Abstract:

This paper develops a model for the analysis and design of a CUSUM chart for monitoring the process mean in short production runs and proposes statistical measures of performance that are appropriate when the process is not operating indefinitely. The model can be used to evaluate the performance of the CUSUM scheme for any given set of chart parameters and thus allows comparisons with Shewhart charts and the determination of the most suitable monitoring scheme in any particular application. The model is flexible enough to be used when the CUSUM chart is applied with Fast Initial Response, as well as when the setup and the restoration of the process after an alarm are imperfect.

Key words: Statistical Process Control; CUSUM chart; Short Runs.

1. Introduction

Control charts are widely used, especially in the manufacturing industry, for monitoring critical quality characteristics of products and processes. When the magnitude of the process disturbance due to the occurrence of assignable causes is small but important, Cumulative Sum (CUSUM) charts are more effective than Shewhart charts in monitoring the process, because they can detect small disturbances more rapidly on average. For this reason, CUSUM charts are becoming increasingly popular among practitioners.

CUSUM charts were introduced by Page (1954, 1961). They can be used both when the quality characteristic is a continuous random variable (for monitoring the mean and the variance) and when it is the fraction defective. In this paper we focus on CUSUM charts used for the monitoring of the mean value \( \mu \) of a continuous quality characteristic \( X \).

There are two ways to implement a CUSUM chart for monitoring the mean; the V-mask and the tabular (algorithmic) approach. The V-mask was originally proposed by Barnard (1959) and is applied to successive values of the CUSUM statistic:

\[
C(t) = \sum_{i=1}^{t} \frac{\bar{x}_i - \mu_0}{\sigma/\sqrt{n}} = \sum_{i=1}^{t} z_i,
\]

where \( \bar{x}_i \) is the mean of sample \( i \), \( \mu_0 \) is the target mean, \( \sigma \) is the standard deviation of \( X \) and \( n \) is the sample size. Statistical properties of CUSUM charts using the V-mask approach have been studied by Johnson (1961) and Goldsmith and Whitfield (1961).
Over the years, the tabular approach has prevailed, mainly because of its simplicity compared to the V-mask. When using the tabular approach to identify upward shifts in the mean, at each sample \( t \) after the initial setup the CUSUM statistic is evaluated by:

\[
C(t) = \max\{0, C(t-1) + z_t - k\}, \quad C(0) = c_0 \geq 0 \tag{1}
\]

where \( z_t \) is again the standardized observation and \( k \) is the reference value (the formulation for downward shifts is analogous and both the CUSUM statistics can be used in case of two-sided charts). The CUSUM chart issues an alarm if and when \( C(t) \) exceeds the control limit \( H \). When the process setup may be imperfect the CUSUM scheme often is used with Fast Initial Response (FIR); the CUSUM statistic is given a head-start by setting \( C(0) = c_0 > 0 \) rather than \( C(0) = 0 \), so as to identify the possible deviation from \( \mu_0 \) more effectively. For the same reason, if the setup of the process after a correct signal may be imperfect, after each attempt to restore the process to the in-control condition, \( C(t) \) is reset to the nonzero value \( c_0 \). However, if the alarm at sample \( t \) proves to be false, \( C(t) \) is set equal to zero.

The run length until the chart triggers a signal is the key measure of the performance of a CUSUM procedure. The mean value of the run length, \( ARL \) (Average Run Length), is used in practice to select the most appropriate CUSUM scheme. There are several papers that deal with the effectiveness of the CUSUM chart. Vance (1986) provides a computer program for evaluating \( ARL \), Hawkins (1992) gives a relatively simple yet very accurate approximating equation for the evaluation of the \( ARL \). Several Markov chain approaches have been used for the computation of the \( ARL \), like the ones by Ewan and Kemp (1960), Brook and Evans (1972) and Fu et al. (2002).

The analysis of CUSUM chart properties is typically based on the often implicit assumption that the process which is monitored will be operating continuously and indefinitely. In practice, most production processes are periodically set up to produce a specific quantity over a specified time period, e.g., an 8-hour shift. In such cases, the limited duration of the production run has to be taken into account in the design of the process control scheme in order for this scheme to be maximally effective. It is not at all obvious that CUSUM schemes will be effective in short runs, since the limited number of samples that can be taken does not fit well with the accumulative character of the scheme. The purpose of this paper is:

- to propose statistical measures of performance of control charts that are appropriate for short runs;
- to present a model for the evaluation of a CUSUM scheme for monitoring the process mean in short runs;
- to compare the effectiveness of CUSUM charts against that of Shewhart charts in short runs.

