Statistical Process Control

MSc: Statistics and Actuarial-Financial Mathematics

Konstantinos Bourazas
kbourazas@aegean.gr

Department of Statistics and Actuarial-Financial Mathematics

Course #7

March 29, 2026

CUSUM charts for detecting persistent shifts

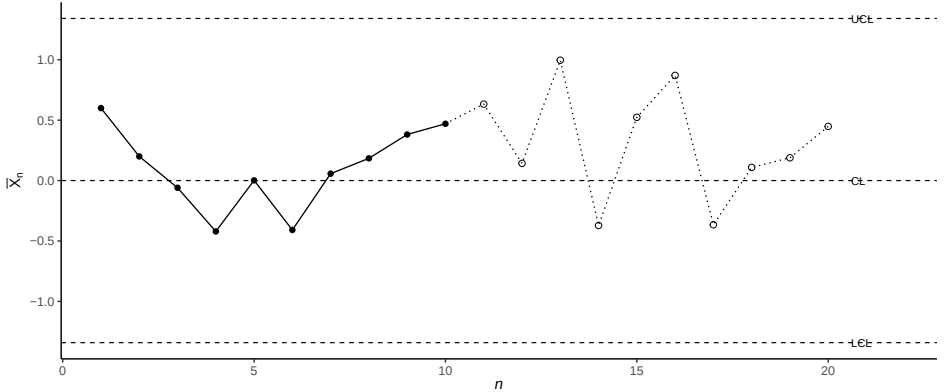
- The Shewhart charts we studied last week make a decision at each time point using **only the current observation**
- This is perfectly fine for **large, abrupt shifts**, but it means that useful information in recent history is thrown away
- When the shift is **small but persistent**, accumulating evidence over time can detect it much faster
- The **Cumulative Sum (CUSUM)** chart, introduced by Page[1], does exactly this
  - It keeps a running total of deviations from the target and signals when the total becomes too large
  - It enjoys **optimality properties** for detecting a shift of a given size
- CUSUM charts are primarily a **Phase II** tool, because persistent shifts are the main concern in online monitoring
- Our discussion follows Qiu (2014), Sections 4.1 and 4.2

[1] Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41, 100–115.

Example 4.1 A small shift that the $\bar{X}$ chart misses

- Generate 10 samples of size $m = 5$ from $N(0, 1)$ followed by 10 samples from $N(0.2, 1)$, simulating a mean shift of $\delta = 0.2$ at the 11th time point
- The $\bar{X}$ chart treats each subgroup **independently**, so at any single time point the shifted mean is still very likely to fall within the 3σ limits
- The shift goes **completely undetected**, both in Phase I (limits estimated from data) and Phase II (known IC parameters)
- The problem is fundamental: a single hypothesis test has low power against a small shift, and the Shewhart chart has no way to combine evidence across consecutive subgroups
- What we need is a chart that **accumulates information** over time

Example 4.1 A small shift that the $\bar{X}$ chart misses



10 subgroups from $N(0,1)$ followed by 10 from $N(0.2,1)$, with known-parameter 3σ limits. The shift is too small for any single subgroup to breach the control limits, illustrating the need for charts that accumulate evidence over time

The cumulative sum statistic C_n

- A natural way to accumulate evidence is to sum the deviations from the target

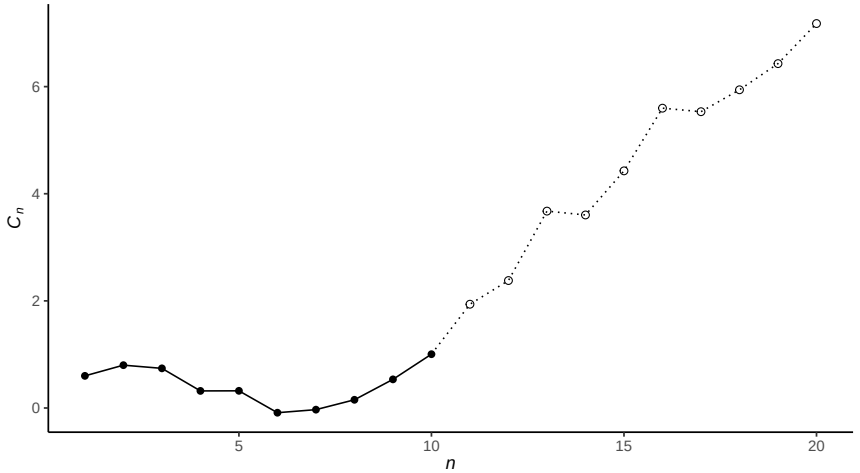
$$C_n = \sum_{i=1}^{n} (X_i - \mu_0), \qquad C_0 = 0$$

