

# Statistical Process Control for Shifts in Mean or Variance Using a Change-point Formulation

Douglas M. HAWKINS

School of Statistics  
University of Minnesota  
Minneapolis, MN 55455  
([doug@stat.umn.edu](mailto:doug@stat.umn.edu))

K. D. ZAMBA

College of Public Health  
Department of Biostatistics  
University of Iowa  
Iowa City, IA 52241  
([gideon-zamba@uiowa.edu](mailto:gideon-zamba@uiowa.edu))

Statistical process control (SPC) involves ongoing checks to ensure that neither the mean nor the variability of the process readings has changed. Conventionally, this is done by pairs of charts—Shewhart  $\bar{X}$  and  $S$  (or  $R$ ) charts, cumulative sum charts for mean and for variance, or exponentially weighted moving average charts for mean and variance. The traditional methods of calculating the statistical properties of control charts are based on the assumption that the in-control true mean and variance were known exactly, and use these assumed true values to set center lines, control limits, and decision intervals. The reality, however, is that true parameter values are seldom if ever known exactly; rather, they are commonly estimated from a phase I sample. The random errors in the estimates lead to uncertain run length distribution of the resulting charts. An attractive alternative to the traditional charting methods is a single chart using the unknown-parameter likelihood ratio test for a change in mean and/or variance in normally distributed data. This formulation gives a single diagnostic to detect a shift in mean, in variance, or in both, rather than two separate diagnostics. Using the unknown parameter formulation recognizes the reality that at best one has reasonable estimates of parameters and not their exact values. This description implies an immediate benefit of the formulation, that the run behavior is controlled despite the lack of a large phase I sample. We demonstrate another benefit, that the change-point formulation is competitive with the best of traditional formulations for detecting step changes in parameters.

KEY WORDS: Control charts; Generalized likelihood ratio; Phase I; Phase II.

## 1. INTRODUCTION

The conceptual model underlying statistical process control (SPC) is that variability in process measurement comes from two basic sources: “common cause” variability, which is all sources of unavoidable random variability that can be removed only by changing the system, and “special cause” variability, which results from some potentially identifiable source that can be removed. A system is said to be in the state of statistical control when the only variability is that due to common causes. When a special cause intervenes, the process is said to be out of control. Operationally, when the system is in control, the process readings appear to be a realization of some random model—in the simplest case, independent observations from some common statistical distribution. SPC is a framework of procedures to detect when a system has gone from in control to out of control. Its objectives may include providing a signal that the process is out of control, an estimate of when it went out of control, and a diagnosis of the way in which it went out of control—for example, whether the mean shifted, the variance jumped, or either of these quantities started a slow drift.

Although process readings can follow any statistical distribution, the most common assumption is the simplest model for SPC: that while the process is in control, the process readings appear to follow a normal distribution and are statistically independent. Our approach uses independent normal readings as the working model, with the note that extending the methodology to more complex settings is a matter of adapting rather than redefining it.

Under this model, while the process is in control, the normal distribution will have some mean  $\mu_1$  and standard deviation  $\sigma_1$ . Important departures from control include the following:

- The mean could shift from  $\mu_1$  to some other value.
- The standard deviation could shift from  $\sigma_1$  to some other value.
- The distribution could shift from normal to some other form.
- The mean or variance could drift from the in-control levels.

We concentrate on the first two of these possibilities, both because they are the most likely and because, at least descriptively, they are able to approximate other departures.

A further dichotomy can be made between settings in which the departure from control is *transient*, or *isolated*, by which we mean that the system goes out of control but then returns to control even in the absence of any intervention, and those in which it is *persistent*, or *sustained*, by which we mean that having left the state of control, the system will remain out of control or even go further from control, until some corrective action is taken. Corresponding to this four-way distinction, the standard tools in SPC are

- The Shewhart  $I$  or  $\bar{X}$  chart for detecting transient shifts in mean
- The Shewhart  $S$  or  $R$  chart for detecting transient shifts in standard deviation
- The location cumulative sum (cusum) or exponentially weighted moving average (EWMA) chart for detecting sustained shifts in mean

© 2005 American Statistical Association and  
the American Society for Quality  
TECHNOMETRICS, MAY 2005, VOL. 47, NO. 2  
DOI 10.1198/00401700400000644

- The variance cusum or EWMA chart for detecting sustained shifts in standard deviation.

We do not dwell on the properties and relative performance of these charts here (see Montgomery 2004; Hawkins and Olwell 1998 for more information).

The performance of control charts is measured by their run length distribution. While the process is in control, the runs should typically be long, but once the process is no longer in control, the response time should be short. The central tendency of the run length distribution is often summarized by the average run length (ARL). The ARL of all the standard charts depends on the distributional form, true mean, and true standard deviation of the process readings, as well as the values of the chart constants (center line and control limits for Shewhart charts; reference value and decision interval for cusums). It can be computed using available software, provided that all of these quantities are known. This is true both while the process is in control and following a persistent step change to a fixed out-of-control level.

The problem with this is that the in-control process mean and standard deviation are seldom known to high precision. If the control chart is designed with its chart constants based on estimates, then the actual run length behavior will be different than these calculations claim. Because the error in the estimates is random, this imparts a systematic distortion in the behavior of the charts. For example, if the estimate of the standard deviation is below the true value, then the chart's run lengths will be systematically shorter than the calculations claim; if it happens to be above the true value, then the run lengths will be systematically longer. For the Shewhart and cusum charts, Quesenberry (1991, 1993) and Hawkins and Olwell (1998) showed that this systematic distortion can be large even if quite large calibration samples are used.

Jones et al. (2004) elaborated on this point and obtained the marginal distribution of the run length of a location cusum by mixing the conditional distribution given the parameter estimates over the distribution of the estimators. Parallel developments for the EWMA were given by Jones (2002) and Jones and Champ (2001). However, it should be borne in mind that this marginal distribution of run length applies to conceptual ensembles of control charts calibrated using different phase I datasets. It does not describe the behavior of any single control chart. Although someone using a chart with estimated parameters could use these results to understand the range of behaviors that a chart like this might reasonably have, the calculations do not describe the characteristics of any particular chart with estimated parameters.

