GROUPED DATA SEQUENTIAL PROBABILITY RATIO TESTS AND CUMULATIVE SUM CONTROL CHARTS

Stefan Steiner
University of Waterloo,
P. Lee Geyer and George Wesolowsky
McMaster University

IIQP Research Report
RR-95-07

May 1995

Grouped Data Sequential Probability Ratio Tests and Cumulative Sum Control Charts

Stefan H. Steiner

Dept. of Statistics and Actuarial Sciences
University of Waterloo
Waterloo, Ontario N2L 3G1
Canada

P. Lee Geyer and George O. Wesolowsky

Faculty of Business
McMaster University
Hamilton, Ontario L8S 4M4
Canada

Abstract

Methodology is proposed for the design of sequential methods when data is obtained by gauging articles into many intervals. While variables data are more efficient than simple attribute data, the use of multiple categories compensates for the loss in efficiency due to imprecise measurement. Exact expressions are obtained for the operating characteristics and average sampling number of Wald tests, and for the average run length properties of Cumulative Sum (CUSUM) schemes based on parametric multinomial data. The methods proposed are simple to implement, and are an economical alternative to variables sequential sampling plans and CUSUM control charts.

Keywords: CUSUM, Fast Initial Response (FIR), Grouped Data, Parametric Multinomial, SPRT.

1. Introduction

It is not always possible or practical to use variables measurement data in quality control. The predominance of attributes type data in industry attests to the economic advantages of collecting go no-go data over exact measurements. Gauging is often preferred over measurement since it takes less skill to gauge properly, is faster, less costly, and is a tradition in certain industries (Schilling, 1982). Ladany (1976) points out that the statistical advantages of variables data may be outweighed by economic considerations since the cost of inspection using a simple go no-go gauge is often much lower than the cost of determining the exact value of a critical characteristic variable.

Work by Stevens (1948), Dudding and Jennett (1944), Mace (1952) and Ott and Mundel (1954) attempts to bridge the gap between variables and attributes procedures by proposing methodology that utilizes go no-go gauges set at artificial levels. The basic idea behind this methodology is that the classification of units as defective or non-defective is inefficient when the proportion of nonconforming units is small. For example, since the sample size required for an attributes plan is inversely related to the size of the proportion nonconforming it is required to detect, a gauge limit that classifies a higher proportion of items as non-conforming (pseudo-nonconforming) will be statistically more efficient, and offer more information about the characteristic of interest. The focus of much of this research has been the testing or control of the mean of a normal distribution.

More recently, Lucas (1985) described design and implementation procedures for counted data which are designed to detect increases and decreases in the count level of data. Schneider and O'Cinneide (1987) proposed a CUSUM scheme for monitoring the mean of a normal distribution with a single compressed limit gauge. They determine solutions based on the normal approximation to the binomial. Geyer, Steiner and Wesolowsky (1995) extended this CUSUM to the use of two compressed limit gauges placed symmetrically about the midpoint between the target mean and the mean that the chart is intended to detect. The Geyer *et al.* (1995) solutions are exact and are derived through the theory of the random walk. Steiner, Geyer and Wesolowsky (1994, 1995) developed methodology for one-sided and two-sided acceptance sampling plans, acceptance control charts and Shewhart type control charts using multiple go no-go type gauges also called step gauges. Step gauges typically consist of pins of various diameters that allow the classification of units into one of many groups. These charts are based upon the ratio of the parametric multinomial sample likelihoods under two simple hypothesis.

In this article we derive SPRTs and CUSUM Procedures based on grouped data. Section 2 introduces the proposed scoring procedure. In Section 3 we consider the design and implementation of Parametric Multinomial Sequential Probability Ratio Tests (PM-SPRT) for testing simple hypotheses about a parameter of interest when data is gauged into multiple intervals and the probability distribution of the quality characteristic is known. We suggest approximating the log-likelihood of the resulting random walk with a simple integer scoring system. The PM-SPRT based on gauging data into several groups bridges the gap between the efficiency of attribute and compressed limit sequential procedures and that of variables sequential sampling plans. Using the theory of sequential analysis, Wald (1947), we derive exact expressions for the Operating Characteristics (OC) and the Average Sampling Number (ASN) of sequential plans utilizing multiple gauges.

The design and implementation of Parametric Multinomial CUSUM quality control schemes (PM-CUSUM) is discussed in Section 4. Following Page (1954), we consider the proposed PM-CUSUM as a sequence of PM-SPRTs and derive the Average Run Length (ARL) properties of the PM-CUSUM using the operating characteristics and average sampling number properties of the individual PM-SPRTs. We also give expressions which are appropriate when the Fast Initial Response (FIR) feature recommended by Lucas and Crosier (1982) is used. A CUSUM plan that utilizes multiple groups will be more efficient than a CUSUM plan utilizing a single compressed limit gauge, and may be more economical than a CUSUM based on exact measurement.

Sections 5 and 6 turn to practical considerations that arise when applying this methodology. Section 5 discusses various design issues, including the choice of group limits, and the performance of these grouped data approaches relative to a variables based approach. Section 6 presents a step by step design procedure and an example drawn from application in a progressive die environment. For simplicity, the analysis in Sections 2-6 assumes a unit sequential implementation of the procedures. Section 7 shows that adapting the procedure to samples of size n is relatively straightforward.

