based on the expected cost is reasonably robust to small changes in the parameters of the prior distribution.

# X-Bar and R Charts for Skewed Populations

최인수 (KAIST 산업공학과)
배도선 (KAIST 산업공학과)

This paper proposes a heuristic method based on a weighted variance (WV) concept in setting up the control limits of X-Bar and R control charts for skewed populations. This method provides asymmetric control limits in accordance with the direction and degree of skewness estimated from the sample data by using different variances in computing the upper and lower control limits. For symmetric populations, however, these control limits are equivalent to those of Shewhart control charts. The new control charts are compared with Shewhart control chart by a Monte Carlo simulation when the underlying populations are normal and Weibull, and are found to provide better performances than Shewhart control charts as skewness increases.

# Simulating the Average Run Length for CUSUM Schemes Using Variance Reduction Techniques

Moon Soo Choi and Chi-Hyuck Jun

Department of Industrial Engineering
Pohang Institute of Science and Technology
Pohang, Korea

## Abstract

본 논문에서는 어떤 공정이 일반적인 확률분포를 따른다는 가정하에, 시뮬레이션에 의한 CUSUM 챠트의 ARL을 추정하는 방법에 관하여 기술하였다. 추정치에 대한 분산을 최소화하기 위하여 TOTAL HAZARD 방법을 적용하였으며, 지수분포를 따르는 공정에 대하여 HAZARD 및 CYCLE 추정치와 분산감소법을 적용하지 않았을 경우의 추정치와 비교분석하였다.

## I. Introduction

Let us consider a process that produces some items sequentially. Suppose that these items have measurable values attached to them and that when the process is "in control" a sequence of these values, $X_1$, $X_2, \cdots$, come from a probability distribution $F(\cdot)$. Let $S_0=0$ and define

$$S_i = max(0, S_{i-1}+X_i-k),\ i \geq 1,$$

where $k$ is given. The quantities $S_i$ are cumulative sums of the successive quantities $X_i-k$ which are not allowed to become negative. That is, the value of the cumulative sum is reset to zero whenever it would fall negative. Usually we choose $k$ so that the values $X_i-k$ have a negative mean when the process is in control. Therefore the cumulative sum will normally have a small value in this case and a large value is an indicator that the process may have gone out of control. However, since large values will eventually occur even when the process

remains in control, we are interested in estimating $\Theta=E[N]$ where

$$N = min(i: S_i \geq h), \tag{1}$$

for some constant $h$. $E[N]$ is the expected time until the process is mistakenly declared out of control when in fact it remains in control throughout. We often call $E[N]$ the average run length (ARL).

Many authors have proposed various methods to evaluate ARL's for CUSUM schemes. Page [4] originally derived an integral equation whose solution gives ARL's for one–sided CUSUM schemes. Goel and Wu [1] obtained approximate ARL's for the normal case using ratios of numerical solutions to two integral equations. Lucas [3] used a Markov chain approach to obtain approximate ARL's for the exponential case. In general, however, obtaining ARL's analytically is a difficult task. Therefore we often use simulation.

Since the raw estimator of $E[N]$ is just $N$, given simulated data from the $r$ runs of simulation $N_1, N_2, \cdots, N_r$, $E[N]$ can be estimated by the sample mean $\overline{N}=\Sigma N_i/r$ and $Var[N]$ by the sample variance $s_N^2=\Sigma(N_i-\overline{N})^2/(r-1)$. The raw simulation is easy to follow but it might need a large number of simulation runs to obtain an accurate estimate.

In this paper, we propose efficient methods to obtain the ARL of the one–sided CUSUM scheme through simulation by using variance reduction techniques.