## 2. THE UNKNOWN-PARAMETER CHANGEPOINT MODEL

To avoid this problem of the dependence on assumed known parameter values, Hawkins, Qiu, and Kang (2003) (abbreviated as HQK hereafter) proposed an alternative to the  $\bar{X}$  and location cusum charts. For ease of understanding, we briefly recap their procedure. It is motivated by modeling a persistent change in the process mean by

$$X_i \sim \begin{cases} N(\mu_1, \sigma^2) & \text{if } i \leq \tau \\ N(\mu_2, \sigma^2) & \text{if } i > \tau. \end{cases} \quad (1)$$

where  $X_1, X_2, \dots, X_i, \dots$  are the successive process readings,  $\mu_1$  is the in-control true mean,  $\mu_2$  is the out-of-control value to which the process mean shifts,  $\tau$  is the changepoint, and  $\sigma$  is the standard deviation of the process readings, assumed to be constant. If all of the parameters except  $\tau$  were known, then the diagnostic of choice would be a cusum using reference value  $(\mu_1 + \mu_2)/2$ . We instead assume that none of these parameters is known a priori. Consider the setting where  $n$  process readings have accrued. For  $0 \leq i < k < n$ , define the summary statistics

$$\bar{X}_{i,k} = \sum_{j=i+1}^k X_j / (k - i) \quad (2)$$

and

$$V_{i,k} = \sum_{j=i+1}^k (X_j - \bar{X}_{i,k})^2. \quad (3)$$

Suppose that it were known that the changepoint was at instant  $\tau = k$ . The conventional estimates of the remaining parameters would then be

$$\hat{\mu}_1 = \bar{X}_{0,k}, \quad (4)$$

$$\hat{\mu}_2 = \bar{X}_{k,n}, \quad (5)$$

and

$$\hat{\sigma}^2 = (V_{0,k} + V_{k,n}) / (n - 2). \quad (6)$$

The likelihood ratio test for the null hypothesis  $H_0: \mu_1 = \mu_2$  is the two-sample  $t$  statistic,

$$T_{k,n} = \sqrt{\frac{k(n-k)}{n}} \frac{\bar{X}_{0,k} - \bar{X}_{k,n}}{\hat{\sigma}},$$

which in the null case, and assuming constant variance, follows a  $t$  distribution with  $n - 2$  degrees of freedom. The changepoint  $\tau$  is unknown, however. The generalized likelihood ratio (GLR) test assuming all four parameters unknown is given by finding  $T_{\max,n}$ , the maximum of  $|T_{k,n}|$  across all possible  $k$  values.

There is a vast literature on the fixed-sample (phase I) changepoint formulation (see, e.g., Hawkins 1977; Worsley 1979, 1982; Carlstein, Müller, and Siegmund 1994). In the phase II SPC setting, however, the sample is not fixed. It continues to grow as long as the process is judged to be in control, and so is allowed to continue without intervention. The use of the changepoint approach in this setting has also received much attention (see, e.g., Lai 2001), though in the context of known in-control parameters. Another stream of phase II application of the changepoint formulation includes the work of Pignatiello and Samuel (2001) and Samuel, Pignatiello, and Calvin (1998a, b). These authors, however, used the changepoint formulation not as a stand-alone procedure, but rather for follow-up estimation after a signal has been given by some other diagnostic, such as a Shewhart chart.

The fixed-sample unknown-parameter changepoint formulation can be adapted to the phase II dynamic setting as follows:

- After observation  $n$  has been added to the total record of the process, compute  $T_{\max,n}$ , the GLR test statistic for a change in mean at some previous instant.

- If  $T_{\max,n} \leq h_n$ , where  $h_n$  is some suitable control limit, then conclude that there is no evidence of a mean shift, and leave the process running uninterrupted.
- If, however,  $T_{\max,n} > h_n$ , then conclude that there is evidence of a mean shift. Complete the diagnosis by noting that the  $k$  value maximizing  $T_{k,n}$  is the maximum likelihood estimator (MLE) of the instant at which the variance changed, and that the MLEs of the before- and after-change means are the within-segment means  $\bar{X}_{0,\hat{\tau}}$  and  $\bar{X}_{\hat{\tau},n}$ .

This framework leaves open the issue of choosing the sequence of control limits,  $h_n$ . HQK proposed that these limits be defined by the property that while the process is in control, the probability of a signal is fixed at some user-selected constant level  $\alpha$ ; in symbols,

$$\Pr[T_{\max,n} > h_n | T_{\max,j} \leq h_j, j < n] = \alpha.$$

This constant probability of a signal parallels the Shewhart  $\bar{X}$  chart, and a proposal by Margavio et al. (1995) to use non-constant control limits for the EWMA chart. It contrasts with the conventional cusum and EWMA charts, where the probability changes from one observation to another. The necessary sequence of constants  $h_n$  does not seem to be amenable to any theoretical calculation, and HQK estimated it by simulation.

As noted in Section 1, control charts have generally come in pairs, one chart for the mean and one for the standard deviation. In line with this, Hawkins and Zamba (2005) developed a procedure parallel to that of HQK for detecting changes in the variance of the process readings without regard to the constancy or otherwise of the mean.

### 3. A COMBINED CHART FOR MEAN OR VARIANCE SHIFTS

Returning to the changepoint formulation, we generalize (1) to

$$X_i \sim \begin{cases} N(\mu_1, \sigma_1^2) & \text{if } i \leq \tau \\ N(\mu_2, \sigma_2^2) & \text{if } i > \tau. \end{cases} \quad (7)$$

In this formulation, the mean, the variance, or both can change when the process crosses the changepoint  $\tau$ . If the changepoint  $\tau$  were known to be  $k$ , then the GLR test statistic would be

$$GLR = k \log \frac{S_{0,n}}{S_{0,k}} + (n - k) \log \frac{S_{0,n}}{S_{k,n}},$$

where we define  $S_{i,j} = V_{i,j}/(j - i)$  to be the MLE (without the usual degrees of freedom bias adjustment of the denominator) of the variance of the sequence  $X_{i+1}, \dots, X_j$ .

In the null case of no shift, this statistic has an asymptotic chi-squared distribution with 2 degrees of freedom. The quality of this approximation can be improved substantially (Lawley 1956) by making the Bartlett correction, dividing by a factor that will make the expectation of the GLR equal to the degrees of freedom. The expectation of the GLR is known but involves the digamma function, one of the less familiar transcendental functions. A standard expansion of the expectation (see, e.g., Kendall and Stuart 1961) shows that to terms of order  $o(n^{-2})$ ,

$$E[GLR] = 2 + \frac{11}{6} \left[ \frac{1}{k} + \frac{1}{n-k} - \frac{1}{n} \right] + 2 \left[ \frac{1}{k^2} + \frac{1}{(n-k)^2} - \frac{1}{n^2} \right].$$

This leads then to the Bartlett-corrected test statistic

$$G_{k,n} = \left( k \log \frac{S_{0,n}}{S_{0,k}} + (n - k) \log \frac{S_{0,n}}{S_{k,n}} \right) / C, \quad (8)$$

$$C = 1 + \frac{11}{12} \left[ \frac{1}{k} + \frac{1}{n-k} - \frac{1}{n} \right] + \left[ \frac{1}{k^2} + \frac{1}{(n-k)^2} - \frac{1}{n^2} \right],$$

where  $C$  is the Bartlett correction factor. If the changepoint is not known a priori but must be estimated along with the testing process, then the two-stage GLR is found by maximizing  $G_{k,n}$  over all possible split points  $k$ , yielding  $G_{\max,n} = \max_k G_{k,n}$ .