The next section describes the problem setting and assumptions. The stochastic model that expresses the operation of the CUSUM scheme is presented in section 3. The measures of performance are developed and presented in section 4. Section 5 presents a comparison between CUSUM and Shewhart charts in short runs through some numerical examples.

2. Problem Setting and Assumptions

We consider a production process that is set up for processing a specific batch of items over a limited time interval (short run). The key measure of the process quality is a continuous random variable \( X \), which is normally distributed with target value \( \mu_0 \) and constant variable \( \sigma^2 \). The setup operation may not always be perfect in the sense that although in the beginning of the process the mean of \( X \) is supposed to be set equal to \( \mu_0 \), there is a probability that the process starts its operation with a mean different from \( \mu_0 \).

The process may be affected by the unobservable occurrence of an assignable cause at some random time. The effect of the assignable cause is a shift in the mean of \( X \) from \( \mu_0 \) to \( \mu_1 = \mu_0 + \delta \sigma \) but no change in
the variance $\sigma^2$. The process remains in that undesirable out-of-control state, until the occurrence of the assignable cause is detected and its effect removed.

The process is monitored by means of an one-sided CUSUM chart for detecting a possible upward shift in the mean using the CUSUM statistic (1). The total number of samples that will be taken till the end of the process is $N$. If the control chart indicates a possible out-of-control condition, the process is stopped for investigation; if the investigation reveals that the assignable cause has indeed occurred, then there is an intervention to restore the process to its in-control condition and operation resumes. This intervention may be imperfect and despite the detection of the cause and the attempt to remove it, the process may continue operating out of control.

There are five parameters that affect the operation and effectiveness of the CUSUM chart during the run:

- the total number of samples $N$,
- the sample size $n$,
- the reference value $k$,
- the control limit $H$ and
- the value $c_0$ of $C(t)$ at the beginning of the process ($t = 0$) and after a true alarm at sample $t$.

The values of $N$, $n$, $k$, $H$, $c_0$ obviously affect both the statistical and economic performance of the monitoring scheme. Therefore, it is necessary to develop a model for evaluating the appropriate measures of performance for any combination of these five parameters, so as to be able to select, at a second stage, the set of parameter values that satisfy our requirements.

3. Stochastic Model

In this section we develop a discrete-time stochastic model for the process and its monitoring scheme, based on the value of the CUSUM statistic $C(t)$ for $t = 0, 1, ..., N$. Although $C(t)$ is theoretically a continuous random variable, for practical computational reasons we discretize it into $m+1$ values following the well-established approach of Brook and Evans (1972). Specifically, we partition the interval from 0 to $H$ into $m$ sub-intervals by first defining $w$, the width of sub-intervals 1 to $m-1$, as follows:

$$w = \frac{2H}{2m-1} \iff m = \frac{H}{w} + \frac{1}{2}.$$  \hfill (2)

Then, the real-valued $C(t)$ is transformed to an integer between 0 and $m$ in the following manner:

$$\begin{align*}
\text{for } & C(t) < \frac{w}{2} \rightarrow C(t)=0 \\
\text{for } & \left( i-\frac{1}{2} \right) w \leq C(t) < \left( i+\frac{1}{2} \right) w \rightarrow C(t)=i \quad i = 1,2, ..., m-1 \\
\text{for } & \left( m-\frac{1}{2} \right) w = H \leq C(t) \rightarrow C(t)=m.
\end{align*}$$  \hfill (3)
If the process is actually in statistical control at sampling instance \( t (\mu = \mu_0) \), \( z_t \) follows the standard normal distribution \( z_t \sim N(0,1) \). If the process is out-of-control when sample \( t \) is taken \( (\mu = \mu_1 = \mu_0 + \delta \sigma) \) then \( z_t \) follows a normal distribution with mean \( \delta \sqrt{n} \) and variance equal to 1: \( z_t \sim N(\delta \sqrt{n},1) \).

The probabilities \( p_{ij} \) of moving from \( C(t-1) = i \) to \( C(t) = j \) may be computed from:

\[
p_{ij} = \begin{cases} 
\int_{-\infty}^{(\frac{i-1}{2})w+k-\delta \sqrt{n}} \phi(z)dz & i = 0,1,\ldots,m-1, \quad j = 0 \\
\int_{(\frac{i-1}{2})w+k-\delta \sqrt{n}}^{(\frac{j-i+1}{2})w+k-\delta \sqrt{n}} \phi(z)dz & i = 0,1,\ldots,m-1, \quad j = 1,2,\ldots,m-1 \\
\int_{(\frac{j-i+1}{2})w+k-\delta \sqrt{n}}^{\frac{m}{2}w+k-\delta \sqrt{n}} \phi(z)dz & i = 0,1,\ldots,m-1, \quad j = m \\
1 & i = m, \quad j = 0,1,\ldots,m-1 \\
0 & i = m, \quad j = m 
\end{cases}
\]