## II. Controlled Estimator

We can improve the raw simulation estimator $N$ by using the method of control variate if we find some variable $Y$ (called a control variate) which is highly correlated with $N$ and whose mean $E[Y]$ is known. For any constant $a$, the controlled estimator is given by

$$N(a) = N + a(Y - E[Y]).$$

Then the best value of $a$ which minimizes $Var[N(a)]$ is

$$a^* = - Cov[N, Y]/Var[Y]$$

and the variance of $N(a)$ at $a^*$ is

$$Var[N(a^*)] = (1 - R_{NY}^2) Var[N] \tag{2}$$

where $R_{NY} = Cov[N, Y]/(Var[N]Var[Y])^{1/2}$ is the correlation coefficient between $N$ and $Y$. The above (2) tells us that we could reduce variance of the raw simulation estimator by using a controlled estimator. For more details on control variates method, see Lavenberg and Welch [2].

Ross [5] suggested using the total hazard as a control variate. The total hazard for a Markov chain $\{S_i, i \geq 0\}$ and a given set of states B is defined by

$$Y = \sum_{i=1}^{N} \Lambda_i$$

where $\Lambda_i = P(N=i \mid S_0, \cdots, S_{i-1})$ and $N = min(i > 0: S_i \in B)$. $\Lambda_i$'s are called random hazards. Ross also showed that

$$E[Y] = P(N < \infty) = 1, \tag{3}$$

which indicates that $Y$ could be utilized as a control variate.

In order to find a control variate (total hazard) for our CUSUM scheme, we first evaluate the random hazard at time i, $\Lambda_i$, by

$$\begin{aligned} \Lambda_i &= P(N=i \mid S_{i-1}) \\ &= P(S_i \geq h \mid S_{i-1}) \\ &= P\{X_i \geq -(S_{i-1} - k - h)\} \\ &= F^c(k + h - S_{i-1}), \end{aligned}$$

where $F^c(\cdot) = 1 - F(\cdot)$.

Then we can use the sum of the random hazards, $Y$,

$$Y = \sum_{i=1}^{N} F^c(k + h - S_{i-1}) \tag{4}$$

as a control variate and propose the following controlled estimator of the ARL:

$$N(a) = N + a(Y-1),$$

where $a=-Cov[N,Y]/Var[Y]$ which needs to be estimated.

Given a sequence of $r$ pairs of simulated output $(N_1, Y_1)$, $(N_2, Y_2), \cdots, (N_r, Y_r)$, $E[N]$ can be estimated by the sample mean of $\hat{N(a)}$:

$$\hat{\Theta}_1 = \bar{N} + \hat{a}(\bar{Y}-1), \tag{5}$$

where $\hat{a}$ is given by

$$\hat{a} = -(\Sigma N_j Y_j - r\bar{N}\bar{Y})/(\Sigma Y_j^2 - r\bar{Y}^2).$$

The variance of the hazard controlled estimator (5) could be estimated by

$$\hat{Var}[\hat{\Theta}_1] = s_N^2(1 - \hat{R}_{NY}^2)/r \tag{6}$$

where $s_N^2$ is the estimate of $Var[N]$ and $\hat{R}_{NY}$ is the sample correlation coefficient between $N$ and $Y$.

## III. Ratio Estimators Using a Cycle

Let us define a cycle to occur either when $S_i$ exceeds $h$ or when it returns to zero, and let $C$ denote the length of a cycle. That is,

$$C = min\{i: S_i=0 \text{ or } S_i \geq h\}. \tag{7}$$

Then the ARL can be obtained by

$$E[N] = E[C]/p,$$

where $p=P\{S_C \geq h\}$. We need to estimate $E[C]$ and $p$ in order to obtain $E[N]$.

In this method we only need to simulate quantities during a cycle. That is, we consider that a single simulation run be completed when a cycle ends.

### 1. Estimation of $E[C]$

The total hazard during a cycle $C$ is obtained by

$$Z = \sum_{i=1}^{C} [P\{S_i \geq h | S_{i-1}\} + P\{S_i=0 | S_{i-1}\}]$$

$$= \sum_{i=1}^{C} \{F^c(k+h-S_{i-1}) + F(k-S_{i-1})\} \qquad (8)$$

Since $E[Z]=1$ (according to the same line with (3)) the following controlled estimator for $E[C]$ are obtained:

$$C(a_1) = C + a_1(Z-1),$$

where $a_1=-Cov[C,Z]/Var[Z]$.

