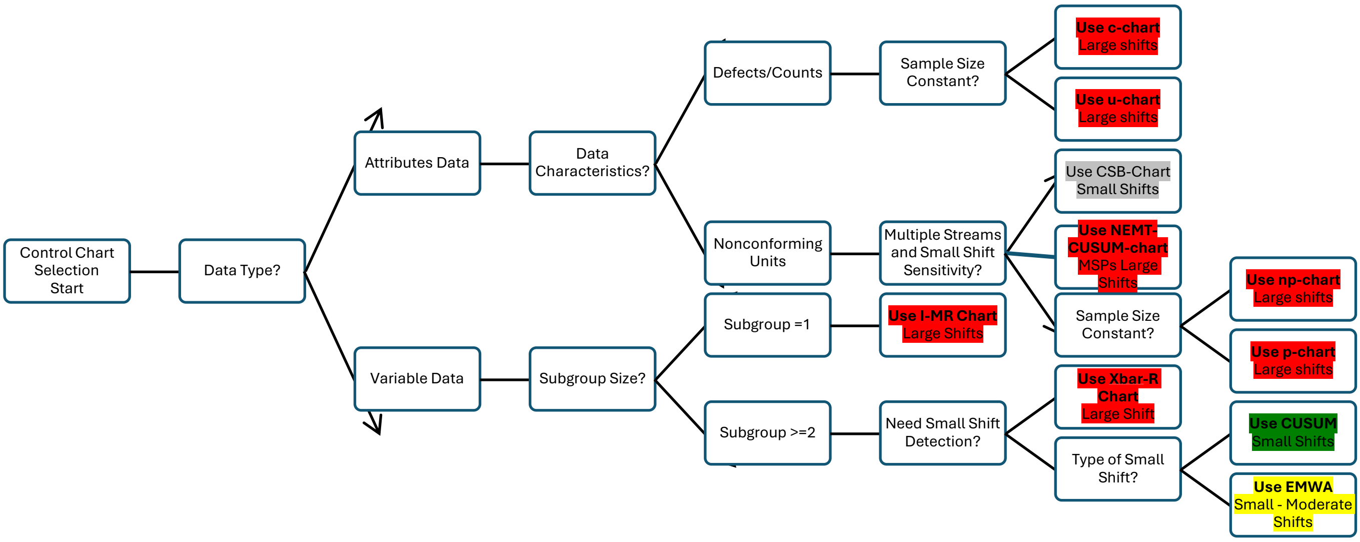
Proposed Methodology: Exact EWMA for Binomial Proportion Monitoring



3.1 Generalized Process Framework

This section introduces the structural and theoretical basis of the Cumulative Standardized Binomial EWMA (CSB-EWMA) control chart. Before exploring its mechanics, it is important to clarify the assumptions that support its application. The charting method presumes that the k data streams operate independently of one another. Within any single stream, observations collected at a given time point are also assumed to be mutually independent. Additionally, the process under surveillance must be measurable on at least an ordinal scale.

Let y_{ij} denote the jth observation randomly sampled from the stream i which has an unknown probability distribution but a known, in-control median, denoted $\tilde{\mu}_0$. Now, a binary indicator variable is then defined as $x_{ij} = I(y_{ij} > \tilde{\mu}_0)$, where $x_{ij} = 1$ if the inequality is true, and 0 otherwise. Provided that the median of stream i remains stable at $\tilde{\mu}_0$, the indicator variable x_{ij} follows a binomial distribution with parameters n = 1 and $p_0 = 0.5$, as shown below:

$$x_{it} \sim BIN(n = 1, p_0 = 0.50). \quad (1)$$

At each sampling time point $j = 1,2, \dots t$, both the raw observations y_{ij} and their corresponding binary indicators x_{ij} are collected across all k streams. These values form the basis for constructing summary tables such as Table 1, which supports further analysis and monitoring.