where \( \phi(z) \) is the density function of the standard normal distribution, \( \delta = 0 \) if the process is under statistical control and \( \delta = (\mu_1 - \mu_0)/\sigma \) if the process operates out of control. In the latter case the transition probabilities \( p_{ij} \) are denoted \( \tilde{p}_{ij} \). To simplify notation, when the indices \( i, j \) of the transition probabilities are equal to 0 or \( m \), they will be denoted 0 or \( m \), e.g., \( p_{0m} \equiv \tilde{p}_{ij} \) for \( i = 0 \) and \( j = m \).

\( C(t) \) evolves as a Markov chain with \( m+1 \) states and transition probabilities \( p_{ij} \) if the process is in statistical control. The transition probability matrix \( P \) may be written as:

\[
P = \begin{bmatrix}
p_{00} & p_{01} & \cdots & p_{0m-1} & p_{0m} \\
p_{10} & p_{11} & \cdots & p_{1m-1} & p_{1m} \\
\vdots & \vdots & \ddots & \vdots & \vdots \\
p_{m-10} & p_{m-11} & \cdots & p_{m-1m-1} & p_{m-1m} \\
0 & 0 & \cdots & 0 & 1
\end{bmatrix} = \begin{bmatrix}
A_{(m\times m)} & B_{(m\times 1)} \\
0_{(1\times m)} & 1_{(1\times 1)}
\end{bmatrix}.
\]

If the process operates in the out-of-control state, then we obtain in a similar way the matrix \( \tilde{P} \), which is the exact analogue of \( P \) with transition probabilities \( \tilde{p}_{ij} \).

The matrixes \( P \) and \( \tilde{P} \) are divided into 4 parts, as shown in (5). Multiplication of \( P \) by itself yields the following form of \( P^t \) (the form for \( \tilde{P}^t \) is analogous):
\[ P^t = \begin{bmatrix} A_{(m,m)} & W_t B_{(m+1)} \\ 0_{(1,m)} & 1_{(1,1)} \end{bmatrix} \quad t = 1, 2, \ldots, N \]  

where

\[ W_t = 1 + A + A^2 + \ldots + A^{t-1} = \sum_{x=0}^{t-1} A^x. \]  

4. **Statistical Measures of Performance**

The most common statistical measure of a control chart’s performance is the Average Run Length (ARL), i.e., the expected number of samples until the chart triggers a signal, given that the process remains in the same condition, either in-control (ARL\(_0\)), or out-of-control (ARL\(_\delta\)). In the case of a short run, where the total number of samples \( N \) is finite and limited, the run may end without the chart having issued any out-of-control signal. Therefore, we define Truncated ARL, denoted by TARL\(_0\) or TARL\(_\delta\), as the mean number of samples until a signal or until the completion of the process, whichever occurs first. If the run ends without any signal in the \( N \) samples, the truncated run length is assigned the value \( N+1 \).

Another useful characteristic of a control chart in short runs is the probability of getting a signal within a certain number of samples \( N \) when operating in control, \( q_0(N) \), and when operating out of control, \( q_\delta(N) \).

Finally another useful measure of the chart’s effectiveness is the average number of false alarms within a certain number of samples \( N \), which we denote by \( F(N) \). The expected percentage of false alarms in \( N \) samples \( F(N)/N \), may be viewed as the analogue of the type I error (false alarm probability).

If the process is monitored with a \( k\)-sigma Shewhart one-sided \( \bar{X} \)-chart, the above performance measures are computed as follows:

\[ TARL_0(N) = \sum_{i=1}^{N} i(1-\alpha)^{i-1} \alpha + (N+1)(1-\alpha)^N = \frac{1-(1-\alpha)^{N+1}}{\alpha} \]

\[ TARL_\delta(N) = \sum_{i=1}^{N} i\beta^{i-1}(1-\beta) + (N+1)\beta^N = \frac{1-\beta^{N+1}}{1-\beta} \]

\[ q_0(N) = 1 - (1 - \alpha)^N \]

\[ q_\delta(N) = 1 - \beta^N \quad \text{and} \]

\[ F(N) = N \alpha \]

where \( \alpha = \Phi(-k) \) is the probability of a type I error and \( \beta = \Phi(k\sigma \sqrt{n}) \) is the probability of a type II error at each sample \( t \). \( \Phi(\cdot) \) is the cumulative distribution function of the standard normal distribution.

