IIE Transactions (1996) **28**, 489–496

Optimal SPRT and CUSUM procedures with compressed-limit gauges

P. LEE GEYER[1], STEFAN H. STEINER[2*] and GEORGE O. WESOLOWSKY[1]

[1]*Faculty of Business, McMaster University, Hamilton, Ontario L8S 4M4, Canada*
[2]*Department of Statistics and Actuarial Sciences, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

Received January 1994 and accepted March 1995

Methodology is presented for the design of single and double compressed-limit sequential probability ratio tests (SPRT) and cumulative sum (CUSUM) control charts to detect one-sided mean shifts in a symmetric probability distribution. We also show how to evaluate the average run length properties with the fast initial response (FIR) feature. The resulting CUSUM plans have a simple scoring procedure, and are extremely simple to derive and implement. The use of two compressed-limit gauges is more efficient than a single compressed-limit gauge. In the case of SPRTs, the use of two compressed limit gauges minimizes the average sampling number required for specified operating characteristics. In the case of CUSUM, the gain in efficiency reduces the out-of-control average run length for a given in-control average run length.

1. Introduction

Control charts are used to detect shifts from the target of a process. Let $X_1, X_2, \ldots$ denote a sequence of independent random variables whose target cumulative distribution function is F_0. Then a one-sided control chart for the detection of an alternate distribution function F_1 consists of an action limit h and a plot of $Y_j = g_j(X_1 X_2, \ldots, X_j)$ against j ($j = 1, 2, \ldots$), where g_j is a function of the previous j random variables. A shift in the process is signaled at time j if $Y_j \geq h$. The economy of a control chart is often evaluated by computing the average run length, where the run length is defined as the number of samples until the chart signals. When the process is at the target value, the run length should be large so that there are few false alarms, and the run length should be small when the process has shifted.

The Shewhart control chart (Shewhart, 1931) can quickly detect large shifts in a process, whereas a cumulative sum (CUSUM) control chart (Page, 1954) is preferred for detecting small shifts in a process. Lucas (1982) recommends combining Shewhart and CUSUM control schemes to obtain efficient control for small and large shifts in a process.

Cumulative sum charts consist of plotting

$$Y_j = \max(0,\ Y_{j-1} + Z_j), \quad j = 1, 2, \ldots, \qquad (1)$$

where Z_j is a function of X_j and $Y_0 = y < h$. The CUSUM chart is a sequence of tests ending in acceptance of $F = F_0$ as long as $Y_j < h$ or with acceptance of $F = F_1$ if $Y_j \geq h$. Moustakides (1986) showed that the CUSUM scheme that minimizes the out-of-control average run length for a given in control run length is always of the form

$$Y_j = \max(0,\ Y_{j-1} + \ln r(x_j)), \quad j = 1, 2, \ldots, \qquad (2)$$

where $r(x_j)$ is the likelihood ratio $L(x_j;F_1)/L(x_j;F_0)$ and $Y_0 = 0$.

Although exact measurement of variables is most efficient, precise measurement can be difficult or expensive. An alternative to the traditional variables-based control chart is to use an attribute control chart (Stevens, 1947; Beattie, 1962; Elder *et al.*, 1981; Lucas and Crosier, 1982). The simplest attribute methods that classify units as conforming or nonconforming are inefficient when the proportion of non-conforming units is very small (Duncan, 1986). Ladany (1976) suggests compensating for the loss in efficiency by using a compressed limit, or a narrow-limit gauge when the underlying distribution of the quality characteristic of interest is known. When a compressed-limit gauge is used, the concept of nonconformity is replaced by one of 'pseudo-nonconformity'. Steiner *et al.* (1994a,b) present methodology for the design of acceptance control and Shewhart-type control charts with multiple-step gauges when the random variable of interest has a normal distribution.

Schneider and O'Cinneide (1987) proposed a CUSUM scheme with a compressed-limit gauge placed at $t = (\mu_0 + \mu_1)/2$, where μ_0 and μ_1 correspond to the acceptable and unacceptable target means of a production process. Beja and Ladany (1974) showed that this choice for the compressed-limit gauge minimizes the sample size for

* Correspondence author

an acceptance sampling plan when the type I and type II error probabilities are equal. This choice was also derived by Sykes (1981) and Evans and Thyregod (1985). Schneider and O'Cinneide (1987) point out that a compressed-limit gauge placed midway between the acceptable and unacceptable mean values maximizes the change in the proportion of pseudo-nonconforming items. Equivalently, this choice maximizes the change in the expected likelihood ratio under μ_0 and μ_1.

Schneider and O'Cinneide's scheme involves setting $Y_0 = 0$, taking samples of size n from a production process and letting Z_j in (1) denote the number of pseudo-nonconforming items in the jth sample minus some reference value k. To obtain the average run length of such a plan, they use the normal approximation to the binomial distribution. This approach is only valid if the sample size is large enough so that $np > 5$, where p denotes the probability of observing pseudo-nonconforming units. This condition must hold when the process is at the target value, and when it is at the unacceptable setting. In most practical applications, this implies that the sample size must be larger than 10 or 15 units. However, many practitioners prefer to take smaller samples more frequently so that the normal approximation to the binomial may not be appropriate. The Schneider and O'Cinneide approach also does not extend easily to the use of two compressed-limit gauges.