- Equivalently we can write the recursion $C_n = C_{n-1} + (X_n - \mu_0)$
- When the process is IC, each term has mean zero, so C_n behaves like a **random walk** around zero
- If a shift of size δ occurs at time τ, then for $n \geq \tau$ the mean of C_n starts to grow linearly with slope δ

$$E[C_n] = (n - \tau + 1)\,\delta, \qquad n \geq \tau$$

- So a sustained upward or downward trend in C_n is a sign that the process mean has shifted

Example 4.2 The cumulative sum for the same data



From the cumulative sum to the CUSUM statistic

The plain cumulative sum $C_n = \sum_{i=1}^{n}(X_i - \mu_0)$ has a problem: its variance grows linearly with n, so even under IC it wanders further and further from zero as time passes. This makes it impractical as a control chart

We fix this in two steps (shown here for detecting an **upward** shift; the downward case is analogous):

1. **Subtract a reference value $k > 0$.** Replace each increment $(X_i - \mu_0)$ by $(X_i - \mu_0) - k$. Under IC the increments now have a *negative* mean $(-k)$, so the sum drifts downward instead of wandering. Only a genuine upward shift of size $\delta > k$ can overcome this pull and push the sum upward
2. **Reset to zero when the sum goes negative.** There is no useful information in a negative sum, it just means the recent data look IC. Resetting discards old IC evidence and lets the chart start fresh whenever new evidence of a shift begins to appear

The upward CUSUM chart

Combining both steps gives the charting statistic

$$C_n^+ = \max\left(0,\ C_{n-1}^+ + (X_n - \mu_0) - k\right), \qquad C_0^+ = 0$$

- The **reference value** k determines which shift size the chart is tuned to detect. For a target shift of δ, the optimal choice is $k = \delta/2$
- The chart signals an upward shift when $C_n^+ > h$, where $h > 0$ is the **decision interval** (control limit). Larger h means fewer false alarms but slower detection
- Under IC the statistic spends most of its time at or near zero, because the negative drift from k and the restart mechanism work together to keep it down. A signal requires a sustained sequence of observations above $\mu_0 + k$

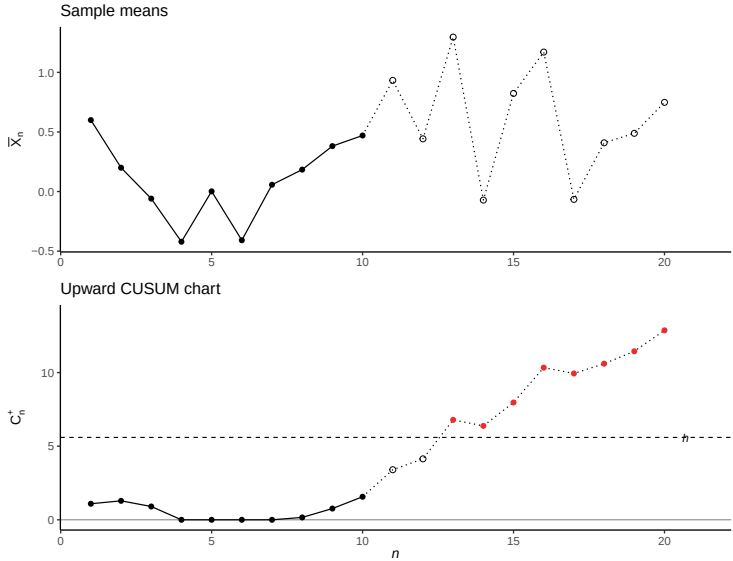
Example 4.2: Building the CUSUM step by step