2. A Sequential Scoring Procedure for Grouped Data

Whenever data are grouped, the need arises to assign the grouped observations a numerical value based on the group they are classified into. For go no/go gauges, observations are usually treated singly as Bernoulli random variables, being either conforming or nonconforming. When observations are grouped into multiple intervals, the likelihood ratio suggests a scoring system. The likelihood ratio is utilized since it has great prominence as a measure of statistical evidence in traditional hypothesis testing, sequential sampling, and in the development of CUSUM control charts. For example, in the area of traditional hypothesis testing, the Neyman-Pearson lemma implies that the likelihood ratio test is the most powerful test for comparing simple hypotheses. Also, Wald (1947) showed that a similar optimality property applies to the use of the likelihood ratio in sequential sampling: the Sequential Probability Ratio Test (SPRT) minimizes the ASN under H_0 and H_1 among all sequential tests for given error probabilities. More recently, Moustakides (1986) proved that the CUSUM procedure based on the likelihood ratio minimizes the Average Run Length (ARL) under H_1 for a given ARL under H_0. This is expected since a CUSUM control chart is a sequence of Wald tests (Page 1954, Johnson 1961, Kemp 1971).

In the case of the simple hypothesis test $H_0: \theta = \theta_0$ versus $H_1: \theta = \theta_1$, the likelihood ratio is given by the ratio of the likelihood of the data under H_1 to the likelihood of the data under H_0. When observations are grouped into k intervals, the likelihood ratio is a ratio of multinomial likelihoods, where the group probabilities depend upon the parameter specifications in the underlying probability distributions under H_0 and H_1. Specifically, let the random variable X have probability distribution $f(x;\theta)$ and cumulative distribution function $F(x;\theta)$, where θ is a parameter of interest. Let $t_1 < t_2 < \ldots < t_{k-1}$ denote the $k-1$ endpoints of the k grouping intervals. We assume for the moment that the $k-1$ group intervals are given. In many applications the grouping criterion are predetermined since it is based on some standard classification device or procedure. In Section 5 this assumption is relaxed and the optimal placement of group limits for detecting shifts in a normal mean is discussed. Defining $t_0 = -\infty$ and $t_k = \infty$, the probability that an observation falls into the jth interval is denoted by

$$\pi_j = F(t_j;\theta) - F(t_{j-1};\theta), \qquad j = 1,2,\ldots,k \qquad (2.1)$$

the dependence of π_j on θ being understood. The contribution to the log-likelihood ratio of an observation that falls into the jth interval is thus given by the weight

$$\ell_j = \ln\left\{\pi_j(\theta_1)/\pi_j(\theta_0)\right\}, \qquad j = 1,2,\ldots,k \qquad (2.2)$$

We assume that for implementation group scores are obtained by first scaling and then rounding off the exact likelihood ratio weights ℓ_j. This is necessary because the properties of sequential procedures based on integer scores can be found exactly through our analysis presented in Sections 3 and 4. To reflect this discretization, let w_j denote the group score applied to any observation in the jth group, i.e. $w_j = \text{round}(q\ell_j)$ where q is the chosen scaling factor, and $j = 1,2,\ldots,k$. Define $\mathbf{w} = (w_1, w_2, \ldots, w_k)$. We assume that all w_j scores are unique; if two or more groups lead to the same score either the scaling factor should be increased or groups should be combined. Due to the rounding of log-likelihood ratio weights, the resulting schemes are only approximately based upon the optimal sequential probability ratio. Notice, however that the properties of the resulting random walk can be made arbitrarily close to that implied by the exact likelihood ratio scores by simply increasing the scale factor. In subsequent sections we utilize the fact that so long as the number of groups is greater than or equal to two and $\theta_0 \neq \theta_1$ at least one individual score is positive and at least one is negative. This implies $\max(\mathbf{w}) > 0$ and $\min(\mathbf{w}) < 0$ and ensures that the SPRTs and CUSUM schemes are capable of concluding either in favor of the null or the alternative hypothesis.

A tradeoff is involved in the appropriate choice of scaling factor q. The solution approach requires integer scores, and is less computational intensive to design (and easier to implement) when the scores are as close to zero as possible. However, on the other hand, we wish to stay as close as possible to the optimal relative weights suggested by the likelihood ratio. We have found that, in most cases, choosing a scaling factor so that the spread in the sample scores ($\max(\mathbf{w}) - \min(\mathbf{w})$) is approximately 50 yields results that are indistinguishable from simulation results using the exact weights, i.e. set $q = 50/[\max(\ell) - \min(\ell)]$. Naturally, smaller scaling factors are also feasible, but may yield slightly inferior results. However, in all cases, it is the relative size of the scores that drives the solution. For this reason, if the resulting group scores have a common factor, all the scores can be divided by this factor without affecting the efficiency of the solution.

3. Sequential Tests with Grouped Data

Consider the sequential test for grouped data which prescribes computing the sample score $s = w_j$ if the sampled unit is classified into group j. Choosing absorbing barriers at log$B$ and log$A$, and denoting the ith sample score as s_i, the sampling terminates on the Nth trial where N is the smallest integer for which either

$$S = s_1 + s_2 + \ldots + s_N \geq \log A$$

or

$$S = s_1 + s_2 + \ldots + s_N \leq \log B$$

where $0 < B < 1 < A < \infty$. If $S \geq \log A$ we conclude θ_1, whereas if $S \leq \log B$ decide in favor of θ_0. Since the observations are all independent and identically distributed, the sequence $S = s_1 + \ldots + s_N$ can be viewed as a random walk with steps $\mathbf{w}$ between absorbing barriers at log$B$ and log$A$. Since the steps can take on only a finite number k of integer values, we may use the theory of Sequential Analysis (Wald, 1947) to derive the exact operating characteristics and average sampling number of the above decision procedure.