We could also obtain an estimator improving upon the raw simulation estimator $C$ by using a stratified sampling. Conditioning on whether $C=1$, or not, $E[C]$ can be evaluated by

$$E[C] = E[C|\, C=1]q + E[C|\, C>1](1-q)$$
$$= q + (1-q)E[C|\, C>1],$$

where $q=P\{C=1\}=F(k)+F^c(k+h)$ and $E[C|\, C>1]$ remains to be evaluated by simulation. Let $C'=C|\, C>1$, then use

$$C'' = q + (1-q)C' \qquad (9)$$

as an estimator of $E[C]$.

Furthermore, we could improve $C''$ by using $Z'$ $(\equiv Z|\, C>1)$ as a control variate when estimating $E[C']$ in (9). That is, use

$$C''(a_2) = q + (1-q)\{C'+a_2(Z'-E[Z'])\} \qquad (10)$$

where $a_2=-Cov[C',Z']/Var[Z']$ and $E[Z']=1+q$.

## 2. Estimation of $p$

Let us define an indicator variable $I$ as follows:

$$I = \begin{cases} 1 & \text{if } S_C \geq h \\ 0 & \text{otherwise.} \end{cases}$$

Since $E[I]=p$, the raw simulation uses just $I$ as an estimator of $p$. In this section, we propose some other estimators which could improve upon the raw simulation estimator $I$.

Since $E[\sum_{i=1}^{C}\Lambda_i]=p$, $p$ can be estimated by the sum of the hazards of $N$ during a cycle, $Q$, which is

$$Q = \sum_{i=1}^{C}\Lambda_i$$
$$= \sum_{i=1}^{C} F^c(k+h-S_{i-1}). \tag{11}$$

We can improve upon the hazard estimator $Q$ by using a stratified sampling. Since

$$E[Q] = E[Q|\,C=1]q + E[Q|\,C>1](1-q)$$
$$= qF^c(k+h) + (1-q)E[Q|\,C>1],$$

$p$ can be estimated by

$$Q'' = qF^c(k+h) + (1-q)Q', \tag{12}$$

where $Q'=Q|\,C>1$.

We could further obtain a better estimator of $p$ by controlling $Q'$ in (12) via the control variate $Z'$ used in (10). In this case, we have

$$Q''(b_1) = qF^c(k+h) + (1-q)\{Q'+b_1(Z'-E[Z'])\} \tag{13}$$

where $b_1=-Cov[Q',Z']/Var[Z']$.

## 3. Ratio Estimator of ARL and Its Variance

We suggest using (10) for estimating $E[C]$ and (13) for $p$. Suppose that $V$ is an estimator for $E[C]$ and $W$ is an estimator for $p$. Suppose also that we have $m$ pairs of simulation output $(V_1,W_1),\cdots,(V_m,W_m)$. Then the classical estimator of the ratio $E[C]/p$ $(=\Theta)$ is

$$\delta = \overline{V}/\overline{W}. \tag{14}$$

Since $\delta$ is biased, the following mean squared error (MSE) will be evaluated for its performance:

$$MSE(\delta) = E[(\delta-\Theta)^2]. \tag{15}$$

We employ the bootstrap approach as described in Ross ([6], PP. 107–108) to estimate the MSE.

## IV. A Comparison of Methods for Exponential Case

The performance of the hazard controlled estimator and the cycle estimator proposed in the previous sections will be compared against the raw simulation estimator when the underlying process follows the exponential distribution with mean 1. Estimated variances of the estimators (MSE for the ratio estimator) will be compared with each other for several combinations of $k$ and $h$.

