How to measure the performance of a control chart? One could the α error and β error, the same as for evaluating any hypothesis test.

But a control chart is actually a series of sample-by-sample hypothesis tests, could we develop something more intuitive and that measure the performance of a chart more directly?

For that, people like to use the average run length (ARL) as a performance measure of control charts.
Control chart performance

- Run length (RL), $L$ \equiv the sample number on which a control chart **first signals**, that is, when a data point is outside the control limits.

- RL is a **random variable**, taking only integer values, \{1, 2, \ldots\}.
- Using RL itself cannot serve the performance benchmarking purpose. We need to use its **expected** value as a performance measure, namely average RL.
- Average RL (ARL) is defined as the expected value of RL, or expected number of samples until a control chart first signals.
Two Types of ARL

- **in-control ARL (ARL₀)**: is the expected number of samples until a control chart signals, *under the condition* that the actual process is truly in control;

- **out-of-control ARL (ARL₁)**: is the expected number of samples until a control chart signals, *under the condition* that the actual process is in fact out-of-control;

- We'd like ARL₀ to be as large as possible and ARL₁ to be as small as possible.
Control chart performance

- **How to calculate ARL?**
  1) **consider in-control ARL (ARL₀).** The actual process is in control (IC), namely H₀ is true.

\[ \alpha = \Pr[\text{a Shewhart chart signals on a given sample } | \ H_0 \text{ is true}] \]

\[
\begin{align*}
\text{Prob}(\bar{x} \in [LCL, UCL] | IC) &= 1 - \alpha \\
\text{Prob}(\bar{x} \notin [LCL, UCL] | IC) &= \alpha
\end{align*}
\]

\[ ARL_0 = E(RL) = \sum_{k=1}^{\infty} k \cdot \Pr(RL = k) = \sum_{k=1}^{\infty} k \cdot (1-\alpha)^{k-1} \alpha = \frac{1}{\alpha} \]
Control chart performance

- **How to calculate ARL?**

  2) **consider out-of-control ARL** ($ARL_1$). The actual process is out of control (OOC), namely $H_1$ is true

  \[
  1 - \beta = \Pr[a \text{ Shewhart chart signals on a given sample } \mid H_1 \text{ is true}]
  \]

\[
\begin{align*}
\text{RL} & \quad \text{Probability} \\
1 & \quad 1 - \beta \\
2 & \quad \beta (1 - \beta) \\
3 & \quad \beta^2 (1 - \beta) \\
\vdots & \\
k & \quad \beta^{k-1} (1 - \beta) \\
\end{align*}
\]

\[
ARL_1 = E(RL) = \sum_{k=1}^{\infty} k \cdot \Pr(RL = k) = \sum_{k=1}^{\infty} k \cdot \beta^{k-1} (1 - \beta) = \frac{1}{1 - \beta}
\]

ARL_1 = \frac{1}{1 - \beta}
Consider an $x$-bar chart with in-control mean and standard deviation $\mu_0 = 30.0$ and $\sigma_0 = 1.5$, and sample size $n = 5$. If the 3-sigma control limits are used, what is the $\alpha$-error? What is the in-control $ARL_0$?

If use the 3-sigma control limits, then $\alpha = 0.0027$.

So $ARL_0 = \frac{1}{\alpha} = \frac{1}{0.0027} \approx 370$.

If the process mean shifts to $\mu_1 = 32.0$, what is the $\beta$ error for the chart to detect such a shift? What is the out-of-control $ARL_1$?

$$\beta = \Phi(z_{\alpha/2} - k \cdot \sqrt{n}) - \Phi(-z_{\alpha/2} - k \cdot \sqrt{n})$$

Because we are using 3-sigma control limits, $z_{\alpha/2} = z_{0.0027/2} = 3$.

Also, $k = \frac{|\mu_1 - \mu_0|}{\sigma_0} = \frac{32 - 30}{1.5} = 1.33$ and $n = 5$

$\Rightarrow \beta = \Phi(3 - 1.33\sqrt{5}) - \Phi(-3 - 1.33\sqrt{5}) = 0.51$

$\Rightarrow ARL_1 = \frac{1}{1 - \beta} = \frac{1}{1 - 0.51} = 2.04$
• **Interpretation**

When the process is in control, $\text{ARL}_0$ suggests that it will take 370 samples, on average, before the chart sets off a false alarm. This is pretty good since 370 is relatively large of samples.

On the other hand, $\text{ARL}_1 = 2.04$ means that it will take two to three samples, on average, to detect the occurrence of a mean shift from 30 to 32. This is quite desirable since two/three samples are considered rather sample number so the delay in detection is not serious.
Three often used tricks in probability

• If \( P(A) \) is the probability that event A occurs, then the probability that A will not occur (i.e., the complementary event) is \( P(A') = 1 - P(A) \).