A derivation of the operating characteristics and average sampling number of this SPRT is possible if we first derive the probability distribution for all possible terminating values of the SPRT. See the Appendix for a derivation of $\xi_j = \Pr(S = c_j)$ where the vector $\mathbf{c} = (c_1, c_2, \ldots, c_d)$ denotes all the possible terminating values of the SPRT including overshoots of the absorbing barriers. Since all w_j are integer, there are a finite number of possible terminating values where the number is based on the range in the weights and the scale of the absorbing barriers. Let $[a]$ be the smallest integer $\geq$ log$A$ and $[b]$ the largest integer $\leq$ log$B$. Then the probability that the random walk terminates with $S \leq [b]$, and thus accepts the null hypothesis is given by

$$P_{accept}(\theta; [a],[b]) = \sum_{j \in \mathbf{c}^-} \xi_j \tag{3.1}$$

where the sum is over all j for which we conclude in favor of the null hypothesis, i.e. all j such that $c_j \leq [b]$. This expression allows the determination of the operating characteristic curve of the sequential test.

Using the probability distribution of S, and Wald's equations, we may derive the exact average sampling number of the sequential test. By Wald's first equation (Wald, 1947 (A:69)) if $E(N) < \infty$ and $E(s) \neq 0$, then

$$E(N; [a],[b]) = E(S)/E(s) = \frac{\sum_{j=1}^{d} \xi_j c_j}{\sum_{j=1}^{k} \pi_j w_j} \tag{3.2}$$

Wald (1947) also showed that if $E(s)=0$ and $E(s^2)<\infty$, then

$$E(N;[a],[b]) = E(S^2)/E(s^2) = \frac{\sum_{j=1}^{d}\xi_j c_j^2}{\sum_{j=1}^{k}\pi_j w_j^2} \tag{3.3}$$

where d represents the number of possible terminating values of the SPRT. The above formulas are valid for an SPRT with initial value zero and absorbing barriers at $[b]<0$ and $[a]>0$. Through a translation this is equivalent to an SPRT with initial value v, $0<v<h$, and absorbing barriers at zero and h. This translation is of interest since the traditional $(0,h)$ CUSUM chart can be modeled as a geometric series of $(0,h)$ SPRTs. For any given values of v and h setting $[a] = h-v$ and $[b]=-v$ in Equations (3.1)-(3.3) yields the desired results. Define $\mathrm{E}(N;v)$ and $P_a(v)$ as the expected sampling number and probability of concluding in favor of the null hypothesis respectively for a $(0,h)$ SPRT with initial score v, $0<v<h$. Then $\mathrm{E}(N;v) = E(N;[a]=h-v,[b]=-v)$ and $P_a(v) = P_{accept}(\theta;[a]=h-v,[b]=-v)$ as given by Equations (3.1) and (3.2) or (3.3).

4. CUSUM Control Charts with Grouped Data

CUSUM control charts exhibit better run length properties than Shewhart Control Charts for the detection of small changes in a parameter, since they accumulate evidence that may not be adequately reflected in a single sample. CUSUM charts consist of plotting $Y_i = \max\left(0, Y_{i-1} + \ln\left(\frac{f(x;\theta_1)}{f(x;\theta_0)}\right)\right)$, where $Y_0=0$. The process is deemed to be in state H_0 as long as $Y_i<h$, and is deemed to be in state H_1 as soon as $Y_i \geq h$. The CUSUM may be seen to be a sequence of Wald tests with initial score zero, and absorbing barriers at zero and h (Page, 1954). It is easy to show that the ARL of a CUSUM chart is given by

$$ARL = \frac{E(N;\, v=0)}{1-\mathrm{P}_a(v=0)}, \tag{4.1}$$

where $\mathrm{E}(N;\ v=0)$ and $P_a(v=0)$ are the average sampling number and probability of acceptance of the component $(0,h)$ Wald tests with starting value zero (Page, 1954). Unfortunately, $\mathrm{E}(N;\ v=0)$ and $P_a(v=0)$ are not directly obtainable from (3.1)-(3.3) since those expressions are derived assuming the SPRT starting value is not equal to lower barrier values, i.e. $v>0$. However, expressions for $\mathrm{E}(N;\ v=0)$ and $P_a(v=0)$ can be derived by conditioning on the value of the first sample score. Notice also that the *ARL* of a CUSUM given by (4.1) is implicitly dependent on the true parameter value θ since changes in the parameter value will change the group probabilities.

Define $\mathbf{w}^+$ and $\mathbf{w}^-$ as the set of all the possible sample scores that are positive and non-positive respectively. Then, remembering from (2.1) that $\pi_j = \Pr(s=w_j)$, we get

$$\mathrm{E}(N;\ v=0) \quad = \quad 1+\sum_{j\in\mathbf{w}^+}\pi_j\,\mathrm{E}(N;\ v=w_j) \tag{4.2}$$

and

$$P_a(v=0) \quad = \quad \sum_{j\in\mathbf{w}^-}\pi_j+\sum_{j\in\mathbf{w}^+}\pi_j P_a(v=w_j) \tag{4.3}$$

where $\mathrm{E}(N;\ v=x)=0$ and $P_a(v=x)=0$ if $x\geq h$. Thus, the *ARL* of the grouped data CUSUM is given by

$$ARL \quad = \quad \frac{1+\sum_{j\in\mathbf{w}^+}\pi_j\,\mathrm{E}(N;\ v=w_j)}{\sum_{j\in\mathbf{w}^+}\pi_j\left(1-P_a(v=w_j)\right)} \tag{4.4}$$

