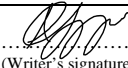




University of
Stavanger

Faculty of Science and Technology

MASTER'S THESIS

Study program/ Specialization: MSc Mathematics and Physics	Spring semester, 2020.. Open / Restricted access
Writer: Alexandra Evangelou	 (Writer's signature)
Faculty supervisor: Jan Terje Kvaløy	
External supervisor(s):	
Thesis title: Handling estimation error in Statistical Process Control with EWMA charts	
Credits (ECTS): 60	
Key words: Process Control, Control Charts, EWMA, Process Analysis Estimation Error	Pages: 76..... + enclosure: Stavanger, June, 2020..... Date/year

Handling estimation error in Statistical Process Control with EWMA charts

Alexandra Evangelou

University of Stavanger
Department of Mathematics and Physics
Submission Date : June 2020
Supervisor : Jan Terje Kvaløy

Preface

I would like to thank my supervisor Jan Terje Kvaløy for the invaluable help and the support during the whole master thesis elaboration period. I would also want to thank Pr. Diko for providing an R-code and for the help he offered on its interpretation. Last but not least, I thank Dr. Hege Langli Ersdal and the SAFER birth team for giving me the data and permitted me to analyze them.

Abstract

This master thesis studies the process control analysis using control charts. It starts with a detailed presentation of the basic types of control charts, with a special focus on the EWMA control chart. An excessive examination is performed on the error arising when the unknown process distribution is being estimated and on ways of eliminating this error. More precisely, the control limits are being adjusted with two different ways and the results are being reviewed. Finally, the theoretical results are applied on data from a study on newborn's birth weight.

Contents

1. Introduction	2
2. Control Charts	4
2.1. Statistical Base for Control Charts	4
2.1.1. General model for Shewhart type control chart	5
2.1.2. Average Run Length	6
2.1.3. Control Charts and hypothesis testing	7
2.2. An example on Shewhart's control chart	8
3. The CUSUM and EWMA Control Charts	11
3.1. The Cumulative Sum (CUSUM) Control Chart	11
3.2. The Exponentially Weighted Moving Average (EWMA) Control Chart	14
4. The EWMA Control Chart for known process distribution	15
4.1. An example of EWMA control chart for known process distribution	17
5. The EWMA Control Chart for unknown distribution	19
5.1. EWMA Control Chart with estimated parameters	19
5.1.1. An example of EWMA control chart with estimated parameters	20
5.2. Control chart's ARL for estimated parameters	21
6. Proposal for handling the estimation error in EWMA Control Chart.	33
6.1. General setting for the adjustment	33
6.1.1. Evaluation of the adjusted control limits	35
6.2. Another threshold adjustment	40
6.3. Adjusted Thresholds for Out Of Control process	46
7. Baby birth weight analysis	52
7.1. EWMA control chart for birth weight	52
7.2. Regression model for birth weight	55
7.2.1. EWMA control chart on the residuals of the predicted data	56
7.2.2. Discussion on birth weight	57
8. Conclusion	59
A. Appendix A	60
A.1. Z distribution table	60
A.2. R codes	61
A.2.1. R-code for the EWMA control chart	61

A.2.2. R code for the distribution of ARL for estimated distribution	62
A.2.3. R code for adjusted ARL distribution	64
A.2.4. R code for regression model	67

1. Introduction

In all processes, regardless of how well they have been designed and how carefully they are being operated, it is common that an amount of variation is present. This variability can be natural or inherent. When this is the case, variation consists of a white noise for the process, meaning that is an uncontrollable (random) variation. In statistical quality control procedures, a process operating under this natural variation is called "stable system of chance causes" and is said to be in statistical control.

On the other hand, a non random variation which consists of a controllable type of variation may be present. Such variability which generally is statistically significant compared to the background noise, causes an unacceptable unstable operating performance. This is the reason why it is crucial to identify and encounter these assignable causes of variation that lead to non random variability.

The following graph is an illustration of the impact of chance and assignable causes of variation on a process.

When a process is under control, its mean and standard variation will

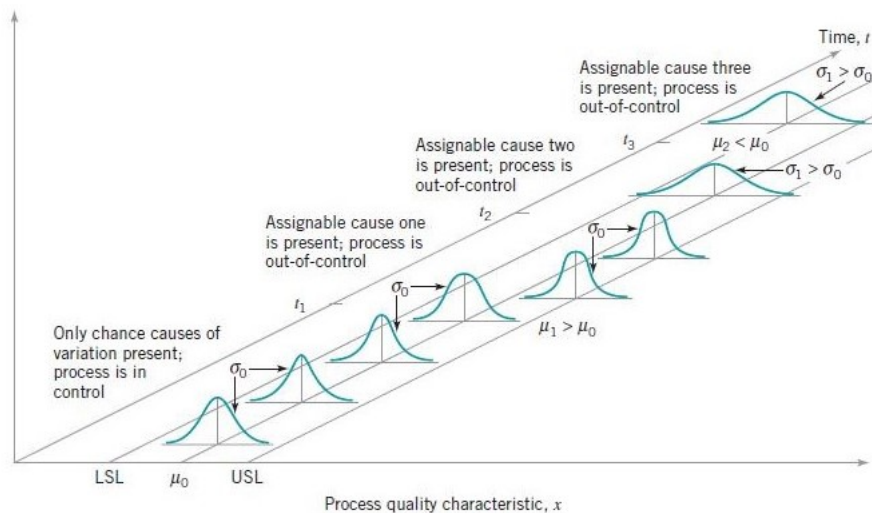


Figure 1.1.: Chance and assignable causes of variation. With permission: "Copyright © 2009 by John Wiley & Sons, Inc. All rights reserved." [7]

be in its in control values μ_0 and σ_0 and most of the observations will lie between two limits the Upper Specification Limit (USL) and Lower Specification Limit (LSL). In this thesis the more standard abbreviation for these limits UCL and LCL will be used, which stand for Upper Control Limit and Lower Control Limit respectively. But when assignable causes occur, they effect violently the mean or the variation. As seen on figure 1.1 an increase on mean to value μ_1 leads the process distribution out of the limits, moving a remarkable amount of the data beyond the USL. While a simultaneous shift on both mean and standard deviation ($\mu_2 < \mu_0$ and $\sigma_1 > \sigma_0$) widens

the sample distribution and moves it further to the left of the LSL. These are cases at which the process is said to be out of control. It is critical then to detect the assignable causes of variation and take action in order the process to be back under control. An effective statistical tool for detecting and defining the assignable causes is the control charts.

2. Control Charts

A control chart is a graphical display of a quality characteristic of a process versus the sample number or time. A basic control chart contains a *central line (CL)* which represent the mean value of the quality characteristic when only white noise occurs. In a basic form of the chart there are also two limit lines *UCL* and *LCL* which represent the *Upper Control Limit* and the *Lower Control Limit* respectively.

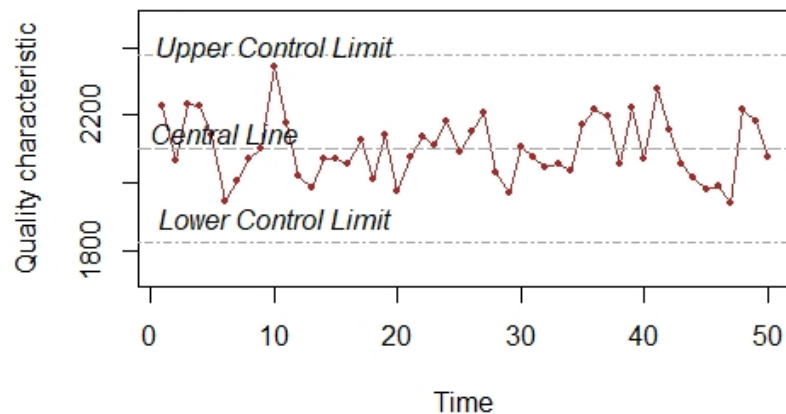


Figure 2.1.: Typical Shewharth's Control Chart

A basic control chart plots both chance and assignment causes of variation. In general, random causes of variability give observations which lie inside the control limits, while under non random variation, the sample data tend to fall off the limits. Control chart's basic operation is to monitor the variations. When an assignable cause of variation exists, the control charts depicts it showing sample data beyond the control limits.

2.1. Statistical Base for Control Charts

A typical type of a control chart (*Shewhart's control chart*) consists of two control lines, the Upper Control Limit and the Lower Control Limit. The basic rule of control charts denotes that if one or more sample points fall out of UCL or LCL, the process may be out of control and further investigation is needed. According to control charts theory, the case in which all points lie inside the control limits does not necessarily mean that the process is in control. When the points do not exceed limits but seem to follow a specific pattern, for example most of the sample is above central line, this

could utmost lead to the suspicion of lack of control. In order to have some indication of a process in control the sample should be in an essential random pattern.

Statistical process control is performed in two phases *Phase I* and *Phase II*. In Phase I or *retrospective phase*, historic data are collected in order to start the process analysis. Firstly, the control charts are being designed by defining the control limits and plotting the collected data. This is when the sample is being checked for any points that lie beyond the control limits or follow any pattern. If such a behavior is spotted, new control limits are being calculated discarding the abnormal data and using the remaining sample. The final scope is to obtain an *in control* process. When this is accomplished, the Phase II or *prospective phase* follows. During this phase, the established control limits are used in order to monitor the process regularly, for spotting any assignable variation.

2.1.1. General model for Shewhart type control chart

Let w denote a sample statistic of a process that describes a quality characteristic with a mean value μ_w and a standard deviation σ_w , for example the mean value \bar{x} . Then the Central Line, the Upper and Lower Control Limits of a Shewharts control chart are as follows

$$\begin{aligned} UCL &= \mu_w + L\sigma_w \\ CL &= \mu_w \\ LCL &= \mu_w - L\sigma_w \end{aligned}$$

where the threshold L is the distance from the central line. Specifying L (width of the control limits), is critical for the forming of a control chart. A common method is to assume that when the process is in control, and by using the Central Limit Theorem, it is accepted that w follows a normal distribution and a $100(1-\alpha)\%$ of the samples w are expected to fall between $\mu_w - Z_{\alpha/2}\sigma_w$ and $\mu_w + Z_{\alpha/2}\sigma_w$. If a $Z_{\alpha/2}$ equal to 3 is chosen then the control chart is a **three sigma control chart**. Generally, a wider L reduces the Type I error which is defined as *a test concludes to an out of control process when actually the process is in control*. However, a more narrow control space can reduce error of Type II *the control chart shows a process to be in control when it is actually out of control*. The plot of Type II error named *operating characteristic curve* is used to illustrate the probability of process shifts in different magnitudes. For example, for the three-sigma control chart assuming that our statistic w is approximately normal and α the Type I error probability, the $1 - \alpha$ is the probability that there is no signal when the process is in control.

$$\begin{aligned} 1 - \alpha &= P(LCL < w < UCL) = P(\mu_w - 3\sigma_w < w < \mu_w + 3\sigma_w) = \\ &= P(-3 < Z < 3) = 0.9973 \end{aligned}$$

This shows that the probability the statistic w to be erroneously beyond control limits for a three-sigma control chart indicating an out of control process (Type I error), is $1 - 0.9973 = 0.0027$, meaning out of 10000 samples, 27 (or 1 in every 370 approximately) will be falsely spotted as out of

control. Alternatively, the Type I error probability can be specified and the corresponding control limits may be calculated. Therefore, for example the probability could be chosen to be 0.001 so that in order to have symmetrical limits, the L is computed by the equations :

$$P(w \geq \mu_w + L\sigma_w) = P(Z \geq L) = 0.0005$$

or

$$P(w \leq \mu_w - L\sigma_w) = P(Z \leq -L) = 0.0005$$

with $Z \sim N(0, 1)$. Following the normal distribution table (see A), $L = 3.9$.

2.1.2. Average Run Length

A control chart monitors a process's stability and ideally should quickly and effectively detect any undesired behavior. Therefore, while designing a control chart it is important to ensure its quality by specifying parameters such as sample size, frequency of sampling, central and control lines.

It is usual for a process evaluation to collect a number of data as a sample in each predetermined time interval and its mean value to be plotted. That indicates that a big number of data in each sample can detect more precisely any abnormality, by spotting smaller shifts. On the other hand, when alterations are larger, a smaller sample is likely to give a more effective control chart.

An important element to be defined, is how to *allocate the sample effort*. It is obvious that ideally a big number of observations would be chosen for each sample point along with a small sampling interval. In practice however, this is not generally feasible. In order to have an indication about a suitable sample size, a combination of the basic control chart parameters ARL and L , as presented in the following is being used. One of the fundamental parameters for a control chart is the *Run Length*, which denotes the number of points plotted until the very first observation to exceed one of the control limits. The Run Length is a discrete random variable with a probability mass function (pmf). Its mean value which represents the expected number of control statistics until the first out of control signal, is named *Average Run Length (ARL)*. In case that the observations are uncorrelated and the process is in control or in the same out of control all the time, it follows the geometric distribution and then the Average Run Length for the Shewharth's control chart is given by the formula

$$ARL = \frac{1}{\alpha}$$

where α is the probability an observation exceeds a control limit. When the process is in control, the ARL_{IC} denotes the average number of points until the first out of control signal is given. On the other hand, ARL_{OOC} , is the Average Run Length until the out of control process is detected. A control chart is efficient when the ARL_{IC} is big and ARL_{OOC} is small. So to ensure a small Type I error, the control chart should have a large ARL_{IC} and a ARL_{OOC} small enough for optimum performance in spotting shifts

of a relevant size. For a discrete process ARL, which from now on will denote the in control ARL_{IC} unless stated otherwise, depends only on the probability that an element lies beyond the defined control limits when the process is in control. If the limits are wide, this probability will be small, giving a big ARL value. For example, as found before, the False Alarm Probability for the three-sigma control chart is 0.0027 and the equivalent $ARL = \frac{1}{0.0027} = 370$. Reducing the value of L to 2.68,

$$\begin{aligned} 1 - a &= P(LCL < w < UCL) = P(\mu_w - 2.68\sigma_w < w < \mu_w + 2.68\sigma_w) = \\ &= P(-2.68 < Z < 2.68) = 0.9264 \end{aligned}$$

and $p = 0.00736$. So the corresponding value of ARL is 136, that is out of 136 one sample will give a false alarm in the long run. By changing therefore the value of L , the Average Run Length can be controlled or vice versa, for a certain value of ARL, the control limits may be adjusted.

Although the ARL is the most widely used performance indicator for the control charts, the significant skewness of the Run Length distribution, render it as a potential misleading operator for the chart analysis. Indeed, the Run Length distribution is considerably right skewed. The last years, some researchers [1] suggested that if some percentiles instead of the average are used, the effects of the skewness could be reduced. The most appropriate from robustness point of view seems to be the 50th percentile (*median*) and the Median Run Length (*MRL*). Similarly to ARL, an efficient control chart has large MRL_{IC} and significant small MRL_{OOC} . In some occasions actually, it is convenient instead of ARL to use the *Average Time to Signal* (ATS), which is the average Run Length when the samples are taken in specified time intervals. That is:

$$ATS = ARL_h$$

2.1.3. Control Charts and hypothesis testing

Control charts and hypothesis testing can easily be considered equivalent as statistical control methods. Let S a process statistic with its standard deviation σ_S to be an unbiased estimator for the process parameter θ (e.g. $E(S) = \theta$ and $Var(S) = \sigma_S^2$). Using a control chart, actually a test is performed on whether the sample's estimator value equals to the value of individual measurements.

$$H_0 : \theta = \theta_0 \vee H_1 : \theta \neq \theta_0$$

In case this statistic is the mean, the control charts tests the hypothesis this mean value (μ) to be equal to the mean of the individual measurements (μ_0), so as to test the following:

$$\begin{aligned} H_0 &: \text{The process is in control} \\ H_1 &: \text{The process is out of control} \end{aligned}$$

The control chart practically tests the hypothesis whether the process is in control. A basic difference between control chart and hypothesis testing is that in a control chart for the evaluation of a process the test is repeated

sequentially, while in hypothesis testing it is usual that a test is applied once. Additionally, comparing hypothesis testing and control chart test for a process the use of the control chart the essential assumptions of an hypothesis testing, such as independence in order to reduce variability, can easily be omitted. In the framework of control charts, the actual use of hypothesis testing is for to evaluate the performance of the control, by calculating the Type I and Type II errors.