In what follows, we extend and improve on the Schneider and O'Cinneide approach. We derive the exact average sampling number and average run length properties for both single and double compressed-limit gauge SPRTs and CUSUM charts. The resulting equations are valid for any sample size as long as observations are entered into the CUSUM unit sequentially.

Following Page (1954), we consider the CUSUM scheme specified by (2) as a sequence of Wald SPRTs with initial score zero and boundaries zero and h. If a fast initial response (FIR) is used, so that $Y_0 \neq 0$, then the first SPRT in the sequence may be considered to have initial score Y_0 (Lucas and Crosier, 1982). Fig. 1 shows the three possible stages in the run length of a CUSUM control chart using the FIR feature. Stage I of the CUSUM consists of a Wald SPRT with absorbing barriers at 0 and h and non-zero initial score. If this test ends in acceptance, a sequence of Wald SPRTs, shown as stage II, with initial score zero, and barriers at zero and h follow. In stage III the CUSUM ends with a rejection Wald SPRT. Note that the FIR CUSUM may terminate in stage I if the cumulative sum reached h before dropping to zero. The traditional CUSUM, without FIR, would consist of stages II and III. The average run length properties of the CUSUM plan are derived by considering the sampling properties of the underlying Wald test.

In Section 2 we show that the scoring procedure suggested by (2) leads to simple random walks. In Section 3 the theory of the random walk (Cox and Miller, 1965) is used to derive the operating characteristics and average sampling number (ASN) for a Wald SPRT with one or two compressed-limit gauges. Average run length properties for the CUSUM specified by (2) are derived in Section 4. In the first sections, we assume that the way in which units are classified is determined either through practical considerations or through some prior knowledge. In Section 5, this assumption is relaxed and optimal compressed limits are derived for both SPRTs and CUSUM procedures. Finally, the ease with which these methods can be used in practice are illustrated with an example.

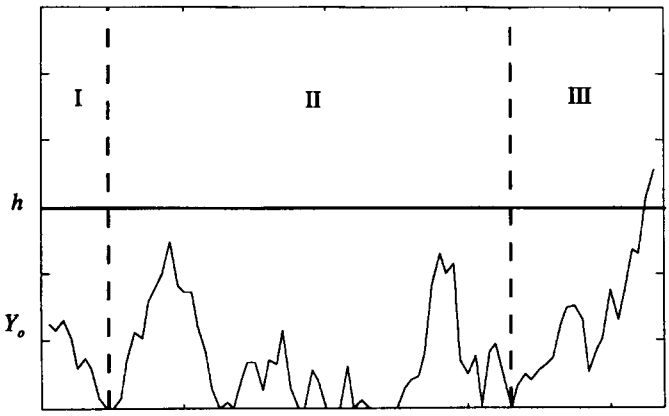

Fig. 1. Three stages of a CUSUM procedure.

2. Notation and definitions

Let μ_0 and μ_1 denote the acceptable and unacceptable process means for a production process. As a result, we wish to test the hypothesis H_0: $\mu = \mu_0$ against H_1: $\mu = \mu_1$. We assume that the quality characteristic is normally distributed with cumulative distribution function $\Phi(x;\mu,\sigma)$. Given that the underlying distribution is known, compressed-limit methods are applicable. Without loss of generality, we assume that $\sigma = 1$. Suppose that compressed-limit gauges are placed at $(\mu_0 + \mu_1)/2 \pm \Delta t$ as suggested by Beja and Ladany (1974). Observations are thus classified as belonging to one of three groups. Three-group data occur frequently in industry through the use of step-gauges and similar classification devices. A step-gauge classifies continuous observations into groups by comparing their dimension with a number of pins of different diameter rather than measuring them precisely. In industry, step-gauges, or other multigroup classification schemes, are used as an alternative when precise measurement is difficult or expensive (see Steiner *et al.*, 1994a).

Let $\pi_1(\mu)$, $\pi_2(\mu)$ and $\pi_3(\mu)$ denote the probabilities of an observation's being classified into each of the three

groups respectively. Then

$$\left.\begin{aligned}\pi_1(\mu) &= \int_{-\infty}^{(\mu_0+\mu_1)/2-\Delta t} \frac{1}{\sqrt{2\pi}}\exp\left(-(x-\mu)^2/2\right)dx,\\ \pi_2(\mu) &= \int_{(\mu_0+\mu_1)/2-\Delta t}^{(\mu_0+\mu_1)/2+\Delta t} \frac{1}{\sqrt{2\pi}}\exp\left(-(x-\mu)^2/2\right)dx,\\ \pi_3(\mu) &= \int_{(\mu_0+\mu_1)/2+\Delta t}^{\infty} \frac{1}{\sqrt{2\pi}}\exp\left(-(x-\mu)^2/2\right)dx.\end{aligned}\right\} \quad (3)$$