• For mutually exclusive events, the probability that either event A or B will occur is the sum of their respective probability: \( P(A \text{ or } B) = P(A) + P(B) \).

• If A and B are independent events, then the probability that both A and B will occur is \( P(A \text{ and } B) = P(A) \times P(B) \).
• An x-bar control chart with the **3-sigma control limits** and sample size of 5 is used for monitoring a process mean shift. In order to enhance detection sensitivity, a **2-sigma limit** is also used as the **warning limit**. Decision of when a chart should signal is based on \( m = 3 \) successive samples using either one of the following rules:

**Rule 1:** If at least one of the three sample means falls outside the 3-sigma control limits, the chart signals;

**Rule 2:** If exactly two out of the three sample means fall outside the 2-sigma warning limits, the chart signals.

(a) What is the Type I error using Rule 1 alone?  
(b) What is the Type I error using Rule 2 alone?
(a) What is the Type I error using Rule 1 alone?

$$\alpha_{rule1} = \Pr[\text{at least one of the three samples falls outside the 3-sigma control limits} \mid \text{IC}]$$

$$= 1 - \Pr[\text{all three samples are inside the 3-sigma control limits} \mid \text{IC}]$$

$$= 1 - (1 - \alpha_{3\text{sigma}})^3 = 1 - (1 - 0.0027)^3 = 0.0081$$

So $$ARL_0 = \frac{1}{\alpha_{rule1}} = \frac{1}{0.0081} \approx 123.$$

(b) What is the Type I error using Rule 2 alone?

First, we need to determine the $$\alpha$$ error for using the 2-sigma warning limit:

$$\alpha_{2\text{sigma}} = \Pr[\bar{x} > UCL \text{ or } \bar{x} < LCL \mid \text{IC}]$$

$$= \Pr[\frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} > \frac{UCL - \mu_0}{\sigma_0/\sqrt{n}} \text{ or } \frac{\bar{x} - \mu_0}{\sigma_0/\sqrt{n}} < \frac{LCL - \mu_0}{\sigma_0/\sqrt{n}} \mid \text{IC}]$$

Using 2-sigma limits implies that $$UCL/LCL = \mu_0 \pm 2 \cdot \frac{\sigma_0}{\sqrt{n}}$$

$$= \Pr[z > 2 \text{ or } z < -2] = 1 - \Phi(2) + \Phi(-2)$$

$$= 1 - 0.9772 + 0.0228 = 0.0455$$
Control chart performance: Example 2.3

(b) What is the Type I error using Rule 2 alone?

\[
\alpha_{rule2} = \Pr[\text{exactly two out of the three samples are outside the 2-sigma limits | IC}]
\]

\[
= \Pr[1^{st} \text{ and } 2^{nd} \text{ samples are outside the 2-sigma limits but the 3^{rd} is inside the 2-sigma limits | IC}] + \Pr[1^{st} \text{ and } 3^{rd} \text{ outside, the } 2^{nd} \text{ inside | IC}] + \Pr[2^{nd} \text{ and } 3^{rd} \text{ outside, the } 1^{st} \text{ inside | IC}]
\]

\[
= \alpha_{2 \text{sigma}} \cdot \alpha_{2 \text{sigma}} \cdot (1 - \alpha_{2 \text{sigma}}) + \alpha_{2 \text{sigma}} \cdot (1 - \alpha_{2 \text{sigma}}) \cdot \alpha_{2 \text{sigma}}
\]

\[
+ (1 - \alpha_{2 \text{sigma}}) \cdot \alpha_{2 \text{sigma}} \cdot \alpha_{2 \text{sigma}}
\]

\[
= 3 \cdot \alpha_{2 \text{sigma}}^2 \cdot (1 - \alpha_{2 \text{sigma}}) = 3(0.0045)^2(1 - 0.0045) = 0.0059
\]

So \( ARL_0 = \frac{1}{\alpha_{rule2}} = \frac{1}{0.0059} \approx 169 \).
• **Sample size effect:** when sample size $n$ increases, both $\alpha$ error and $\beta$ error decrease, so $\text{ARL}_0$ is getting larger while $\text{ARL}_1$ is getting smaller.

• But a large sample size $n$ will delay the detection of a change because collecting samples takes time. In the meanwhile, taking more samples increases sampling costs.

• Typically, $n = 4, 5, \text{ or } 6$ strikes a good balance between obtaining desirable average run lengths (both $\text{ARL}_0$ and $\text{ARL}_1$) and keeping the sampling cost low.
One drawback of using a sample size of \( n=5 \) is its ineffectiveness in detecting a small magnitude in process change.