An earlier proposal by Sullivan and Woodall (1996) was to use the maximum GLR for the analysis of fixed-sample phase I data. These authors also suggested a correction to bring the statistics computed for different  $k$  values to a common distribution—analogueous to the Bartlett correction—but estimated the correction from simulation rather than deriving it analytically.

We have not explicitly defined the range of  $k$ . Note that because  $S(0, 1) = S(n - 1, n) = 0$ , the GLR can be made infinite by making either of the segments of length 1. These degenerate solutions are not interesting, however, and so we restrict the lengths of both segments to be at least 2, giving strictly positive values for the two segment variances.

Finally, adapting this formulation to use in the SPC setting where the sample size is not fixed but grows indefinitely, we set up the ongoing SPC phase II procedure:

- After observation  $n$  has been added to the total record of the process, compute  $G_{\max,n}$ .
- If  $G_{\max,n} \leq h_n$ , where  $h_n$  is some suitable control limit, then conclude that there is no evidence of a shift in either mean or variance, and leave the process running uninterrupted.
- If, however,  $G_{\max,n} > h_n$ , then conclude that there is evidence of a shift in the mean, the variance, or both.

If there is a signal, then the question arises of exactly what changed between the two segments. A thorough investigation of this question would best be done by splitting the process history at the estimated changepoint and carrying out a two-sample comparison between the two resulting segments, using graphical methods, such as comparative boxplots, and more formal parametric or nonparametric tests. The GLR approach is designed for sustained shifts in mean and/or variance, but the trigger could be outliers, a false alarm resulting from nonnormality of the process readings, or a number of other possibilities. A useful starting point, however, is to take the normal distribution changepoint model at face value and test for a shift in mean and for a shift in variance using conventional normal distribution methods. If the changepoint  $k$  had been chosen ahead of time, then the conventional parametric tests for these possibilities would be as follows:

- The two-sided  $F$  test for a variance change using degrees of freedom  $k - 1$  and  $n - k - 1$  and test statistic  $F = V_{0,k}(n - k - 1) / ((k - 1)V_{k,n})$ .
- The approximate  $t$ -test for a change in mean using the Satterthwaite–Welch approximate  $t$  statistic

$$t = \frac{\bar{X}_{0,k} - \bar{X}_{k,n}}{\sqrt{S_{0,k}/(k-1) + S_{k,n}/(n-k-1)}}$$

which follows an approximate  $t$  distribution with  $r$  degrees of freedom, where

$$r = \left( \frac{S_{0,k}}{k-1} + \frac{S_{k,n}}{n-k-1} \right)^2 / \left[ \frac{1}{k-1} \left( \frac{S_{0,k}}{k-1} \right)^2 + \frac{1}{n-k-1} \left( \frac{S_{k,n}}{n-k-1} \right)^2 \right]$$

The searching over  $k$ , and conditioning on the large  $G_{\max,n}$  value that triggers these follow-up tests, invalidate the assumptions that lead to the null  $F$  and approximate  $t$  distributions of these statistics, and indeed also those of any other two-sample test procedures that might be substituted for them. However, they do not detract from their value in providing practical guidance on whether the cause of the signal is more plausibly a variance shift, a mean shift, or a shift in both.

### 3.1 The Choice of the Control Limits

We have not specified how the control limit sequence  $h_n$  is to be chosen. The hazard function, familiar from life testing, is the probability that a unit will fail at time  $n$ , conditional on its not having failed before time  $n$ . Analogously, the hazard function of an SPC scheme may be defined as the conditional probability of a signal at process reading  $n$ , conditional on there having been no signal up to that instant. The Shewhart chart has a constant hazard function, but most other SPC approaches do not. An attractive choice for the changepoint approach, paralleling the proposals of HQK and Margavio et al. (1995), would be to have a constant hazard function while the process is in control. In symbols,

$$P[G_{\max,n} > h_{n,\alpha} | G_{\max,j} \leq h_{j,\alpha}, j < n] = \alpha, \tag{9}$$

where  $\alpha$  is the specified probability of an erroneous signal, and we rewrite the control limit as  $h_{n,\alpha}$  to emphasize this dependence. Because this probability is constant, it corresponds to an in-control ARL of  $1/\alpha$ . It does not seem possible to solve for these  $h_{n,\alpha}$  values theoretically, and so they were estimated using simulation. This used 10 million random samples of  $N(0, 1)$  data series of length up to 500, and covering  $\alpha$  values of .05, .02, .01, .005, .002, and .001. The resulting estimated fractiles have standard errors of approximately .02.

Because there must be a minimum of two observations in each segment, it is logically possible to start testing right from the fourth observation, but this is unlikely in most applications. Although it is an attraction of the formulation that it does not require a large (and thus expensive) phase I data-gathering exercise to estimate the in-control parameters accurately, standard practice is likely to involve gathering at least some carefully monitored observations before the formal SPC is set in place. Our simulations incorporated this possibility, generating 17 tables of  $h_{n,\alpha}$  values corresponding to gathering 3, 4, ..., 19 observations without applying the  $G_{\max,n}$  screen and then starting with the next reading. The full table is available on the website [www.stat.umn.edu/hawkins](http://www.stat.umn.edu/hawkins).

A reasonable approach to the monitoring might be to gather nine observations without formal monitoring, and start monitoring with the tenth observation. Table 1 is a table of cutoffs for this setting. We stress that this is not a prescription, but simply

Table 1. Control Limit  $h_{n,\alpha}$  for Sample Size  $n$ , Hazard  $\alpha$  Starting at  $n = 10$

$n$	Threshold $h_{n,\alpha}$					
	.05	.02	.01	.005	.002	.001
10	10.128	12.237	13.795	15.330	17.352	18.840
11	9.213	11.389	12.996	14.556	16.609	18.173
12	8.854	11.083	12.719	14.313	16.397	17.965
13	8.690	10.961	12.631	14.265	16.353	17.950
14	8.616	10.917	12.610	14.249	16.361	17.978
15	8.588	10.909	12.618	14.277	16.423	18.015
16	8.582	10.928	12.637	14.324	16.464	18.083
17	8.586	10.942	12.664	14.345	16.501	18.113
18	8.590	10.964	12.692	14.386	16.550	18.192
19	8.593	10.973	12.709	14.394	16.567	18.187
20	8.599	10.989	12.734	14.419	16.614	18.219
22	8.611	11.015	12.763	14.462	16.653	18.261
24	8.629	11.043	12.790	14.497	16.682	18.343
26	8.644	11.060	12.819	14.536	16.722	18.368
28	8.661	11.089	12.846	14.559	16.754	18.375
30	8.669	11.097	12.877	14.591	16.785	18.436
35	8.683	11.133	12.910	14.636	16.845	18.478
40	8.692	11.151	12.932	14.657	16.884	18.572
45	8.712	11.164	12.947	14.676	16.908	18.572
50	8.714	11.181	12.971	14.713	16.946	18.617
60	8.734	11.197	12.990	14.746	16.980	18.616
70	8.745	11.210	13.003	14.751	16.985	18.634
80	8.748	11.224	13.022	14.766	16.989	18.679
90	8.743	11.238	13.035	14.785	17.043	18.710
100	8.770	11.231	13.029	14.771	17.052	18.708
125		11.257	13.061	14.799	17.033	18.697
150		11.253	13.037	14.812	17.069	18.756
175		11.283	13.060	14.822	17.069	18.729
200		11.277	13.078	14.803	17.065	18.738
250			13.040	14.846	17.096	18.748
300			13.155	14.826	17.091	18.723
350			13.130	14.857	17.097	18.772
400			13.148	14.863	17.127	18.761
500				14.904	17.089	18.770

a guess at what practitioners might wish to do. Those who prefer some different number of initial familiarization values can replace the Table 1 figures with others from the website.