Table 1: Recoded Multiple Stream Process Data Structure

	Sample Number			
Stream	j=1	j=2	...	j=t
1	x_{11}	x_{12}		x_{1t}
2	x_{21}	x_{22}	...	x_{2t}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
k	x_{k1}	x_{k2}		x_{kt}
Column Totals	C_1	C_2	...	C_t

Note, because it is assumed that for a given time point, t, each of the observations between the streams are mutually independent Bernoulli random variables, then it is straightforward to see that:

$$C_j = \sum_{i=1}^{k} x_{ij} \sim BIN(n = k, p_0 = 0.50). \tag{2}$$

Our methodology addresses binomial proportion monitoring across diverse applications in equation (2). Since we represent the count of events at time t, where k is the number of streams (or opportunities for an event) and $p_0 = P(y_{ij} > \tilde{\mu}_0)$ is the true proportion (i.e. the true in-control probability that an observation exceeds the process median $\tilde{\mu}_0$). The cumulative count is

$$Q_t = \sum_{j=1}^{t} C_j \sim BIN(n = kt, p_0 = 0.50) \tag{3}$$

With:

$$E[C_j] = kp_0, \qquad Var[C_j] = kp_0(1 - p_0)$$

And let $\mu = E[C_j] = kp_0$ and $\sigma^2 = Var[C_j] = kp_0(1 - p_0)$ denote the expected value and variance of the count at any individual time point j. Then, with expectation $E[Q_t]$ and variance $Var[Q_t]$ of the cumulative count Q_t are:

$$E[Q_t] = tkp_0 = \mu t, \qquad Var[Q_t] = tkp_0(1 - p_0) = t\sigma^2 \tag{4}$$

The standardized statistic W_t is defined as:

$$W_t = \frac{Q_t - \mu t}{\sqrt{t\sigma^2}} \quad (5)$$

This represents the standardized deviation of the cumulative count from its expected value.

3.2 Exact Mean and Variance EWMA Statistic (r_t)

The EWMA statistic is defined by the recursive equation:

$$r_t = \lambda W_t + (1 - \lambda) r_{t-1} \quad (6)$$

with $r_0 = E[W_t]$ as a constant initial value, where $0 < \lambda \leq 1$ is the smoothing parameter.

3.2.1 Exact Mean EWMA Statistic (r_t)

The recursive relation $r_t = \lambda W_t + (1 - \lambda) r_{t-1}$, can be expressed as

$$r_t = \lambda \sum_{i=0}^{t} (1 - \lambda)^{t-i} W_i \quad (7)$$

we find the expectation of $E[r_t]$;

$$E[r_t] = \lambda \sum_{i=1}^{t} \left[(1 - \lambda)^{t-i} E[W_i] \right] + (1 - \lambda)^t E[r_0] \quad (8)$$

But,

$$E[W_i] = E\left[\frac{Q_i - \mu_i}{\sqrt{i\,\sigma^2}}\right]$$

$$E[W_i] = E\left[\frac{Q_i - \mu_i}{\sqrt{i\,\sigma^2}}\right] = \frac{E[Q_i] - \mu_i}{\sqrt{i\,\sigma^2}} = \frac{\mu_i - \mu_i}{\sqrt{i\,\sigma^2}} = 0 \qquad (9)$$

Substituting equation (9) (i.e. $E[W_i] = 0$), back into the expectation, we have

$$E[r_t] = \lambda \sum_{i=1}^{t} \left[(1-\lambda)^{t-i} * 0\right] + (1-\lambda)^t r_0$$

Thus,

$$E[r_t] = (1-\lambda)^t r_0 \qquad (10)$$

3.2.2 Exact Variance EWMA Statistic (r_t)

Since we express r_t explicitly as

$$r_t = \lambda \sum_{i=1}^{t} (1-\lambda)^{t-i} W_i + (1-\lambda)^t r_0$$

Where r_0 is constant, the variance is

$$Var(r_t) = \lambda^2 \sum_{i=1}^{t} \sum_{j=1}^{t} (1-\lambda)^{2t-i-j} Cov(W_{i,} W_j) \qquad (11)$$

We need to find $Cov(W_{i,} W_j) = \frac{1}{\sigma^2 \sqrt{ij}} Cov(Q_i - \mu i, Q_j - \mu j)$, without loss of generality, assume $i \leq j$. Then:

	$$Cov(W_{i,}W_j) = \frac{1}{\sigma^2\sqrt{ij}} Cov(Q_i - \mu i, Q_j - \mu j) = \frac{1}{\sigma^2\sqrt{ij}} Cov(Q_i, Q_j)$$	12