For example, use \( n = 5 \) for an \( \bar{X} \) chart to detect a mean shift \( \mu_1 = \mu_0 + 0.5\sigma_0 \) (namely \( k = 0.5 \)). Suppose \( \alpha = 0.0027 \).

\[
\beta = \Phi(z_{\alpha/2} - 0.5 \cdot \sqrt{5}) - \Phi(-z_{\alpha/2} - 0.5 \cdot \sqrt{5}) = 0.9701
\]

\[
\Rightarrow ARL_1 = \frac{1}{1-\beta} \approx 33, \text{ which is rather slow in detection.}
\]

One reason that a Shewhart chart does not respond to small changes quickly is because it has not "memory," namely that a Shewhart chart always use the latest update of process measurements instead of utilizing the information embedded in the entire sequence of measurements.
Control charts with memory

• By considering previous samples (i.e., let a chart to have memory), a chart can achieve faster detection of small mean shifts.

• In essence, giving a control chart memory is similar to increasing its sample size.

• Two primary methods:
  - CUSUM (Cumulative Sum)
  - EWMA (Exponentially Weighted Moving Average)
CUSUM

- CUSUM method was first proposed by Page (1954)
- Basic setting

Let \( \{x_1, x_2, \ldots\} \) be a sequence of observations.

\[ \mu_0 \equiv \text{in-control mean} \]

\[ \mu \equiv \text{the current mean (or the mean after a process change)} \]

\[ \sigma \equiv \text{standard deviation of } x_i \text{'s} \]

Note that CUSUM could be constructed using either individual \( x_i \)
or sample average \( \bar{x} \). For now we show how it is applied toindividual \( x_i \).
A simple approach of CUSUM

chart at observation $i$, such that,

$$C_i \equiv \sum_{j=1}^{i}(x_j - \mu_0) \quad \text{i.e., the cumulative sum of } x_j - \mu_0$$

$$= \sum_{j=1}^{i-1}(x_j - \mu_0) + (x_i - \mu_0)$$

$$= C_{i-1} + (x_i - \mu_0)$$

memory current value

**Note:** this simple approach is not the CUSUM that is used in practice. But it can help understand the mechanism behind the actual CUSUM.
Effect of mean shift on $C_i$

$$E(C_i) = E[C_{i-1} + x_i - \mu_0] = E[C_{i-1}] + E[x_i] - \mu_0 = E[C_{i-1}] + \mu_i - \mu_0$$  here $\mu_i = E[x_i]$

**CASE I:** $\mu_i = \mu_0$, no shift,
$E(C_i) = E(C_{i-1})$, $C_i$ will stay where it was.

**CASE II:** $\mu_i > \mu_0$, upward shift,
$E(C_i) = E(C_{i-1}) + (\mu_i - \mu_0) > E(C_{i-1})$, $C_i$ will also drift upward.

**CASE III:** $\mu_i < \mu_0$, downward shift,
$E(C_i) = E(C_{i-1}) + (\mu_i - \mu_0) < E(C_{i-1})$, $C_i$ will also drift downward.

Then, the trend in $C_i$ is an indicator of the process change.

Then, the trend in $C_i$ is an indicator of changes in the process.
- CUSUM example versus the Shewhart chart

![CUSUM example versus the Shewhart chart](image)
CUSUM

- Standard CUSUM formulation

\[ C_i^+ \equiv \max\{0, C_{i-1}^+ + [x_i - (\mu_0 + K)]\} \]  upper CUSUM statistic

\[ C_i^- \equiv \max\{0, C_{i-1}^- + [-x_i + (\mu_0 - K)]\} \]  lower CUSUM statistic

with \( C_0^+ = C_0^- = 0 \) and \( K \) is a user specified constant.

- Why not use the original simple CUSUM approach?

  (a) The inclusion of \( K \), an offset value, will make the CUSUM statistic, \( C_i^+ \) or \( C_i^- \), stay where it was, when

  \[ \mu_0 - K < \mu < \mu_0 + K. \]

  This \( K \) functions like a safeguard threshold, which prevents a lot of small fluctuations from accumulating and causing a CUSUM chart to signal.

  (b) The new formulation uses two statistics \( C_i^+ \) and \( C_i^- \) in order to decouple the negative and positive mean shifts so that they are not canceling out or disturbing each other’s previous effect.
CUSUM

- To implement a CUSUM method:
  
  1. Select $K$ and a decision interval $H$
  
  2. if $C_i^+ > H$ or $C_i^- > H \Rightarrow$ chart signals
     $\Rightarrow$ conclude that the process has changed.

- How to choose $K$ and $H$

  $K$ is often chosen about the halfway between $\mu_0$ and the out-of-control value $\mu_1$ that we mean to detect:

  $\Rightarrow \mu_1 = \mu_0 \pm \delta \cdot \sigma \Rightarrow \text{set } K = \frac{|\delta \cdot \sigma|}{2} = \frac{|\mu_1 - \mu_0|}{2}$

  A common choice of $H$ is $5\sigma$. 
Example 2.4

System parameters: $\mu_0 = 20$, $\sigma = 1$, $n = 1$. Try to detect a mean shift to $\mu_1 = \mu_0 + \sigma$ (i.e., $\delta = 1$).