If a Fast Initial Response (FIR) feature is used (Crosier and Lucas, 1982), then the average run length is determined by conditioning on the outcome of the first Wald test in the sequence. Only the initial Wald test is unique, if the initial test does not signal then all subsequent Wald tests start at zero. Denote $ARL(\omega)$ as the average run length of a FIR CUSUM with initial value ω. Then, with *ARL* given by (4.4),

$$ARL(\omega) \quad = \quad \mathrm{E}(N;v=\omega)+P_a(v=\omega)ARL. \tag{4.5}$$

5. Gauge Limit Placement and Relative Efficiency

In practice the placement of group or gauge limits is often predetermined through the use of standard gauges. However, in some circumstances design of the step gauge is possible and thus we may wish to determine the optimal grouping criterion. In any event, it is of interest to compare the efficiency of utilizing various grouping criterion relative to the traditional variables based approaches. Clearly, the efficiency of a grouped data approach will be less than that of standard variables based methods since some information is lost due to the grouping. However, as will be shown, this loss of information is small for well chosen group limits, and, as a result, may be more than compensated by lower data collection costs. The methodology presented thus far in Sections 2-4 is applicable for any distribution and parameter of interest, however efficiencies and optimal limits depend on the distributional assumptions. In this section, we derive optimal gauge limits for sequential tests and analyze the relative efficiency of grouped data CUSUM Procedures when detecting mean shifts for a normal distribution. In the optimization problems considered we assume, without the loss of generality, that in-control the process has mean $\mu_0 = 0$ and standard deviation $\sigma = 1$.

The goal of our SPRT is to distinguish between μ_0 and μ_1. As a result, we may maximize the SPRT's ability to differentiate between the two parameter values by setting gauge limits so as to maximize the difference between the expected weight under μ_0 and under μ_1. With this goal in mind we solve the following maximization problem:

$$\text{maximize} \quad E(\ell \mid \mu = \mu_1) - E(\ell \mid \mu = \mu_0)$$

where ℓ the log-likelihood ratio is given by Equation (2.2). In words, we maximize the difference in the expected log-likelihood ratio under μ_0 and μ_1. We use ℓ rather than **w** to ensure that the optimal gauge limits do not depend on scaling factor used. Strictly speaking the above optimization problem is appropriate only if we are equally interested the parameter values μ_0 and μ_1. If not we should consider a weight difference of the expected log-likelihood ratio. However, the solution to the above problem will provide guidance as to the best group limits choices in any event. This maximization problem can be solved through Nelder-Mead simplex algorithm (Press et al., 1988), and results are given in Table 1 for various μ_1 values. Optimal solution are all symmetric about $(\mu_0 + \mu_1)/2$. To save space the optimal solutions in Table 1 are given in terms of Δt, where Δt is the deviation from $(\mu_0 + \mu_1)/2$. The results in Table 1 can be translated into optimal gauge limits by adding $(\mu_0 + \mu_1)/2$ to all values. For example, when μ_1 = 1.5 and the number of groups equals four we get gauge limits (–.3703 (=–1.1203+.75), .75, 1,8703), and for five groups we get (–.6701, .3042, .75, 1.11958, 2.1701). Notice that the effect of μ_1 on the optimal gauge limits written in terms of Δt is small. Thus, from Table 1, close to optimal gauge limits can be easily determined for a wide variety of μ_1 values.

Table 1: Optimal SPRT Gauge Limits

assume $\mu_0 = 0$ and $\sigma = 1$

add $(\mu_0 + \mu_1)/2$ to get optimal gauge limits

	μ_1 = 0.5	μ_1 = 1	μ_1 = 1.5
# groups	Δt	Δt	Δt
2	0	0	0
3	±.6209	±.6487	±.6976
4	0, ±.9972	0, 1.0439	0, ±1.1203
5	±.3893, ±1.2652	±.4104, ±1.3259	±.4458, ±1.4201
6	0, ±.6716, ±1.4721	0, ±.7091, ±1.5437	0, ±.7696, ±1.6503
7	±.2861, ±.8918, ±1.6397	±.3032, ±.9423, ±1.7201	±.3313, ±1.0211, ±1.836

CUSUMs are typically evaluated in terms of their ARL at the alternate mean value when the CUSUM is designed to achieve a given ARL at the null mean. However, for a PM-CUSUM this comparison is difficult to make due to the discreteness inherent in the problem. For any given gauge limit design ***t*** it is usually impossible to set the (integer) absorbing barrier h in such a manner to achieve precisely the $ARL(\mu_0)$ value desired. Thus, in order to make such a comparison we utilize the fact that as h varies a plot of $\ln(ARL(\mu_0))$ and $ARL(\mu_1)$ forms approximately a straight line. Using interpolation between the two values of h that yield $ARL(\mu_0)$ values closest to that desired we can determine a theoretical $ARL(\mu_1 | ARL(\mu_0) = ARL_0)$. We find optimal limits that minimize this theoretical ARL by utilizing the Nelder-Mead simplex algorithm. Table 2 shows the results of this optimization problem for the case: $\mu_1 = 1$, $ARL(\mu_0)$ = ARL_0 = 1000. For different values of ARL_0 these results differ little, and for different values of μ_1 scaling the results in Table 2 by μ_1 yields near optimal gauge limits. Not surprisingly the optimal gauge limits for CUSUM charts is quite different from the optimal limits for SPRTs. This difference is due to the different goals of SPRTs and CUSUMs.