We compute $Cov(Q_i, Q_j)$, since $i \leq j$. We can write:

$$Q_j = Q_i + \sum_{k=i+1}^{j} C_k$$

Then:

$$Cov(Q_i, Q_j) = Cov\left(Q_i, Q_i + \sum_{k=i+1}^{j} C_k\right)$$

$$= Cov(Q_i, Q_i) + \ Cov\left(Q_i, \sum_{k=i+1}^{j} C_k\right)$$

Since, Q_i and $\sum_{k=i+1}^{j} C_k \ \ are\ independent$:

$$Cov(Q_i, Q_j) = var(Q_i) + \ 0 = i\,\sigma^2$$

	$$Cov(Q_i, Q_j) = i\,\sigma^2$$	13

Substitute equation (13) back into covariance expression $of\ Cov(W_{i,}W_j)$

$$Cov(W_{i,}W_j) = \frac{1}{\sigma^2\sqrt{ij}} Cov(Q_i, Q_j) = \frac{1}{\sigma^2\sqrt{ij}} * i\,\sigma^2$$

	$$Cov(W_{i,}W_j) = \frac{\sqrt{i}}{\sqrt{j}}$$	14

By symmetry, for $i \geq j$, we get $\sqrt{j}/\sqrt{i}$. Therefore, in general:

$$Cov(W_{i,}W_j) = \frac{\sqrt{\min{(i,j)}}}{\sqrt{\max(\,i,j)}} \tag{15}$$

Substitute $Cov(W_{i,}W_j)$ into $Var(r_t)$ expression in equation (11), we have

$$Var(r_t) = \lambda^2 \sum_{i=1}^{t}\sum_{j=1}^{t}(1-\lambda)^{2t-i-j}\frac{\sqrt{\min{(i,j)}}}{\sqrt{\max(\,i,j)}} \tag{16}$$

Simplify the double summation, let:

$$S = \sum_{i=1}^{t}\sum_{j=1}^{t}(1-\lambda)^{2t-i-j}\frac{\sqrt{\min{(i,j)}}}{\sqrt{\max(\,i,j)}} \tag{17}$$

We can split this sum into three regions: $i < j$, $i >$ j, and $i = j$

$$S = \sum_{i=1}^{t}\sum_{j=1}^{i-1}(1-\lambda)^{2t-i-j} * \frac{\sqrt{\mathrm{j}}}{\sqrt{i}} + \sum_{i=1}^{t}\sum_{j=i+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{\mathrm{i}}}{\sqrt{j}} + \sum_{i=1}^{t}(1-\lambda)^{2t-i-i} * \frac{\sqrt{\mathrm{i}}}{\sqrt{i}}$$

Note by symmetry, the first two sums are equal. Simplifying further:

$$S = 2\sum_{i=1}^{t}\sum_{j=1}^{i-1}(1-\lambda)^{2t-i-j} * \frac{\sqrt{\mathrm{j}}}{\sqrt{i}} + \sum_{i=1}^{t}(1-\lambda)^{2t-2i}$$

Let's change the order of summation in the first term. For fixed j, i runs from $(j+1)$ to t:

$$\sum_{i=1}^{t}\sum_{j=1}^{i-1}(1-\lambda)^{2t-i-j} * \frac{\sqrt{\mathrm{j}}}{\sqrt{i}} = \sum_{j=1}^{t-1}\sum_{i=j+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{\mathrm{j}}}{\sqrt{i}}$$

Then we have:

$$S = 2\sum_{j=1}^{t-1}\sum_{i=j+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{j}}{\sqrt{i}} + \sum_{i=1}^{t}(1-\lambda)^{2t-2i} \quad (18)$$

Substituting equation 18 back into equation 14 ($Var(r_t) = \lambda^2 S$), we have:

$$Var(r_t) = \lambda^2 \sum_{i=1}^{t}\sum_{j=1}^{t}(1-\lambda)^{2t-i-j}\frac{\sqrt{\min(i,j)}}{\sqrt{\max(i,j)}} = \lambda^2 S$$

$$\lambda^2\left[2\sum_{j=1}^{t-1}\sum_{i=j+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{j}}{\sqrt{i}} + \sum_{i=1}^{t}(1-\lambda)^{2t-2i}\right]$$

Therefore, for Finite t, the variance

$$Var(r_t) = \lambda^2\left[2\sum_{j=1}^{t-1}\sum_{i=j+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{j}}{\sqrt{i}} + \sum_{i=1}^{t}(1-\lambda)^{2t-2i}\right]$$

$$Var(r_t) = \lambda^2\left[2\sum_{j=1}^{t-1}\sum_{i=j+1}^{t}(1-\lambda)^{2t-i-j} * \frac{\sqrt{j}}{\sqrt{i}} + \sum_{i=1}^{t}(1-\lambda)^{2t-2i}\right] \quad (19)$$

The computational implementation of this exact variance formulation, along with validation code, is provided in Appendix A.