The number of simulation replications is 1000, where each replication lasts until $N$ is realized (i.e. $S_i \geq h$). Obviously each replication consists of one or more cycles defined earlier, and therefore the number of observations for the cycle estimator would be much larger than that for the raw or hazard estimators. To estimate MSE for the cycle method, we generate 200 bootstrap samples. The simulation procedures are following:

1) Generate simulated output: $N_1, \cdots, N_r$ from (1), $Y_1, \cdots, Y_r$ from (4), $C_1, \cdots, C_m$ ($m$ is the total number of cycles) from (7), $Z_1, \cdots, Z_m$ from (8), and $Q_1, \cdots, Q_m$ from (11).
2) Compute each estimator: i) hazard estimator by (5), ii) to obtain the cycle estimator, first calculate (10) and (13) (here $a_2$ and $b_1$ should be estimated first). Let $\overline{V}$ and $\overline{W}$ be the average of (10) and (13) over $m$ values, respectively. Then the cycle estimator is obtained by (14).
3) Estimate variance of each estimator: use (6) for the hazard estimator and (15) with bootstrapping for the cycle estimator. (We also have experiments to estimate the MSE of the hazard estimator using bootstrapping and find that its MSE is almost identical to the estimated

variance from (6).)

The results of simulation are summarized in the Table 1. Variance ratios (last two columns) are obtained by dividing the variance of the raw estimator by that of the hazard (or cycle) estimator. As shown in the table, the hazard and the cycle estimators give us much smaller variance than the raw simulation does. In general, the cycle estimator performs better than the hazard estimator.

## V. Concluding Remarks

We may have some computational burden when the cumulative distribution function $F(\cdot)$ cannot be easily evaluated. In this case, we suggest using numerical tables instead of evaluating $F(\cdot)$ by integration or summation. On the other hand, if the process follows some empirical distribution, our techniques can be easily applied.


## Acknowledgements

This work was partially supported by KOSEF Engineering Research Center at POSTECH.