Basing an implementation on cutoffs in a table is cumbersome, and so it is helpful to have some more compact, even if approximate, way of computing suitable cutoff values. Based on this reasoning, we suggest the following approach for obtaining control limits. Compare the first five test statistics  $G_{\max,n}$  (i.e., for  $n = 10, \dots, 14$ ) to the first five entries of the table; then for  $n > 14$ , use the approximation

$$h_{n,\alpha} = \begin{cases} 1.58 - 2.52 \log(\alpha) + \frac{.094 + .33 \log(\alpha)}{\sqrt{n-9}} & \text{if } .001 \leq \alpha < .05 \\ 8.43 + .074 \log(n-9) & \text{if } \alpha = .05. \end{cases}$$

This reproduces Table 1 with a maximum absolute deviation of .09.

### 3.2 Rational Groups

The discussion so far has been in terms of charting individual observations. In some circumstances, sampling economies of scale may lead to the use of rational groups of size larger than 1. Simply unravelling the successive rational groups into individual observations and allowing a changepoint only at the last observation of a rational group brings the rational group setting within the framework of individual observations. This approach differs from the usual analysis of rational groups in that

the scale is estimated from a single common mean, rather than from the individual rational groups' means. It is certainly possible to recast the changepoint model to allow the more conventional summaries of rational groups by their individual means and variances, but it is less than clear that any performance advantage can be gained by this adaptation.

#### 4. COMPUTATIONAL DETAILS AND WINDOWS

All of the computations necessary for the testing can be found from two arrays—one array of the running total of the data  $W_n = \sum_1^n X_i$ , and the other of  $V_{0,n}$ —the running sum of squared deviations from the running mean. These have fast, simple updates,

$$W_{n+1} = W_n + X_{n+1}$$

and

$$V_{0,n+1} = V_{0,n} + n(X_{n+1} - W_n/n)^2/(n + 1).$$

From these two tables, all quantities needed to perform the test (and to perform the follow-up parametric analyses) are easy to compute,

$$\bar{X}_{i,k} = (W_k - W_i)/(k - i)$$

and

$$V_{i,k} = V_{0,k} - V_{0,i} - i(k - i)/k(\bar{X}_{0,i} - \bar{X}_{i,k})^2.$$

Note that because these formulas have some potential for loss of precision due to subtractive cancellation, the tables and the arithmetic should be in double precision.

The updates of the  $W_n$  and  $V_{0,n}$  are fast, but searching for the  $k$  maximizing the split statistic involves computing  $n - 3$  test statistics. Although the computations themselves are fast, for very large  $n$  or in settings where many process characteristics are to be monitored, this computation could become a burden. A remedy for this that comes to mind is to restrict the searching to a "window" of the most recent observations.

The obvious way to do this is using the Willsky-Jones (1976) method of retaining only the  $M$  most recent observations and using only these observations in the testing procedure, successively applying a fixed-sample-size phase I analysis to the observations in the window. A less obvious but better method flows from implementation based on the summary tables  $W_n$  and  $V_{0,n}$ . This method involves retaining only the  $M$  most recent values of  $W$  and  $V$ . When a new observation accrues, the  $W$  and  $V$  arrays are moved down one place, dropping the oldest entry and adding the newly created one. The  $k$  search is then confined to the entries in the table and so does not grow with  $n$  but rather remains bounded by  $M - 2$ .

Note that this approach is fully statistically efficient. Observations to the left of the window are not lost; they are still included in all summary numbers. All that is lost is the ability to declare a split point more than  $M$  time intervals in the past. Because estimated split points in the remote past are unhelpful in practice, and are common in neither the in-control nor out-of-control setting, this loss is unlikely to cause much concern or to invalidate using a window of moderate width (perhaps in the low hundreds).

If a window restriction is desired, then we recommend this approach rather than the "ignore everything outside the window" approach. Not only is it preferable statistically, but it is also computationally much faster.

#### 5. PERFORMANCE

The control limits were chosen to make the in-control hazard function constant, and the in-control run length therefore follows a geometric distribution. This raises the question of the form of the out-of-control run length distribution. Qualitatively, following a shift, the two-sample statistic splitting at the true changepoint will change from an approximate central chi-squared distribution to a noncentral distribution. The noncentrality parameter becomes nonzero with the first out-of-control observation, then increases to a limit that depends on the magnitude of the parameter shift and on the length of the in-control history. To illustrate this, suppose that the variance remains at the in-control level but the mean shifts by  $\delta$  standard deviations immediately after observations  $\tau$ . Then at observation  $n > \tau$ , the Student  $t$  component of the two-sample GLR statistic for a split at  $\tau$  will have a noncentrality parameter that can be written as

$$\tau \left( 1 - \frac{\tau}{n} \right) \delta^2,$$

which increases to a maximum of  $\tau\delta^2$ , implying that the unconditional probability of a signal at observation  $n$  would increase up to some limit. This suggests that the hazard function will increase initially after the shift, but then may stabilize at some level. To see where this leads in terms of the run length distribution, consider Figure 1, which shows the hazard functions for the changepoint formulation and a small-shift cusum (details of which are provided later) when a one standard deviation shift in mean follows 50 in-control readings. The hazard functions were estimated from a simulation of 200,000 normal series. Note that both hazard functions rise to a maximum then drop off, but the changepoint hazard function is relatively steady. This suggests that the distribution of the run length is not heavy-tailed, an impression confirmed by the mean and standard deviation (SD) of the run lengths:

	Mean	SD
Changepoint	25.5	30.5
Cusum	16.8	20.2

The performance of quality control charts is commonly measured by their ARL. If the hazard function were constant, then the ARL would completely characterize the run length distribution. Although the run length distribution of the changepoint approach is not geometric, neither does it appear to have tail weights far out of line with those of the geometric distribution. We therefore believe that the ARL provides an adequate basis for comparing the performance of the changepoint method with a cusum approach.

This finding motivated more extensive simulations, involving each of 10,000 sequences of length up to 10,000 and using  $\alpha = .002$  so as to get an in-control ARL of 500. Two other settings need to be specified: the number of initial familiarization observations gathered before the start of the testing and the instant at which the change in the parameter(s) occurs. We set the first of these at 9, so that formal testing began with the tenth observation of the sequence.