The symmetry of the normal distribution gives the following relations:

$$\left.\begin{aligned}\pi_1(\mu_0) &= \pi_3(\mu_1)\\ \pi_2(\mu_0) &= \pi_2(\mu_1)\\ \pi_3(\mu_0) &= \pi_1(\mu_1)\end{aligned}\right\} \quad (4)$$

Suppose that we sample one observation at a time. Relaxation of this assumption is discussed at the end of Section 4. Moustakides' result implies that we should use a CUSUM of the form

$$Y_j = \max(0,\ Y_{j-1} + z_j), \quad j = 1, 2, \ldots, \quad (5)$$

where

$$z_j = \ln(\pi_i(\mu_1)/\pi_i(\mu_0)) \quad (6)$$

if the jth observation falls into the ith group.

Using (4) we notice that $\ln(\pi_1(\mu_1)/\pi_1(\mu_0)) = -\ln(\pi_3(\mu_1)/\pi_3(\mu_0))$, and $\ln(\pi_2(\mu_1)/\pi_2(\mu_0)) = 0$, so that after suitable rescaling, the optimal CUSUM scheme has $z_j = 1$ if the jth observation is classified as belonging to group three, and $z_j = -1$ if the jth observation is classified as belonging to group one. If the jth observation falls into the second group, it offers no information regarding the relative merits of the two hypothesis and, using (4), $z_j = 0$.

Notice that the equivalent scoring system for the single compressed-limit CUSUM of Schneider and O'Cinneide (1987) may be obtained as a special case of the above by setting $\Delta t = 0$. Throughout the following, equivalent results for the single compressed-limit plans may be obtained by letting $\Delta t = 0$ so that the simple random walk with steps $\{-1, 0, 1\}$ becomes a random walk with steps -1 and 1 and with $\pi_2 = 0$.

3. Compressed-limit Wald tests

Consider a traditional Wald Sequential Probability Ratio Test with absorption boundaries at $-B$ and A, with A and $B > 0$, a starting value of zero and the scoring system proposed in Section 2. The test terminates at the nth trial, where n is the smallest integer for which either

$$\left.\begin{aligned}z_1 + \ldots + z_n &\le -B,\\ z_1 + \ldots + z_n &\ge A,\end{aligned}\right\} \text{ where } z_j = \left\{\begin{array}{ll}-1 & \text{with probability } \pi_1(\mu)\\ 0 & \text{with probability } \pi_2(\mu)\\ 1 & \text{with probability } \pi_3(\mu)\end{array}\right\}$$

Define $Z_n = \sum_{i=1}^{n} z_i$ as the terminating value of the SPRT, and assume that $\pi_1(\mu)$ and $\pi_3(\mu)$ are not equal to zero. Then, since the possible steps are $\{-1,0,1\}$, choosing the absorbing barriers as integer values means the SPRT cannot overshoot the absorbing barriers. Thus, the SPRT termination value Z_n must equal either A or $-B$. If Z_n terminates at A, then we decide in favor of μ_1, whereas if $Z_n = -B$, then we decide μ_0. We wish to determine the average sampling number (ASN) and operating characteristics for this test. Let $\xi_{-B} = \Pr(Z_n = -B)$ and $\xi_A = \Pr(Z_n = A)$ denote the probabilities of absorption at $-B$ and A respectively. Thus the probability of accepting the null hypothesis is ξ_{-B} and the probability of rejecting the null hypothesis is ξ_A.

The probabilities of absorption at A and $-B$ and the average number of steps to the boundary are derived by using the theory of the random walk (Cox and Miller, 1965). In particular, for a simple random walk with steps $\{-1, 0, 1\}$ and corresponding probabilities $\{\pi_1(\mu), \pi_2(\mu), \pi_3(\mu)\}$, the probabilities of absorption at A and $-B$ are given by

$$\xi_A = \left\{\begin{array}{ll}\pi_3(\mu)^A \dfrac{\pi_3(\mu)^B - \pi_1(\mu)^B}{\pi_3(\mu)^{A+B} - \pi_1(\mu)^{A+B}} & \text{for } \pi_1(\mu) \ne \pi_3(\mu),\\ \dfrac{B}{A+B} & \text{for } \pi_1(\mu) = \pi_3(\mu),\end{array}\right.$$

and

$$\xi_{-B} = \left\{\begin{array}{ll}\pi_1(\mu)^B \dfrac{\pi_3(\mu)^A - \pi_1(\mu)^A}{\pi_3(\mu)^{A+B} - \pi_1(\mu)^{A+B}} & \text{for } \pi_1(\mu) \ne \pi_3(\mu),\\ \dfrac{A}{A+B} & \text{for } \pi_1(\mu) = \pi_3(\mu).\end{array}\right.$$