Table 2: Optimal CUSUM Gauge Limits
assume μ_0 = 0, $\mu_1 = 1$, σ = 1 and $ARL(\mu_0)$ = 1000

# groups	t_1	t_2	t_3	t_4	t_5
2	.8861				
3	.3958	1.5637			
4	.0252	.9947	1.9090		
5	–.2945	.5720	1.3013	2.1194	
6	–.5591	.1787	.8415	1.5017	2.2019

Given a gauge limit design, it is of interest to evaluate the loss in efficiency that must be expected when articles are gauged into groups rather than measured precisely. A direct comparison of the PM-CUSUM with various number of groups and the traditional variables based approach is difficult due to the discreteness inherent in any scheme that utilizes categorical data. However, using interpolation, as discussed above, a comparison of multi-group PM-CUSUM Procedures and the variables based approach can be made. Using the solution approach suggested by Brook and Evans (1972) for a variables CUSUM when H_0: $\mu_0 = 0$, H_1: $\mu_1 = 1$, with $\sigma = 1$ and $h = 5$ ($k = \mu_1/2 = 1/2$) we obtain average run lengths of 904.81 and 10.39 at the null and alternate mean values. We consider the two to six group cases. The log-likelihood ratios presented in Table 3 are derived using the optimal gauge limits suggested in Table 2. The group scores were obtained from the likelihood ratio weights by multiplying by $q = 50/[\max(\ell) - \min(\ell)]$ and rounding off the weights to the nearest integer.

Table 3: Optimal Log-likelihood Ratio and Scores for PM-CUSUM

# groups	log-likelihood ratio (2.2)	assigned group scores **w**
2	−.5802, 1.0661	−18, 32
3	−.8740, .4282, 1.5811	−18, 9, 32
4	−1.1296, .0092, .8883, 1.8653	−19, 0, 15, 31
5	−1.3688, −.3393, .4177, 1.1455, 2.0440	−20, −5, 6, 17, 30
6	−1.5774, −.6596, .0097, .6477, 1.2983, 2.1150	−21, −9, 0, 9, 18, 29

To conduct our comparison we would like to determine $ARL(\mu_1)$ when $ARL(\mu_0)$ = 904.81. For completeness we also determine $ARL(\mu_0)$ that correspond to $ARL(\mu_1)$ = 10.39. In order to compare the efficiency of PM-CUSUMs with the variables CUSUM we use the methodology presented in Section 4 to determine the two values of h and their corresponding ARLs that yield $ARL(\mu_1)$ values closest to 10.39 and the two h values that yield $ARL(\mu_0)$ values closest to 904.81. Based on those two points, we interpolate to determine approximations for the ARLs that match the variables based chart at one ARL.

Table 4: Comparison of the Efficiency of Variables based CUSUM versus Grouped Data CUSUM

	match $ARL(\mu_0)$ = 904.81		match $ARL(\mu_1)$ = 10.39	
# Groups	$ARL(\mu_1)$ estimate	Efficiency	$ARL(\mu_0)$ estimate	Efficiency
2	14.37	72.3	240.0	26.5
3	12.04	86.3	458.0	50.4
4	11.28	92.1	592.6	65.5
5	10.96	94.9	694.0	76.7
6	10.78	96.4	756.0	82.5
variables	10.39	100	904.81	100

Table 4 presents the results where percent efficiency is derived by comparing the ARL estimates obtained for each number of groups with that obtained in the variables case.

Table 4 shows that the gain in efficiency going from the dichotomous case to 3 or more groups in significant.

6. Design of Grouped Data SPRTs and CUSUM Procedures

To assist the practitioner implement grouped data SPRTs or CUSUMs we present the following two iterative design procedures. A PM-SPRT with starting value 0 and absorbing barriers at $[a] > 0$ and $[b] < 0$ can be designed following steps S1-S8. Notice that an SPRT with starting value v and absorbing barriers at 0 and h can be determined using these steps by applying the transformations $h = [a] - [b]$ and $v = -[b]$. The key problem in the design is determining appropriate values for $[a]$ and $[b]$.

S1. Determine, based on application, the null and alternate mean values μ_0 and μ_1, and the maximum desired type I and II error rates α and β respectively.

S2. Set group limits (t_i's) either at optimal values as discussed in Section 5, or at predetermined values that are based on the application.

S3. Use Equation (2.2) to calculate the likelihood ratio for each group.

S4. Scale the weights and round to integer values. Aim for $\max(\mathbf{w}) - \min(\mathbf{w}) \approx 50$ unless for ease of implementation the weights must be very close to zero. Note that common factors in the resultant scores can be removed without affecting the efficiency of the procedure.

S5. Choose initial values for the absorbing barriers $[a]$ and $[b]$. For the continuous variable problem, Wald (1947) suggests choosing $A = (1-\beta)/\alpha$ and $B = \beta/(1-\alpha)$. This choice of absorbing barriers is derived by ignoring the possibility of overshooting the barriers, but is useful as a guide. We have found good initial values are the Wald approximations appropriately scaled, i.e. let $[b]$ equal the largest integer smaller than $q * \ln[\beta/(1-\alpha)]$ and $[a]$ equals the smallest integer larger than $q * \ln[(1-\beta)/\alpha]$, where q is the scaling factor used to scale the log-likelihood ratios as discussed in Section 2.

S6. Using the current $[a]$ and $[b]$, and the methodology in the appendix derive the probability function $\xi_j = \Pr(S = c_j)$ for S the terminating value of the SPRT.