2.2. An example on Shewhart's control chart

This example consists a demonstration of a Shewhart's control chart. For the presentation of this chart data are simulated from a normal distribution and their mean value is plotted. The control limits are being calculated by a desired predefined value for ARL and by using the formula,

$$\alpha = \frac{1}{ARL}$$

the α is derived and subsequently the factor L in order Upper and Lower Control Limits to be computed. More precisely, a set of 5 observation from a normal distribution with mean value 1500 and standard deviation 50 are simulated and their mean value is being calculated. This way the first element of the variable X_i , $i = 1, \dots, 50$ is being obtained. This is repeated until a total of 50 simulated mean values is gathered and plotted. The mean value of the sample equals to $\hat{\mu}_x = 1499.434$ and its sample standard deviation is $\hat{\sigma}_x = 17.0537$. Choosing an ARL equal to 370 gives $L = 3$. The control limits UCL, CL and LCL are then respectively 1550.59, 1499.43 and 1448.28.

Then, the derived control chart is the following:

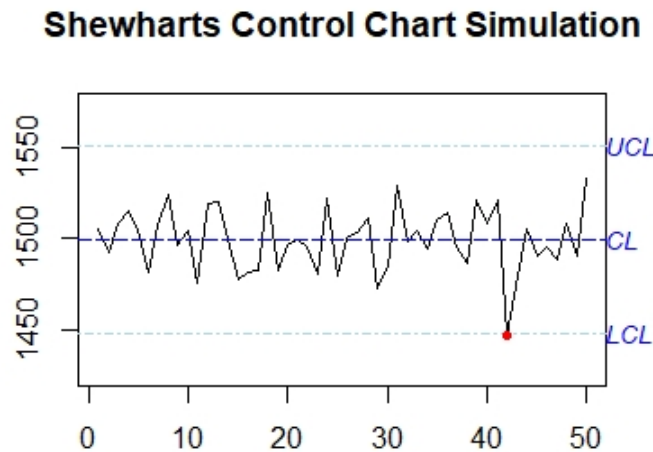


Figure 2.2.: Control chart for grouped data. The control limits have been calculated on estimated $\hat{\mu}$ and $\hat{\sigma}$.

The chart spots an extreme value equal to 1447.16. This would be natural because since there is a point that exceeds one of the control limits, it easily could be claimed that the chart signals our sample to be out of control. But as it is proved also by the following, this is a typical example of false alarm of a point exceeding the control limits while the process in control (Type I error). In this first simulation the control limits are calculated on the sample's mean and standard deviation. If instead in the simulation the same factors are preserved but instead UCL CL and LCL are based on the true values of $\mu_x = 1500$ and $\sigma_x = 22.4$, the control chart is as follows.

Shewharts Control Chart Simulation

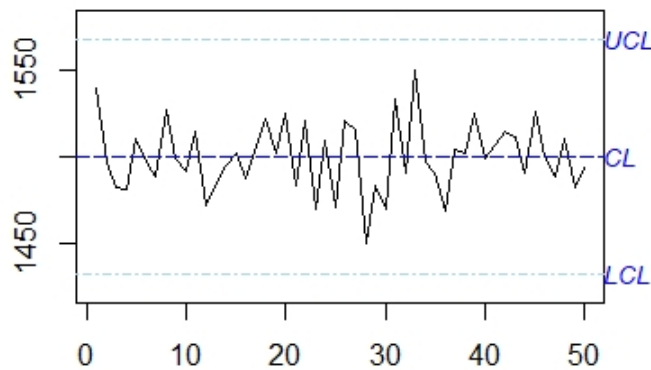


Figure 2.3.: Control chart for grouped data. Control limits are calculated on true values of mean and standard deviation.

In this case the limit space $(LCL, CL, UCL) = (1567.1915001432.81)$, is so big that if a decent value for ARL is needed (e.g the 3-sigma ARL which equals to 370), even if the process is out of control will be difficult to be tracked (big value for Type II error).

In the following example the Shewhart control chart is used in order to examine its behavior in small shifts of the mean value. For this case two samples of 30 and 50 equivalently observations each of size 5 are being generated. The first 30 are in control ($\mu_0 = 0$ and $\sigma_0 = 1$) while the next 50 have a shift in the mean equal to $\delta = 0.4$ ($\mu = 0.4$ and $\sigma = 1$). For this chart $ARL = 370$. As it is clear although the points seem to have a trend to be over CL line, all the data remain inside the control limits so the process control fails to spot assignable variation of the process for such a small shift.

Shewharts Control Chart Simulation

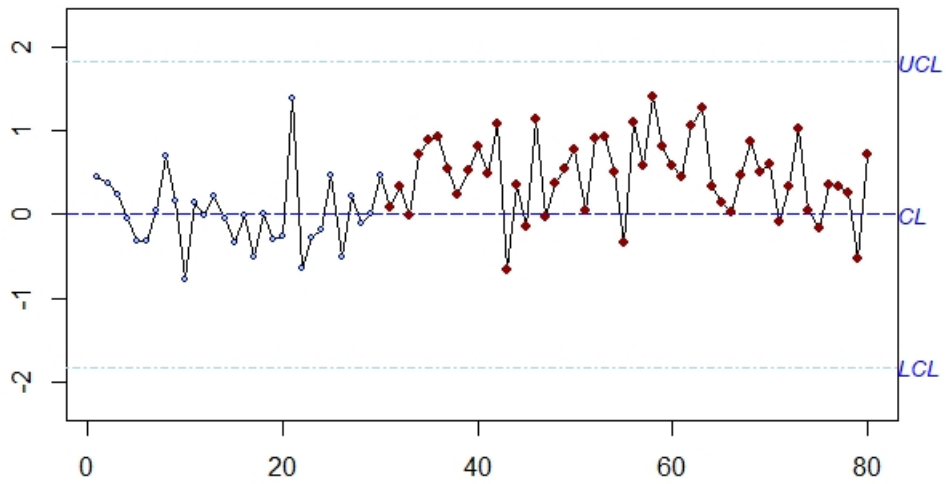


Figure 2.4.: Shewhart's Control Chart for small shift

In real life it is more convenient to use the standardized form of a control chart. This form plots the data in standard deviation units. For the statistic w the standardized mode is

$$W_x = \frac{x - \mu_x}{\sigma_x} \quad (2.1)$$

If x follows the normal distribution, the W_x follows the standard normal distribution which gives a constant value of L since standard deviation is absorbed in the standardized statistic.

3. The CUSUM and EWMA Control Charts

3.1. The Cumulative Sum (CUSUM) Control Chart

The example of figure 2.4 demonstrated in the previous chapter, is characteristic of the drawbacks a typical Shewhart's control chart may have. This type of control charts could be inefficient for small shifts of the process. Additionally, in its implementation only the last observation of the process is being used, ignoring in reality the whole process. These deficiencies make Shewhart's control chart less useful on Phase II control when small shifts occur.

So, although Shewhart's control chart is a convenient and easy to implement method and is usually preferred for a statistical process control for large shifts, when the process itself displays small shifts the method could be proved rather inefficient. This is where other alternative type of control charts can be used. Two very good choices are the Cumulative Sum (CUSUM) control chart and the Exponentially Weighted Moving Average (EWMA) control chart.

Let $\{X_i\}, i = 1, 2, \dots, n$ be a process following a normal distribution with known in control mean μ_0 and standard deviation σ_0 . Assuming subsequently that $\{X_{ij}\}, i = 1, 2, \dots, n$ and $j = 1, 2, \dots, k$ is the i^{th} sample group of the process in the $1 \leq i \leq n$ time interval. Now let w be the main quality statistic which e.g. will be the estimator of the process mean[1]

$$w = \psi\{X_1, X_2, \dots, X_k\}$$

where ψ a function. A very common function ψ is the mean value $x_i = \{\overline{X_{ij}}\}$. For the CUSUM control chart the aggregate distance of each sample mean value from the process mean value C_i for $i = 1 \dots, n$ is plotted where

$$C_i = \sum_{j=1}^i (x_j - \mu_0) \tag{3.1}$$

The C_i is the *cumulative sum* of the sample and in practice is the sum of the preceding x_i distance from the in control mean μ_0 . So the (3.1) becomes

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

When a process is in control, the cumulative sum fluctuates around zero, while a different value indicates an out of control process. A typical CUSUM control chart has the following form.[9]

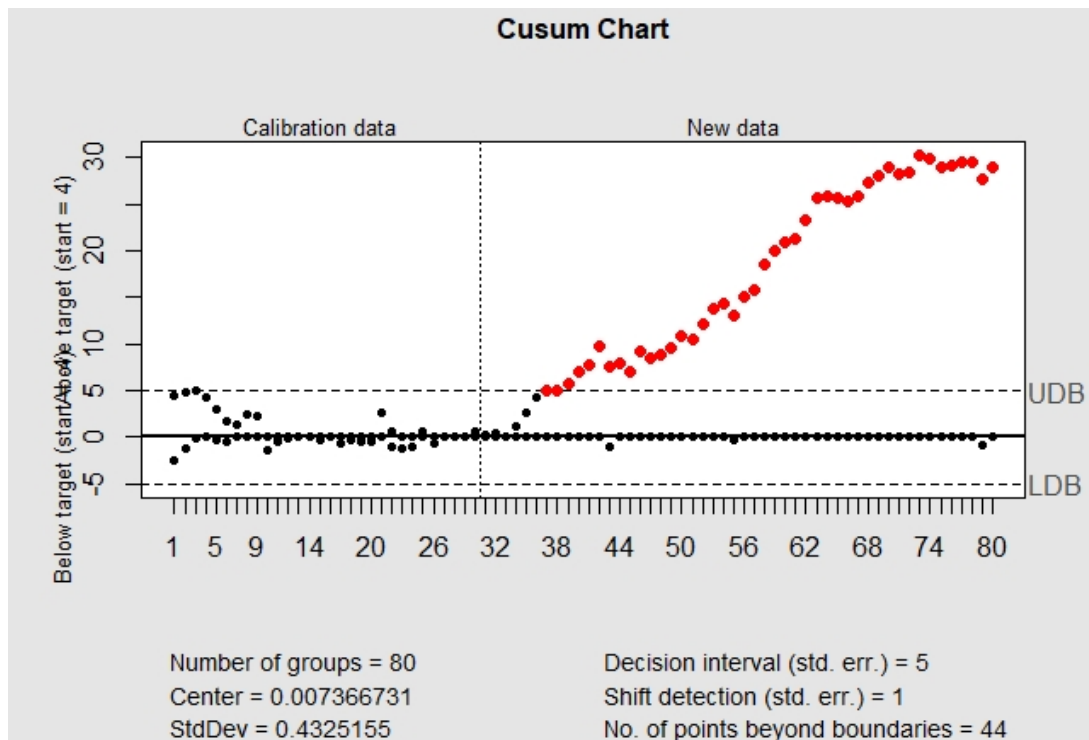


Figure 3.1.: Cusum Control Chart for out of control process

The chart depicts the quantities $x_i - \mu_s$, $s = 0, 1$ using the data used for depicting the Shewharth control chart of the figure 2.4. The first 30 observations therefore have mean value $\mu_0 = 0$ and its C_i values are close to 0, so there seems to be in control. For the following 50 points the mean for the sample shifted to $\mu_1 = 0.4$ giving an out of control process with $C_i > 0$. The two processes plotted in 3.1 represent the negative and positive deviation of the in control CUSUM. There are two methods for representing CUSUM control charts. The *tabular* or *algorithmic method* and the *v-mask method*. The last throughout the years it was proved to be less appropriate so in practice it is not used.

Let x_i be the i^{th} observation of the distribution as defined above. The value μ_0 is often called *target value*[7]. The CUSUM control chart monitors and signals the shift from this target value. The CUSUM chart may have three forms. The *Upper One-Sided CUSUM*, the *Lower One-Sided CUSUM* and the *Two-Sided CUSUM*.

The Upper One-sided CUSUM identifies the quantities C_i that are greater than the statistic

$$C_i^+ = \max[0, x_i - \mu_0 - K + C_{i-1}^+] \text{ for } i = 1, 2, 3, \dots \quad (3.2)$$

where the starting value of $C_0^+ = 0$ and K is called *reference value* (or *slack value* or *allowance*). The parameter K is usual of the form $K = k\sigma_0$ and according to Montgomery [7] a convenient form for K would be

$$K = \frac{\delta}{2} \cdot \sigma_0 = \frac{|\mu_1 - \mu_0|}{2}$$

The test signals when the first sample point $C_i^+ > H$, where H the *decision interval*. For the Lower one-sided CUSUM control chart the statistic is given by the formula

$$C_i^- = \max[0, (\mu_0 - K) - x_i + C_{i-1}^-] \quad (3.3)$$

or by

$$C_i^{-*} = \min[0, x_i - \mu_0 + K + C_{i-1}^{-*}] \quad (3.4)$$

with initial values $C_0^- = 0$ and $C_0^{-*} = 0$. The test then signals when the first $C_i^- > H$ if (3.3) and when $C_i^{-*} < -H$ if (3.4). The two-sided CUSUM signals the sample element which either $C_i^+ > H$ or $C_i^- > H$ if (3.3) and when $C_i^{-*} < -H$ if (3.4). Which ever of these types of CUSUM control chart is used, when there is a signal, an assignable cause of variation is present, so there is a shift in the mean of the sample. This new mean is estimated by the formula

$$\hat{\mu} = \begin{cases} \mu_0 + K + \frac{C_i^+}{N^+}, & \text{if } C_i^+ > H \\ \mu_0 - K - \frac{C_i^{-*}}{N^-}, & \text{if } C_i^- > H \end{cases}$$

The counters N^+ and N^- compute the time interval until the first signal for statistics C_i^+ , C_i^- or C_i^{-*} .

Specifying K and H

In the designing of the CUSUM control chart two basic parameters need to be specified. The reference value K and the decision interval H . Regarding the allowance K , as mentioned before it is a product of the process standard variation σ_0 with a constant k . ($K = k\sigma_0$). This is a choice based on the shift of the mean. For example for $\frac{1}{2}\sigma_0$, the $k = \frac{1}{2}$ and the $K = \frac{1}{2}\delta$, where δ is the absolute difference between the mean of the sample and the fixed process mean. Equivalently, H can be written as product of σ with a constant h . In order the cusum chart to be appropriate for the statistical control it is essential that the ARL_{IC} is as large as possible. So an optimal combination of K and H that could obtain this is necessary. Since the values of k can be predefined, the proper h are to be specified to reach this desired ARL. For example, if ARL needs to be 370 (3-sigma Shewhart control chart ARL), then for different values of k in a two-sided CUSUM control chart the combination of k and l is : [6]

	$ARL=370$			
k	0.25	0.50	0.75	1.25
l	8.01	4.77	3.34	1.99

Table 3.1.: Combination of k and l for given ARL

There are numerous ways of calculating ARL in the cusum method. Siegmund's approximations for one-sided cusum control chart is one of them [7]

$$ARL = \frac{\exp(-2\Delta b) + 2\Delta b - 1}{2\Delta^2}$$

where $\Delta = \delta^* - k \neq 0$ for C_i^+ and $\Delta = -\delta^* - k \neq 0$ for C_i^- , $b = h + 1.116$ and $\delta^* = \frac{\mu_1 - \mu_0}{\sigma_0}$. For two-sided CuSum, the ARL can be computed by the formula

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

3.2. The Exponentially Weighted Moving Average (EWMA) Control Chart

Besides the CUSUM control chart in 1959 Roberts introduced Exponentially Weighted Moving Average (*EWMA*) Control Chart [7]. The EWMA method is gaining ground in statistical process analysis because it is remarkable effective both in detecting small shifts and also for securing an effective Phase II process analysis. The EWMA control chart is using the sample's exponential weighted moving average as a reference statistic which is calculated gradually on the previous observation. The EWMA control chart can be used also on non normal distributed processes. This thesis is going to deal with the EWMA control chart in detail. Firstly, there will be an introduction of the design of the EWMA control chart when the process distribution is known. Afterwards the thesis will focus on unknown process distribution which will be estimated and then will define and deal with the impact of the estimation errors arising.

4. The EWMA Control Chart for known process distribution

Let $\{X_{ij}\}$, $i = 1, \dots, n$ and $j = 1, \dots, k$ a set of n samples from a process with a known distribution which have mean μ_0 and standard deviation σ_0 . From each sampling k elements are derived. Let now $\{x_i\}$, $i = 1, \dots, n$ be the mean of each sampling $x_i = \overline{X_{ij}}$. The Exponentially Weighted Moving Average core statistic is then defined as

$$z_i = \lambda x_i + (1 - \lambda)z_{i-1}, \text{ for } i = 1, 2, \dots, n \quad (4.1)$$

where the smoothing parameter $0 < \lambda \leq 1$ is a constant and which satisfies the initial condition $z_0 = \mu_0$. Using the (4.1) and substituting z_{i-1} with $\lambda x_{i-1} + (1 - \lambda)z_{i-2}$

$$\begin{aligned} z_i &= \lambda x_i + (1 - \lambda)[\lambda x_{i-1} + (1 - \lambda)z_{i-2}] \\ &= \lambda x_i + \lambda(1 - \lambda)x_{i-1} + (1 - \lambda)^2 z_{i-2} \\ &= \lambda x_i + \lambda(1 - \lambda)x_{i-1} + (1 - \lambda)^2 [\lambda x_{i-2} + (1 - \lambda)z_{i-3}] \\ &= \lambda x_i + \lambda(1 - \lambda)x_{i-1} + \lambda(1 - \lambda)^2 x_{i-2} + (1 - \lambda)^3 z_{i-3} \end{aligned}$$

This eventually gives

$$z_i = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j x_{i-j} + (1 - \lambda)^i z_0 \quad (4.2)$$

By the equation (4.2) is derived that the statistic z_i is a weighted average of the previous observations. Additionally, the weights $\lambda(1 - \lambda)^j$ decrease geometrically towards zero.