These equations can be adjusted to reflect a SPRT starting at w with absorbing barriers at 0 and h, where $0 < w < h$. Let $P_{\text{reject}}(w)$ and $P_{\text{accept}}(w)$ equal the probabilities of the SPRT's ending at the rejection barrier h and acceptance barrier zero respectively, when the initial score of the SPRT is w. These acceptance and rejection probabilities follow directly from the above equations

$$P_{\text{reject}}(w) = \left\{\begin{array}{ll}\pi_3(\mu)^{h-w} \dfrac{\pi_3(\mu)^w - \pi_1(\mu)^w}{\pi_3(\mu)^h - \pi_1(\mu)^h} & \text{for } \pi_1(\mu) \ne \pi_3(\mu),\\ \dfrac{w}{h} & \text{for } \pi_1(\mu) = \pi_3(\mu),\end{array}\right. \quad (7)$$

and

$$P_{\text{accept}}(w) = \begin{cases} \pi_1(\mu)^w \dfrac{\pi_3(\mu)^{h-w} - \pi_1(\mu)^{h-w}}{\pi_3(\mu)^h - \pi_1(\mu)^h} & \text{for } \pi_1(\mu) \neq \pi_3(\mu) \\ \dfrac{h-w}{h} & \text{for } \pi_1(\mu) = \pi_3(\mu) \end{cases} \tag{8}$$

With this formulation, the possible termination values, denoted Z_n, are $-w$ and $h - w$.

Thus, suppressing the dependence on μ, the moments of Z_n are

$$E(Z_n^k) = \begin{cases} \dfrac{(h-w)^k \pi_3^{h-w}(\pi_3^w - \pi_1^w) + (-w)^k \pi_1^w(\pi_3^{h-w} - \pi_1^{h-w})}{\pi_3^h - \pi_1^h} & \text{for } \pi_1 \neq \pi_3, \\ \dfrac{w(h-w)^k + (h-w)(-w)^k}{h} & \text{for } \pi_1 = \pi_3. \end{cases}$$

Wald (1947) gave a general result to find the average sampling number of SPRTs:

$$\text{ASN} = \begin{cases} E(Z_n)/E(z) & \text{for } E(z) \neq 0, \\ E(Z_n^2)/\text{Var}(z) & \text{for } E(z) = 0. \end{cases}$$

In our application, $E(z) = 0$ only if $\pi_1 = \pi_3$. Thus, the average number of steps to absorption for an SPRT starting at w is

$$\text{ASN}(w, h, \mu) = \begin{cases} \dfrac{h(\pi_3^h - \pi_1^w \pi_3^{h-w})}{(\pi_3 - \pi_1)(\pi_3^h - \pi_1^h)} - \dfrac{w}{\pi_3 - \pi_1} & \text{for } \pi_1 \neq \pi_3, \\ \dfrac{w(h-w)}{2\pi_1} & \text{for } \pi_1 = \pi_3. \end{cases} \tag{9}$$

4. Compressed-limit CUSUM control charts

We derive the average run length properties of the CUSUM plan by noticing that a CUSUM scheme is a sequence of Wald sequential probability ratio tests with initial score zero and boundary h (Fig. 1). The average run length of the CUSUM scheme is given by

$$\text{ARL} = \text{ASN}(w = 0)/(1 - P_{\text{accept}}(0)), \tag{10}$$

where $\text{ASN}(w = 0)$ and $P_{\text{accept}}(0)$ are the average sampling number and probability of acceptance of the corresponding Wald test with initial score zero (Page, 1954). This follows from the fact that, at termination of the CUSUM at h, a geometric number of Wald tests have been observed.

To obtain the average sampling number and the probability of acceptance of a single Wald test when the starting value is zero, we condition on the outcome of the first observation. If the first step is -1 or 0, the Wald test ends in acceptance; if, on the other hand, the first step is $+1$, then the probability of acceptance is given by (8). Therefore the probability of acceptance with a starting value of zero is given by

$$P_{\text{accept}}(0) = \pi_1 + \pi_2 + \pi_3 \, P_{\text{accept}}(1);$$

$$P_{\text{accept}}(0) = \begin{cases} \dfrac{\pi_3^h(1 + \pi_1 - \pi_3) - \pi_1^h}{\pi_3^h - \pi_1^h} & \text{for } \pi_1 \neq \pi_3, \\ 1 - \dfrac{\pi_1}{h} & \text{for } \pi_1 = \pi_3. \end{cases} \tag{11}$$

and the average sampling number when starting at zero is

$$\text{ASN}(w = 0) = \pi_1 + \pi_2 + \pi_3 \, (\text{ASN}(w = 1) + 1);$$

$$\text{ASN}(w = 0) = \begin{cases} \dfrac{h\pi_3^h}{\pi_3^h - \pi_1^h} - \dfrac{\pi_1}{\pi_3 - \pi_1} & \text{for } \pi_1 \neq \pi_3, \\ \dfrac{h+1}{2} & \text{for } \pi_1 = \pi_3. \end{cases} \tag{12}$$

By using the results given in (8), (9), (11) and (12) it is possible to solve for the exact ARL of any two- or three-step CUSUM scheme through (10). This gives

$$\text{ARL} = \begin{cases} \dfrac{\pi_1^{h+1} - \pi_1 \pi_3^h}{\pi_3^h(\pi_3 - \pi_1)^2} + \dfrac{h}{\pi_3 - \pi_1} & \text{for } \pi_1 \neq \pi_3, \\ \dfrac{h(h+1)}{2\pi_1} & \text{for } \pi_1 = \pi_3. \end{cases} \tag{13}$$