S7. Using ξ_j and Equation (3.1) calculate P_{accept} at $\mu = \mu_0$ and $\mu = \mu_1$. Denote the actual error rates obtained by $\alpha_a = 1 - P_{accept}(\mu_0)$ and $\beta_a = P_{accept}(\mu_1)$.

S8. Consider the four possible cases:

- If $\alpha_a > \alpha$ and $\beta_a > \beta$ decrement $[b]$ and increment $[a]$ by one unit.
- If $\alpha_a < \alpha$ and $\beta_a > \beta$ decrement $[b]$ by one unit.
- If $\alpha_a > \alpha$ and $\beta_a < \beta$ increment $[a]$ by one unit.
- If $\alpha_a < \alpha$ and $\beta_a < \beta$ increment $[b]$ and decrement $[a]$ by one unit.

Repeat Steps S6-S8 until we obtain the values of $[a]$ and $[b]$ closest to zero that satisfy both error constraints $1 - P_{accept}(\mu_0) > \alpha$ and $P_{accept}(\mu_1) > \beta$.

The design approach for the PM-CUSUM Procedure is similar to the design of a PM-SPRT. However, typically CUSUM Procedures are evaluated based on their *ARL* at the null and alternate mean values. Also, the design of a CUSUM is easier since, once the group limits and null and alternate mean values are given, the only design parameter is h. This means, however, that generally we can not match both *ARL* criteria closely, since one *ARL* constraint will be more difficult to satisfy. Steps C1-C5 can be followed to design a PM-CUSUM.

C1. Determine, based on application, the null and alternate mean values μ_0 and μ_1, and specify the desired minimum *ARL* at the null and maximum *ARL* at the alternate. Denote the minimum and maximum as ARL_0 and ARL_1 respectively.

C2. Perform Steps S2-S4.

C3. Choose an initial value for the absorbing barrier h. For typical CUSUM Procedures $h = q * \ln(ARL_0 / ARL_1)$ is a good choice for the initial guess, where q is the scaling factor discussed in Section 2.

C4. Use Equations (4.1), (4.2) and (4.3) and subsequently (3.1) and (3.2) to determine $ARL(\mu_0)$ and $ARL(\mu_1)$.

C5. Adjust h as necessary to satisfy the given ARL constraints.

- If either $ARL(\mu_0) > ARL_0$ or $ARL(\mu_1) < ARL_1$ increase h by one unit.
- Otherwise decrease h by one unit.

If the initial guess is not close to the correct value the adjustment of h can be made larger. For example, we may adjust h by five units at a time. Repeat steps C4 and C5 until the smallest h that satisfies both constraints is found.

Consider the following simple example to illustrate the solution procedure. We first design an appropriate SPRT. In control the process of interest is normal with mean $\mu_0 = 74.3$ and standard deviation $\sigma = 1.3$ and we wish to detect mean shifts to $\mu_1 = 75.6$ (i.e. mean shifts of one standard deviation unit). Using a standardized gauge the group limits are $\mathbf{t}$ = (74, 75, 76). We assume the shift in the mean does not effect the normality or standard deviation of the process. This implies group probabilities $\mathbf{p}(\mu = \mu_0)$ = (0.4087, 0.2961, 0.1997, 0.0955) and $\mathbf{p}(\mu = \mu_1)$ = (0.1092, 0.2130, 0.2986, 0.3792). The log-likelihood ratios (2.2) are thus (–1.3200, –0.3294, 0.4096, 1.3791). Based on the application we determine that the error rates α and β should both be less than 0.1. This completes Steps S1-S3. Scaling the weights as recommended in Step S4 yields scores $(-24, -6, 8, 26)$. These scores have a common factor of two, thus without any loss in efficiency, we use $\mathbf{w} = (-12, -3, 4, 13)$. Since the scaling factor needed to yield $\mathbf{w}$ is q = 9.26 the recommended initial estimates for the absorbing barriers are –21 and 21. Table 5 shows the results of the iteration S6-S8. The best solution found occurred at iteration 8. Based on $[a] = 18$ and $[b] = -16$, from Equation (3.2) we get average sampling numbers of 4.7767 and 4.7616 when $\mu = \mu_0$ and $\mu = \mu_1$ respectively.

Table 5: Design Iterations for SPRT Example

iteration	$[b]$	$[a]$	α_a	β_a
1	–21	21	.0710	.0561
2	–20	20	.0743	.0624
3	–19	19	.0787	.0662
4	–18	18	.0876	.0758
5	–17	17	.1088	.0792
6	–17	18	.0869	.0811
7	–16	17	.1082	.0835
8*	–16	18	.0864	.0856
9	–15	17	.1050	.1112

To illustrate the design of a CUSUM, consider the example given in Steiner et al. (1994). That example concerned the manufacture of metal fasteners in a progressive die environment. In control the process mean is 74 thousands of an inch with a standard deviation of 1.3. Using a 6-step gauge with group intervals defined by: 73, 73.75, 74.35, 74.94, 75.55, and 76.3 a fixed sample size control chart to detect a mean shift to 75.3 thousands of an inch with type I and II errors rates of 0.005 required a sample size of 27 units. Thus, the average run length of this fixed sample size chart, in terms of average number of units examined before a signal are 27/.005 = 5400 and 27.1 units when the process is "in control" and "out of control" respectively. Since this example involves detection of a fairly small mean shift, i.e. a shift of only one standard deviation unit, a CUSUM chart would be expected to outperform a Shewhart type control chart.