If the short run is monitored with the one-sided CUSUM chart described in the previous sections, then TARL\(_0\) is the expected value of the following random variable \( U \):
Using an approach similar to Fu et al (2002), we find for $C(0) = c_0 = 0$, that

$$\text{TARL}_0 = \left[ \begin{array}{ccc} 1 & 0 & \ldots & 0 \\ m \text{ elements} \end{array} \right] W_{N+1} \cdot \left[ \begin{array}{c} 1 \\ \ldots \\ 1 \end{array} \right]^T = 1 + \sum_{i=0}^{N} \sum_{j=0}^{m-1} P_{ij}^{(0)} . \quad (9)$$

In a similar way we can evaluate $\text{TARL}_0$ for a CUSUM chart with Fast Initial Response, i.e., starting with $C(0) = c_0 > 0$. Note that the value of $\text{TARL}_0$ in (9) is essentially the first element of the $m \times 1$ vector $W_{N+1}[1 \ 1 \ldots \ 1]^T$; in the case of an FIR-CUSUM with $C(0) = c_0 > 0$, $\text{TARL}_0$ is the element located in the $i_0$-th row of this column vector, where $i_0 = c_0 / w$, i.e., $i_0$ is the state of the transition probability matrix $P$ that corresponds to $C(t) = c_0$. $\text{TARL}_0$ is computed in the same way but using the elements of matrix $\tilde{P}$.

The probability that the CUSUM chart issues a false alarm within $N$ samples, when the process stays in control is:

$$q_0(N) = P_{im}^{(N)} . \quad (10)$$

Similarly, the probability that the CUSUM chart correctly indicates a problem (true alarm) within $N$ samples after the mean shifts to $\mu_0 + \delta \sigma$ is:

$$q_0(N) = \tilde{P}_{im}^{(N)} . \quad (11)$$

In order to evaluate $F(N)$ for the CUSUM chart, the bottom row of matrix $P$ in (5) must be modified and made identical to the top row (state 0) so that state $m$ stops being absorbing. Then $F(N)$ is computed from

$$F(N) = \sum_{i=1}^{N} P_{im}^{(i)} \quad (12)$$

using the modified $P$.

5. Numerical Illustration – Comparison with Shewhart Charts

We investigate the properties of a CUSUM chart designed to identify upward mean shifts of size $\delta = 1$, with $m = 1$, $k = 0.5$ and $C(0) = 0$, for various values of $N$ and $H$. The width $w$ of the discretization interval $(0,H)$ is set equal to $0.1$. In addition, we compare its performance against that of an one-sided Shewhart chart with 3-sigma limit ($k_s = 3$) and $n = 1$.

Figures 1 and 2 show that $\text{TARL}_0$ and $\text{TARL}_\delta$ of the CUSUM are increasing in $N$, but $\text{TARL}_\delta$ tends to stabilize quickly, implying that $\text{TARL}_\delta$ approaches $A\text{RL}_\delta$ of the respective “infinite” process. As expected, $\text{TARL}_0$ and $\text{TARL}_\delta$ are increasing in $H$. $\text{TARL}_0$ of the Shewhart chart with $k_s = 3$ has a very similar behavior with $\text{TARL}_0$ of the CUSUM chart with control limit $H = 5.05$. The $\text{TARL}_\delta$ of the Shewhart chart, though, is considerably larger than the respective one of the CUSUM chart with $H = 5.05$, especially for $N > 10$ (since the CUSUM chart needs some time to accumulate the observations in order to be effective).
Figures 3 and 4 tell a similar story from the point of view of the false and true alarm probabilities $q_0(N)$ and $q_δ(N)$. The value of $q_0(N)$ of the Shewhart chart lies between the values of $q_0(N)$ of the CUSUM charts with $H = 4.05$ and $H = 5.05$. On the other hand, for $N > 5$, the detection probabilities $q_δ(N)$ of the CUSUM charts are much larger than those of the Shewhart chart. For example $q_δ(N = 10) = 0.601$ for the CUSUM with and $H = 5.05$ while $q_δ(N = 10) = 0.206$ for the Shewhart chart with $k_s = 3$.

Figures 5 and 6 show how the FIR feature helps to significantly shorten the average time required to identify an out-of-control condition, but at the same time increases the proportion of false alarms, $F(N)/N$, especially for low $N$. 
The main conclusion of this brief numerical comparison is that the statistical properties of the CUSUM scheme in short runs are superior to those of the Shewhart chart with 3-sigma limit when each sample contains a single measurement. More extensive and detailed experimentation is required to determine the degree to which this finding can be generalized for larger sample sizes and for different magnitudes of shift in the process mean.

References