The ARL of a CUSUM that utilizes the FIR feature is also easily obtainable if the different stages shown in Fig. 1 are considered. If the CUSUM is set to have a FIR initial value of w the ARL of the FIR CUSUM can be obtained based on the previous results if we consider the outcome of stage I. If stage I ends in rejection, the CUSUM consists only of a single SPRT starting at w and ending at absorbing barrier h. If, on the other hand, the initial stage ends with acceptance, the remaining CUSUM is identical to the standard CUSUM with starting value zero. Thus, the ARL of the FIR CUSUM with head start w can be written in terms of (7)–(9) and (13), namely

$$\text{ARL}_{\text{FIR}} = (\text{ARL} + \text{ASN}(w))P_{\text{accept}}(w) + \text{ASN}(w)P_{\text{reject}}(w);$$

$$\text{ARL}_{\text{FIR}} = \text{ASN}(w) + P_{\text{accept}}(w)\text{ARL};$$

$$\mathrm{ARL_{FIR}} = \begin{cases} \dfrac{\pi_1^{h+1} - \pi_1^{w+1}\pi_3^{h-w}}{\pi_3^h(\pi_3 - \pi_1)^2} + \dfrac{h-w}{\pi_3 - \pi_1} & \text{for } \pi_1 \neq \pi_3, \\ \dfrac{h(h+1) - w(w+1)}{2\pi_1} & \text{for } \pi_1 = \pi_3. \end{cases} \tag{14}$$

Notice that if ARL is the average run length of a single-observation compressed-limit CUSUM, then the average run length of such a scheme when observations are sampled n units at a time is merely ARL/n as long as observations are entered into the CUSUM one at a time. Clearly, curtailment within the samples is feasible because the scoring system is simple.

5. Optimal symmetric double compressed limits

In this article we assume that the two gauge limits are placed symmetrically about the midpoint between the acceptable and unacceptable means of the process, namely at $(\mu_0 + \mu_1)/2 \pm \Delta t$. As shown by Beja and Ladany (1974), this symmetric placement is optimal when the type I and II error rates of our hypothesis test, denoted α and β respectively, are equal. However, the question of the best Δt remains. Clearly, if the value of Δt is too small, there will be little additional information and only a small advantage over the two-group plan. Similarly, if Δt is too large, there will effectively be only one group, and the resultant tests and charts may perform poorly compared with when $\Delta t = 0$. Below we derive the optimal Δt value for SPRTs and CUSUM procedures.

An SPRT is characterized by its probability of making a type I or II error, and its ASN for various parameter values. In our application, the goal of the SPRT is distinguish between the null ($\mu = \mu_0$) and alternate ($\mu = \mu_1$) hypotheses. Eqns (7) and (8) give the actual error rates $\alpha_{act} = P_{reject}(w;\mu = \mu_0$ and $\beta_{act} = P_{accept}(w,\mu = \mu_1)$, and (9) yields the ASN. The SPRT will have relatively short ASN values when the mean value is at either the null or alternate mean value. Typically, these ASN values are significantly smaller than the sample size of the corresponding fixed sample size procedure, hence the rationale for considering sequential methods. However SPRTs can be criticized because the sample size required is not known with certainty, and may be large. Based on maximizing (9), the SPRT has its largest ASN at $\mu = (\mu_0 + \mu_1)/2$. At this value, the random walk has no trend. As a result, it makes sense to place the gauge limits so as to minimize the ASN when $\mu = (\mu_0 + \mu_1)/2$ subject to certain restrictions on the error rate.

Using (7)–(9), and defining α and β as the desired maximum error rates, the optimization problem is to find the Δt, h and w that satisfy:

$$\underset{\Delta t,\, h,\, w}{\text{Minimize}}\ \mathrm{ASN}\left(\mu = \tfrac{1}{2}(\mu_0 + \mu_1)\right) = w(h-w)/2\pi_1$$

$$\text{subject to } P_{reject}(\mu = \mu_0) \leq \alpha \text{ and } P_{accept}(\mu = \mu_1) \leq \beta,$$

where h is the upper absorbing barrier of the SPRT, and w is the starting value.

Notice that as h and/or Δt increase, ASN increases; however, both the actual error rates decrease. Also, both h and w must be integers with $h \geq 2$, $h > w \geq 1$, and $\Delta t \geq 0$. This problem can be solved by considering all feasible h and w combinations and incrementing h from $h = 2$. Since we assume $\alpha = \beta$, the search can be restricted to combinations of (h,w) where $w = \lfloor h/2 \rfloor$ ($w = \lceil h/2 \rceil$ would also do, but yields smaller α_{act} values). Fortunately, the number of h values that need to be considered can be bounded. For large h, the error rate constraints already hold when $\Delta t = 0$. Because further increases in h result in a further decrease in the error rates and an increase in the ASN there is no need to consider any h values that are larger. The Δt, h and w combination that yields the lowest ASN at $\mu = (\mu_0 + \mu_1)/2$ is the optimal solution.