Recall that

$$\sum_{j=0}^{i-1} \lambda(1 - \lambda)^j = \lambda \sum_{j=0}^{i-1} (1 - \lambda)^j = \lambda \frac{1 - (1 - \lambda)^i}{1 - (1 - \lambda)} = 1 - (1 - \lambda)^i \quad (4.3)$$

The expected value of the statistic z_i is

$$\begin{aligned} E(z_i) &= E\left[\sum_{j=0}^{i-1} (\lambda(1 - \lambda)^j x_{i-j}) + (1 - \lambda)^i z_0\right] \\ &= \sum_{j=0}^{i-1} \lambda(1 - \lambda)^j E(x_{i-j}) + (1 - \lambda)^i E(z_0) \\ &= \sum_{j=0}^{i-1} \lambda(1 - \lambda)^j \mu_0 + (1 - \lambda)^i \mu_0 \\ &= (1 - (1 - \lambda)^i) \mu_0 + (1 - \lambda)^i \mu_0 = \mu_0 \end{aligned}$$

Its variance σ^2

$$\begin{aligned}
Var(z_i) &= Var\left[\sum_{j=0}^{i-1} (\lambda(1-\lambda))^j x_{i-j} + (1-\lambda)^i z_0\right] \\
&= \lambda^2 \sum_{j=0}^{i-1} (1-\lambda)^{2j} Var(x_{i-j}) + (1-\lambda)^{2i} Var(z_0) \\
&= \lambda^2 \frac{1 - (1-\lambda)^{2i}}{1 - (1-\lambda)^2} \sigma_0^2 \\
&= \lambda^2 \frac{1 - (1-\lambda)^{2i}}{2\lambda - \lambda^2} \sigma_0^2 \\
&= \lambda \frac{1 - (1-\lambda)^{2i}}{2 - \lambda} \sigma_0^2 \\
&= \sigma_0^2 \left(\frac{\lambda}{2 - \lambda}\right) [1 - (1-\lambda)^{2i}], \text{ for } i = 1, 2, \dots, n
\end{aligned}$$

Respectively to Shewhart's and the CuSum control chart, EWMA control chart is the plot of the statistics z_i against time. The control limits of the method are

$$\begin{aligned}
UCL &= \mu_0 + L\sigma_0 \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1-\lambda)^{2i}]} \\
CL &= \mu_0 \\
LCL &= \mu_0 - L\sigma_0 \sqrt{\frac{\lambda}{2 - \lambda} [1 - (1-\lambda)^{2i}]}
\end{aligned}$$

The constant L is a parameter of the width of the control limits (threshold). When the smoothing parameter equals to 1, the EWMA control chart reduces to Shewhart's control chart since $(LCL, CL, UCL) = (\mu_0 - L\sigma_0, \mu_0, \mu_0 + L\sigma_0)$. For big values of i the weight $(1 - (1 - \lambda))^{2i}$ approaches to unity so steady state for the EWMA control limits is

$$\begin{aligned}
UCL &= \mu_0 + L\sigma_0 \sqrt{\frac{\lambda}{2 - \lambda}} \\
LCL &= \mu_0 - L\sigma_0 \sqrt{\frac{\lambda}{2 - \lambda}}
\end{aligned} \tag{4.4}$$

In practice the control limits are often set to have width equal to L times $\sigma_0 \sqrt{\frac{\lambda}{2-\lambda}}$. For a two-sided EWMA control chart an indication of an out of control chart is when an observation exceeds one of the two control limits. For the one-sided version of this control chart the charting statistics are as follows

$$\begin{aligned}
z_i^+ &= \max[\mu_0, \lambda x_i + (1-\lambda)z_{i-1}] \\
z_i^- &= \min[\mu_0, \lambda x_i + (1-\lambda)z_{i-1}]
\end{aligned} \tag{4.5}$$

with starting value $z_0 = \mu_0$. The parameters L and λ are the core design parameters for the EWMA control chart. The specification of the values of

both L and λ consists the essence for a reliable control chart. The procedure which is normally used is to first choose a proper λ to the shift magnitude desired to be spotted. The parameter λ in practice determines how much the weight will be between old and new observations. The λ depends on the length of memory. When λ is close to zero needs longer memory for smaller shift while a greater λ signals bigger shifts using shorter memory. After the choice of λ a convenient for the process ARL is picked and afterwards L is calculated on λ and ARL . The following table demonstrates a combination of λ s and L s that give ARL to be equal to 100, 370 and 450 respectively [3].

	<i>ARL=100</i>				<i>ARL=370</i>				<i>ARL=450</i>			
λ	0.01	0.05	0.20	0.40	0.01	0.05	0.20	0.40	0.01	0.05	0.20	0.40
L	1.15	1.88	2.36	2.50	1.82	2.49	2.86	2.96	1.92	2.57	2.93	3.02

Table 4.1.: Combination of λ and L for various values of ARL for IC process

It is common to use standardized data

$$y_i = \frac{x_i - \mu_0}{\sigma_0/\sqrt{n}} \quad (4.6)$$

(2.1)which gives the EWMA control chart statistic

$$Z_i = \lambda y_i + (1 - \lambda)Z_{i-1} \quad (4.7)$$

and if x_i is normally distributed, Z_i follows the standard normal distribution giving constant control limits

$$\begin{aligned} UCL &= L\sqrt{\frac{\lambda}{2-\lambda}} \\ CL &= 0 \\ LCL &= -L\sqrt{\frac{\lambda}{2-\lambda}} \end{aligned}$$

From this point unless it is stated otherwise the standardized statistic will be used.

4.1. An example of EWMA control chart for known process distribution

In this section there will a presentation of an example on the EWMA statistical analysis control chart. In this example a sample of 80 points of size 5, $\{X_{i,j}\}$, $i = 1, \dots, 80$, $j = 1, \dots, 5$ from a normal distribution is simulated. Then their mean $x_i = \{\overline{X_{i,j}}\}$ is derived and the standardized EWMA statistic of this mean

$$W_i = \frac{x_i - \mu_0}{\sigma_0/\sqrt{k}}$$

is computed, where μ_0 and σ_0 are the reference mean and standard deviation for an in control process. For an illustration of the method the 30 first points generated from the standard normal distribution use the assumption that the sample simulates an in control process and then the next 50 points are derived from a normal distribution with mean 0.4 and standard deviation 1. In this way we demonstrate an out of control process caused by a small shift in the mean. For the example we chose an ARL equal to the 3-sigma Shewhart's control chart ($ARL = 370$) and a smoothing parameter $\lambda = 0.05$. The EWMA chart we eventually get is the following.

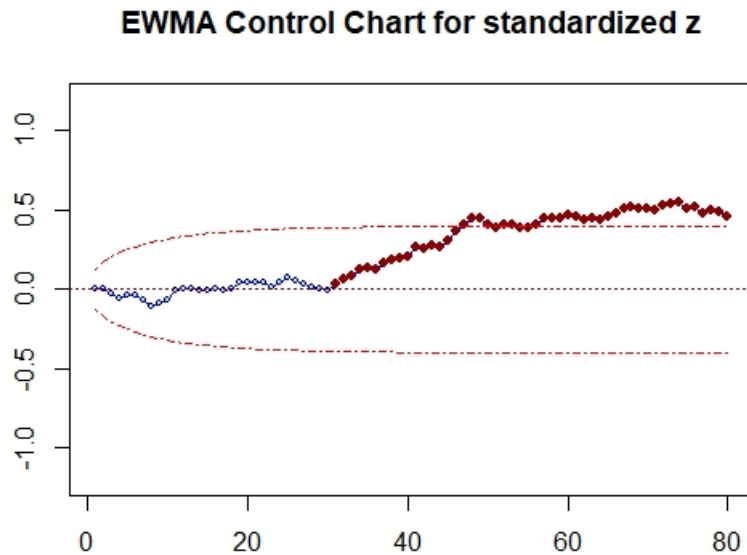


Figure 4.1.: EWMA control chart for standardized data. The first 30 first observations are generated from the standard normal distribution and the next 50 from a normal distribution with mean 0.4 ($\delta = 0.4$) and standard deviation 1

From the figure 4.1 it is very clear that the first 30 observations are in control since all of them lie around the $CL = 0$ central line. When the shift in the mean occurs, the exponentially weighted moving average for the standardized data increase and after a while the control statistic exceeds the UCL control limit confirming the fact that the process is out of control. This example actually performs a process control of the same parameters as the Shewhart's control chart 2.4 of chapter 2. Indeed for such a small shift EWMA control chart is successful in spotting it in contrast to the Shewhart control chart.

5. The EWMA Control Chart for unknown distribution

The EWMA control chart that was presented in the previous chapter could be seen as a theoretical basis for the statistical process control using the exponentially weighted moving average. In reality the cases that the distribution of a process is known are really rare to find. In the majority of applications of a statistical control chart the real mean value μ_0 and standard deviation σ_0 are unknown. In such cases the only option is to estimate the distribution parameters. However this might have a large impact on the performance of the chart.

5.1. EWMA Control Chart with estimated parameters

Let as in the previous chapter assume a k element sample $\{X_{ij}\}$, $i = 1, \dots, n$ and $j = 1, \dots, k$ of size n and $\{x_i\} = \{\bar{X}_{ij}\}$ the set of their mean value. If the sample follows a normal distribution with a mean value μ_0 and a standard deviation σ_0 , the EWMA standardized data would be

$$y_i = \frac{x_i - \mu_0}{\sigma_0/\sqrt{k}} \quad (5.1)$$

the control chart statistic

$$z_i = \lambda y_i + (1 - \lambda)z_{i-1}$$

and the control limits

$$(LCL, CL, UCL) = (-L\sqrt{\frac{\lambda}{2-\lambda}}, 0, L\sqrt{\frac{\lambda}{2-\lambda}})$$

But when the process' distribution is unknown, in the equation (5.1) the estimated values from the Phase I data of μ_0 and σ_0 should be used. Thus the (5.1) takes the form

$$\hat{y}_i = \frac{x_i - \hat{\mu}}{\hat{\sigma}/\sqrt{k}} \quad (5.2)$$

or [8]

$$\hat{y}_i = \frac{1}{Q}(v_i + \gamma - \frac{Y}{\sqrt{n}}) \quad (5.3)$$

where $Q = \hat{\sigma}/\sigma_0$ is the ratio of the estimated standard deviation to the real in control standard deviation, $v_i = \sqrt{k}(x_i - (\mu_0 + \delta))/\sigma_0$ is the standardized sample mean for a δ shift in the mean, $\gamma = \sqrt{k}\delta/\sigma$ is the standardized mean

shift and $Y = \sqrt{mk}(\hat{\mu} - \mu_0)/\sigma_0$ is the standardized difference between the estimated and the real in control mean. A good estimator of the process mean is

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \quad (5.4)$$

Regarding the standard deviation there are a number of estimators. Two of them are [8]

$$\begin{aligned} \hat{\sigma}_1 &= \frac{S_{pooled}}{c_4(v+1)}, \\ \hat{\sigma}_2 &= c_4(v+1)S_{pooled} \end{aligned} \quad (5.5)$$

where $S_{pooled} = \sqrt{(\sum_{i=1}^n S_i^2)/n}$, with $S_i^2 = \frac{\sum_{j=1}^k (X_{ij} - x_i)^2}{k-1}$, $v = n(k-1)$ and $c_4(v+1)$ is a control constant which is given by the equation

$$c_4(v+1) = \frac{\sqrt{2}\Gamma(\frac{n(k-1)+1}{2})}{\sqrt{n(k-1)}\Gamma(\frac{n(k-1)}{2})}$$

and fulfills the condition $E[S_{pooled}] = c_4(v+1)\sigma_0$. In [7] there is a table with values $c_4(v+1)$. Their main differences are that $\hat{\sigma}_1$ estimator is unbiased in contrast with the $\hat{\sigma}_2$, while the latter has the smallest mean squared error. [8]

Impact of estimation error on in control ARL

The estimation of the process distribution apparently affects the control chart performance influencing the value of *ARL*. For an IC process, a $\hat{\mu} > \mu_0$ reduces the value of the statistic \hat{y}_i . That gives shorter *ARL*. In parallel, the underestimated mean $\hat{\mu} < \mu_0$ increases also the statistic and returns smaller RLs and shorter *ARL*. Additionally, and underestimation of $\hat{\sigma} < \sigma_0$ increases the \hat{y}_i and returns shorter *ARLs*, while on the other hand, the $\hat{\sigma} > \sigma_0$ returns smaller \hat{y}_i and gives longer *ARL*.

5.1.1. An example of EWMA control chart with estimated parameters

In order to demonstrate a simple EWMA control chart with estimated parameters a simulation of a sample of $N = 100$ of single observations ($k = 1$) from the standard normal distribution $N(\mu_0 = 0, \sigma_0^2 = 1)$ was performed in order to represent an IC process. The $ARL_0 = 370$ and the smoothing parameter $\lambda = 0.2$. The estimation of the parameters were made on a sample of size $n_{est} = 100$ from a standard normal distribution and for the standard deviation estimator the standard deviation of the sample was used. The plot depicts the standardized statistic with estimated parameters $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

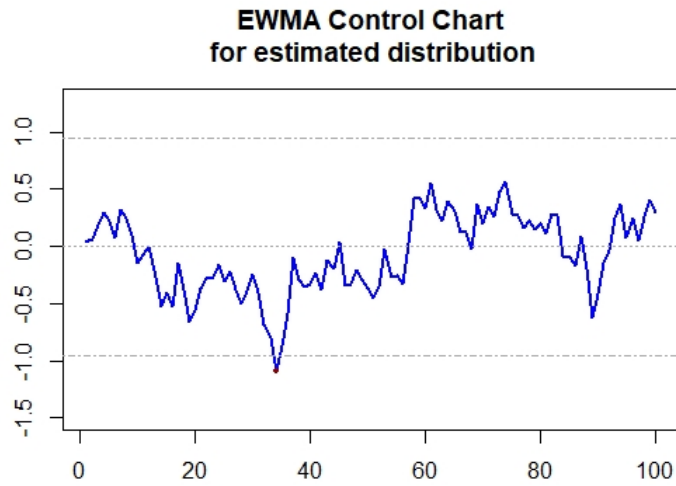


Figure 5.1.: EWMA control chart for an IC estimated distribution for a X_i , $i = 1, \dots, 100$ sample from standard normal distribution and parameters $ARL_0 = 370$ and $\lambda = 0.2$. The estimated $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

5.2. Control chart's ARL for estimated parameters

In the previous example the EWMA chart with estimated parameters although the process is in control, seems to spot a shift in the mean. If nevertheless two EWMA control charts for the same process but with the first to standardized the EWMA statistic on estimated distribution while the second one uses the known process distribution for standardization are compared, the derived combined plot of the figure 5.2 of both EWMA control charts with the blue line to depict the estimated chart and the purple one the process control based on known process distribution is obtained.

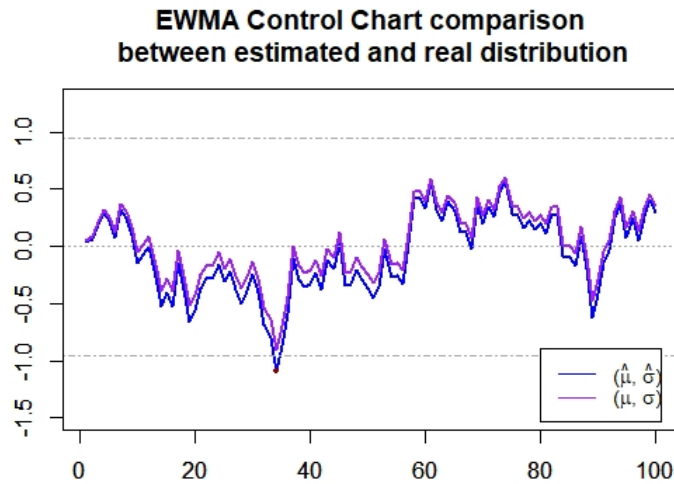


Figure 5.2.: Comparison of EWMA control chart for an IC estimated distribution for a X_i , $i = 1, \dots, 100$ sample from standard normal distribution and parameters $ARL_0 = 370$ and $\lambda = 0.2$. The estimated $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

In figure 5.2 the control chart statistic for the estimated parameters as mentioned exceeds the lower control limit in a point, while the one with the true parameters lies inside the control limits for the whole process. This leads to the assumption that there exists an estimation error on the run length of the control chart generated by the unknown distribution of Phase I in the process analysis. In this example increases the Type I error of the control chart.