$Z_n = \bar{X}_n \cdot \sqrt{m}$ is the standardised mean, $C_n = \sum_{i=1}^{n} Z_i$ the plain cumulative sum, and $C_n^+ = \max(0,\ C_{n-1}^+ + Z_n - k)$ the CUSUM with $k = 0.25$ and $h = 5.597$

n	Z_n	C_n	C_n^+	n	Z_n	C_n	C_n^+
1	1.34	1.34	1.09	11	2.09	4.33	3.40
2	0.45	1.79	1.29	12	0.99	5.32	4.14
3	-0.13	1.65	0.90	13	2.90	8.22	**6.79**
4	-0.94	0.71	0.00	14	-0.16	8.06	**6.38**
5	0.00	0.72	0.00	15	1.84	9.90	**7.97**
6	-0.91	-0.20	0.00	16	2.62	12.52	**10.34**
7	0.13	-0.07	0.00	17	-0.15	12.37	**9.94**
8	0.41	0.34	0.16	18	0.91	13.28	**10.60**
9	0.85	1.19	0.76	19	1.09	14.38	**11.45**
10	1.05	2.25	1.57	20	1.67	16.05	**12.87**

C_n wanders with growing variance. C_n^+ resets to zero whenever evidence vanishes, and signals only when sustained deviations accumulate past h.

Example 4.2 (continued): The upward CUSUM chart



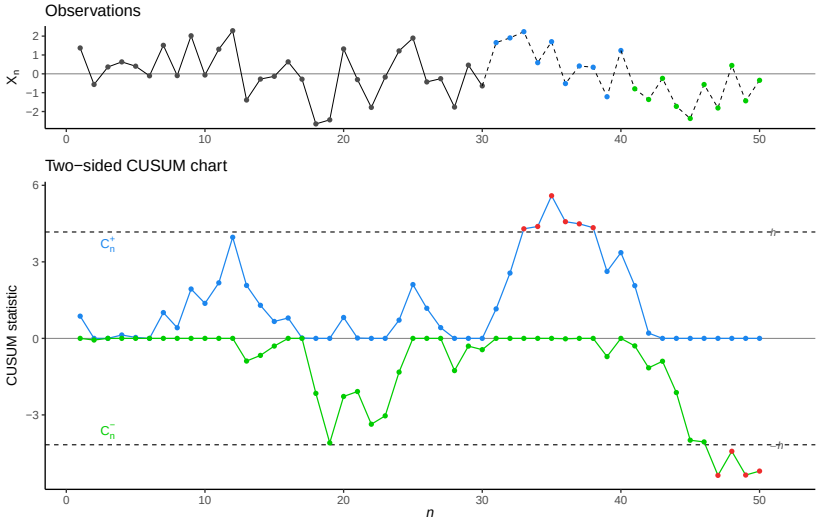
The downward CUSUM and the two-sided version

- To detect a **downward** mean shift, define

$$C_n^- = \min\big(0,\ C_{n-1}^- + (X_n - \mu_0) + k\big), \qquad C_0^- = 0$$

  and signal when $C_n^- < -h$
- The charts C_n^+ and C_n^- are both **one-sided**
  - C_n^+ is designed to detect upward shifts
  - C_n^- is designed to detect downward shifts
- The **two-sided CUSUM** combines the two one-sided charts and delivers a signal whenever **either** $C_n^+ > h$ or $C_n^- < -h$
- This is the version most commonly used in practice, since we usually want to detect shifts in both directions
- If subgroup data of size m are used instead of individual observations, simply replace X_n by $\bar{X}_n$ in the formulas

Two-sided CUSUM in action



50 observations: IC $N(0,1)$ for $n = 1$–30, upward shift to $N(1,1)$ at $n = 31$, downward shift to $N(-1,1)$ at $n = 41$. C_n^+ (blue) and C_n^- (green) track each direction; red points mark alarms.

Choosing the reference value k

- The reference value k determines which shift size the chart is **tuned to detect**
- For a target upward shift of size δ, the optimal choice is $k = \delta/2$. This comes from the CUSUM's connection to the Sequential Probability Ratio Test (discussed shortly)
- **Small** k is ideal for small shifts but reacts slowly to large ones; **large** k is the opposite
- The chart with $k = \delta/2$ is **optimal**: its ARL_1 for detecting shift δ is the shortest among all charts with the same ARL_0
- A common default in practice is $k = 0.5$ (in σ units), targeting a shift of $\delta = 1\sigma$
- When working with standardised observations $Z_n = (X_n - \mu_0)/\sigma$, the parameters k and h are in units of σ, so the design tables are universal