A simple extension of this optimization methodology can be used to find the optimal symmetric gauge limit placement when $\alpha \neq \beta$. For each h value there are $(h-1)$ possible values for w. For each feasible (h,w) combination we perform two simple one-dimensional searches to find the smallest Δt value that satisfies $P_{reject}(\mu = \mu_0) \leq \alpha$, and the smallest Δt value that satisfies $P_{accept}(\mu = \mu_1) \leq \beta$. The larger of these two Δt values is the smallest Δt that satisfies both error rate constraints for the given h and w values. Thus, this Δt value also yields the best ASN($\mu = (\mu_0 + \mu_1)/2$) for the given combination of h and w. The number of h and w combinations that need to be considered can be reduced by noticing that, owing to symmetry, if $w < h/2$ then $\beta_{act} > \alpha_{act}$, and if $w > h/2$ then $\alpha_{act} > \beta_{act}$. As a result, if we desire error rates such that $\alpha \leq \beta$, then the best value for w must be less than or equal to $h/2$ and vice versa for $\alpha \geq \beta$. Notice that when $\alpha \neq \beta$ there may be a better non-symmetric gauge limit placement, although unless α and β are very different the optimal symmetric placement should be close to the global optimum. When $\alpha = \beta$, the optimal symmetric gauge limit placement is also the global optimal gauge limit placement.

Letting Δt^*, h^* and w^* denote the optimal values, the results of this optimization for various mean shifts and desired error rate values are shown in Table 1. In Table 1, ASN* denotes the optimal ASN of the SPRT at $\mu = (\mu_0 + \mu_1)/2$. The corresponding ASN at mean values of μ_0 or μ_1 are significantly lower. The optimal Δt values vary considerably. However, much of the variation is due to the discreteness inherent in the problem because h and w must be integers. When $\alpha = \beta$ and the values for h and w are fairly large, the optimal values for Δt range between

Table 1. Optimal symmetric gauge limit design, SPRT barriers and starting values for various mean shifts and type I and II error rates

		Mean shift (in standard deviation units)											
		0.5σ				*σ*				*1.5σ*			
α	β	Δt^*	h^*	w^*	ASN^*	Δt^*	h^*	w^*	ASN^*	Δt^*	h^*	w^*	ASN^*
0.001	0.001	0.6521	22	11	235.2	0.5044	12	6	58.6	0.4926	8	4	25.7
0.001	0.005	0.5469	21	9	184.8	0.8099	9	4	47.8	0.5286	7	3	20.1
0.0025	0.0025	0.5747	20	10	176.8	0.5681	10	5	43.9	0.7355	6	3	19.5
0.0025	0.0125	0.4341	19	8	132.5	0.5665	9	4	35.0	0.2725	7	3	15.3
0.005	0.005	0.5457	18	9	138.4	0.7352	8	4	34.6	0.5273	6	3	15.1
0.005	0.025	0.3778	17	7	99.2	0.7305	7	3	25.8	0.5964	5	2	10.9
0.01	0.01	0.5083	16	8	104.7	0.5012	8	4	26.0	0.3082	6	3	11.9
0.01	0.05	0.5711	13	5	70.4	0.9858	5	2	18.5	0.2957	5	2	7.8

0.5 and 0.7.

For CUSUM procedures, the best choice for Δt depends upon the size of the mean shift that one is trying to detect, and the desired average run length properties. In general, given a minimum average run length when the mean is at the acceptable level, we wish to find an h and a Δt that minimize the average run length when the mean is at the unacceptable level. Symbolically, we have the following optimization problem.

$$\underset{\Delta t,\, h}{\text{Minimize}}\ \text{ARL}(\mu_1)$$

$$\text{subject to } \text{ARL}(\mu_0) = \text{ARL}_0.$$

The optimal values of Δt and h may be easily obtained by using the fact that the average run length of the simple random walk CUSUM is monotonically increasing in both Δt and h, and that only a finite number of integer values for h need be considered. Given a CUSUM barrier h, the minimum average run length is obtained when $\Delta t = 0$. Begin by letting $\Delta t = 0$, and find the maximum feasible h for which $\text{ARL}(\mu_0) \leq \text{ARL}_0$. For each feasible integer h, find the value of Δt for which the constraint achieves equality. Finally, compute $\text{ARL}(\mu_1)$ for each of these $(h,\Delta t)$ combinations and select that which leads to a minimum $\text{ARL}(\mu_1)$.

Let Δt^* and h^* denote the optimal solution. Table 2 gives the values of Δt^* and h^* and the corresponding $\text{ARL}_1{}^*$ obtained for various mean shifts and desired in-control average run lengths. For moderate mean shifts, Δt^* will be between 0.5σ and 1σ and the optimal CUSUM barrier h^* will be between 2 and 10, depending upon the required in-control average run length. In general, we notice that the optimal gauge limits will be outside the zone of indifference between μ_0 and μ_1. For fixed h, smaller average run lengths are obtained by bringing the gauge limits closer to the acceptable and rejectable mean levels. Optimal Δt values for the CUSUM procedure are, for the most part, quite similar to the optimal values obtained for the SPRTs. For both procedures the optimal values are affected to a significant degree by the discreteness required in h. However, choosing Δt between 0.5 and 0.65 is never far from optimal.