3.2.3 Asymptotic Behavior of EWMA Statistic (r_t)

Asymptotic Expectation

From equation (10):

$$\lim_{t\to\infty} E[r_t] = \lim_{t\to\infty}(1-\lambda)^t r_0 = 0 \quad (20)$$

Asymptotic Variance

For large t $(as\ t \to \infty)$, the dominant contributions come from terms where i and j are close to t, because the exponential weights $(1-\lambda)^{2t-i-j}$ decay rapidly for smaller $i\ ,j$, then $\max(i\ ,j) \approx t$.

In this region, $\frac{\sqrt{\min\ (i,j)}}{\sqrt{\max(i,j)}} \approx 1$.

For the approximate, $(1-\lambda)^{2t-i-j} \approx\ (1-\lambda)^{s+u}$, where $s = t - i, u = t - j$

So, from (16)

$$Var(r_t) = \lambda^2 \sum_{i=1}^{t} \sum_{j=1}^{t} (1-\lambda)^{2t-i-j} \frac{\sqrt{\min\ (i,j)}}{\sqrt{\max(\,i,j)}}$$

$$Var(r_t) \approx \lambda^2 \sum_{i=1}^{t} \sum_{j=1}^{t} (1-\lambda)^{2t-i-j} * (1) = \lambda^2 \sum_{s=0}^{\infty} \sum_{u=0}^{\infty} (1-\lambda)^{s+u}$$

But the double sum is geometric:

$$\sum_{s=0}^{\infty} \sum_{u=0}^{\infty} (1-\lambda)^{s+u} = \left(\sum_{u=0}^{\infty} (1-\lambda)^{s} \right)^2 = \left(\frac{1}{\lambda}\right)^2$$

Therefore:

$$Var(r_t) \approx \lambda^2 * \frac{1}{\lambda^2} = \left(\frac{\lambda}{\lambda}\right)^2 = 1$$

For large t, the variance of the EWMA statistic approaches 1:

$$\lim_{t \to \infty} Var[r_t] = 1 \tag{21}$$

3.3 Adaptive Control Limits

The time-varying control limits are:

$$UCL_t = (1-\lambda)^t r_0 + L\sqrt{Var[r_t]},\ \ LCL_t = (1-\lambda)^t r_0 - L\sqrt{Var[r_t]}$$

where L denotes the half-width of control limits, typically $L = 3$ for 3-sigma limits.

For large t, since $Var[r_t] \to 1$, the control limits approach:

	$$UCL_t = L,\ LCL_t = -L$$	22

3.4 Validation and Verification of Theoretical Derivations

Prior to evaluating the chart's performance for shift detection, a critical step is to validate the computational implementation and verify the accuracy of the theoretical derivations presented in 3.2.1 and 3.2.2. This validation was conducted by comparing the theoretical mean $E[r_t]$ and variance $Var[r_t]$ of the EWMA statistics against their empirical counterparts obtained from $N_{sim} = 10{,}000$ Monte Carlo replications under the in-control process assumption, with parameters $\boldsymbol{k = 10},\ p_0 = 0.5$ and $\boldsymbol{\lambda = 0.2}$. All analyses were conducted using R version 4.3.2 (R Core Team, 2023), with Code availability in Appendix A to ensure complete reproducibility of all reported results.

Table 3.4.1 presents the validation metrics at selected time points, demonstrating the convergence behavior of both the mean and variance statistics.