Set $K = \frac{|\mu_1 - \mu_0|}{2} = 0.5\sigma = 0.5$ and $H = 5\sigma = 5$.

<table>
<thead>
<tr>
<th>observation # i</th>
<th>$x_i$</th>
<th>$C_i^+$</th>
<th>$N^+$</th>
<th>$C_i^-$</th>
<th>$N^-$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>19.32</td>
<td>0</td>
<td>0</td>
<td>0.18</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>20.55</td>
<td>0.05</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>18.14</td>
<td>0</td>
<td>0</td>
<td>1.36</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>18.96</td>
<td>0</td>
<td>0</td>
<td>1.9</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>21.17</td>
<td>0.67</td>
<td>1</td>
<td>0.23</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>20.77</td>
<td>0.93</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>18.73</td>
<td>0</td>
<td>0</td>
<td>0.77</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>22.27</td>
<td>1.77</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td>19.13</td>
<td>0.4</td>
<td>2</td>
<td>0.37</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>19.72</td>
<td>0</td>
<td>0</td>
<td>0.15</td>
<td>2</td>
</tr>
</tbody>
</table>

In the above table, $N^+$ and $N^-$ are the counts of the consecutive periods that the upper or the lower CUSUM statistic has been nonzero, respectively.
CUSUM: Example 2.4

- Interpretation of CUSUM:
  - Locate the data point $i_{out}$ at which $C^+$ or $C^-$ exceeds $H$.
  - The location of the last in-control data is

$$i_{in} = i_{out} - N_{out}^+ \text{ or } i_{in} = i_{out} - N_{out}^-$$

where $N_{out}^+$ or $N_{out}^-$ are the $N^+$ or $N^-$ at the point of $i_{out}$, respectively.
CUSUM: Example 2.4

- Interpretation of CUSUM:
  
  - Estimate the process mean after the change

  \[
  \hat{\mu} = \begin{cases} 
  \mu_0 + K + \frac{C_i^+}{N_{out}^+} & \text{if } C_i^+ > H \\ 
  \mu_0 - K - \frac{C_i^-}{N_{out}^-} & \text{if } C_i^- > H 
  \end{cases}
  \]

- In this example:

  The first \( i_{out} = 35 \) in the \( C^+ \) chart, and the corresponding \( N_{out}^+ = 4 \) \( \Rightarrow \) the last in-control data is at \( i_{in} = 35 - 4 = 31 \).

  The new process mean is

  \[
  \hat{\mu} = \mu_0 + K + \frac{C_i^+}{N_{out}^+} = 20 + 0.5 + \frac{5.02}{4} = 21.75.
  \]

  The true mean is \( \mu = 21 \).
CUSUM: chart design

• CUSUM chart design is to choose K and H, two parameters used by any CUSUM method.

• The previous choice of K and H is one of the typical guidelines.

• More general CUSUM chart design is based on proper choice of average run length, namely that we should select appropriate K and H so that we have a large enough $ARL_0$ when there is no process change and a small enough $ARL_1$ when there is a process change.

• These two parameters K and H are often expressed in terms of $\sigma$, i.e.,

$$K = k \cdot \sigma \quad \text{and} \quad H = h \cdot \sigma$$

So choosing K and H is equivalent to choosing k and h.
The impact of $K$ and $H$ on ARL is complicated. People have to rely on computational procedures to evaluate an ARL for a given combination of $K$ and $H$. Some results are available in literature: the combination of $k=0.5$ and $h=5$ produces a good overall result.

### CUSUM: chart design

#### Table 7-3 ARL Performance of the Tabular Cusum with $k = \frac{1}{2}$ and $h = 4$ or $h = 5$

<table>
<thead>
<tr>
<th>Shift in Mean (multiple of $\sigma$)</th>
<th>$h = 4$</th>
<th>$h = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$ARL_0$</td>
<td>168</td>
<td>465</td>
</tr>
<tr>
<td>$ARL_1$</td>
<td>74.2</td>
<td>139</td>
</tr>
<tr>
<td>0.25</td>
<td>26.6</td>
<td>38.0</td>
</tr>
<tr>
<td>0.50</td>
<td>13.3</td>
<td>17.0</td>
</tr>
<tr>
<td>0.75</td>
<td>8.38</td>
<td>10.4</td>
</tr>
</tbody>
</table>