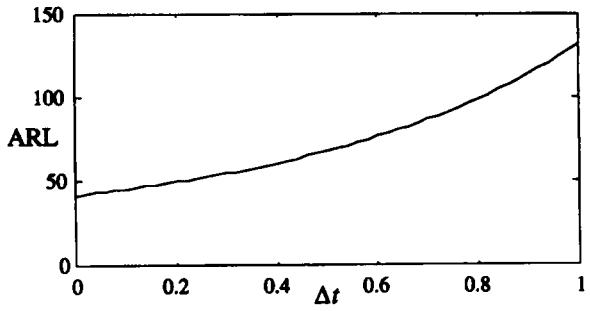

Fig. 2. ARL versus Δt for $\mu = (\mu_0 + \mu_1)/2$ and $h = 6$.

6. Example

Let us compare the ARLs of CUSUM schemes using two

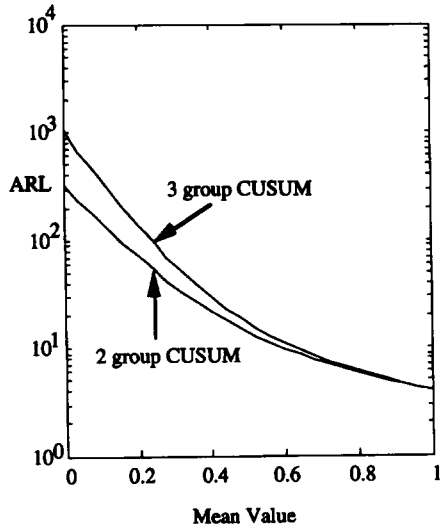

Fig. 3. ARL on log scale versus true mean value.

Table 2. Optimal symmetric gauge limit design and CUSUM barriers for various mean shifts and in-control average run lengths

	Mean shift (in standard deviation units)								
	0.5σ			σ			1.5σ		
ARL_0	Δt^*	h^*	$ARL_1{}^*$	Δt^*	h^*	$ARL_1{}^*$	Δt^*	h^*	$ARL_1{}^*$
250	0.6283	5	23.88	0.7688	3	9.10	0.9414	2	4.97
500	0.6315	6	29.95	0.5643	4	10.79	0.5365	3	5.76
750	0.5434	7	33.77	0.6858	4	11.74	0.6478	3	6.14
1000	0.6430	7	36.35	0.7681	4	12.51	0.7236	3	6.45
2000	0.6569	8	42.97	0.6073	5	14.19	0.8973	3	7.34
3000	0.5912	9	46.90	0.7062	5	15.20	0.5373	4	7.82
4000	0.6706	9	49.72	0.7739	5	16.00	0.5989	4	8.09
5000	0.5645	10	51.95	0.5392	6	16.58	0.6457	4	8.31

and three groups. Choosing $\mu_0 = 0$, $\mu_1 = 1$, and $\sigma = 1$ we will attempt to detect a 1σ shift in the mean of a normal distribution with a sample of size 4. We consider two cases: a two-group example with a gauge limit placed at 0.5, and a three-group example with gauge limits placed symmetrically about 0.5. Using (13) we can derive the ARL of the CUSUM procedure at various parameter values. Fig. 2 illustrates the effect of various Δt values for the three-group case when the actual mean value is $(\mu_0 + \mu_1)/2$ and the absorbing barrier is $h = 6$. Notice that the ARL for the two-group case is given when $\Delta t = 0$. For subsequent analysis on the three-group example we use gauge limits placed at zero and unity (i.e. $\Delta t = 0.5$).

Table 3 shows the in-control ARL, denoted ARL_0, and the out-of-control ARL, denoted ARL_1, derived from (13), for the two-group and three-group case for various values of h. For example, using three groups and $h = 6$ gives $ARL(\mu = \mu_0, n = 4) = 4183.2/4 = 1045.8$ and $ARL(\mu = \mu_1, n = 4) = 16.2/4 = 4.05$. Similarly, with a single gauge limit at 0.5 and $h = 7$ the corresponding average run lengths are $ARL(\mu = \mu_0, n = 4) = 328.9$ and $ARL(\mu = \mu_1, n = 4) = 4.05$. These two particular example cases have identical ARL out of control, but the three-group CUSUM has a significantly better average run length in control than the two-step scheme. Fig. 3 plots the ARL for the given two- and three-group examples for various true mean values. The results for the three-group CUSUM can be improved by considering the optimal placement of the group limits. Using group limits placed at (−0.2895, 1.2895), i.e. $\Delta t = 0.7895$, with a sample of size 4 leads to the optimal result that gives the largest in-control ARL with an out-of-control ARL of 16.2/4 = 4.05. The corresponding in-control ARL, $ARL(\mu = \mu_0, n = 4)$, is 4278.5/4 = 1069.6. This value is slightly larger than that obtained when using group limits at (0, 1). Notice also that the optimal result is not dependent on the sample size chosen.