Table 1. Simulation Variances of ARL Estimators

| h | k | ARL | Variance of Estimator | | | Variance Ratio (vs Raw Estimator) | |
|---|---|---|---|---|---|---|---|
| | | | Raw | Hazard | Cycle | Hazard | Cycle |
| 0.5 | 0.5 | 2.54 | 0.00408 | 0.00005 | 0.00002 | 86.8 | 204.6 |
| | 1.0 | 4.31 | 0.01565 | 0.00005 | 0.00002 | 300.5 | 716.9 |
| | 1.5 | 7.21 | 0.04654 | 0.00007 | 0.00002 | 675.8 | 1886.5 |
| | 2.0 | 12.01 | 0.11066 | 0.00006 | 0.00003 | 1738.9 | 4244.8 |
| | 2.5 | 19.91 | 0.33168 | 0.00007 | 0.00002 | 4688.7 | 16667.3 |
| | 3.0 | 32.94 | 1.02665 | 0.00006 | 0.00002 | 17964.0 | 50325.7 |
| 1.0 | 0.5 | 3.51 | 0.00745 | 0.00051 | 0.00031 | 14.6 | 23.7 |
| | 1.0 | 6.39 | 0.03357 | 0.00077 | 0.00036 | 43.6 | 93.5 |
| | 1.5 | 11.18 | 0.09580 | 0.00080 | 0.00044 | 120.2 | 218.4 |
| | 2.0 | 19.09 | 0.29419 | 0.00087 | 0.00041 | 339.4 | 711.3 |
| | 2.5 | 32.11 | 0.89840 | 0.00086 | 0.00036 | 1039.6 | 2495.2 |
| | 3.0 | 53.60 | 3.10396 | 0.00092 | 0.00039 | 3364.9 | 8061.0 |
| 1.5 | 0.5 | 4.50 | 0.01137 | 0.00157 | 0.00115 | 7.2 | 9.9 |
| | 1.0 | 8.97 | 0.06053 | 0.00392 | 0.00211 | 15.4 | 28.7 |
| | 1.5 | 16.84 | 0.23017 | 0.00416 | 0.00260 | 55.3 | 88.5 |
| | 2.0 | 29.87 | 0.77224 | 0.00476 | 0.00264 | 162.3 | 292.0 |
| | 2.5 | 51.36 | 2.83776 | 0.00520 | 0.00267 | 545.3 | 1064.7 |
| | 3.0 | 86.78 | 7.21764 | 0.00541 | 0.00226 | 1333.9 | 3193.9 |
| 2.0 | 0.5 | 5.50 | 0.01571 | 0.00340 | 0.00267 | 4.6 | 5.9 |
| | 1.0 | 12.06 | 0.09895 | 0.01174 | 0.00896 | 8.4 | 11.0 |
| | 1.5 | 24.76 | 0.52061 | 0.01675 | 0.01273 | 31.1 | 40.9 |
| | 2.0 | 46.21 | 2.17266 | 0.02074 | 0.01212 | 104.8 | 179.3 |
| | 2.5 | 81.63 | 6.45036 | 0.02200 | 0.01422 | 293.2 | 453.7 |
| | 3.0 | 140.00 | 21.54582 | 0.02470 | 0.01064 | 872.5 | 2025.0 |
| 2.5 | 0.5 | 6.50 | 0.01953 | 0.00719 | 0.00608 | 2.7 | 3.2 |
| | 1.0 | 15.64 | 0.16336 | 0.02891 | 0.02315 | 5.7 | 7.1 |
| | 1.5 | 35.68 | 1.25932 | 0.05864 | 0.04808 | 21.5 | 26.2 |
| | 2.0 | 70.77 | 4.74412 | 0.06463 | 0.04539 | 73.4 | 104.5 |
| | 2.5 | 129.10 | 18.16911 | 0.08127 | 0.04942 | 223.6 | 367.7 |
| | 3.0 | 225.40 | 49.70991 | 0.08395 | 0.06192 | 592.2 | 802.8 |
| 3.0 | 0.5 | 7.50 | 0.02246 | 0.00969 | 0.00909 | 2.3 | 2.5 |
| | 1.0 | 19.72 | 0.32321 | 0.06372 | 0.06338 | 5.1 | 5.1 |
| | 1.5 | 50.65 | 2.50344 | 0.18998 | 0.15733 | 13.2 | 15.9 |
| | 2.0 | 107.60 | 11.95971 | 0.22912 | 0.22778 | 52.2 | 52.5 |
| | 2.5 | 203.60 | 41.55061 | 0.25744 | 0.19022 | 161.4 | 218.4 |
| | 3.0 | 362.30 | 135.20432 | 0.22402 | 0.18683 | 603.5 | 723.7 |

## References


[1] Goel, A. L.; and Wu, S. M. (1971). "Determination of A.R.L. and a Contour Nomogram for CUSUM Charts to Control Normal Mean", *Technometrics*, 13, pp. 221–230.

[2] Lavenberg, S. S.; and Welch, P.D. (1981). "A Perspective On the Use of Control Variables To Increase the Efficiency of Monte Carlo Simulations", *Management Science*, 27, 3, pp. 322–335.

[3] Lucas, J.M. (1985). "Counted Data CUSUM's", *Technometrics*, 27, pp. 129–144.

[4] Page, E.S. (1954). "Continuous Inspection Schemes", *Biometrika*, 41, pp. 100–114.

[5] Ross, S.M. (1990a). "Variance Reduction in Simulation Via Random Hazards", *Probability in the Engineering and Informational Sciences*, 4, pp. 299–309.

[6] Ross, S.M. (1990b). *A Course In Simulation*, Macmillian Publishing Co., New York.

[7] Vardeman, S.; and Ray, D. (1985). "Average Run Lengths for CUSUM Schemes When Observations Are Exponentially Distributed", *Technometrics*, 27, 2, pp. 145–150.