Measuring performance with the Average Run Length

- For a given k, the value of h controls the ARL_0
  - Larger h means fewer false alarms (larger ARL_0)
  - Smaller h means faster detection but more false alarms
- A useful approximation due to Siegmund[2] is

$$\text{ARL}_0 \approx \frac{\exp\big(2k(h+1.166)\big) - 2k(h+1.166) - 1}{2k^2}$$

  which is quite accurate for small to moderate k (say $k \leq 1$)
- In practice, we can also use **Monte Carlo simulation** or the `spc` package in R to compute exact ARL values
- Note: the ARL values in the tables are **zero-state**, meaning they start from $C_0^+ = 0$. A **steady-state** ARL measured after the chart has reached equilibrium is slightly different, but the distinction is small for ARL_0 and usually negligible in practice

[2] Siegmund, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer.

Table of h values and how to choose them (Qiu, Table 4.1)

	k						
ARL_0	0.10	0.25	0.50	0.75	1.00	1.25	1.50
50	4.567	3.340	2.225	1.601	1.181	0.854	0.570
100	6.361	4.418	2.849	2.037	1.532	1.164	0.860
200	8.520	5.597	3.502	2.481	1.874	1.458	1.131
300	9.943	6.324	3.892	2.745	2.073	1.624	1.282
370	10.722	6.708	4.095	2.882	2.175	1.709	1.359
500	11.890	7.267	4.389	3.080	2.323	1.830	1.466
1000	14.764	8.585	5.071	3.538	2.665	2.105	1.708

- As expected, h increases with ARL_0 and decreases with k
- When the value is not in the table, find it with a **bisection search** on h until the computed ARL_0 matches the target to a tolerance ρ

Computing the OOC performance (ARL_1)

- Suppose the process shifts from $N(\mu_0, \sigma^2)$ to $N(\mu_0 + \delta,\ \lambda^2\sigma^2)$
- Then the ARL_1 of the chart with parameters (k, h) equals the **ARL_0 of the same chart** with modified parameters

$$k^* = \frac{k - \delta}{\lambda\sigma}, \qquad h^* = \frac{h}{\lambda\sigma}$$

  applied to $N(0, 1)$ data
- This relationship lets us reuse the ARL_0 tables and the Siegmund formula to obtain ARL_1 values without additional simulation
- In the pure mean-shift case ($\lambda = 1$, $\sigma = 1$), this simplifies to $k^* = k - \delta$ and $h^* = h$

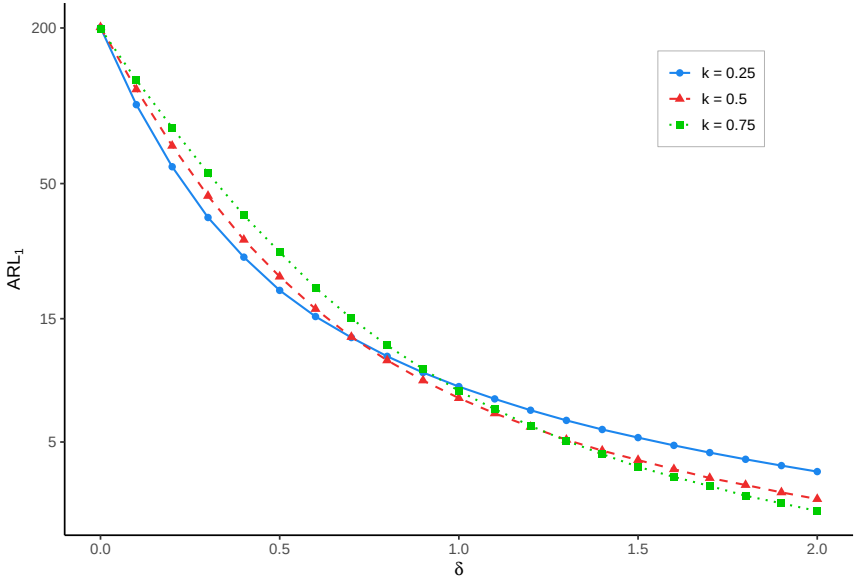
Example 4.3 ARL_1 under different types of shift

- Consider the upward CUSUM with $k = 0.5$, $h = 3.502$, and $\text{ARL}_0 = 200$ for an IC distribution $N(0, 1)$
- Using the relationship on the previous slide and the Siegmund formula