Table 3. In-control and out-of-control ARLs for two-group and three-group CUSUM procedures when $n = 4$

h	*Two groups*		*Three groups*	
	ARL_0	ARL_1	ARL_0	ARL_1
2	3.4	0.88	8.1	1.16
3	10.1	1.48	30.3	1.88
4	25.9	2.11	101.8	2.59
5	62.2	2.75	328.8	3.32
6	144.2	3.40	1045.8	4.05
7	328.9	4.05	3307.0	4.79
8	743.5	4.70	10434.9	5.52
9	1673.5	5.35	32900.9	6.25
10	3756.5	6.00	103707.0	6.98

7. Conclusions

Singly and doubly compressed sequential probability ratio tests (SPRTs) and cumulative sum (CUSUM) schemes to detect one-sided shifts in the mean of a normal distribution are presented. Explicit formulas for the average run lengths of the SPRTs and CUSUM schemes are determined using conditioning arguments and the theory of the random walk. These results are also applicable for detecting shifts in the mean of any symmetric distribution. The three-group (doubly compressed group limits) SPRTs and CUSUM procedures have a significant advantage in terms of average run length over the two-group schemes. In addition, a discussion of the optimal group limits for the three-group SPRTs and CUSUM procedures to detect shifts in a normal mean is presented.

Acknowledgments

This research was supported, in part, by the Natural Sciences and Engineering Research Council of Canada. The authors also thank two anonymous referees. The

comments and suggestions they made greatly improved the presentation of the material.

References

Beattie, D.W. (1962) A continuous acceptance sampling procedure based upon a cumulative sum chart for the number of defectives. *Applied Statistics*, **11**, 137–147.

Beja, A. and Ladany, S.P. (1974) Efficient sampling by artificial attributes. *Technometrics*, **16**, 601–611.

Cox, D.R. and Miller, H.J. (1965) *The Theory of Stochastic Processes*, Chapman & Hall, London.

Duncan, A.J. (1986) *Quality Control and Industrial Statistics*, 5th edn, Richard D. Irwin, Homewood, IL.

Elder, R.S., Provost, L.P. and Ecker, O.M. (1981) United States Department of Agriculture CUSUM acceptance sampling procedure. *Journal of Quality Technology*, **13**, 59–64.

Evans, I.G. and Thyregod, P. (1985) Approximately optimal narrow limit gauges. *Journal of Quality Technology*, **17**, 63–66.

Ladany, S.P. (1976) Determination of optimal compressed limit gaging sampling plans. *Journal of Quality Technology*, **8**, 225–231.

Lucas, J.M. (1982) Combined Shewhart–CUSUM quality control schemes. *Journal of Quality Technology*, **14**, 51–59.

Lucas, J.M. and Crosier, R.B. (1982) Fast initial response for CUSUM quality control schemes: give your CUSUM a head start. *Technometrics*, **24**, 199–205.

Moustakides, G.V. (1986) Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, **14**, 1379–1387.

Page, E. (1954) Continuous inspection schemes. *Biometrika*, **41**, 100–115.

Schneider, H. and O'Cinneide, C. (1987) Design of CUSUM control charts using narrow limit gauges. *Journal of Quality Technology*, **19**(April), 63–68.

Shewhart, W.A. (1931) *Economic Control of Quality of Manufactured Product*, D. Van Nostrand, New York.

Steiner, S.H., Geyer, P.L. and Wesolowsky, G.O. (1994a) Control charts based on grouped data. *International Journal of Production Research*, **32**, 75–91.

Steiner, S.H., Geyer, P.L. and Wesolowsky, G.O. (1994b) Shewhart control charts to detect mean shifts based on grouped data. *Journal of Quality and Reliability International* (submitted).

Stevens, W.L. (1947) Control by gauging. *Journal of the Royal Statistical Society*, **B 1**, 54–108.

Sykes, J. (1981) A nomogram to simplify the choice of a sampling plan using a single gauge. *Journal of Quality Technology*, **13**, 36–41.

Wald, A. (1947) *Sequential Analysis*, John Wiley & Sons, New York.

Biographies

P. Lee Geyer is a Ph.D. student at McMaster University in the Management Science/Systems Area. He also consults extensively in the area of health care funding and utilization analysis. His research interests include process control and statistical modelling.

Stefan H. Steiner is an assistant professor in the department of statistics and actuarial sciences at the University of Waterloo. He obtained his Ph.D. in Management Science from McMaster University and M.Sc. and BMATH degrees from the University of British Columbia and the University of Waterloo respectively. His research interests included industrial statistics and applications of statistics in operations research. Dr Steiner is a member of ASQC and INFORMS.

George O. Wesolowsky is a professor of management science at McMaster University in Hamilton Ontario Canada. His research interests include quality control and location models.