Several instants of change were tested (observations 10, 20, 50, 150, and 250), but we report only the results for observations 10, 50, and 250, because the two intermediate values

added  
N(0, 1)  
for  $\delta$   
w  
viati  
ering t  
deviati

Tabl  
of app  
represe  
they ar

Exam  
to dete  
for 250  
fast, bu  
much s  
would  
proced  
mean s  
period

The  
ter the

$\delta$	$\tau$
0	
	2
.5	
	2
1.0	
	2
1.5	
	2
2.0	
	2

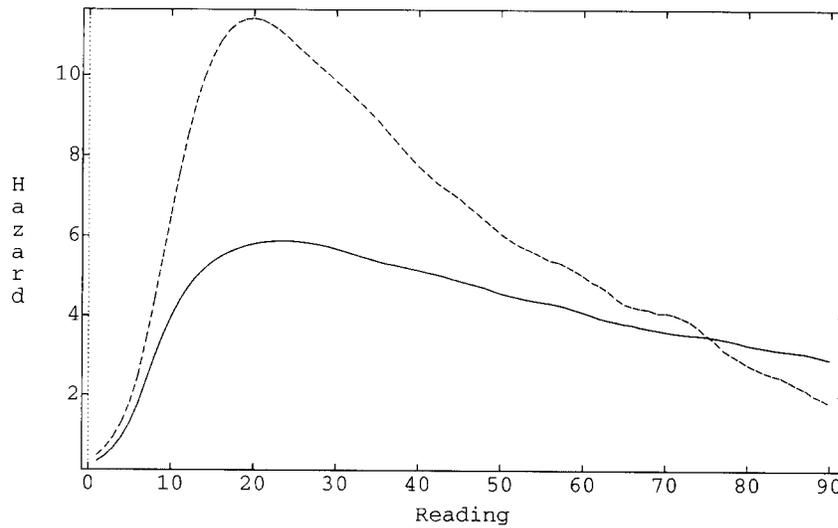


Figure 1. Percentage Hazard Function of Changepoint and (solid line) Cusum (dotted) Following Shift.

added little insight. At the instant of change, the in-control  $N(0, 1)$  distribution was changed to  $N(\delta, \sigma^2)$ . The values used for  $\delta$  were 0, .5, 1, 1.5, and 2.0. The out-of-control standard deviation  $\sigma$  was set to the  $-3, -2, \dots, 2, 3$  powers of 1.25, covering the range from a halving to a doubling of the standard deviation in constant proportion steps.

Table 2 gives the resulting ARLs. All have standard errors of approximately 1% of their value. The setting  $\delta = 0, \sigma = 1$  represents the in-control setting, so the ARLs should be 500, as they are to within random sampling errors.

Examining the  $\sigma = 1$  column reveals the procedure's ability to detect shifts in the mean only. If the process has run in control for 250 periods, then detection of even modest shifts is quite fast, but short in-control periods coupled with modest shifts are much slower to detect, as the earlier comments on noncentrality would lead one to expect. The rows with  $\delta = 0$  demonstrate the procedure's ability to detect pure variance shifts. As with pure mean shifts, there is quite low power if the initial in-control period is very short, but much higher sensitivity if it is even 50.

The modest ability to detect small shifts occurring soon after the start of monitoring may cause dismay. This dismay is

misplaced, however. It is a strength of the unknown-parameter changepoint formulation that it can control its ARL while monitoring short runs. It would be too much to expect it to do so with the same sensitivity as methods that require long, carefully controlled phase I studies.

Moving off the first row-fourth column "T"-shaped portion of the table shows that the performance of the procedure for detecting shifts in both mean and variance is generally better than the performance in detecting either departure alone. This is to be expected, both on intuitive grounds and because of the fact that the noncentrality of the GLR statistic combines the separate noncentralities of its  $t$  and  $F$  components.

The second performance issue relates to the follow-up diagnoses made after a signal to decide whether this was due to a change in mean, a change in variance, or a change in both mean and variance. Table 3 follows the layout of Table 2 and gives the percentage of times the follow-up approximate  $t$ -test for equality of mean attained significance at the nominal 1% level and the percentage of times the  $F$  test attained nominal 1% significance. The figures shown are for shifts at sample number 50; the other sample size gave qualitatively similar pictures. The three entries for  $\delta = 0, \sigma = 1$  show that in this null case, the follow-up diagnoses a mean shift in some 80% of cases and a variance shift in some 90% of cases. Note that this is not the

Table 2. In-Control and Out-of-Control ARL

$\delta$	$\tau + 1$	ARL						
		$\sigma$						
		.51	.64	.80	1.00	1.25	1.56	1.95
0	10	287.6	415.9	480.4	496.6	508.7	500.0	471.9
	50	32.3	123.2	393.1	498.4	458.1	205.4	30.5
	250	23.2	44.6	186.6	491.1	204.0	34.2	14.2
.5	10	180.0	321.1	422.3	473.7	492.7	477.7	448.1
	50	20.2	39.3	123.1	265.1	250.4	105.4	22.8
	250	16.8	25.2	41.4	63.7	47.6	23.3	12.4
1.0	10	47.2	120.5	228.3	355.0	410.6	419.0	407.2
	50	11.0	13.9	18.8	25.0	28.6	22.1	13.6
	250	10.0	12.0	14.3	16.2	15.7	12.8	9.2
1.5	10	12.3	21.0	55.2	132.7	230.9	298.6	316.9
	50	7.0	8.0	9.0	10.1	10.7	10.4	8.8
	250	6.5	7.1	7.8	8.2	8.2	7.7	6.7
2.0	10	7.2	8.8	13.6	26.9	65.8	132.2	178.9
	50	5.0	5.4	5.8	6.3	6.6	6.5	6.1
	250	4.6	4.9	5.2	5.4	5.4	5.3	5.0

Table 3. Percentage of Signals Diagnosed to Mean and to Variance Shift

$\delta$		Percent nominal significance in follow-up tests						
		$\sigma$						
		.51	.64	.80	1.00	1.25	1.56	1.95
0	Mean	8	25	66	82	77	40	13
	Variance	100	99	94	90	86	88	90
.5	Mean	65	71	84	88	82	48	18
	Variance	95	82	61	57	64	77	84
1.0	Mean	99	99	99	97	89	62	28
	Variance	66	43	23	14	21	49	70
1.5	Mean	100	100	99	96	85	62	35
	Variance	34	21	12	10	14	33	56
2.0	Mean	99	98	96	89	77	56	33
	Variance	17	12	8	8	12	27	48

probability of a type I error; that error already occurred in the signal being given.

The patterns in the table are quite complex. Moving down the  $\sigma = 1$  column, the proportion of variance signals decreases, as one would expect. The proportion of mean signals increases but then, counterintuitively, decreases. This decrease comes about because the short run before detection leads to the combined test giving a signal before sufficient observations have accrued for its  $t$  component to reach statistical significance.

