, N. Brenner (IBM Research) for essential help in the preparation of the software, D. Cooper, M. Stein, D. Withers (IBM Research), and S. Wheeler (IBM Austin) for useful comments, and B. Newman for careful preparation of this manuscript for SCRIPT processing.

References and note

1. E. Lehmann, *Theory of Point Estimation*, John Wiley & Sons, Inc., New York, 1983.
2. E. Page, "Continuous Inspection Schemes," *Biometrika* **41**, 100–115 (1954).
3. N. Johnson, "A Simple Theoretical Approach to Cumulative Sum Control Charts," *J. Amer. Statist. Assoc.* **56**, 835–840 (1961).
4. S. W. Roberts, "A Comparison of Some Control Chart Procedures," *Technometrics* **8**, 411–430 (1966).
5. R. Khan, "On Cumulative Sum Procedure and the SPRT with Applications," *J. Roy. Statist. Soc. B* **46**, 79–85 (1984).
6. M. R. Reynolds, "Approximations to the Average Run Length in Cumulative Sum Control Charts," *Technometrics* **17**, 65–71 (1975).
7. G. Lorden, "Procedures for Reacting to a Change in Distribution," *Ann. Math. Statist.* **42**, 1897–1908 (1971).
8. M. Bagshaw and R. A. Johnson, "The Influence of Reference Values and Estimated Variance on the ARL of Cusum Tests," *J. Roy. Statist. Soc. B* **37**, 413–420 (1975).
9. R. Woodward and P. L. Goldsmith, "Cumulative Sum Techniques," *Imperial Chemical Industries Monograph No. 3*, Oliver and Boyd, London, 1964.
10. D. S. van Dobben de Bruyn, *Cumulative Sum Tests: Theory and Practice*, Hafner Publishing Co., New York, 1968.
11. A. Bissell, "Cusum Techniques for Quality Control," *Appl. Statist.* **18**, 1–30 (1969).
12. British Standards Institution, *Guide to Data Analysis and Quality Control Using Cusum Techniques, Parts 1–4* (1980–1983).
13. J. M. Lucas and R. B. Crosier, "Fast Initial Response for Cusum Quality Control Schemes: Give Your Cusum a Head Start," *Technometrics* **24**, 199–205 (1982).
14. J. M. Lucas, "Combined Shewhart-Cusum Quality Control Schemes," *J. Qual. Technol.* **14**, No. 2, 51–59 (1982).
15. (a) W. H. Woodall, "The Statistical Design of Quality Control Charts," contributed paper, Annual Meeting of the American Statistical Association, Philadelphia, 1984.
(b) W. H. Woodall, "On the Markov Chain Approach to the Two-Sided CUSUM Procedure," *Technometrics* **26**, 41–46 (1984).
16. D. Brook and D. A. Evans, "An Approach to the Probability Distribution of Cusum Run Length," *Biometrika* **59**, 539–549 (1972).
17. W. H. Woodall, "The Distribution of the Run Length of One-Sided Cusum Procedures for Continuous Random Variables," *Technometrics* **25**, 295–300 (1983).
18. J. N. Darroch and E. Seneta, "On Quasi-Stationary Distributions in Absorbing Discrete-Time Finite Markov Chains," *J. Appl. Probabil.* **2**, 88–100 (1965).
19. R. Khan, "A Note on Page's Two-Sided Cumulative Sum Procedure," *Biometrika* **68**, 717–719 (1981).
20. E. Yashchin, "On a Unified Approach to the Analysis of Two-Sided Cumulative Sum Schemes with Headstarts," *Adv. Appl. Probabil.* **17** (1985, in press).
21. S. Karlin and H. M. Taylor, *A First Course in Stochastic Processes*, Academic Press, Inc., New York, 1975.
22. E. Yashchin, "Design, Analysis and Running of CUSUM-Shewhart Control Schemes," *Research Report RC-10869*, IBM Thomas J. Watson Research Laboratory, Yorktown Heights, NY, 1984.
23. K. Kemp, "The Average Run Length of the Cumulative Sum Chart When a V-Mask Is Used," *J. Roy. Statist. Soc. B* **23**, 149–153 (1961).
24. G. Barnard, "Control Charts and Stochastic Processes," *J. Roy. Statist. Soc. B* **21**, 239–257 (1959).
25. A. J. Duncan, *Quality Control and Industrial Statistics*, R. D. Irwin, Inc., Homewood, IL, 1974.
26. A. L. Goel and S. M. Wu, "Determination of ARL and a Contour Nomogram for Cusum Charts to Control Normal Mean," *Technometrics* **13**, 221–230 (1971).
27. R. A. Johnson and M. Bagshaw, "The Effect of Serial Correlation on the Performance of CUSUM Tests," *Technometrics* **16**, 103–122 (1974).
28. M. Bagshaw and R. A. Johnson, "The Effect of Serial Correlation on the Performance of CUSUM TESTS II," *Technometrics* **17**, 73–80 (1975).
29. P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, Inc., New York, 1968.
30. J. M. Lucas and R. B. Crosier, "Robust Cusum: A Robustness Study for Cusum Quality Control Schemes," *Commun. Statist.–Theor. Meth.* **11**, 2669–2687 (1982).
31. The author thanks Dr. D. Cooper (IBM Research) for providing this example.

Received December 20, 1984; revised February 13, 1985

Emmanuel Yashchin *IBM Research Division, P.O. Box 218, Yorktown Heights, New York 10598.* Dr. Yashchin joined IBM at the Thomas J. Watson Research Center in 1983. His research interests include quality control, applied statistics, extreme value theory, and operations research. He received a Diploma in applied mathematics from Vilnius State University (U.S.S.R.) in 1974 and an M.Sc. in operations research and a D.Sc. in statistics from the Technion–Israel Institute of Technology in 1977 and 1981, respectively. In 1982 he was a Visiting Assistant Professor at Iowa State University, Ames. He is a member of the American Society for Quality Control, the American Statistical Association, and the Institute of Mathematical Statistics.