Shewhart chart $ARL_1=43.96$

<table>
<thead>
<tr>
<th>Shift in Mean (multiple of $\sigma$)</th>
<th>$h = 4$</th>
<th>$h = 5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.00</td>
<td>4.75</td>
<td>5.75</td>
</tr>
<tr>
<td>1.25</td>
<td>3.34</td>
<td>4.01</td>
</tr>
<tr>
<td>2.50</td>
<td>2.62</td>
<td>3.11</td>
</tr>
<tr>
<td>3.00</td>
<td>2.19</td>
<td>2.57</td>
</tr>
<tr>
<td>4.00</td>
<td>1.71</td>
<td>2.01</td>
</tr>
</tbody>
</table>

#### Table 7-4 Values of $k$ and the Corresponding Values of $h$ That Give $ARL_0 = 370$ for the Two-Sided Tabular Cusum [from Hawkins (1993a)]

<table>
<thead>
<tr>
<th>$k$</th>
<th>0.25</th>
<th>0.5</th>
<th>0.75</th>
<th>1.0</th>
<th>1.25</th>
<th>1.5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h$</td>
<td>8.01</td>
<td>4.77</td>
<td>3.34</td>
<td>2.52</td>
<td>1.99</td>
<td>1.61</td>
</tr>
</tbody>
</table>
CUSUM: approximation of ARL

- People also developed some approximation formula for calculating the average run length for the CUSUM method. The most widely used one is the Siegmund's approximation.

For one-sided CUSUM (namely, either $C_i^+$ or $C_i^-)$:

$$A R L_{\pm} = \begin{cases} \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2} & \text{if } \Delta \neq 0 \\ \frac{1}{b^2} & \text{if } \Delta = 0 \end{cases}$$

where

$$\Delta = \begin{cases} \delta^* - k & \text{for } C_i^+ \\ -\delta^* - k & \text{for } C_i^- \end{cases}$$

$$b = h + 1.166$$

$$\delta^* = \frac{\mu - \mu_0}{\sigma}$$

- For a two-sided CUSUM, ARL is given by $\frac{1}{A R L} = \frac{1}{A R L^+} + \frac{1}{A R L^-}$
- When $\delta^* = 0 \Rightarrow$ we get $A R L_0$
- When $\delta^* \neq 0 \Rightarrow$ we get $A R L_1$
CUSUM: approximation of ARL

- A numerical example

Suppose that $\mu_0 = 40$, $\sigma = 2$. Try to detect a mean shift to $\mu_1 = 43$. Also attempt to maintain the in-control $ARL_0 = 370$, the same as an $\bar{x}$ chart with a 3-sigma control limit.

Choose $k$ s.t. $k = \frac{|\mu_1 - \mu_0|}{2\sigma} = \frac{3}{2.2} = 0.75$. How to choose $h$? ⇒ use the Siegmund’s approximation.

(1) Try $h = 3.25$

Set $\delta^* = 0$ to verify if $ARL_0$ is close enough to 370.

Then, $\Delta = -k = -0.75$ for both $C_i^+$ and $C_i^-$; $b = h + 1.166 = 4.4166$;

⇒ $ARL_0^+ = ARL_0^- = \frac{\exp(-2 \cdot (-0.75) \cdot 4.4166)}{2(-0.75)^2} = 662$

⇒ $\frac{1}{ARL} = \frac{1}{ARL^+} + \frac{1}{ARL^-} = 0.0003 \Rightarrow ARL_0 = 331 < 370$

$ARL_0$ is not large enough so we will increase $h$ a little and try again.
CUSUM: approximation of ARL

- A numerical example

(2) Try \( h = 3.33 \). It is easy to verify that \( ARL_0 = 374 \) now, which is close enough to 370.

So \( H = h\sigma = (3.33)\times 2 = 6.66 \)

\[ K = k\sigma = (0.75)\times 2 = 1.5. \]

Then, need to verify \( ARL_1 \) for detecting \( \mu_1 \)

\[ \delta^* = \frac{\mu_1 - \mu_0}{\sigma} = 1.5 \]

For \( C_i^+ \), \( \Delta = \delta^* - k = 0.75 \);
for \( C_i^- \), \( \Delta = -\delta^* - k = -2.25 \);

\[ b = h + 1.166 = 4.496 \]

\[ \Rightarrow ARL_1^+ = 5.11 \text{ and } ARL_1^- = 6 \times 10^7 \Rightarrow \frac{1}{ARL_1^-} \approx \frac{1}{ARL_1^+} \]

\[ \Rightarrow ARL_1 \approx 5.11 \]

As a comparison with a Shewhart individual chart (i.e., \( n = 1 \)) with 3-sigma control limits: its \( ARL_0 = 370 \) but \( ARL_1 = 81 \) for detecting a process change of the same magnitude, which is 16 times slower than the CUSUM method.
CUSUM: more discussion

- We can also apply CUSUM to a sample average $x$-bar.
  
  - replace $x_i$ by $\bar{x}_i$
  
  - replace $\sigma_x$ by $\sigma_{\bar{x}} = \sigma_x / \sqrt{n}$

- We can only use a one-sided CUSUM (either $C^+$ or $C^-$) if we are only interested in process changes in one specific direction.