Now consider a PM-CUSUM chart for the same example. The group weights given by (2.2) are (–1.7492, –0.9553, –0.4503, 0, 0.4503, 0.9553, 1.7492). Assuming we wish to do at least as well as the fixed sample size approach set $ARL_0 = 5400$ and $ARL_1 = 27.1$. Scaling the weights as recommended in S4 gives scores $\mathbf{w}$ = (–25, –14, –6, 0, 6, 14, 25). Based on our iteration approach C3-C5, shown in Table 6, we obtain the solution $h = 98$ with $ARL(\mu_0) = 5646$ and $ARL(\mu_1) = 14.6$.

Table 6: Design Iterations for CUSUM Example

iteration	h	$ARL(\mu_0)$	$ARL(\mu_1)$
1	76	1200.4	11.45
2	81	1692.1	12.17
3	86	2470.3	12.97
4	91	3576.4	13.76
5	96	5026.3	14.49
6	97	5336.9	14.49
7*	98	5646.5	14.62

Clearly it is desirable for a control chart to have a long ARL when the process "in control" and a short ARL when the process is "out of control." The given PM-CUSUM chart has approximately the same performance as the Shewhart approach at the null, but is dramatically better when the mean is at the alternate value.

7. Extension to Samples of Size *n*

In this article we have focused on the unit sequential implementation of PM-SPRTs and PM-CUSUMs. However, the same methodology is appropriate when using samples of size n. For a sample of size n, where n_j observations fall into the jth interval, the sample weight, defined as the sum of the individual log-likelihood ratios, is $\sum_{j=1}^{k} n_j \log\{\pi_j(\theta_1)/\pi_j(\theta_0)\}$. Scaling the group weights to get scores as in Section 2 we may defining our sample score as $y = \sum_{j=1}^{k} n_j w_j$. Thus, increasing the sample size simply increases the number and spread of the possible sample scores. By the multinomial distribution we have $\Pr\left(y = \sum_{j=1}^{k} n_j w_j\right) = \frac{n!}{\prod_{j=1}^{k} n_j!} \prod_{j=1}^{k} \pi_j^{n_j}$. However, different combinations of n_j values may lead to the same sum, thus we define $\mathbf{z} = (z_1, z_2, \ldots, z_m)$ as the m possible values of the sample scores. Define

$$\Pr(y = z_i) = p_i \tag{7.1}$$

where the probability $\Pr(y = z_i)$ equals the sum of the all the above multinomial probabilities where the n_j's are such that the sample score is z_i. The number of possible sample score values m grows exponential with the number of groups and polynomially with the sample size. Fortunately, in our application, we need only consider moderate sample sizes and number of groups. If the sample size is large, a normal approximation solution is appropriate, and if the number of groups is very large, the problem can be accurately modeled assuming variables data. To derive solutions for samples of size n, make the following substitutions in the analysis of the previous Sections: $k = m$, $w_i = z_i$ and $\pi_i = p_i$. Notice that if the desired sample size is large then the scaling factor q may need to be reduced since the spread in the sample scores is now $n[\max(\mathbf{w}) - \min(\mathbf{w})]$. Note that N refers to the number of samples until absorption whereas n is the sample size.

8. Conclusions

The prevalence of step gauges in applied quality control work attests to their convenience and economic advantages relative to exact measurement. Despite this prevalence, few statistical quality control techniques explicitly take into account the inherent grouping of the data which occurs. We propose using a simple integer scoring procedure for sequential tests with grouped data. The scoring procedure is an integer approximation of the parametric multinomial likelihood ratio. The resulting procedures are convenient to implement on the shop floor. We show how to derive the probabilities of errors of the first and second kind and the average sampling number of unit-sequential and batch sequential procedures based upon the integer scoring system. For repeated tests of hypothesis, CUSUM procedures based upon the integer scoring system may be used. The average run length properties of the parametric multinomial CUSUM procedure may be derived using the operating characteristics and average sampling properties of the component parametric sequential tests.

Appendix: Derivation of the Probability Distribution of S

We shall derive the exact probability distribution of the cumulative sum S at the termination of the random walk. Based on (2.2), the moment generating function of the group score s is given by $E(e^{st}) = \sum_{j=1}^{k} \pi_j u^{w_j} = \phi_s(t)$, say, where $u = e^t$. Thus, the moment generating function of $S = \sum_{j=1}^{N} s_j$ is given by $\phi_S(t) = [\phi_s(t)]^N$. Consider

$$\phi_s(t) = \sum_{j=1}^{k} \pi_j u^{w_j} = 1 \qquad \text{(A.1)}$$

This is a polynomial in u, which has degree $d = \max(\mathbf{w}) - \min(\mathbf{w})$. Let $u_1, \ldots, u_d$ denote the d roots of (A.1), and assume that $u_i \neq u_j$ for $i \neq j$. The roots are unique so long as $E(w) \neq 0$. When the underlying process parameter θ is such that $E(w) = 0$ then $u = 1$ is a double root. Now consider Wald's Fundamental Identity (Wald 1947, A:16) for a random walk between two absorbing barriers:

$$\mathrm{E}\left\{e^{St}[\phi_s(t)]^{-N}\right\} = 1$$

which holds for any t in the complex plane such that the moment generating function of s exists. Since, by definition $\phi_s(u_i) = 1$, substituting u_i for e^t in the fundamental identity gives the d equations

$$E\left(u_i^{\,S}\right) = 1 \qquad \text{for } i = 1, \ldots, d. \qquad \text{(A.2)}$$