Table 3.4.1: Validation Metrics at Selected Time Points

Time (t)	Theoretical Mean	Simulated Mean	Theoretical Variance	Simulated Variance	Relative Bias (Variance)
10	0.000	-5.54×10^{-3}	0.6369	0.6385	0.26%
50	0.000	-5.23×10^{-3}	0.9512	0.9359	-1.61%
100	0.000	-1.95×10^{-3}	0.9768	0.9758	-0.10%
500	0.000	-6.06×10^{-4}	0.9955	0.9967	0.12%
1000	0.000	-3.09×10^{-3}	0.9978	0.9996	0.19%

The validation results demonstrate good agreement between theoretical predictions and empirical simulations:

The theoretical expectation $E[r_t] = (1-\lambda)^t r_0$ correctly predicts values effectively equal to zero across all time points. The simulated means show negligible deviations from zero, with a root-mean-square bias of $2.89 * 10^{-3}$ and the maximum absolute bias of $9.00 * 10^{-3}$. These minor fluctuations are consistent with Monte Carlo sampling variation and confirm the unbiasedness of the EWMA estimator under in-control conditions.

The exact theoretical variance formula shows remarkable accuracy when compared against simulated values. The relative bias remains below 3% across all time points, decreasing to approximately 0.19% as the process approaches steady-state. The variance converges rapidly to its asymptotic value of 1, reaching 99% of the asymptotic variance by $\boldsymbol{t = 227}$**.** The root-mean-square bias for variance is $7.92 * 10^{-3}$, indicating high precision in the theoretical predictions.

Figure 3.1 visually confirms the close alignment between theoretical and simulated values for both mean and variance across the entire monitoring period.

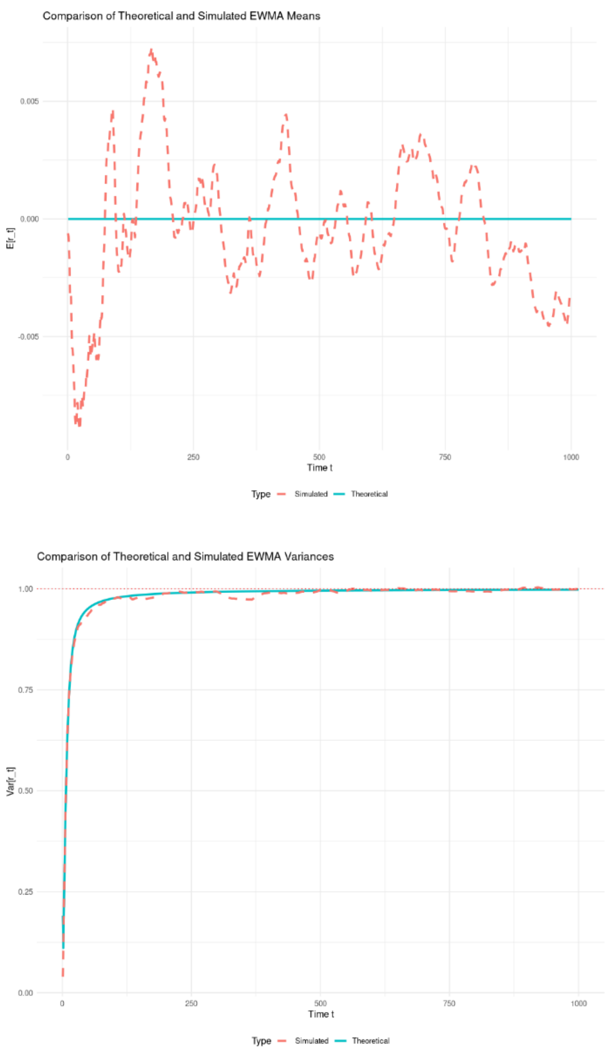


Figure 3.1: Comparison of theoretically proposed EMWA Mean and Variance with Monte Carlo simulation

The convergence behavior observed in Table 3, visualized by Figure 3.1, aligns with theoretical expectations: the variance increases monotonically from approximately 0.637 at $t = 10$ to nearly 1.000 at $t = 100$, following the exact variance derivation in Equation (19). The close agreement across all time points, particularly the 0.2% relative bias at steady-state, provides strong empirical evidence for the correctness of both the mathematical derivations and computational implementation.

This rigorous validation establishes that the proposed CSB-EWMA chart's statistical properties are fully characterized and that the algorithm is implemented correctly. The successful verification ensures that any subsequent performance comparisons against asymptotic methods will be based on a correctly specified exact model, providing a solid foundation for the performance evaluation study that follows.