- CUSUM is good at detecting process changes of small to moderate magnitude, while Shewhart chart is good at detecting process changes of large magnitude. In practice, one can combine both methods for the purpose of effective detections.
• EWMA = Exponentially Weighted Moving Average; an alternative to CUSUM, another control chart with memory

• Basic idea: chart the EWMA statistic

\[ z_i = \lambda x_i + (1 - \lambda)z_{i-1}, \text{ with } z_0 = \mu_0 \]
where \( \lambda \) is a user specified constant and \( 0 < \lambda \leq 1 \).

• Why is this statistic \( z_i \) called EWMA?

If substitute \( z_{i-1} \) using \( z_{i-2} \) and then substitute \( z_{i-2} \) using \( z_{i-3} \), and so on, we can have the following expression:

\[ z_i = \lambda \sum_{j=0}^{i-1}(1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0 \]

So \( z_i \) is a weighted moving average of the past data of \( x_{i-j} \) with the weight being \( (1 - \lambda)^j \). This weight decreases exponentially as one goes back in time. That is why it is called “Exponentially Weighted Moving Average.”
How to choose UCL/LCL for an EWMA chart?

Assume $x_i \sim \text{NID}(\mu, \sigma^2)$, then $z_i$ also follows a normal distribution because $z_i$ is a linear combination of $x_i$.

$\Rightarrow$ need to find the mean and standard deviation of $z_i$ and then use the general control chart model (i.e., now $w = z_i$).

Mean of $z_i$

$$
\mu_i \equiv E[z_i] = E[\sum_{j=0}^{i-1} \lambda(1-\lambda)^j x_{i-j} + (1-\lambda)^i z_0] = \sum_{j=0}^{i-1} \lambda(1-\lambda)^j E[x_{i-j}] + (1-\lambda)^i \mu_0
$$

$$
= \sum_{j=0}^{i-1} \lambda(1-\lambda)^j \mu + (1-\lambda)^i \mu_0 \rightarrow \mu \text{ as } i \rightarrow \infty \text{ because } (1-\lambda)^i \rightarrow 0 \text{ and } \lim_{i \rightarrow \infty} \sum_{j=0}^{i-1} \lambda(1-\lambda)^j = \frac{\lambda}{1-(1-\lambda)} = 1.
$$

The above equation also implies that if $\mu = \mu_0$, then $\mu_i = \mu_0$.
• **Variance of** \( z_i \)

\[
\sigma_i^2 \equiv \text{var}[z_i] = \text{var}[\sum_{j=0}^{i-1} \lambda(1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0] \\
= \sum_{j=0}^{i-1} \lambda^2 (1 - \lambda)^{2j} \text{var}[x_{i-j}] \\
= \sigma^2 \lambda^2 \sum_{j=0}^{i-1} (1 - \lambda)^{2j} = \sigma^2 \lambda^2 \frac{1 - (1 - \lambda)^{2i}}{1 - (1 - \lambda)^2} \\
= \sigma^2 \frac{\lambda}{2 - \lambda} [1 - (1 - \lambda)^{2i}]
\]

Here we utilize the formula \( \sum_{j=0}^{k} a^j = \frac{1 - a^{k+1}}{1 - a} \).

\[
\sigma_i^2 \rightarrow \sigma^2 \frac{\lambda}{2 - \lambda} \text{ when } i \rightarrow \infty, \text{ and this is called } \text{steady state} \text{ value.}
\]

• **Time-varying (or sample-varying) control limits**

Under \( H_0, \mu_w = \mu_0 \) and \( \sigma_w = \sigma_i \), so

\[
\text{UCL}_i/LCL_i = \mu_0 \pm L \cdot \sigma \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1 - \lambda)^{2i}]}
\]

• **Steady-state control limits (namely, when } i \rightarrow \infty \text{ or large enough)**

\[
\text{UCL}_i/LCL_i = \mu_0 \pm L \cdot \sigma \sqrt{\frac{\lambda}{2 - \lambda}} \\
\text{since } (1 - \lambda)^{2i} \rightarrow 0 \text{ as } i \rightarrow \infty
\]
Example 2.5: Consider the data in the table of next slide. The first 20 of these observations were drawn at random from a normal distribution with mean $\mu = 10$ and standard deviation $\sigma = 1$. The last 10 observations were drawn from a normal distribution with mean $\mu = 11$ and standard deviation $\sigma = 1$. 