Case	OOC distribution	k^*	Approx. ARL_1
(i)	$N(0.25, 1)$	0.25	55.9
(ii)	$N(-0.25, 1)$	0.75	969.6
(iii)	$N(0, 4)$	0.25	14.7
(iv)	$N(0.25, 4)$	0.125	11.0

- Case (ii) illustrates the **biasedness** of one-sided charts. When the shift goes in the opposite direction, the ARL_1 becomes much **larger** than ARL_0
- Cases (iii) and (iv) show that an increase in variance also triggers the upward CUSUM, because observations become more spread out

ARL_1 for different values of k and shift size δ



ARL of the two-sided CUSUM

- The two-sided CUSUM combines the upward chart (C_n^+, h) and the downward chart $(C_n^-, -h)$
- Let ARL^+ and ARL^- be the ARL values of the two one-sided charts. Van Dobben de Bruyn (1968) showed

$$\frac{1}{\text{ARL}^*} = \frac{1}{\text{ARL}^+} + \frac{1}{\text{ARL}^-}$$

- This formula holds for both IC and OOC ARL values
- Practical consequence for design: if we want $\text{ARL}_0^* = 200$ for the two-sided chart, each one-sided chart must have $\text{ARL}_0 = 400$
- For example, with $k = 0.5$ and $\text{ARL}_0 = 400$, the table gives $h = 4.171$, so the two-sided chart uses limits at ± 4.171

The CUSUM from the SPRT perspective

- Testing H_0: $\mu = \mu_0$ vs. H_1: $\mu = \mu_1$ with $\mu_1 > \mu_0$, based on i.i.d. observations from $N(\mu, \sigma^2)$
- The likelihood ratio after n observations is

$$\Lambda_n = \frac{L(\mu_1;\, X_1, \ldots, X_n)}{L(\mu_0;\, X_1, \ldots, X_n)} = \prod_{i=1}^{n} \frac{f(X_i; \mu_1)}{f(X_i; \mu_0)}$$

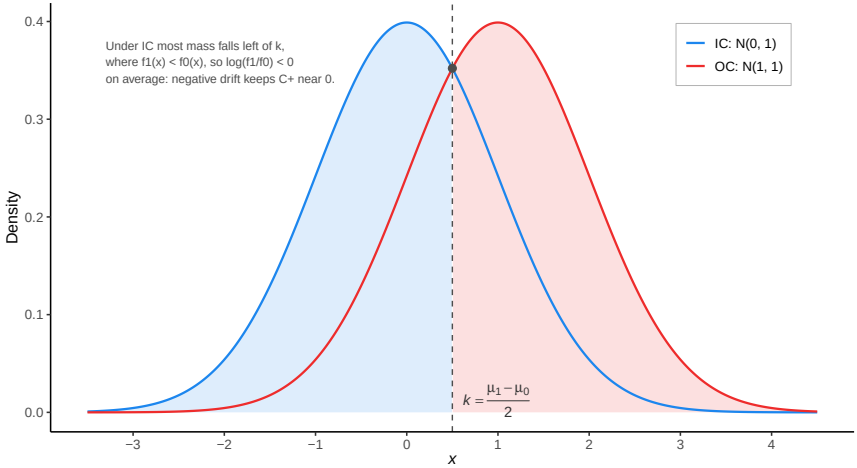
- Taking logarithms and substituting the normal density

$$\log \Lambda_n = \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{i=1}^{n} \big[(X_i - \mu_0) - k \big], \qquad k = \frac{\mu_1 - \mu_0}{2}$$

- The CUSUM statistic C_n^+ adds a **restart to zero** whenever $\sum_{i=1}^{n} [(X_i - \mu_0) - k]$ becomes negative, discarding old IC evidence
- Moustakides[3] proved that the CUSUM with $k = \delta/2$ has the **shortest $\mathbf{ARL_1}$** for shift δ among all charts with a given ARL_0

[3] Moustakides, G. V. (1986). Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, 14, 1379–1387.

Why the CUSUM stays near zero under IC



Under IC, most observations fall where $f_1(x) < f_0(x)$, so $\log(f_1/f_0)$ is negative on average and G_n drifts towards zero. The two densities intersect exactly at $k = (\mu_1 - \mu_0)/2$.