Looking at the  $\delta = 0$  rows shows that, perhaps surprisingly, recognition of a variance decrease is more reliable than that of a variance increase of the same size, although all show a signal rate of at least 85%.

Moving into the non-T portion of the table shows that the simultaneous occurrence of a shift in mean reduces the ability to diagnose a shift in variance. An increase in variance hurts the ability to diagnose in mean, but a variance reduction improves it. This indeterminacy in the cause of the signal comes from the short run lengths to signal in these settings. Although it might seem a drawback that one can have a signal without a clear diagnosis, it is rather a tribute to the way the combined method is able to detect a problem before its individual component tests do.

## 6. COMPARISON WITH THE CUSUM

In advancing a new methodology, one wishes to compare it to an existing standard. A suitable benchmark is the self-starting cusum proposal, which is also intended for the situation of unknown parameters. The method was explained by Hawkins and Olwell (1998). Briefly, each successive observation  $X_n$  is Studentized using the running mean and standard deviation of all preceding observations, and the Studentized deviation (which follows a scaled  $t$  distribution) is transformed to a  $N(0, 1)$  variate  $U_n$  using the probability integral transform. A cusum of the sequence  $U_n$  then provides a test for shifts in location, and a cusum of the  $\chi_1^2$  quantity  $U_n^2$  gives a simultaneous test for shifts in variance.

A cusum is "tuned" to shifts of a specific magnitude. We used two benchmark cusum schemes, a "small-shift" scheme and a "large-shift" scheme. Each scheme comprised two pairs of cusum charts, a location chart for an upward shift in mean and another for a downward shift, and a scale chart for an upward shift in variance and another for a downward shift (see Hawkins and Olwell 1998 for details of cusum design). Each of the four charts was calibrated for an in-control ARL of 2,000, so that the combined cusum scheme would have an in-control ARL of approximately 500.

The small-shift scheme, "tuned" for a location shift of .5 standard deviations and for a scale shift of 25%, had reference value  $K = .25$  and decision interval  $H = 9.93$  for the two location cusums. The variance cusums used  $K = 1.24$ , and  $H = 22.5$  for upward shifts and  $K = .79$  and  $H = 15.37$  for downward shifts. The large-shift scheme was "tuned" for a 2 standard deviation shift in the mean and for a doubling or halving of the standard deviation. The location cusums used  $K = 1$  and  $H = 3.01$ . The upward variance cusum had  $K = 1.85$  and  $H = 13.36$ , and the downward variance cusum used  $K = .46$  and  $H = 3.96$ . Normal sequences were then simulated, and shifts were introduced after 50 in-control readings. Any sequence in which the changepoint or cusums would have given a signal before this point was discarded. The remaining sequences were used to estimate the ARL of each scheme.

Figure 2 shows the resulting ARLs in response to shifts in the mean. The performance agrees with what one would expect—the "small-shift" cusum scheme is the best of the three where the shift in mean is relatively small (below approximately 1.5 standard deviations), and the "large-shift" cusum scheme is best when the shift in mean is large (above 2 standard deviations). At all values of the shift, the changepoint is best or a close second-best method.

For scale shifts (shown in Fig. 3), the changepoint method is attractive in settings where the variance decreases but is less powerful than the cusums for variance increases. However, the good performance of the cusum schemes for variance increases is not quite what it seems; a variance increase reduces the ARL of the location cusum, whereas a variance decrease increases

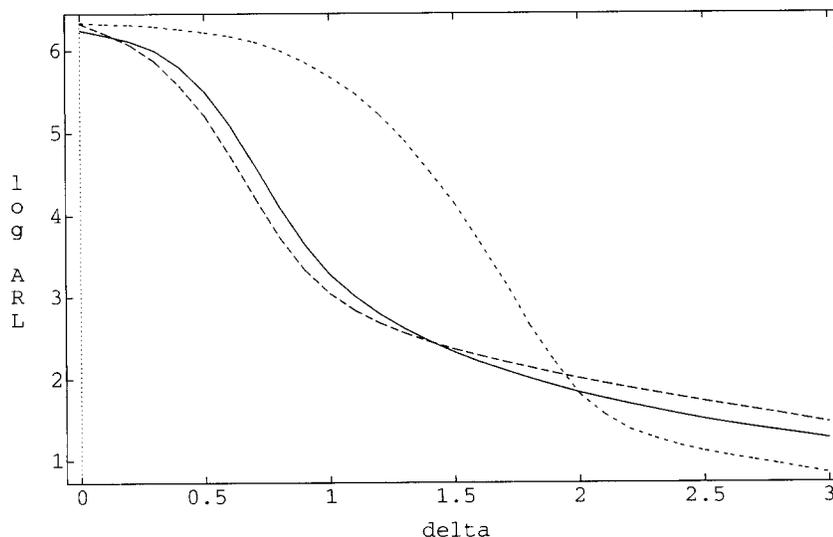


Figure 2. Comparing Changepoint (solid line) With Cusums (dash-lines).

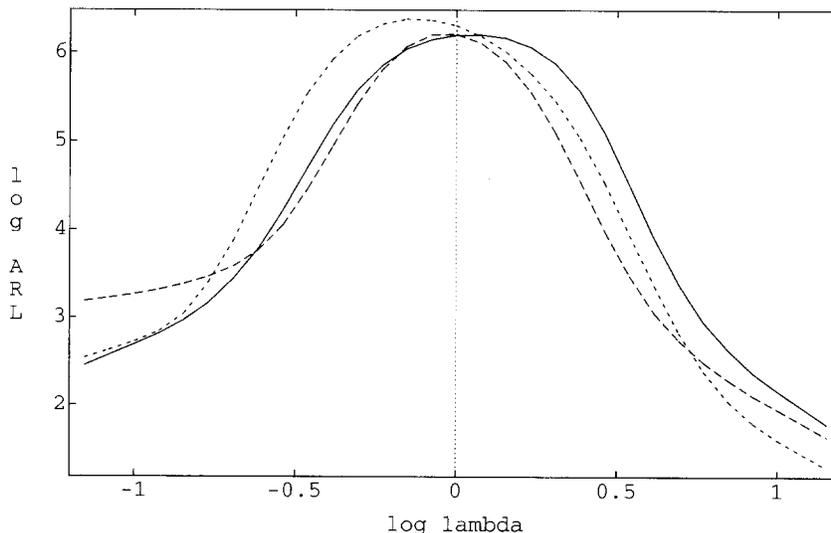


Figure 3. Comparing Changepoint (solid line) With Cusums (dash-lines).

it. A large part of the cusums' ARL reduction following small variance increases comes not from correct signals from the scale cusums, but rather from wrong signals given by the location cusums.

Overall, the changepoint approach is competitive in that, despite their potential appeal to optimality properties, neither cusum dominates it for either location or scale shifts. This bodes well for the changepoint formulation.