[Graph of EWMA chart]
### EWMA: Example 2.5 data table

<table>
<thead>
<tr>
<th>Subgroup, $i$</th>
<th>$x_i$</th>
<th>EWMA, $z_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9.45</td>
<td>9.945</td>
</tr>
<tr>
<td>2</td>
<td>7.99</td>
<td>9.7495</td>
</tr>
<tr>
<td>3</td>
<td>9.29</td>
<td>9.70355</td>
</tr>
<tr>
<td>4</td>
<td>11.66</td>
<td>9.8992</td>
</tr>
<tr>
<td>5</td>
<td>12.16</td>
<td>10.1253</td>
</tr>
<tr>
<td>6</td>
<td>10.18</td>
<td>10.1307</td>
</tr>
<tr>
<td>7</td>
<td>8.04</td>
<td>9.92167</td>
</tr>
<tr>
<td>8</td>
<td>11.46</td>
<td>10.0755</td>
</tr>
<tr>
<td>9</td>
<td>9.2</td>
<td>9.98796</td>
</tr>
<tr>
<td>10</td>
<td>10.34</td>
<td>10.0232</td>
</tr>
<tr>
<td>11</td>
<td>9.03</td>
<td>9.92384</td>
</tr>
<tr>
<td>12</td>
<td>11.47</td>
<td>10.0785</td>
</tr>
<tr>
<td>13</td>
<td>10.51</td>
<td>10.1216</td>
</tr>
<tr>
<td>14</td>
<td>9.4</td>
<td>10.0495</td>
</tr>
<tr>
<td>15</td>
<td>10.08</td>
<td>10.0525</td>
</tr>
<tr>
<td>16</td>
<td>9.37</td>
<td>9.98426</td>
</tr>
<tr>
<td>17</td>
<td>10.62</td>
<td>10.0478</td>
</tr>
<tr>
<td>18</td>
<td>10.31</td>
<td>10.074</td>
</tr>
<tr>
<td>19</td>
<td>8.52</td>
<td>9.91864</td>
</tr>
<tr>
<td>20</td>
<td>10.84</td>
<td>10.0108</td>
</tr>
<tr>
<td>21</td>
<td>10.9</td>
<td>10.0997</td>
</tr>
<tr>
<td>22</td>
<td>9.33</td>
<td>10.0227</td>
</tr>
<tr>
<td>23</td>
<td>12.29</td>
<td>10.2495</td>
</tr>
<tr>
<td>24</td>
<td>11.5</td>
<td>10.3745</td>
</tr>
<tr>
<td>25</td>
<td>10.6</td>
<td>10.3971</td>
</tr>
<tr>
<td>26</td>
<td>11.08</td>
<td>10.4654</td>
</tr>
<tr>
<td>27</td>
<td>10.38</td>
<td>10.4568</td>
</tr>
<tr>
<td>28</td>
<td>11.62</td>
<td>10.5731</td>
</tr>
<tr>
<td>29</td>
<td>11.31</td>
<td>10.6468*</td>
</tr>
<tr>
<td>30</td>
<td>10.52</td>
<td>10.6341*</td>
</tr>
</tbody>
</table>
EWMA: chart design

- Design of an EWMA chart boils down to choosing two parameters: \( \lambda \) and \( L \).

- \( \lambda \) controls how much "memory" the chart has.
  - when \( \lambda = 1 \), \( z_i = \lambda x_i + (1-\lambda)z_{i-1} = x_i \) (memoryless)
  - thus, EWMA chart \( \equiv \) Shewhart chart
  - smaller \( \lambda \) \( \Rightarrow \) longer memory \( \Rightarrow \) good for detecting small changes
  - larger \( \lambda \) \( \Rightarrow \) shorter memory \( \Rightarrow \) good for detecting large changes

- Typically, use \( 0.05 \leq \lambda \leq 0.25 \)
- Generally, \( L = 3 \). but for \( \lambda \leq 0.1 \), \( L = 2.6 \sim 2.8 \).
- More general approach relies on the impact of \( \lambda \) and \( L \) combination on the average run length. Some results from the literature.
EWMA: chart design

- Average run length under different combinations of $\lambda$ and L.

<table>
<thead>
<tr>
<th>Shift or Mean (multiple of 5)</th>
<th>$L = 3.054$</th>
<th>$\lambda = 0.40$</th>
<th>2.998</th>
<th>2.962</th>
<th>2.814</th>
<th>2.615</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>0.25</td>
<td>224</td>
<td>170</td>
<td>150</td>
<td>106</td>
<td>84.1</td>
<td></td>
</tr>
<tr>
<td>0.50</td>
<td>71.2</td>
<td>48.2</td>
<td>41.8</td>
<td>31.3</td>
<td></td>
<td>28.8</td>
</tr>
<tr>
<td>0.75</td>
<td>28.4</td>
<td>20.1</td>
<td>18.2</td>
<td>15.9</td>
<td></td>
<td>16.4</td>
</tr>
<tr>
<td>1.00</td>
<td>14.3</td>
<td>11.1</td>
<td>10.5</td>
<td>10.3</td>
<td></td>
<td>11.4</td>
</tr>
<tr>
<td>1.50</td>
<td>5.9</td>
<td>5.5</td>
<td>5.5</td>
<td>6.1</td>
<td></td>
<td>7.1</td>
</tr>
<tr>
<td>2.00</td>
<td>3.5</td>
<td>3.6</td>
<td>3.7</td>
<td>4.4</td>
<td></td>
<td>5.2</td>
</tr>
<tr>
<td>2.50</td>
<td>2.5</td>
<td>2.7</td>
<td>2.9</td>
<td>3.4</td>
<td></td>
<td>4.2</td>
</tr>
<tr>
<td>3.00</td>
<td>2.0</td>
<td>2.3</td>
<td>2.4</td>
<td>2.9</td>
<td></td>
<td>3.5</td>
</tr>
<tr>
<td>4.00</td>
<td>1.4</td>
<td>1.7</td>
<td>1.9</td>
<td>2.2</td>
<td></td>
<td>2.7</td>
</tr>
</tbody>
</table>