Estimation error on the RL distribution

In order to start studying the estimation error for the EWMA control chart the boxplot in the figure 5.3 is presented in which there is a demonstration (not showing the extreme values) of the simulated RLs for a process. The simulation size is $N_{RL} = 2500$, the IC $ARL_0 = 200$, the smoothing parameter $\lambda = 0.2$ and the estimation of the distribution was made on a $n_{est} = 100$ observations of the standard normal distribution. In this chart it can be seen that when the EWMA control chart is designed on the real distribution and for an IC process the statistical control gives RL distribution with median equal to 135 and mean value 196.2. A small deviation of the $ARL_0 = 200$ which can be justified by the simulation error. On the other hand, when the distribution is being estimated with parameters $\hat{\mu} = 0.1205$ and $\hat{\sigma} = 0.9499$, the control chart's mean RL decreases to 111.14 and median to 79, adding therefore an estimation error to the simulation error, increasing here the Type I error.

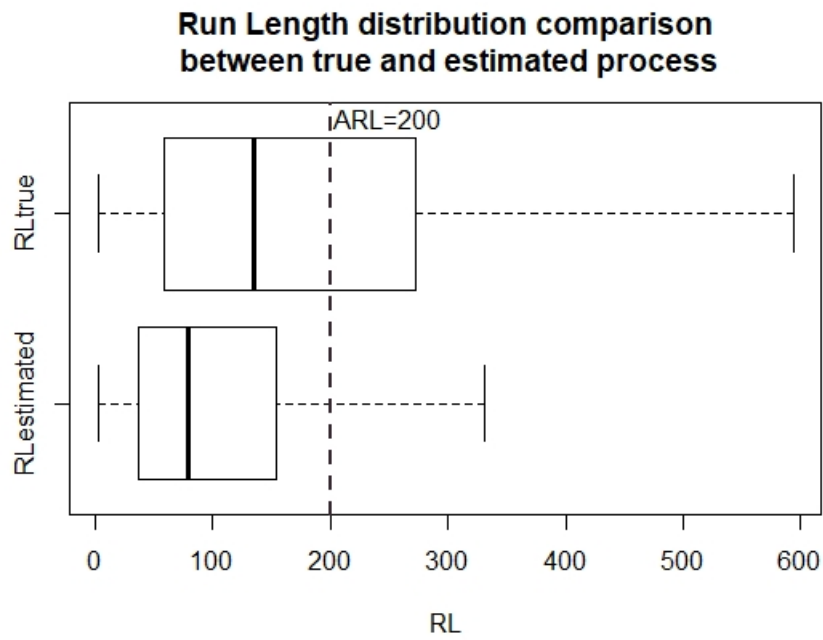


Figure 5.3.: Comparison of IC RL distributions for an EWMA statistical control process both true and estimated, for a $X_i, i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200, \lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = 0.1205$ and $\hat{\sigma} = 0.9499$ were estimated on a $n_{est} = 100$ sample of standard normal distribution.

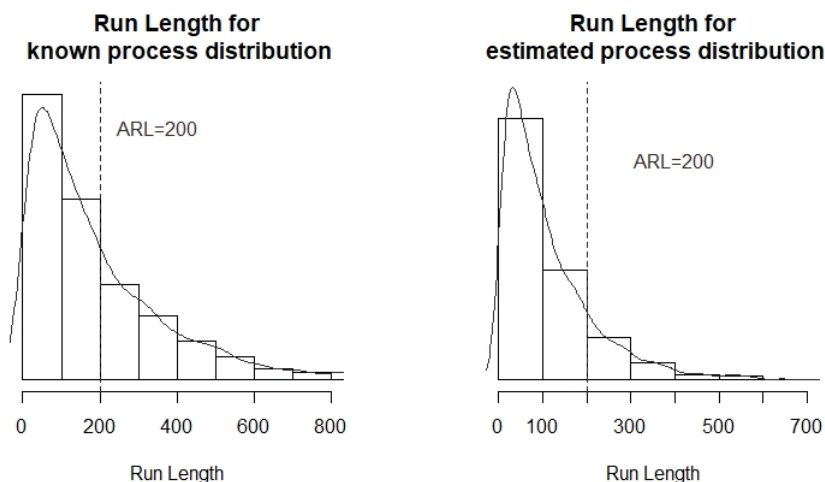


Figure 5.4.: Comparison histogram of IC RL distributions for an EWMA statistical control process both true and estimated, for a $X_i, i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200, \lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = -0.0056$ and $\hat{\sigma} = 0.9352$ were estimated on a $n_{est} = 100$ sample of standard normal distribution. The limits of the plot have been adjusted so as not to show the extreme values for both distributions.

The histogram in figure 5.4 confirms the trend of error in the Run Length of the EWMA control chart when mean and variance are being estimated. The nature of the error depends on the estimation itself. In the previous estimation for example, where $\hat{\sigma} < \sigma_0$ and $\hat{\mu} > \mu_0$, the distribution is steeper, but less skewed and less long tailed with the majority of RL values to be concentrated around $RL = 80$ considerable smaller than the desired $ARL_0 = 200$ (Type I error). For another estimation as presented in figure 5.5 with $\hat{\mu} > \mu_0$ but $\hat{\sigma} > \sigma_0$, the EWMA control chart generates bigger RLs while it has a steeper, more skewed and long tailed distribution.

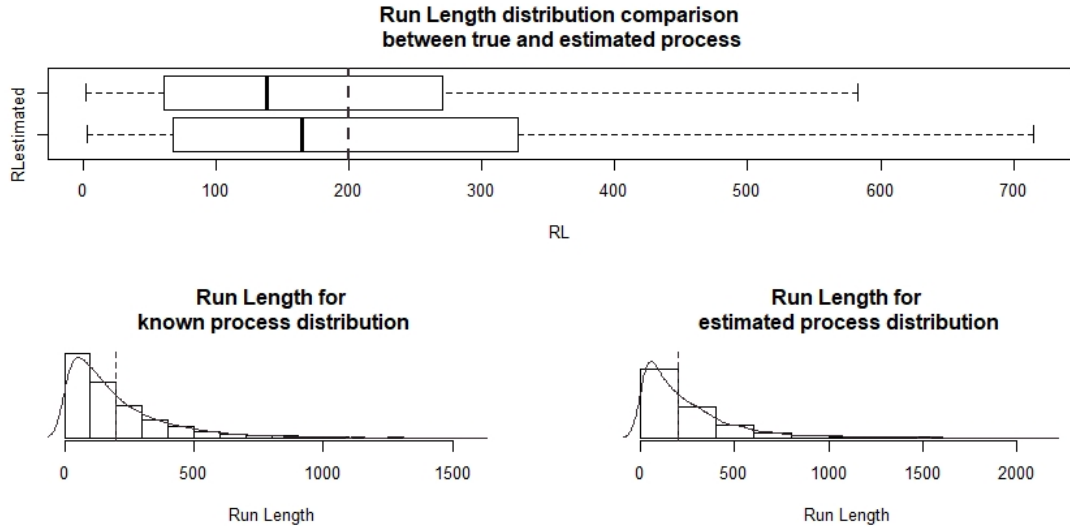


Figure 5.5.: Comparison of IC RL distributions for an EWMA statistical control process both true and estimated, for a X_i , $i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200$, $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = 0.0354$ and $\hat{\sigma} = 1.0272$ were estimated on a $n_{est} = 100$ sample of standard normal distribution.

The study so far indicates that the estimation of a process causes an error on the statistical process control. In order to examine further this assumption, EWMA control charts for an In Control process are simulated altering a number of parameters. For this purpose an R-code has been designed which firstly estimates the distribution parameters $\hat{\mu}$ and $\hat{\sigma}$ in a $n_{est} = 100$ set of standard normal distribution observation set. Then, the code generates a sample of $N_{RL} = 2500$ from standard normal distribution (to simulate an IC process) and calculates the standardized statistic z_i , $i = 1, \dots, nrl$. It examines when the process exceeds the control limits and returns the smaller value. It repeats the procedure for $N_{ARL} = 5000$ and take the mean value of RLs to return the ARL . The code finally, simulates $narl = 500$ repeating the above loop. The standard parameters used are $\lambda = 0.2$, and $ARL_0 = 200$. To start a plot depicting a simple ARL distribution is presented.

Estimation error on the ARL distribution

The estimation of the process distribution, as seen so far, shifts the run length distribution from the desired ARL . This inevitably leads to the assumption that the process estimation generates error on the ARL distribution.

It is interesting therefore to compare the EWMA control chart for the same sample's statistic z computed on real and estimated distribution. Then the statistical error derived by the estimation can more clearly be identified and distinguished it from the less significant simulation error. According to the figure 5.6 when the control chart is designed on an estimated distribution,

it has a range of 953 with a median of 161.5, while the true EWMA statistic for IC process has a 35 range (simulation error) and a median equal to 200). This is in total agreement with the theory presented in paragraph 5.1. The estimation error of the mean $\hat{\mu}$, gives too short ARL , while on the other hand, the estimation error of the standard variation $\hat{\sigma}$ works both ways causing the big range and the long tails. But the combination of the estimation of the two parameters, return more shorts $ARLs$.

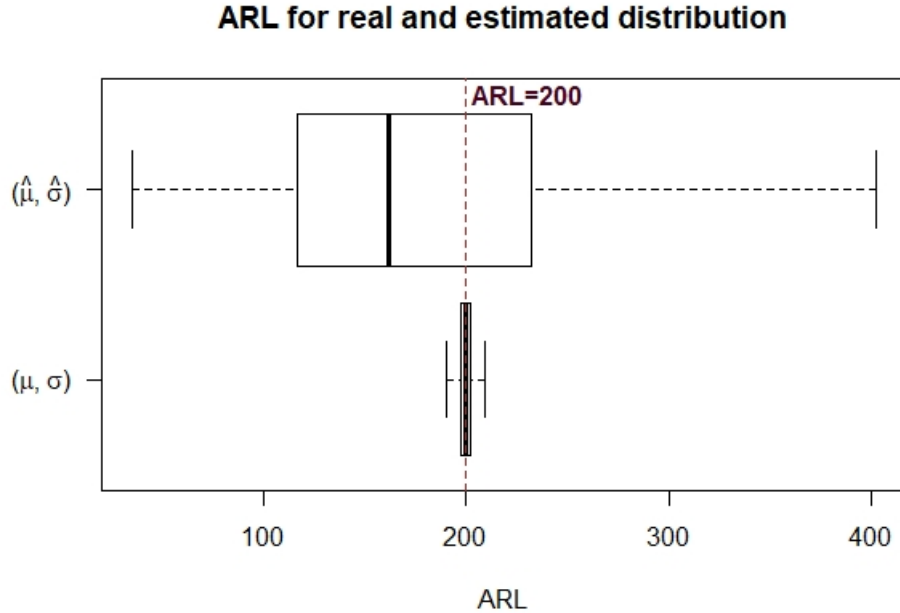


Figure 5.6.: Comparison of ARL distribution for simulated $ARLs$. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

The histogram of both simulated $ARLs$ illustrates that while for the true parameters the ARL is distributed almost normally with mean 200 (again as a result of the simulation), for estimated parameters the ARL distribution is steep, skewed and long tailed. (For presentation reason the extreme values of ARL for estimated process are not shown in the histogram).

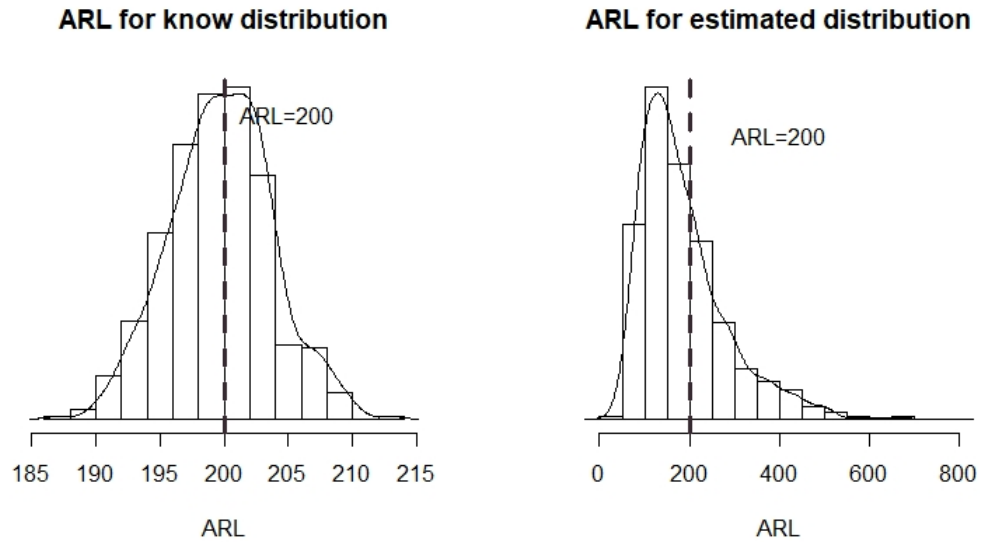


Figure 5.7.: ARL distribution histogram for estimated process distribution. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL equals to $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

In a boxplot for the ARL error distribution for the above EWMA control charts, is clear that the EWMA control chart designed on true distribution has insignificant simulation errors. But if therefore the EWMA control chart is designed on estimated parameters, the corresponding error has a range of 918 with median to be -38.5 , meaning that spots OOC observation in an earlier time (Type I error).

**ARL error distribution
for true and estimated distribution**

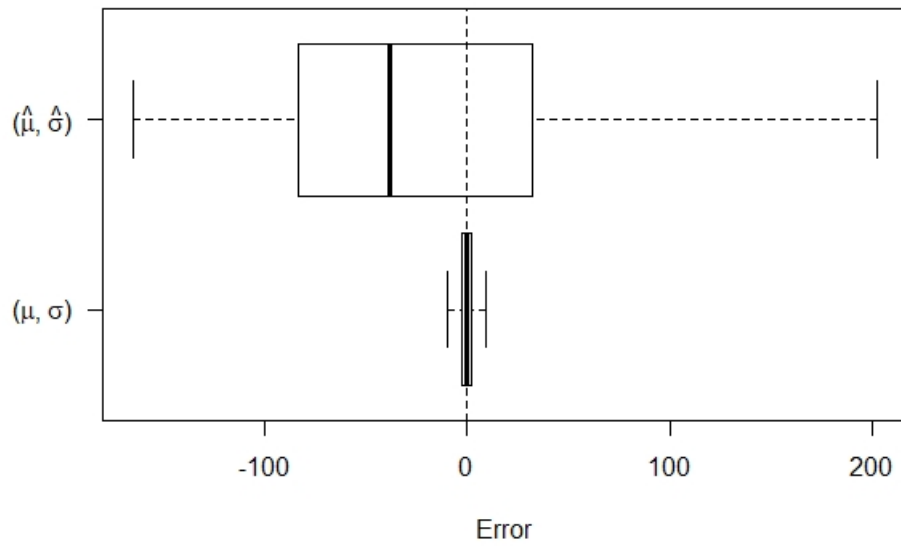


Figure 5.8.: Error distribution for estimated EWMA Control Chart. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

The following histogram confirms that the errors when EWMA control chart is designed on unknown process, are in majority negative while on the other hand have also a big range giving long-tailed and skewed distribution.

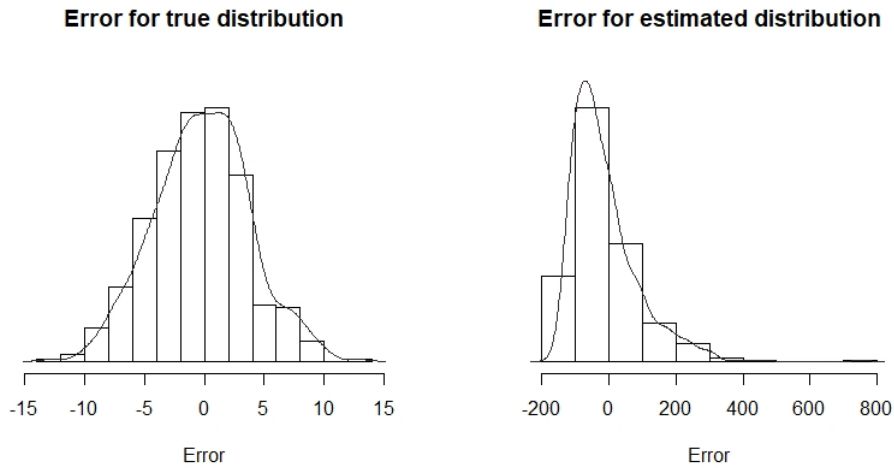


Figure 5.9.: Error distribution histogram for estimated EWMA Control Chart. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

The presence of this error makes it crucial to examine the impact various statistical control parameters have on it. To begin, the effect the desired IC ARL_0 has on the distribution of the estimated ARL will be examined. In the following boxplot therefore, which was derived by a 500 times process simulation using the code 5.2 with $\lambda = 0.2$ but for four different ARL i.e. $ARL_1 = 100$, $ARL_2 = 200$, $ARL_3 = 370$ and $ARL_4 = 500$ (figure 5.10), the median ARL for each case deviates from the IC ARL (table (refrangevariousarl)).