### 7. EXAMPLE: GOLD MINE SAMPLING QUALITY CONTROL

Samplers in gold mines extract samples of the face at regular spacings and submit them for chemical assay for their gold content. As a quality check, supervisors cut fresh samples at some of the locations already sampled. This gives rise to pairs of samples and of gold content, one by the original sampler and one by the supervisor. The log of the ratio of the gold content of the sampler to that of the supervisor has an approximately

normal distribution. This distribution should have a zero mean (or else the sampler is biased) and a small variance (or else the sampler is erratic). Figure 4, one of a number of examples given by Rowland and Sichel (1961), shows a sequence of such log ratios for a junior sampler.

Changepoints in both mean and variance are possible and important. A change in mean leads to a bias in mine valuation. An increase in variance may warn of loss of motivation to sample carefully, whereas a decrease in variance indicates a highly desirable improvement in measuring quality, perhaps as a result of learning skills. We address these questions using our changepoint formulation to look for changes in this sampler's output.

Although the log ratio was described as approximately normal, it has a potentially troublesome feature—that, due to rounding of the gold assays, the distribution is "grainy," with 10% of the log ratios being recorded as exactly 0 and some tied nonzero values. To avoid the problem of the log of a zero variance, we preclude two-point segments with identical  $X$  values.

Another potential modeling concern has to do with possible dependence. Because the entire data sequence reflects the work

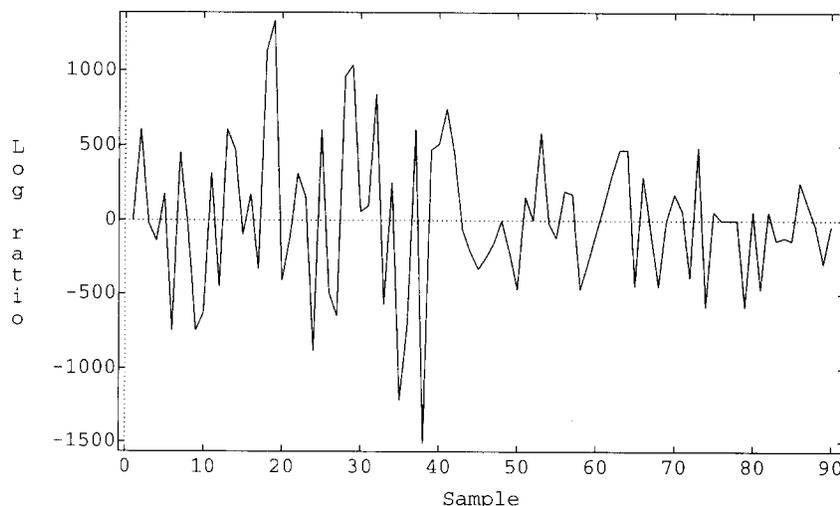


Figure 4. Sample vs. Log Ratio of Sampler to Supervisor Assay.

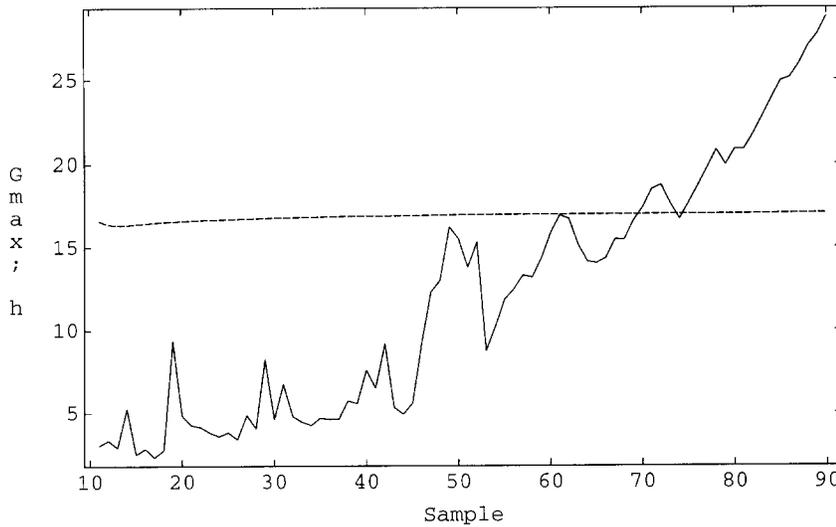


Figure 5.  $G_{max,n}$  and  $h_{.002,n}$ .

of a single sampler and supervisor, there are no cluster effects in the data. Serial correlation is possible from the sampler having “good” and “bad” days (perhaps through patches of unusually hard or friable rock), but the sequence did not demonstrate any such autocorrelation. We thus feel justified in proceeding with the changepoint analysis as described.

Figure 4 suggests that the variability in the portion up to around observation 40 is higher than that for the remainder of the sequence. The mean, however, appears to center on 0 for the whole sequence. Figure 5 gives the chart of the GLR along with the control limit,  $h_{n,.002}$ , corresponding to an in-control ARL of 500 for normally distributed data.

$G_{max,n}$  first crosses the control limit at reading 70, and continues upward except for a brief dip below at reading 74.

Right from the first hint of a changepoint, the GLR gave the left-segment estimates  $\hat{\tau} = 42$ ,  $\hat{\mu}_1 = .053$ , and  $\hat{\sigma}_1 = .42$ . Table 4 lists the summary statistics of the right segments, the GLR, the separate test statistics for identity of mean and of variance,

and the base-10 log of their normal-distribution  $p$  values. For all values in this range,  $h_{.002,n}$  is close to 17.

These values confirm that there is no indication of a shift in mean, but that the standard deviation is substantially smaller after the changepoint than before. By the time of the first signal at sample 70, the estimated postchange standard deviation has stabilized close to its ultimate value of .283—about two-thirds of its value in the first portion of the series. This reduction of variance is of substantial value to the mine; it means that each sample now produced by this junior sampler is more informative than two of his previous samples were. This has obvious implications for the improvement in mine valuation and selection of which ore to extract.

### 8. CONCLUSION

Unknown-parameter changepoint models are attractive for a number of reasons, of which not needing values of the in-control parameters is the most immediately visible. The formulation presented here for detecting changes in mean, in variance, or in both mean and variance has the further attraction of parsimony—of using a single chart rather than separate charts to monitor both mean and variance. The combined chart has good performance against process shifts.