Properties of the CUSUM statistic

- More generally, for any IC density f_0 and OOC density f_1, the optimal CUSUM has charting statistic

$$G_n = \max\left(0,\ G_{n-1} + \log \frac{f_1(X_n)}{f_0(X_n)}\right), \qquad G_0 = 0$$

  and signals when $G_n > h$. All CUSUM charts in this course are special cases of this formula

- **Supermartingale property.** When the process is IC, $E[\log(f_1(X_n)/f_0(X_n))] < 0$, so the increments of G_n have a negative drift on average
  - Intuitively, under IC the evidence for H_1 tends to **shrink towards zero** at each step
  - The restart mechanism reinforces this by resetting G_n to zero whenever it would go negative
  - This is why false alarms are rare: the statistic has to overcome a persistent downward pull to reach h

- **Change point estimation.** When the chart signals at time n, the shift location τ can be estimated as the time point **immediately after the last zero** of the charting statistic, i.e. $\hat{\tau} = \max\{i < n : G_i = 0\} + 1$

Shewhart charts are a special case of the CUSUM

- The $\bar{X}$ chart with known parameters signals when

$$\frac{\bar{X}_n - \mu_0}{\sigma/\sqrt{m}} > Z_{1-\alpha/2}$$

  which is equivalent to an upward CUSUM with $k = Z_{1-\alpha/2}$ and $h = 0$

- With $Z_{1-\alpha/2} = 3$, the Shewhart chart is therefore **optimal for detecting a shift of size** 6σ, confirming that it is designed for large shifts
- This connection clarifies when each chart is most useful
  - Shewhart catches **large, abrupt** shifts immediately
  - CUSUM catches **small, persistent** shifts that build up
- In practice, **both charts are often run together** so that both types of shift are covered

Effect of autocorrelation and the residual approach

- All ARL results so far assume **independent** observations. Autocorrelation can severely distort the actual ARL values
- With $(k, h) = (0.5, 3.502)$ (nominal $\text{ARL}_0 = 200$) and an AR(1) model $X_n = \mu_0 + \phi(X_{n-1} - \mu_0) + e_n$

ϕ	-0.5	0	0.5
Actual ARL_0	743	200	38

- Positive autocorrelation **inflates** the false alarm rate; negative autocorrelation makes the chart **too conservative**
- The standard remedy is the **residual approach**: fit a time series model to the IC data, compute residuals $\hat{e}_n = X_n - \hat{X}_n$, and apply the CUSUM to those
- The residuals are approximately i.i.d. $N(0, \hat{\sigma}_e^2)$ when the process is IC, so the standard tables apply again

CUSUM charts for monitoring the process variance

- Shifts in process variability affect product quality just as much as shifts in the mean, so dedicated charts are needed
- The CUSUM for the mean has **some** ability to detect upward variance shifts, but it is not designed for this purpose
- Using the general CUSUM formula with $f_0 = N(\mu_0, \sigma_0^2)$ and $f_1 = N(\mu_0, \sigma_1^2)$, the optimal chart for an **upward** variance shift is

$$C_n^+ = \max\left(0,\ C_{n-1}^+ + \left(\frac{X_n - \mu_0}{\sigma_0}\right)^2 - k^+\right), \qquad k^+ = \frac{2\log(\sigma_0/\sigma_1)}{(\sigma_0/\sigma_1)^2 - 1}$$

  and signals when $C_n^+ > h_U$
- A corresponding downward chart replaces max with min and signals when $C_n^- < h_L$

CUSUM charts for discrete distributions

- The general CUSUM formula applies to **any** distribution, including discrete ones
- **Binomial case**: if $X_n \sim \text{Binomial}(m, \pi)$ and we want to detect an upward shift from π_0 to π_1

$$C_n^+ = \max(0,\ C_{n-1}^+ + X_n - k^+), \qquad k^+ = \frac{-m \log\big((1-\pi_1)/(1-\pi_0)\big)}{\log\Big[\big(\frac{\pi_1}{1-\pi_1}\big)/\big(\frac{\pi_0}{1-\pi_0}\big)\Big]}$$

- **Poisson case**: if $X_n \sim \text{Poisson}(\lambda)$ and we want to detect a shift from λ_0 to λ_1

$$C_n^+ = \max(0,\ C_{n-1}^+ + X_n - k^+), \qquad k^+ = \frac{\lambda_1 - \lambda_0}{\log \lambda_1 - \log \lambda_0}$$

- Important caveat: because the charting statistic takes **discrete values**, only certain ARL_0 levels are achievable for a given k^+