Distribution of ARL for various IC ARL

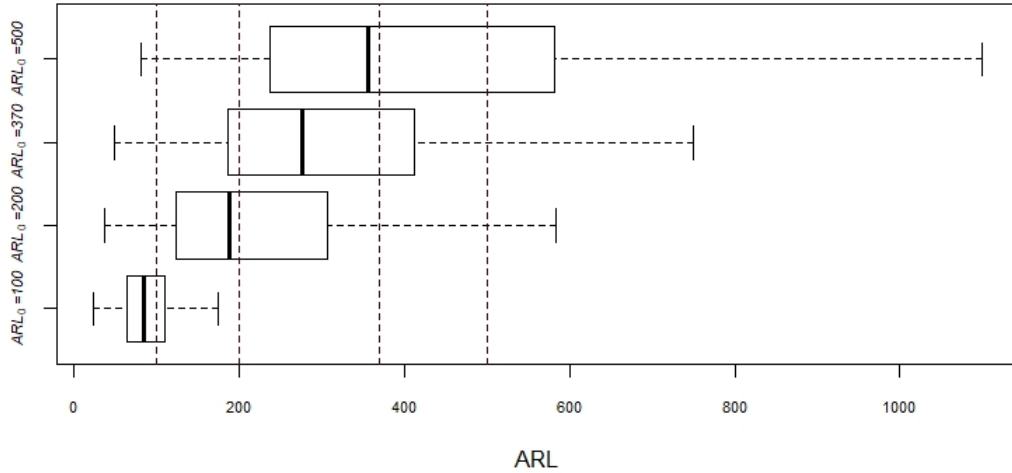


Figure 5.10.: Distribution of ARL for $ARL_0 = 100$, $ARL_0 = 200$, $ARL_0 = 370$ and $ARL_0 = 500$. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The process sample has been derived from standard normal distribution. The simulation parameter $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

ARL_0	100	200	370	500
median	84.00	187.50	277.00	357.00
median- ARL_0	16.00	12.50	93.00	143
range	333.00	1361.00	2082.00	2321

Table 5.1.: Median and range for ARL distribution derived from simulation for $ARL_0 = 100$, $ARL_0 = 200$, $ARL_0 = 370$ and $ARL_0 = 500$

The range of the ARL distribution appears therefore to correlate to the value of the IC ARL_0 . For bigger ARL_0 the distribution median has larger deviation from ARL_0 . Additionally, the range of the simulated ARL s is also affected by the estimation of the process. In the figure 5.10 the smaller IC ARL_0 the smaller this estimation error Table 5.1. In this table are presented the median and its difference ARL_0 of the corresponding distributions as well as the range of the distribution in order to study more precisely the impact the ARL_0 has to ARL 's range. It is now more clear that too large ARL_0 gives a right tailed distribution reducing the Type I error and increasing the effectiveness of the control chart.

So far the effect of simulation size and IC ARL_0 on estimated ARL has been studied. Further on, the estimated error distribution variation for different values of the smoothing parameter λ and then for different sample sizes will be audited. The figure 5.11 depicts the ARL when λ varies.

Distribution of ARL for simulations of various lambda

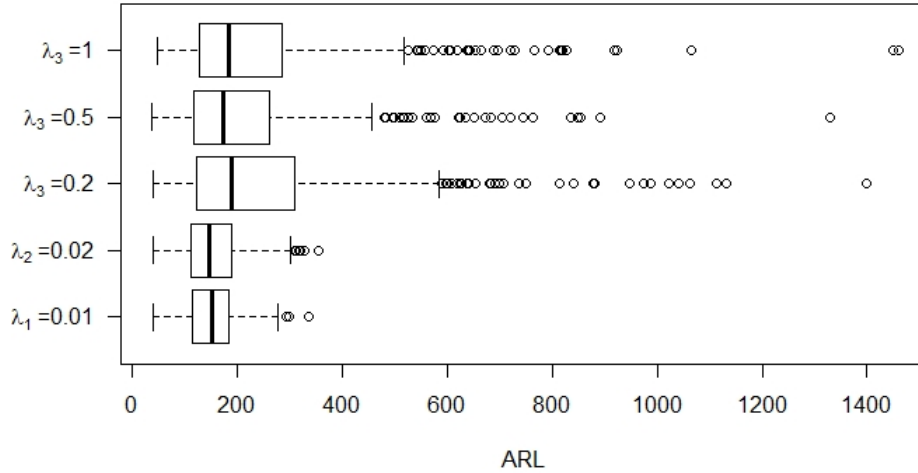


Figure 5.11.: Distribution of ARL for $(\lambda_1 = 0.01, L_1 = 1, 5)$, $(\lambda_2 = 0.02, L_2 = 1.83)$, $(\lambda_2 = 0.2, L_2 = 2.64)$, $(\lambda_2 = 0.5, L_2 = 2.78)$ and $(\lambda_3 = 1, L_3 = 2.81)$. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The process sample was generated by a standard normal distribution with $ARL_0 = 200$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

According to figure 5.11 smaller smoothing parameter has smaller ARL . This means that for small λ the control chart spots OOC observations sooner (Type I error). On the other hand, for bigger λ the ARL distribution seem to give a median more close to the IC ARL_0 , but the range is very big (analogous to the value of λ), while the extreme cases increase both in number and value. This outcome is in absolute line with the study presented in table 1 of the paper of Jones & al. (2001). [5] Finally the ARL distribution for three different samples sizes $n_{est1} = 100$, $n_{est2} = 1000$ and $n_{est3} = 16000$ is :

Distribution of ARL for simulations of various estimation size

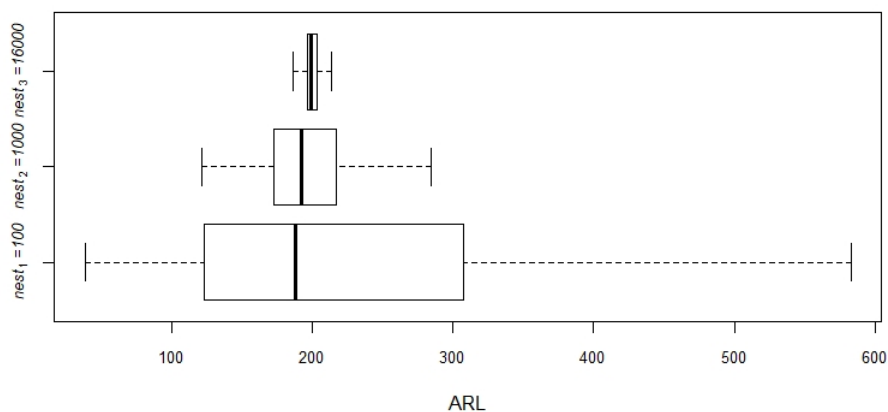


Figure 5.12.: ARL distribution estimation on a 100 process simulation where the process parameters have been estimated on a $n_{est1} = 100$, $n_{est2} = 1000$ and $n_{est3} = 16000$ sample size from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs.

It is clear as seen in figure 5.12 that the sample size defines the effectiveness of the EWMA control chart. Smaller sample size increases the estimation error. It is then obvious that in order to achieve an efficient process control when the process is being estimated, a really big sample size for the estimation of the parameters should be used.

6. Proposal for handling the estimation error in EWMA Control Chart.

As seen so far, it is very rare for a process distribution to be known. So, when trying to apply a statistical control chart it is essential that the process should be estimated using Phase I data. This estimation eventually transforms ARL to a random variable. When the process is estimated, as already seen, the EWMA control chart often tends to give smaller ARL s than the desired ARL_0 . This deviation from the ARL_0 is also affected by a number of other parameters (see section 5.2). This chapter deals with a study on methods for taking into account the estimation error with focus on adjustment of the threshold.

6.1. General setting for the adjustment

The estimation of process parameters, affects the ARL distribution. Wrongly estimated $\hat{\mu}$ (both over or underestimated) will tend to imply too small in control ARL . On the other hand a miss-estimated $\hat{\sigma}$ may increase ARL (when $\hat{\sigma} > \sigma$) or reduce it (when $\hat{\sigma} < \sigma$). That means that the statistic z_i exceeds sooner one of the control limits. By adjusting therefore the control limits the false alarm error can be reduced. The method introduced by Gandy and Kvaløy (2013)[4] suggests the construction of an bootstrapped approximate confidence interval for the control limits. The bootstrap method is a re-sampling method for the estimation of a distribution [8].

For an EWMA control chart, let P be the real in control distribution of the process and \hat{P} an estimation of it. The parameters of the process are the $\xi = (\mu_0, \sigma_0)$ and equivalently $\hat{\xi} = (\hat{\mu}, \hat{\sigma})$ is their estimation. Let q be a function denoting the in control condition the statistical chart holds which depends on P and ξ ($q(P; \xi)$) or their estimates (e.g. $q(P; \hat{\xi})$). For instance, this function q could be the threshold to achieve a certain ARL . Since the process distribution is unknown, the function $q(P; \hat{\xi})$ is random and thus for the bootstrap estimation of the confidence interval the $q(\hat{P}; \hat{\xi})$ is used. Let $\alpha \in (0, 1)$ constant. The subject is to determine an one-sided confidence interval which will guarantees that $(1 - \alpha)\%$ of the $q(\hat{P}; \hat{\xi})$ to be equal or bigger to ARL_0 . So, if p_α is a constant that represents the α percentile of the $q(\hat{P}; \hat{\xi}) - q(P; \hat{\xi})$ (assuming that it actually exists),

$$P(q(\hat{P}; \hat{\xi}) - q(P; \hat{\xi}) > p_\alpha) = 1 - \alpha$$

gives

$$P(q(\hat{P}; \hat{\xi}) - p_\alpha > q(P; \hat{\xi})) = 1 - \alpha$$

. The interval $(-\infty, q(\hat{P}; \hat{\xi}) - p_\alpha)$ is therefore a lower limit for the CI of $q(\hat{P}; \hat{\xi})$.

The constant p_α is unknown, and Gandy and Kvaløy proposed the use of the following bootstrap method to estimate the $(-\infty, q(\hat{P}; \hat{\xi}) - p_\alpha)$.

- First estimate the \hat{P} and the $\hat{\xi}$
- Then take a bootstrap sample for \hat{P} to estimate \hat{P}^* and $\hat{\xi}^*$.
- Repeat the bootstrap B times to get the two sets of estimations $\hat{P}_i^*, i = 1, \dots, B$ and $\hat{\xi}_i^*, i = 1, \dots, B$.
- Then p_α^* is an estimation of p_α which represents the empirical α quantile of $q(\hat{P}_i^*; \hat{\xi}_i^*) - q(\hat{P}; \hat{\xi}), i = 1, \dots, B$.
- The adjusted threshold therefore is $q(\hat{P}; \hat{\xi}) - \hat{p}_\alpha^*$

To demonstrate how the adjustment of the control limits works on an EWMA control chart, follows an application on the data from figure 5.1 in order to obtain an one-sided confidence interval that ensures that 90% of the $ARLs$ will be at greater than or equal to desired ARL_0 . (For all the simulations the threshold is being estimated using the R package *spcadjust* [3].)

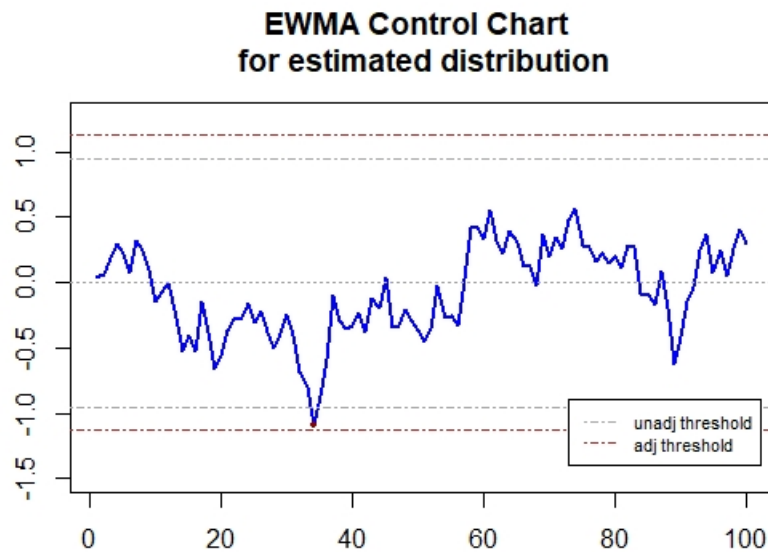


Figure 6.1.: EWMA control chart for an IC estimated distribution for a $X_i, i = 1, \dots, 100$ sample from the standard normal distribution and parameters $ARL = 370$ and $\lambda = 0.2$. Both the adjusted and non-adjusted Control Limits are included. The estimated $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

The adjusted control limits are wider and thus prevent the control chart from spotting falsely an abnormal observation as the regular plot did. Although it follows a further study on this adjustment at first sight it seems

that it has the potentials to reduce the Type I error the process estimation caused. For this purpose there will be a try to replicate some of the Section's 5 plots comparing this time the estimated EWMA control chart for both control limits.

6.1.1. Evaluation of the adjusted control limits

In the figure 5.3 the estimated EWMA control chart gives RLs smaller than the desired $ARL_0 = 200$. In order to increase the effectiveness of the control chart, these small RLs should be eliminated. So, if an adjustment on the threshold was applied, the control limits would be broadened and thus both the alarm time and the RL values would be increased.

Run Length IC distribution comparison when process is estimated for adjusted and unadjusted threshold

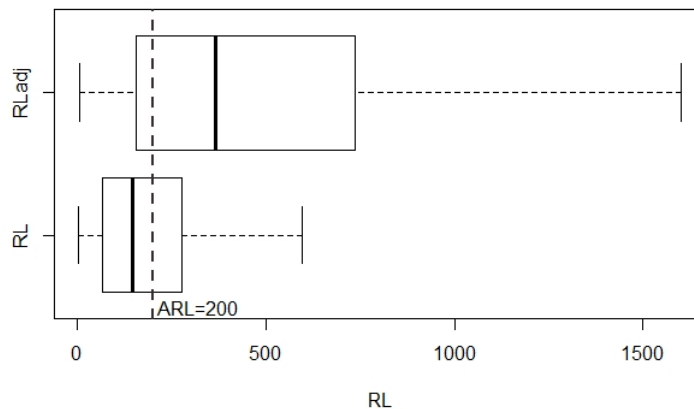


Figure 6.2.: Comparison of IC RL distributions for an EWMA statistical control process for adjusted and non-adjusted CL, for a $X_i, i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200, \lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = -0.0599$ and $\hat{\sigma} = 1.0154$ were estimated on a $n_{est} = 100$ sample of standard normal distribution.

The new control limits allow the EWMA control chart to avoid in control observations to be labeled as out of control. This is what the boxplot 6.2 indicates. That for adjusted threshold, the run length of the control chart returns bigger values e.g in the example the median of the RL is 365.5 for 145 when the unadjusted threshold is used. This may be easily confirmed by the equivalent histogram (figure 6.3) in which for the adjusted threshold the RL distribution is more smooth with bigger predominant RL value.

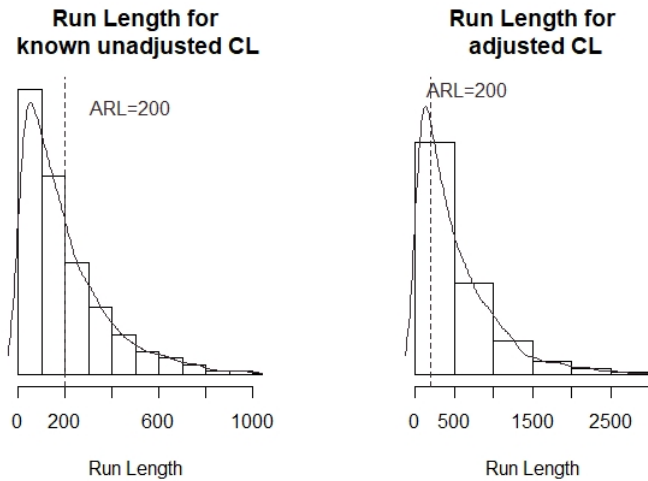


Figure 6.3.: Comparison of IC RL distributions for an EWMA statistical control process both unadjusted and adjusted threshold, for a $X_i, i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200, \lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = -0.0599$ and $\hat{\sigma} = 1.0154$ were estimated on a $n_{est} = 100$ sample of the standard normal distribution.