We may obtain the exact probability distribution of S by conditioning each of the d left hand sides of the fundamental identity on the terminating value of the process. Let $[a]$ be the smallest integer $\geq \log A$ and $[b]$ the largest integer $\leq \log B$. Then, since all w_j are integer, the possible terminating values of the random walk are $([b] + \min(\mathbf{w}) + 1)$, $([b] - \min(\mathbf{w}) + 2)$, ..., $[b]$ for acceptance of $\theta = \theta_0$ and $[a], [a+1], \ldots, ([a] + \max(\mathbf{w}) - 1)$ for acceptance of $\theta = \theta_1$. Let $c_1, \ldots, c_d$ denote these d terminating values of the random walk. Then, the d equations of the fundamental identity (A.1) may be written as

$$\sum_{j=1}^{d} \xi_j u_i^{c_j} = 1 \qquad \text{for } i = 1, ..., d \qquad \text{(A.3)}$$

where
$$\xi_j = \Pr(S = c_j) \qquad \text{for } j = 1, ..., d \qquad \text{(A.4)}$$

The d equations in (A.2) are linear in ξ_j. Using a method to solve a system of linear equations such as LU decomposition (Press, et al., 1988) we can determine all the ξ_j's that give the exact probability distribution of S.

To illustrate all the calculations required to determine exact probability distribution of S consider a simpler version of the SPRT example introduced in Section 6. Rescaling the ratios through multiplication by 1.75 followed by rounding to the nearest integer results in the most compact distinct group scores, namely $\mathbf{w} = (-2, -1, 1, 2)$. Solving equation (A.1) for the four roots yields **u** = (–4.516, –0.494, 1, 1.9188) when $\mu = \mu_0$, and **u** = (–2.1239, –0.2367, 0.5729, 1) when $\mu = \mu_1$. If the absorbing barriers are chosen at 4 ($[a]$) and –4 ($[b]$) the only possible terminating values of this SPRT, i.e. possible values for S, are $\mathbf{c} = (-5, -4, 4, 5)$. Subsequently solving the system of 4 equations in 4 unknowns given by equation (A.3) for the probabilities of the SPRT terminating at each of these values gives ξ = (0.2912, 0.6489, 0.0494, 0.0104) when $\mu = \mu_0$ and ξ = (0.016, 0.0706, 0.6367, 0.2767) when $\mu = \mu_1$. Based on equations (3.1) and (3.2) this SPRT thus has a probability of acceptance and average sampling number respectively of 0.9402 and 5.26 when $\mu = \mu_0$, and 0.0866 and 5.70 when $\mu = \mu_1$.

References

Brook, D., Evans, D.A. (1972), "An approach to the probability distribution of cusum run length," *Biometrika*, Vol. 59, 539-549.

Dudding, B.P. and Jennett, W.J. (1942), *Quality Control Charts*, *British Standard 600R*.

Geyer, P.L., Steiner, S.H., Wesolowsky, G.O. (1995), "Optimal SPRT and CUSUM Procedures Using Compressed Limit Gauges," to appear in the *IIE Transactions*.

Kemp, K.W. (1971), "Formal Expressions Which can be Applied to CUSUM Charts, With Discussion," *Journal of the Royal Statistical Society*, Series B, Vol. 33, 331-360.

Johnson, N. L. (1961), "A Simple Theoretical Approach to Cumulative Sum Control Charts," *Journal of the American Statistical Association*, Vol. 56, 835-840.

Ladany, S.P, (1976), "Determination of Optimal Compressed Limit Gaging Sampling Plans," *Journal of Quality Technology*, Vol. 8, No. 4, 225-231.

Lucas, J.M. and Crosier, R.B. (1982), "Fast Initial Response for CUSUM Quality Control Schemes," *Technometrics*, Vol. 24, No. 3, 199-205.

Lucas, J. M. (1985), "Counted Data CUSUMs," *Technometrics*, Vol. 27, No. 2, 129-144.

Mace, A.E. (1952), "The Use of Limit Gages in Process Control," *Industrial Quality Control*, January, 24-31.

Moustakides, G.V. (1986), "Optimal Stopping Times for detecting changes in distributions," *Annals of Statistics,* Vol. 14, 1379-1387.

Ott, E.R., Mundel, A.B. (1954), "Narrow-Limit Gaging," *Industrial Quality Control*, March, 221-28.

Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika,* Vol. 41, 100-114.

Press, W.H., Flannery, B.P., Teukolosky, S.A., Vetterling, W.T. (1988), *Numerical Recipes*, Cambridge University Press, Cambridge.

Schilling, E.G., (1981), *Acceptance Sampling in Quality Control*, Marcel Dekker, New York.

Schneider, H. and O'Cinneide, C. (1987). "Design of CUSUM Control Charts Using Narrow Limit Gauges," *Journal of Quality Technology*, Vol. 19, No. 2, 63-68.

Steiner, S.H., Geyer, P.L., Wesolowsky, G.O. (1994), "Control Charts based on Grouped Data," *International Journal of Production Research* Vol. 32, No. 1, 75-91.

Steiner, S.H., Geyer, P.L., Wesolowsky, G.O. (1995), "Shewhart Control Charts to Detect Mean and Standard Deviation Shifts based on Grouped Data," revision currently with the *Journal of Quality and Reliability Engineering International.*

Stevens, W.L., (1948), "Control by Gauging," *Journal of the Royal Statistical Society: Series B*, No. 1, 54-108.

Wald, A. (1947), *Sequential Analysis*, Dover, New York