There is a tension between the objective of starting process monitoring as soon as possible and that of gathering sufficient preliminary information to have confidence in one’s baseline data. Although we presented the changepoint formulation as one that in principle can start phase II monitoring after gathering only three phase I readings, we do not see it being used in any such extreme fashion. In some settings, such as short-run job-shop problems, there may be a considerable history supporting the expectation that the process readings will follow a normal distribution. In this case, the changepoint charting could indeed be started with the minimal accumulation of baseline information. At the other end of the spectrum are settings in which there is no advance reason to expect normally distributed process readings and there is no particular obstacle to gathering a reasonably sized phase I dataset. In this setting, we see

Table 4. The Gold Mine Summary Statistics

$i$	$\bar{X}_{\hat{\tau},i}$	$sd_{\hat{\tau},i}$	$G_{max}$	$t$	$\log P$	$F$	$\log P$
69	-.035	.297	16.60	.85	-.40	4.74	-4.39
70	-.027	.294	17.36	.79	-.36	4.83	-4.59
71	-.024	.289	18.47	.77	-.35	4.99	-4.85
72	-.037	.292	18.72	.88	-.42	4.90	-4.88
73	-.020	.303	17.60	.72	-.33	4.57	-4.68
74	-.037	.314	16.70	.87	-.41	4.24	-4.44
75	-.034	.310	17.66	.86	-.40	4.36	-4.66
76	-.033	.305	18.70	.85	-.40	4.50	-4.91
77	-.032	.301	19.76	.85	-.40	4.63	-5.15
78	-.032	.296	20.84	.85	-.40	4.77	-5.40
79	-.047	.306	19.94	.98	-.48	4.47	-5.16
80	-.044	.302	20.87	.96	-.47	4.58	-5.38
81	-.055	.306	20.86	1.06	-.53	4.47	-5.34
82	-.052	.303	21.76	1.03	-.52	4.57	-5.55
83	-.054	.299	22.79	1.06	-.53	4.67	-5.78
84	-.056	.296	23.85	1.08	-.54	4.79	-6.01
85	-.057	.292	24.91	1.10	-.56	4.89	-6.24
86	-.050	.293	25.11	1.03	-.52	4.88	-6.32
87	-.047	.290	25.94	1.01	-.50	4.97	-6.53
88	-.047	.287	27.01	1.01	-.50	5.08	-6.77
89	-.052	.286	27.71	1.06	-.53	5.11	-6.90
90	-.052	.283	28.79	1.06	-.53	5.22	-7.15

the cha  
tools, a  
of this  
Ther  
mal dis  
normal  
against  
expecte  
data to  
of the n  
In se  
other t  
distribu  
gaining  
the dev  
feasible

The  
careful  
script.  
Founda

Carlstein  
Proble  
Hawkins  
Locati  
Hawkins  
ing for  
Hawkins  
Statist

the changepoint as a valuable adjunct to other phase I analysis tools, and leading to the possibility of an earlier transition out of this precursor mode into ongoing production monitoring.

There is also the question of how appropriate it is to use a normal distribution framework if the data do not conform to the normal distribution very well. The GLR statistic is not robust against heavy tails and is likely to give more false alarms than expected. Here perhaps the easiest approach is to transform the data to approximate the tail weight (if not the full distribution) of the normal distribution.

In settings where the working distribution is known but is other than normal (e.g., when working with strength or life distributions where the Weibull distribution is the standard), gaining the full power of the changepoint approach will require the development of distribution-specific models, a nontrivial but feasible task.

### ACKNOWLEDGMENTS

The authors thank the editorial staff and the referees for their careful reading and their constructive comments on the manuscript. This work was partially supported by National Science Foundation grant DMS-03-06304.

[Received August 2003. Revised June 2004.]

### REFERENCES

- Carlstein, E. G., Müller, H.-G., and Siegmund, D. (eds.) (1994), *Change-Point Problems*, Hayward, CA: Institute for Mathematical Statistics.
- Hawkins, D. M. (1977), "Testing a Sequence of Observations for a Shift in Location," *Journal of the American Statistical Association*, 72, 180–186.
- Hawkins, D. M., and Olwell, D. H. (1998), *Cumulative Sum Charts and Charting for Quality Improvement*, New York: Springer-Verlag.
- Hawkins, D. M., Qiu, P., and Kang, C. W. (2003), "The Changepoint Model for Statistical Process Control," *Journal of Quality Technology*, 35, 355–365.
- Hawkins, D. M., and Zamba, K. D. (2005), "A Change Point Model for a Shift in Variance," *Journal of Quality Technology*, 37, 21–31.
- Jones, L. A. (2002), "The Statistical Design of EWMA Control Charts With Estimated Parameters," *Journal of Quality Technology*, 34, 277–288.
- Jones, L. A., and Champ, C. W. (2001), "The Performance of Exponentially Weighted Moving Average Charts With Estimated Parameters," *Technometrics*, 43, 156–167.
- Jones, L. A., Champ, C. W., and Rigdon, S. E. (2004), "The Run Length Distribution of the CUSUM With Estimated Parameters," *Journal of Quality Technology*, 36, 95–108.
- Kendall, M. G., and Stuart, A. (1961), *The Advanced Theory of Statistics*, Vol. 2, New York: Hafner.
- Lai, T. L. (2001), "Sequential Analysis: Some Classical Problems and New Challenges," *Statistica Sinica*, 11, 303–350.
- Lawley, D. N. (1956), "A General Method for Approximation to the Distribution of Likelihood Ratio Criteria," *Biometrika*, 43, 295–303.
- Margavio, T. M., Conerly, M. D., Woodall, W. H., and Drake, L. G. (1995), "Alarm Rates for Quality Control Charts," *Statistics and Probability Letters*, 24, 219–224.
- Montgomery, D. C. (2004), *Introduction to Statistical Quality Control* (5th ed.), Canada: Wiley.
- Pignatiello, J. J., Jr., and Samuel, T. R. (2001), "Estimation of the Change Point of a Normal Process Mean in SPC Applications," *Journal of Quality Technology*, 33, 82–95.
- Quesenberry, C. P. (1991), "SPC Q-Charts for Start-up Processes and Short or Long Runs," *Journal of Quality Technology*, 33, 3.
- (1993), "The Effect of Sample Size on Estimated Limits for  $\bar{X}$  and  $X$  Control Charts," *Journal of Quality Technology*, 25, 237–247.
- Rowland, R. St. J., and Sichel, H. S. (1961), "Statistical Quality Control of Routine Underground Sampling," *Journal of the South African Institute of Mining and Metallurgy*, 60, 251–284.
- Samuel, T. R., Pignatiello, J. J., Jr., and Calvin, J. A. (1998a), "Identifying the Time of a Step Change With Xbar Control Charts," *Quality Engineering*, 10, 521–527.
- (1998b), "Identifying the Time of a Step Change in a Normal Process Variance," *Quality Engineering*, 10, 529–538.
- Sullivan, J. H., and Woodall, W. H. (1996), "A Control Chart for Preliminary Analysis of Individual Observations," *Journal of Quality Technology*, 28, 3.
- Willsky, A. S., and Jones, H. L. (1976), "A Generalized Likelihood Ratio Approach to Detection and Estimation of Jumps in Linear Systems," *IEEE Transactions in Automatic Control*, 21, 108–112.
- Worsley, K. J. (1979), "On the Likelihood Ratio Test for a Shift in Location of Normal Populations," *Journal of the American Statistical Association*, 74, 365–367.
- (1982), "An Improved Bonferroni Inequality and Applications," *Biometrika*, 69, 297–302.