Indeed when it comes to ARL the adjustment of the threshold gives a skewer distribution which median is 200 compared to non adjusted ARL which have median 163. Additionally, by the figure 6.4 it is confirmed that this threshold guaranteed that 90% of the $ARLs$ are bigger that $ARL_0 = 200$.

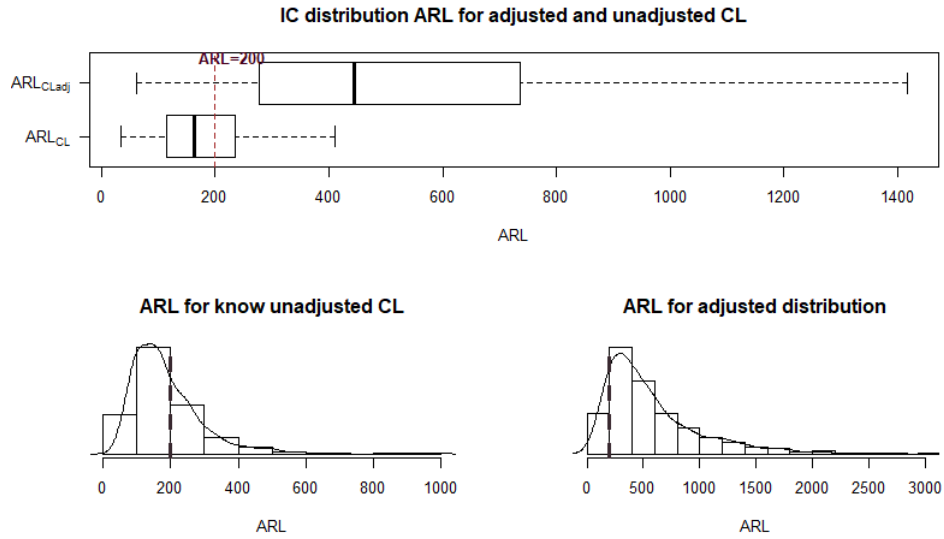


Figure 6.4.: Comparison of ARL distribution for simulated ARL s for adjusted and unadjusted threshold. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The extreme values have been excluded from the plot.

Accordingly to the study of the previous chapter follows a simulation of the process for various ARL_0 (i.e $ARL = 100, 200, 370$ and 500) to conclude through the figure 6.5 that when the CLs are adjusted the probability of a very early false alarm is being remarkable reduced. The new control chart gives bigger ARL s, the distribution range is being increased with the ARL_0 value and the 90% seems to hold.

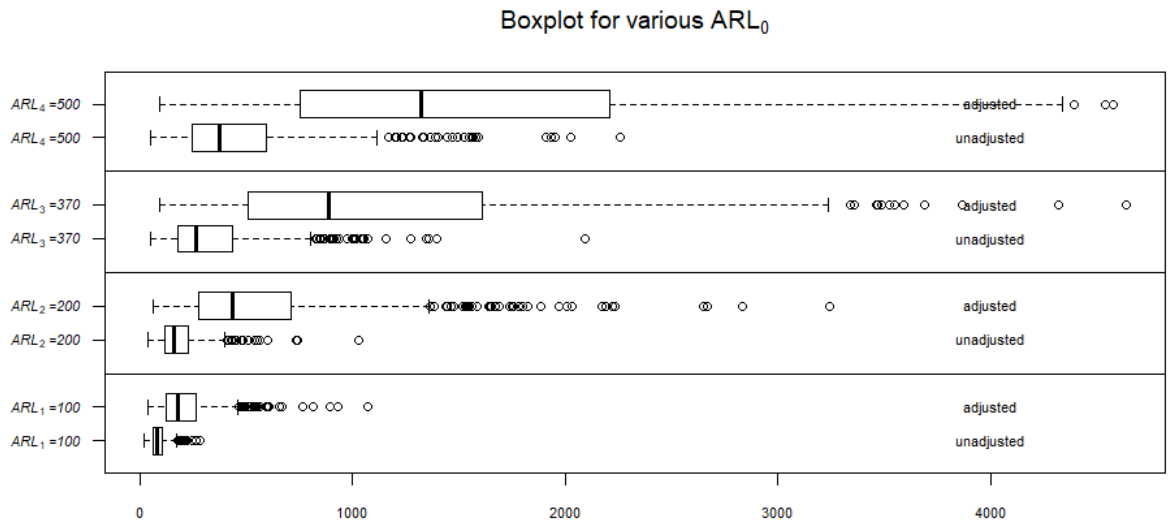


Figure 6.5.: Comparison of ARL distribution for simulated ARL s for adjusted and unadjusted threshold for various ARL_0 (e.g. $ARL_1 = 100$, $ARL_2 = 200$, $ARL_3 = 370$ and $ARL_4 = 500$). The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameter $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$.

Furthermore, the impact of the adjustment of the threshold can be seen also while the estimation size for the process alters. For the reference value of $ARL_0 = 200$, the smaller the n_{est} , the most efficient the adjustment seem to be. But also for the pretty large $n_{est} = 16000$, although the conventional threshold can be efficient the adjusted one is even better as seen in figure 6.6. Actually, the larger the estimation sample the less adjustment is needed.

Distribution of ARL for simulations of various estimation sizes

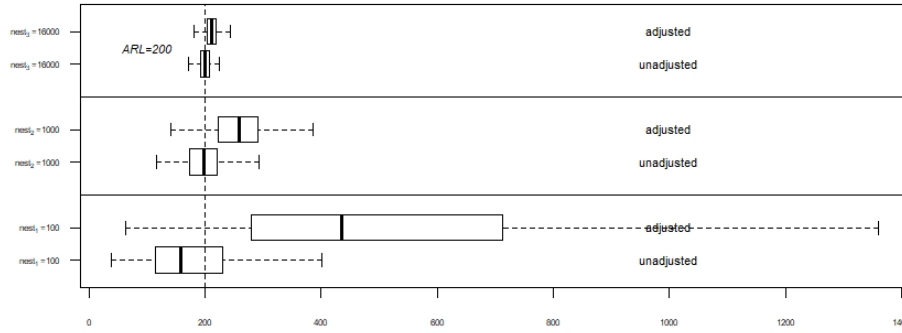


Figure 6.6.: Comparison of ARL distribution for simulated ARL s for adjusted and unadjusted threshold for various sizes for estimation of the parameters $n_{est} = 100$, $n_{est} = 1000$ and $n_{est} = 16000$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The extreme values have been excluded from the plot.

The influence of the smoothing parameter λ has been shown to be important for the EWMA control chart. Regarding the interaction between the smoothing parameter and the adjusted threshold the new control limits seem to give indeed larger ARL s while for smaller λ s the ARL distribution appears more right skewed than for distribution with bigger λ s ($\lambda > 0.1$) (figure 6.7).

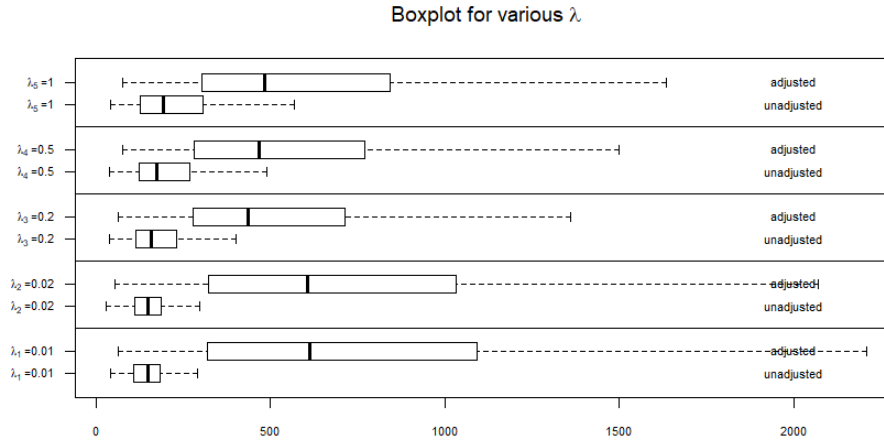


Figure 6.7.: Comparison of ARL distribution for simulated ARL s for adjusted and unadjusted threshold for various λ (e.g. $\lambda_1 = 0.01$, $\lambda_2 = 0.02$, $\lambda_3 = 0.2$, $\lambda_4 = 0.5$ and $\lambda_5 = 1$). The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameter $ARL_0 = 200$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The extreme values have been excluded from the plot.

6.2. Another threshold adjustment

The threshold adjustment proposed by Gandy and Kvaløy (2013)[4], and applied above, ensures a 90% of the ARL to be bigger than ARL_0 . This particular estimation is based on a bootstrap method which in practice estimates the already estimated process.

Another interesting threshold adjustment is the one Diko, Chakraborti and Does proposed [2]. This adjustment method estimates the conditional ARL distribution ($CARL$) for various Phase I sample size. The implementation of the method suggests to trial various L from the interval (L_0, ∞) , where L_0 is the starting L each time. The method seeks for the value of L that gives $CARL_{IC,p} > ARL_0 (1 - \epsilon)$ for a known probability p . Enhancing figure 6.1 with control limits derived for the new threshold, it can be seen that these control limits may also prevent false alarms, but tend to be more narrow than the one studied above.

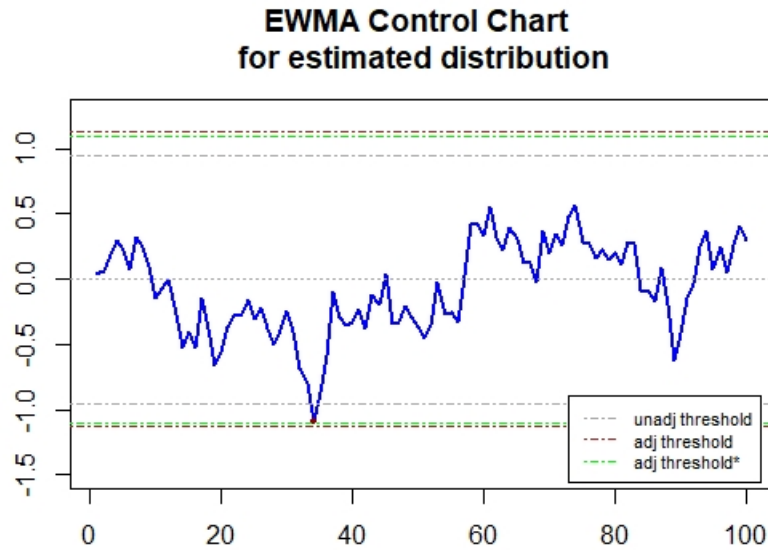


Figure 6.8.: EWMA control chart for an IC estimated distribution for a X_i , $i = 1, \dots, 100$ sample from the standard normal distribution and parameters $ARL = 370$ and $\lambda = 0.2$. The adjusted and two type of unadjusted Control Limits are included (adj. threshold holds for Gandy & Kvaløy [3] adjustment while threshold* for Diko, Chakraborti & Does approximation for $\epsilon = 0$ [2]). The estimated $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

By depicting the RL distribution for Diko, Chakraborti & Does (L^*) threshold it is interesting to note that the RLs given using this threshold are a little smaller than the ones Gandy & Kvaløy's threshold (L) gives. It has a smaller range, but overall seem to handle nicely the unwanted smaller RL values. The run length is by means smaller e.g. \bar{L}^* is 400.95 while \bar{L} is 440.27, with median for L^* to be 281 and for L to be 311.

Run Length IC distribution comparison when process is estimated for adjusted and unadjusted threshold

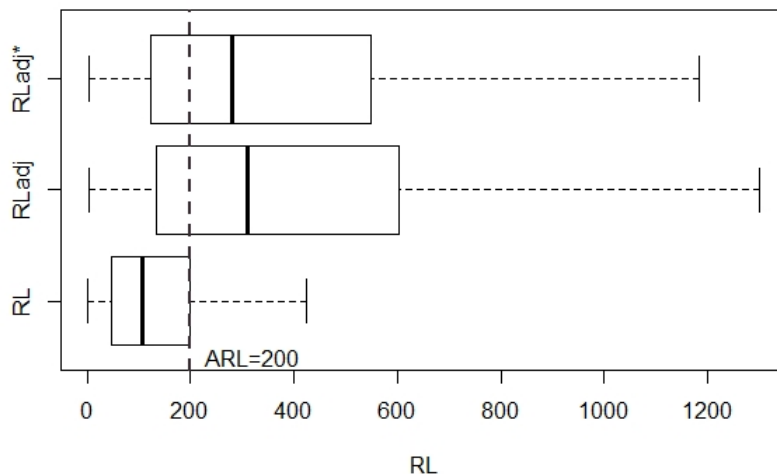


Figure 6.9.: Comparison of IC RL distributions for an EWMA statistical control process for adjusted and unadjusted thresholds L and L^* (for $\epsilon = 0$), for a X_i , $i = 1, \dots, 100$ sample from the standard normal distribution and estimated process for $ARL_0 = 200$, $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = 0.022$ and $\hat{\sigma} = 1.04$ were estimated on a $n_{est} = 100$ sample of the standard normal distribution.

The boxplot for the ARL for all the three thresholds, shows as expected that the two adjustments increase the Average Run Length reducing the false alarm probability. In fact, the ARL has a mean of 589.196 with a range of 3091 for the ARL^* mean of 605.53 and range of 3078. It seems that the two different adjustment tend to produce similar results, but judging by the following plots, in general the Diko, Chakraborti and Does threshold (L^*) gives slightly bigger $ARLs$. This is not the case, nevertheless, when the smoothing parameter λ is very small. Then L^* are much larger than L . So, L^* seems to reduce estimation error more effectively when λ is relatively small. Indeed for $\lambda = 0.02$ and for $\lambda = 0.1$, the L^* gives ARL closer to the ARL_0 according to the Tables 6.1 and 6.2. But for larger λ the differences are reduced.

	ARL_L	ARL_{L^*}
10%	183.80	208.00
5%	149.95	167.95

Table 6.1.: Table of the 10% and 5% of ARL distribution for adjusted thresholds. The L denotes the Gandy & Kvaløy's threshold while the L^* the Diko, Chakraborti and Does threshold. The $\lambda = 0.02$ while the $n_{ARL} = 500$, the $N_{RL} = 2500$ and the $n_{est} = 100$.

	ARL_L	ARL_{L^*}
10%	191.90	205.90
5%	156.00	168.90

Table 6.2.: Table of the 10% and 5% of ARL distribution for adjusted thresholds. The L denotes the Gandy & Kvaløy's threshold while the L^* the Diko, Chakraborti and Does threshold. The $\lambda = 0.1$ while the $n_{ARL} = 500$, the $N_{RL} = 2500$ and the $n_{est} = 100$.

The group of all the following plots compare the two threshold adjustments when the control chart parameters varied. As said both adjustments seem pretty much similar with adjusted L^* to give slightly bigger ARL s. What seem to be different in the two adjustments is the range of the ARL distribution. The adjusted L^* has larger range and actually this range is wider when ARL_0 is larger, figure 6.11, and when either of λ and estimation sample is smaller as seen in figures 6.12 and 6.10 respectively.

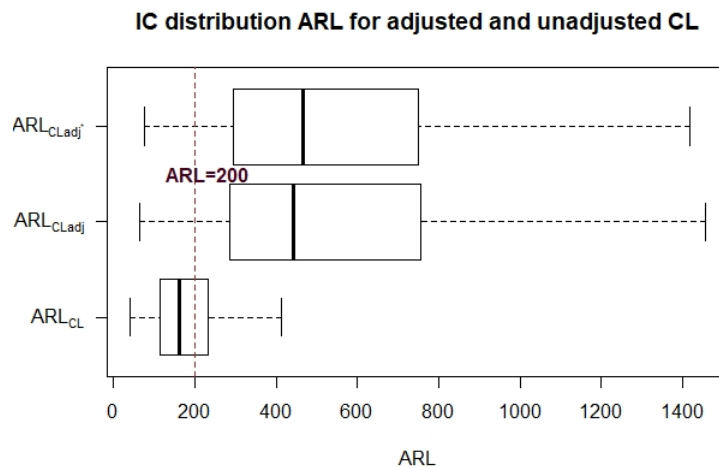


Figure 6.10.: Comparison of ARL distribution for unadjusted and adjusted thresholds. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does's (L^*).