Table 7-10: Average Run Lengths for Several EWMA Control Schemes [Adapted from Lucas and Saccucci (1990)]
EWMA versus CUSUM

- Compared with CUSUM, EWMA can achieve a very similar performance.

<table>
<thead>
<tr>
<th>EWMA</th>
<th>CUSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda = 0.1$</td>
<td>$k = 0.5$</td>
</tr>
<tr>
<td>$L = 2.814$</td>
<td>$h = 5$</td>
</tr>
<tr>
<td>$\text{ARL}_0 = 500$</td>
<td>$\text{ARL}_0 = 465$</td>
</tr>
<tr>
<td>$\text{ARL}_1 (1\sigma) = 10.3$</td>
<td>$\text{ARL}_1 (1\sigma) = 10.4$</td>
</tr>
</tbody>
</table>
EWMA: more discussion

- One can apply EWMA to sample averages x-bar:
  - replace $x_i$ with $\bar{x}_i$ and $\sigma$ with $\sigma_{\bar{x}} = \sigma / \sqrt{n}$.

$$UCL_i = \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2-\lambda)n} \left[ 1 - (1 - \lambda)^2i \right]}$$

$$CL = \mu_0$$

$$LCL_i = \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2-\lambda)n} \left[ 1 - (1 - \lambda)^2i \right]}$$

- Even though we assume $x$ follows a normal distribution when deriving the mean and variance of the EWMA statistic, the EWMA statistic is in fact insensitive to the normality assumption. So it is more suitable than a Shewhart chart for individual observations since individual observations may not necessarily follow a normal distribution very well.
MA: a special form of EWMA

- Moving average (MA) chart is a special form of EWMA – it is an unweighted moving average.

\[ M_i \equiv \frac{\sum_{j=i-w+1}^{i} x_j}{w} \], where \( w \) is called window size.

- Moving average is to average all data falling into a moving window of size \( w \). This moving window incorporates some of the memory of the past data by continuously dropping the oldest data and adding the newest data.

\[
\begin{align*}
\text{Moving Average} & \quad \text{EWMA} \\
\text{weight} & \quad \text{weight} \\
\text{time} & \quad \text{time} \\
i-w+1 & \quad \text{moving front} \\
i & \quad i \\
\end{align*}
\]
MA: control limits

- Let $\mu_0$ be the in-control mean and $\text{var}(M_i) = \sigma^2/w$

- when $i \geq w$

\[
UCL = \mu_0 + L \frac{\sigma}{\sqrt{w}}
\]

\[
CL = \mu_0
\]

\[
LCL = \mu_0 - L \frac{\sigma}{\sqrt{w}}
\]

- At the beginning of a detection process, when $i < w$, the number of data in the window is less than $w$, then use

\[
UCL_i = \mu_0 + L \frac{\sigma}{\sqrt{i}}
\]

\[
CL = \mu_0 \quad \text{i = 1, 2, ..., w-1}
\]

\[
LCL_i = \mu_0 - L \frac{\sigma}{\sqrt{i}}
\]
Example 2.6: Apply a MA chart to the data in Example 2.5

\[ w = 5, \mu_0 = 10, \sigma = 1, n = 1, L = 3 \]

\[ i = 1, M_1 = x_1 = 9.45, UCL_1/LCL_1 = \mu_0 \pm 3 \cdot \frac{1}{\sqrt{1}} = [13, 7] \]

\[ i = 2, M_2 = \frac{x_1 + x_2}{2} = 8.72, UCL_2/LCL_2 = \mu_0 \pm 3 \cdot \frac{1}{\sqrt{2}} = [12.12, 7.87] \]

... 

\[ i = 5 \text{ and onward, } M_5 = \frac{x_{i-4} + x_{i-3} + \ldots + x_i}{5}, UCL/LCL = \mu_0 \pm 3 \cdot \frac{1}{\sqrt{5}} = [11.34, 8.66] \]