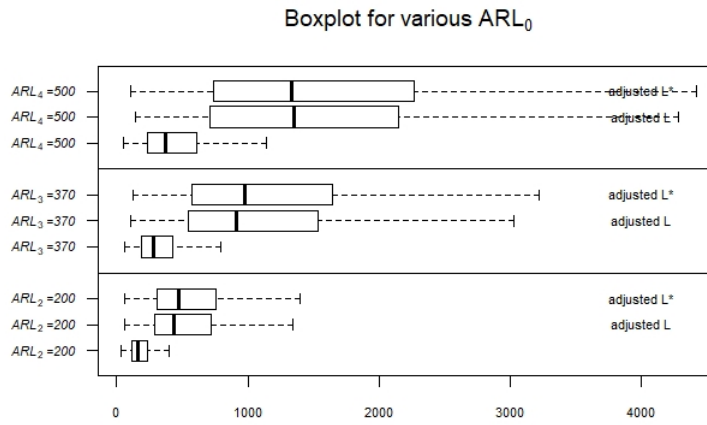


Figure 6.11.: Distribution of ARL for $ARL_0 = 200$, $ARL_0 = 370$ and $ARL_0 = 500$ for unadjusted and adjusted thresholds. The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The process sample has been derived from standard normal distribution. The simulation parameter $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does 's (L^*).

Boxplot for various λ

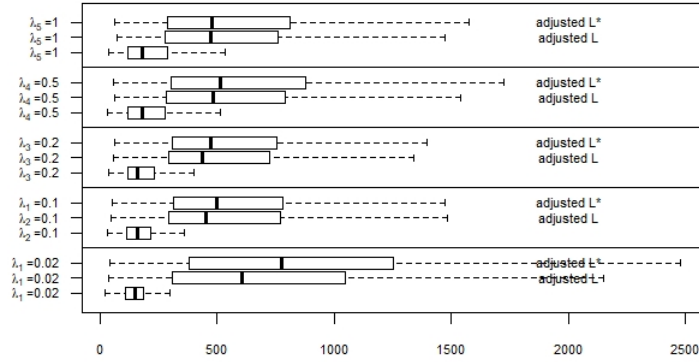


Figure 6.12.: Distribution of ARL for $\lambda_1 = 0.02$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$, $\lambda_4 = 0.5$ and $\lambda_5 = 1$ for unadjusted and adjusted thresholds.

The estimation of the parameters has been made on a $n_{est} = 100$ sample from the standard normal distribution. The process sample has been derived from standard normal distribution. The $ARL_0 = 200$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does 's (L^*).

Distribution of ARL for simulations of various estimation sizes

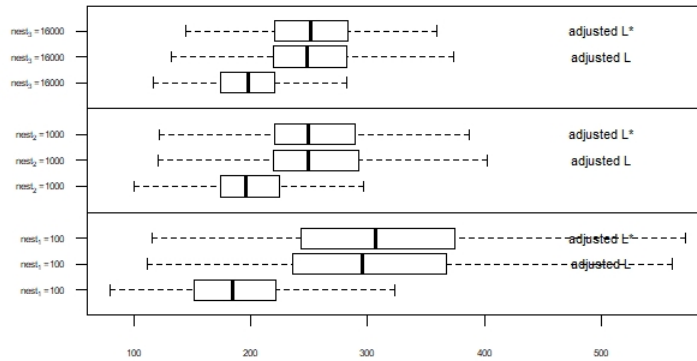


Figure 6.13.: Distribution of ARL for estimation of the parameters to be made on a $n_{est1} = 100$, $n_{est2} = 1000$ and $n_{est3} = 16000$ sample from the standard normal distribution. The process sample has been derived from standard normal distribution. The $ARL_0 = 200$ and the $\lambda = 0.5$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does 's (L^*).

6.3. Adjusted Thresholds for Out Of Control process

So far, the error the process estimation causes to an EWMA control chart has been widely examined and there have been proposed ways of handling it using adjusted thresholds. For this purpose two different threshold adjustments have been applied, the one Gandy and Kvaløy [3] and the one Diko, Chakraborti and Does [2] proposed. Also a number of simulations have been performed for an in control process to examine the behavior the adjusted control limits on the EWMA check statistic.

In this section the simulations will depict an out-of-control process for evaluating the adjusted thresholds when they must indeed deal with an abnormal observation. For an OOC process the plot 6.14 shows that the two adjusted tests find the out of control observations in the same time slot. This is also depicted in a boxplot for the RL distribution. The run length given by the unadjusted threshold has a mean of 18.77 while the adjusted ones are 27.87 for L and 29.13 for L^* .

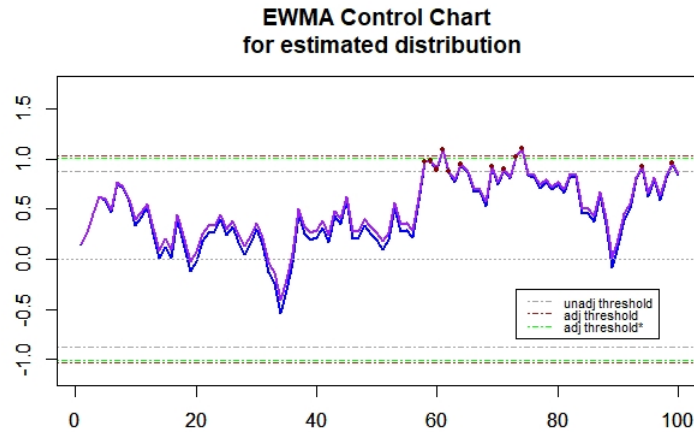


Figure 6.14.: EWMA control chart for an OOC estimated distribution for a X_i , $i = 1, \dots, 100$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$ and parameters $ARL = 370$ and $\lambda = 0.2$. The adjusted and two type of unadjusted Control Limits are included (adj. threshold holds for Gandy & Kvaløy [3] adjustment while threshold* for Diko, Chakraborti & Does approximation [2]). The estimated $\hat{\mu} = 0.0904$ and $\hat{\sigma}_1 = 0.9128$.

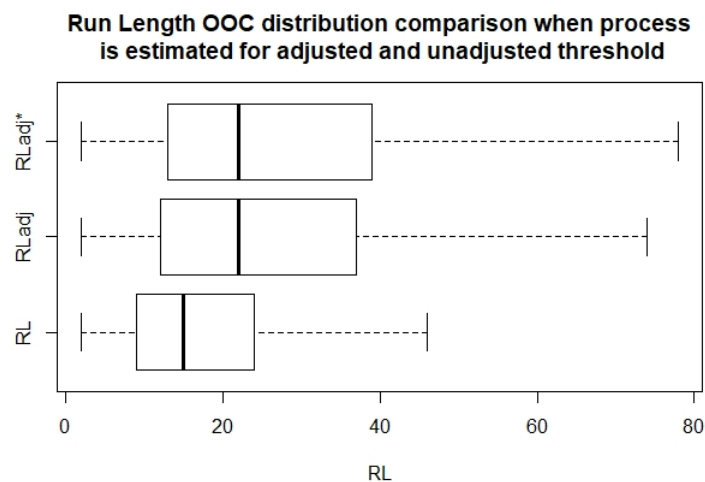


Figure 6.15.: Comparison of OOC RL distributions for an EWMA statistical control process for adjusted and non-adjusted CL for an OOC process e.g for a X_i , $i = 1, \dots, 100$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$ and estimated process for $ARL_0 = 200$, $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$. The $\hat{\mu} = -0.078$ and $\hat{\sigma} = 0.964$ were estimated on a $n_{est} = 100$ sample of standard normal distribution.

Regarding the ARL distribution in the basic simulation where desired value of ARL is 200, and with a $\delta = 0.5$, the two adjusted thresholds give equivalent distributions with a median of 43 and 44 for L and L^* respectively which can be satisfying in process control with such characteristics. For different starting values of ARL_0 , plot 6.17 shows that when the thresholds are being adjusted, the control chart may variate more effectively in the increase of the ARL_0 . The adjustment itself may potentially come with a cost. It could detect abnormal observations later. This is the cost for avoiding too short IC ARL_0 . But according to figure 6.15 this cost is not in practice so big.

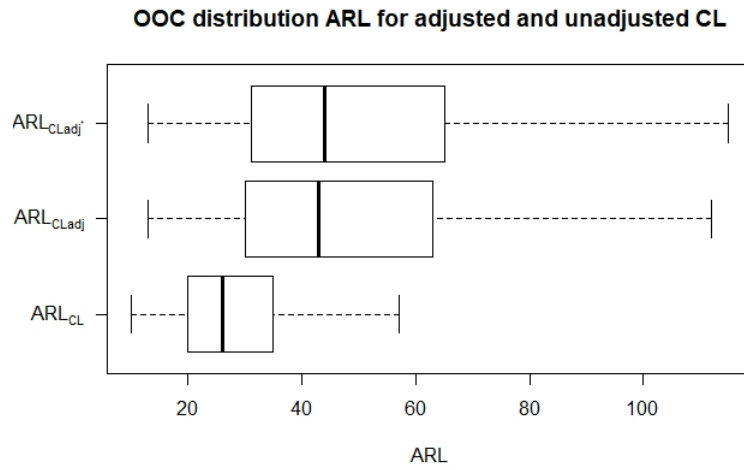


Figure 6.16.: Comparison of ARL distribution for unadjusted and adjusted thresholds for an OOC process. The estimation of the parameters has been made on a $n_{est} = 100$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$. The simulation parameters are $ARL_0 = 200$ and $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL was calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does's (L^*).

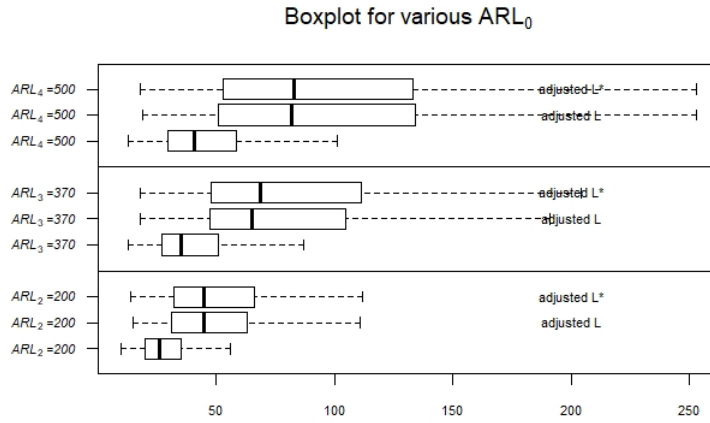


Figure 6.17.: Distribution of ARL for an OOC process and for $ARL_0 = 200$, $ARL_0 = 370$ and $ARL_0 = 500$ for unadjusted and adjusted thresholds. The estimation of the parameters has been made on a $n_{est} = 100$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$. The process sample has been derived from standard normal distribution. The simulation parameter $\lambda = 0.2$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does 's (L^*).

An equivalent result can be retrieved in the figure 6.18 for different λ s. It is more than obvious that for larger λ , the ARL distribution is wider and has bigger values. This of course has to do with the fact that $\delta = 0.5$ a fair small mean deviation especially for $\lambda = 1$.

Boxplot for various λ

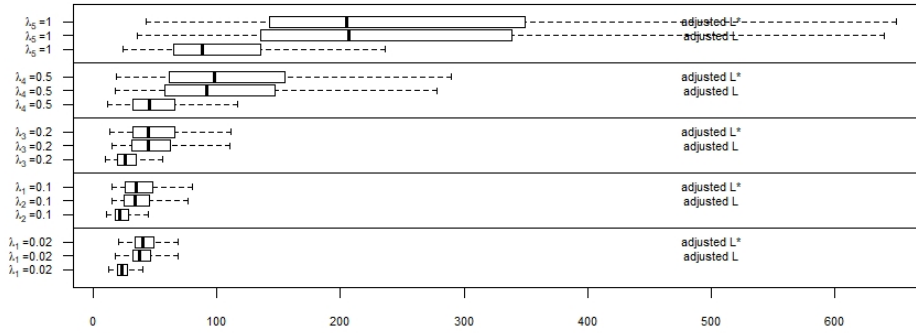


Figure 6.18.: Distribution of ARL for an OOC process and $\lambda_1 = 0.02$, $\lambda_2 = 0.1$, $\lambda_3 = 0.2$, $\lambda_4 = 0.5$ and $\lambda_5 = 1$ for unadjusted and adjusted thresholds. The estimation of the parameters has been made on a $n_{est} = 100$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$. The process sample has been derived from standard normal distribution. The $ARL_0 = 200$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does's (L^*).

As expected, the variation of estimation sample just confirms what said above. For confronting estimation error, two different threshold adjustment methods were practiced. Both of them seem to respond satisfactory to handling of the Type I error. More precisely, the L^* of Diko, Chakraborti & Does gives slightly larger ARL distribution reducing the probability of false alarm and because of the fact that produces values for the parameter L it is more flexible to different types of control limits and can also be used for derivation of non-steady state control limits. Additionally it is a code that although time consuming it has to be calculated only once. On the other hand, the Gandy and Kvaløy's threshold refers just to the steady state limits (4.5) which could potential restrict the process control and it needs to run a bootstrap for each data set. In contrary, it is more easily to interpret and more user friendly .

Distribution of ARL for simulations of various estimation sizes

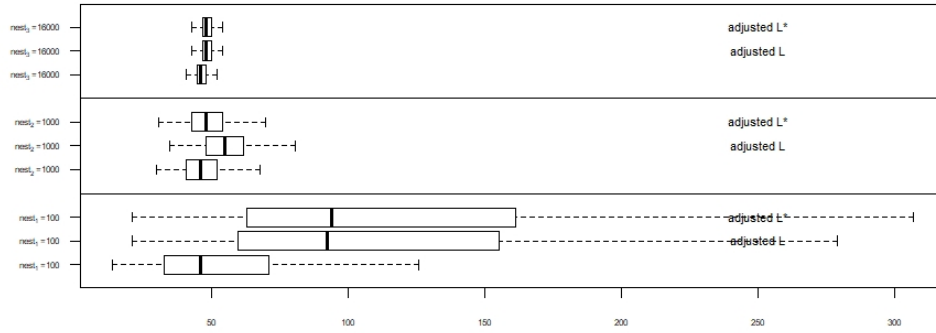


Figure 6.19.: Distribution of ARL for an OOC process and for estimation of the parameters to be made on a $n_{est1} = 100$, $n_{est2} = 300$ and $n_{est3} = 16000$ sample from a normal distribution with $\mu_0 = 0.5$ and $\sigma_0 = 1$. The process sample has been derived from standard normal distribution. The $ARL_0 = 200$ and the $\lambda = 0.5$. The simulation size for RL is $N_{RL} = 2500$ and each ARL has been calculated on $N_{ARL} = 5000$ RLs. The ARL simulation size is $n_{ARL} = 500$. The thresholds used are the Gandy-Kvaløy's estimation (L) and Diko, Chakraborti & Does 's (L^*).

7. Baby birth weight analysis

So far, a wide study on EWMA process control analysis has been made focusing mainly on the error arising from the process estimation. In this chapter the EWMA chart will be applied for evaluation of real data. The data used are from the *Helping Babies Breath (HBB)* project at *Haydom Lutheran Hospital* in Northern Tanzania. The data consist of measurements of important variables of newborn e.g birth weight, height, heart rate. The data spread over 7 years of measurements (February 2010-February 2017). The analysis following will concentrate on the evaluation of the impact the application of two new fees charged, one on ambulance use after July 2013 and one delivery fee after January 2014, have on the hospital's newborn weight.

On 1st of July 2013 the Haydom Lutheran Hospital applied a fee for the ambulance transfer of women in labor. Additionally, next year's (2014) January an extra fee has been charged on delivery services provided by the hospital. The main subject of the analysis will be to evaluate how these fees affect the decision of women to deliver in the hospital in relation with their pregnancy progress and overall situation and how this eventually depicts on the newborn's babies birth weight. For this an EWMA process control chart will be used evaluating the control statistic with both adjusted and unadjusted control limits. For the Phase I estimation, the data before the ambulance fee will be used, i.e the data up until 30 of June 2013 which is a set of 16164 measurements. Phase II data for the process control will be 11025 observations after the application of the hospital fee (1st of January 2014). The in between data will be omitted because the control concerns the birth weight depending on both fees. Later on a regression model will be applied in order to specify which particular variables are related to the baby's birth weight.

7.1. EWMA control chart for birth weight

Using the EWMA control chart studied in the above chapters, first there will be an estimation of the birth weight distribution from 1st February 2010 to 30th June 2013 to approximate the birth weight distribution. The sample of these data although at the *Anderson-Darling* test for normality seem not to be normal distributed, marginally the deviation the normal distribution as seen in figure 7.1 can be accepted. Actually the rejection of the normality hypothesis at the test may be attributed to the extreme cases as seen in the qqplot. This estimation therefore gives $\hat{\mu} = 3099$ and $\hat{\sigma} = 484.55$. For the EWMA control chart the λ will be equal to 0.02 and the desired $ARL_0 = 40000$. The unadjusted control limits are $(-0.357, 0.357)$ while the adjusted one, computed using the Gandy and Kvaløy's bootstrap

are $(-0.391, 0.391)$. These two control limits are not very different, but this due to large Phase I sample.

Anderson-Darling normality test

```
data: hbbdataphaseI$BIRTH_WEIGHT
A = 44.927, p-value < 2.2e-16
```

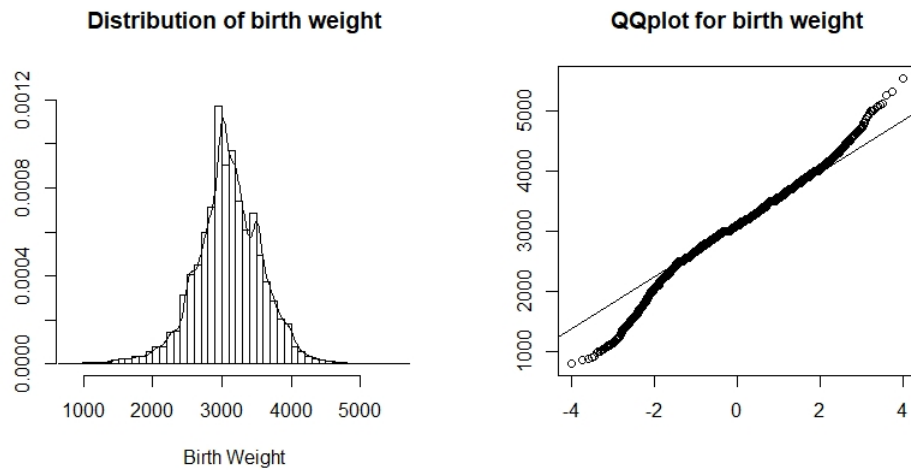


Figure 7.1.: Test for normality for the distribution of birth weight for the data from 1st February 2010 to 30th June 2013 using an histogram and a qqplot

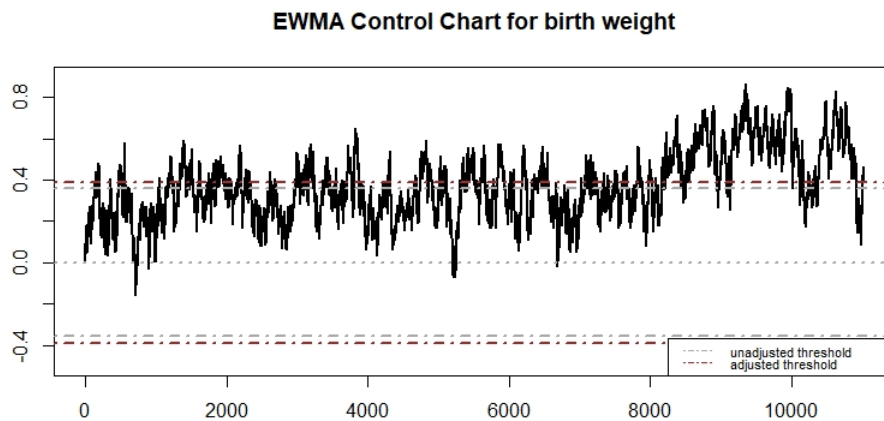


Figure 7.2.: EWMA control chart for birth weight of newborns of HBB data. The birth weight distribution estimation has been made on the measurements before 1st July 2013 giving $\hat{\mu} = 3099$ and $\hat{\sigma} = 484.55$ and the control chart applied on data after 1st January 2014 with $\lambda = 0.02$ and $ARL_0 = 40000$.

The plot 7.2 used as parametric factors $\lambda = 0.02$ and $ARL_0 = 40000$. The choice of λ to be small depends on the fact that it is expected the

birth weight to have a small persistence change in the mean and the process analysis seeks for small changes in longer time while the choice of ARL_0 is based on the big data set, e.g the data size which is 31122 and a big ARL_0 suppresses the false alarms into one out of 40000 and this reference to one false alarm every 10 years. Following the figure 7.2, the birth weight distribution tends to increase leading to an out of control process which in practice means that after the application of the two fees the mean birth weight increases from the estimated $\hat{\mu} = 3099$. I.e mothers to be of average or smaller income with no seen complications and with a healthy normal weight fetus decide to avoid giving birth in the hospital. So, the number of babies born with significantly different weight from the mean values in the hospital increases. There is actually the tendency for a slight increase of the birth weight which could indicates that after the extra expenses more healthier women with higher weighted fetus choose to give birth in the hospital. This is an expected result as when the expenses raise and if the prenatal control secures a healthy baby to be born and an easy delivery, mothers decide to avoid additional expenses. Another interesting result arising from the control chart is that towards the end of the observations, EWMA statistic tends to lessen. This potentially means that after years of applied fees, people take the extra expenses for granted or that they do not have experience of fee free hospital so and the fees will eventually stop being a disincentive for giving birth in the hospital. As said the smoothing parameter is being chosen on the size of the shift and on the research's interest. In order to study the effect the choice of λ has on the EWMA control chart for the birth weight two more control charts for different smoothing parameters are applied. The first has a smaller smoothing parameter $\lambda = 0.01$, which means that the shift which is desired to be spotted is smaller in a more wider time frame then the plot 7.3 indicates that indeed the process goes out of control due to the fees introduce but also that towards the last observations this variance is being reduced. The second is derived for a larger $\lambda = 0.05$ for finding larger shifts in a smaller time space. The plot 7.4 fails actually to spot this birth weight abnormality. This actually shows that the birth weight alters with smaller shifts which need some time to pick up.

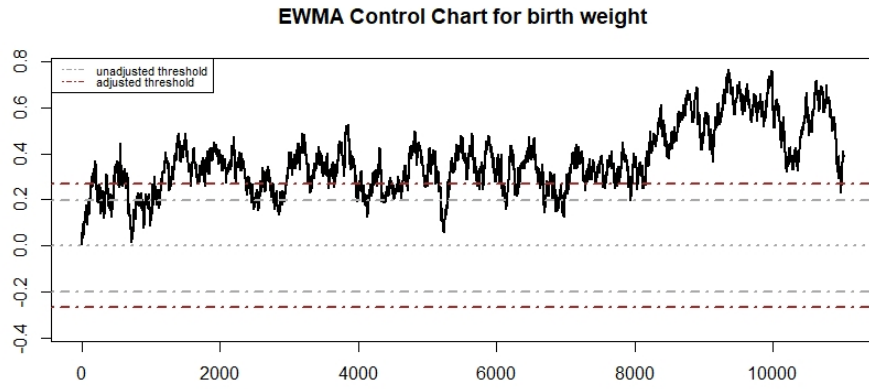


Figure 7.3.: EWMA control chart for birth weight of newborns of HBB data. The birth weight distribution estimation has been made on the measurements before 1st July 2013 giving $\hat{\mu} = 3099$ and $\hat{\sigma} = 484.55$ and the control chart applied on data after 1st January 2014 with $\lambda = 0.01$ and $ARL_0 = 40000$.

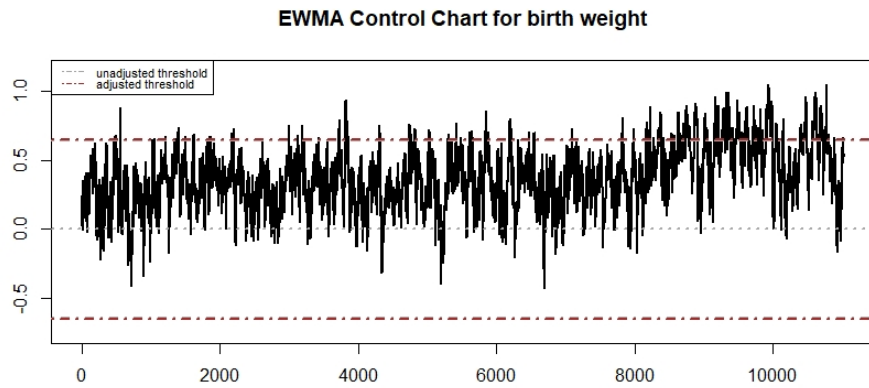


Figure 7.4.: EWMA control chart for birth weight of newborns of HBB data. The birth weight distribution estimation has been made on the measurements before 1st July 2013 giving $\hat{\mu} = 3099$ and $\hat{\sigma} = 484.55$ and the control chart applied on data after 1st January 2014 with $\lambda = 0.05$ and $ARL_0 = 40000$.

7.2. Regression model for birth weight

The process control showed that the application of two extra fees in the birth procedure, affects the weight of newborns in the hospital. In the following, a linear model will be fitted on the Phase I data in order to be determined which variables also influence the birth weight. The variables which are going to be considered are *PREG_COMP*, *INFECTION*, *CS_indication*, *PREECLAMP*, *BLEEDING*, *GEST_AGE* which represent equivalently *Compilations during pregnancy*, *Presence of infection during pregnancy*, *Indication for cesarean section*, *Pregnancy's preeclampsia*,

Bleeding before labor, Gestational age in weeks. Firstly, a model including all the variables gives

Call:

```
lm(formula = BIRTH_WEIGHT ~ PREG_COMP + INFECTION + CS_indication +
    PREECLAMP + BLEEDING + GEST_AGE, data = hbbdataphaseI)
```

Residuals:

Min	1Q	Median	3Q	Max
-2084.21	-284.88	-24.21	285.71	2216.46

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-3553.160	197.804	-17.963	< 2e-16 ***
PREG_COMP	109.200	37.668	2.899	0.00375 **
INFECTION	117.270	25.408	4.615	3.95e-06 ***
CS_indication	-34.113	7.078	-4.820	1.45e-06 ***
PREECLAMP	367.922	69.529	5.292	1.23e-07 ***
BLEEDING	196.678	48.359	4.067	4.78e-05 ***
GEST_AGE	139.331	2.563	54.359	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 443 on 16157 degrees of freedom

Multiple R-squared: 0.1643, Adjusted R-squared: 0.164

F-statistic: 529.6 on 6 and 16157 DF, p-value: < 2.2e-16

This linear model suggests that all the variables are significant. So the model which will be chosen will include all the variables. The suitability of this regression model confirmed also by the residuals, it is somewhat expected as all the variables chosen have generally an impact on the way of labor. The model fitting therefore, gives the incentive to apply a new EWMA control chart on the residuals of the predicted data.

7.2.1. EWMA control chart on the residuals of the predicted data

An EWMA control chart on the residuals of the predicted data is [3]

$$Z_i = \lambda(Y_i - X_i\beta) + (1 - \lambda)Z_{i-1}, \quad Z_0 = 0 \quad (7.1)$$

where Z is the control chart statistic, Y is the response variable, X the vector of the independent variables and β the coefficient vector. The smoothing parameter λ as usual determines how to weight the recent data versus the past ones. The coefficient vector β has been estimated on Phase I data, while the chart used the Phase II data. Using the regression model, an EWMA control chart is derived

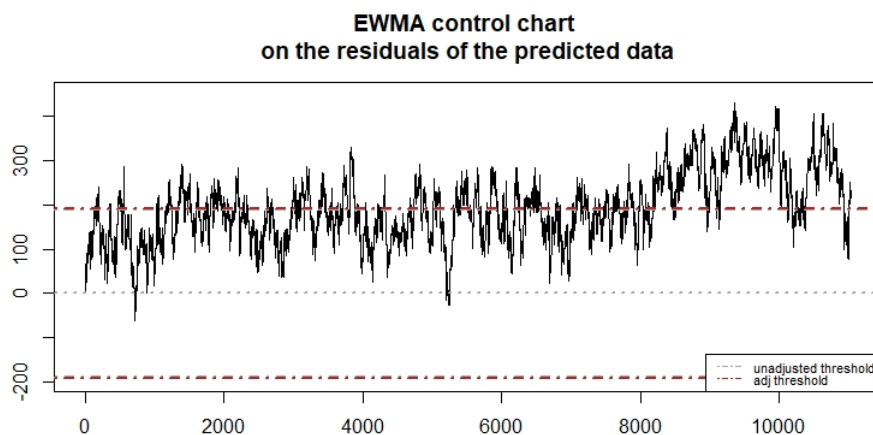


Figure 7.5.: EWMA control chart for predicted birth weight of newborns using HBB data. The control chart applied on data after 1st January 2014 with $\lambda = 0.02$ and $ARL_0 = 40000$. The regression model uses the *Compilations during pregnancy*, *Presence of infection during pregnancy*, *Indication for cesarean section*, *Pregnancy's preeclampsia*, *Bleeding before labor*, *Gestational age in weeks* and the fitting has been made on the Phase I data (measurements before 1st July 2013).

The EWMA chart on the fitted residuals shows that the prediction of the birth weight confirms the fact that the process actually will be out of control after the application of the extra fees. The statistic on the residuals is greater than zero and exceeds soon the control limit. This means that the fitted model fails to fully explain the birth weight for the period of interest i.e. after the application of the fees.

7.2.2. Discussion on birth weight

The analysis performed on the data of 7 years birth statistics provided by the Haydom Lutheran Hospital in Northern Tanzania focused mainly on the evaluation of birth weight value of newborns. The critical point that made this variable of particular interest was the fact that after July 2013 and January 2014 two consequent fees where applied by the hospital. A process control therefore was conducted on birth weight data and more precisely an EWMA control chart. For the unknown birth weight distribution an estimation was made on the data before July 2013 while the EWMA analysis was performed on data after January 2014. The control chart showed that indeed the application of the extra fees altered the birth weight deviating it from the estimated mean value and more precisely showed the upward tendency of the variable. In order to evaluate the other factors influencing the birth weight, linear regression was used. The study showed that compilations during pregnancy, presence of infection during pregnancy, indication for cesarean section, pregnancy's preeclampsia, bleeding before labor and gestational age in weeks were critical in determination of birth weight values. The regression model failed to fully explain change in the birth weight

when the explanatory variables vary after the hospital fees are applied.

8. Conclusion

In this thesis the aim was to determine and reduce the error arising from the estimation of the distribution when performing a process analysis using the EWMA control chart. Firstly this error was defined and reviewed both individually and in relation with other also important control chart factors, the smoothing parameter λ , the desired ARL_0 , the process estimation size and the number of simulations for the determination of ARL . Later on, two different threshold adjustment for reducing this error were prescribed, tested and evaluated.

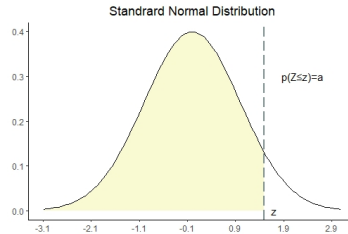
The first one was the one Gandy and Kvaløy [3] proposed. This threshold consists a bootstrap approximation of the threshold that guarantees a certain $\alpha\%$ amount of $ARLs$ to lie inside the control limits. This adjusted threshold indeed provides the control chart with wider control limits something that prevents Type I error. This was confirmed with proper comparison plots for adjusted and unadjusted thresholds.

The second was the one Diko, Chakraborti and Does proposed [2]. This particular threshold calculates the parameter L successively from a starting value L_0 . This method of adjusting the threshold provides a matrix which contains values of L for a combination of process estimation size, a desired λ and an specific ARL_0 . Once again the suitability of this adjustment was examined and tested with proper simulations subject to the two other thresholds. The study of the two adjustment confirmed that both of them are almost equal effective on reducing false alarms. But the second one was pretty inconvenient regarding its use. Because of the fact that the calculation is made on particular combination of three different parameters a unique matrix had to be calculated for each case. Additionally, this whole simulation was very time consuming that prevent its use in the real data study carried out in the next chapter.

In the final chapter an application of the so far study methods was made on real data. The data involve measurements made on mothers in labor and their newborns for a time interval of 7 years and were provided by the *Haydom Lutheran Hospital* in Northern Tanzania. The EWMA control chart with unadjusted and adjusted control limits was applied in order to confirm a change of newborn babies weight after an extra charge on ambulance use and hospitalization as this appears in the newborns birth weight. After the out of control process specification a linear model was applied in order to specify the other parameters that also affect birth weight. The outcome of the data examination was that indeed the birth weight varied from the mean value determined from the data before the fee application and for the future prediction of birth weight a model was constructed and a new EWMA control chart was applied on the prediction residuals.

A. Appendix A

A.1. Z distribution table



	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0	0.50	0.50	0.51	0.51	0.52	0.52	0.52	0.53	0.53	0.54
0.1	0.54	0.54	0.55	0.55	0.56	0.56	0.56	0.57	0.57	0.58
0.2	0.58	0.58	0.59	0.59	0.59	0.60	0.60	0.61	0.61	0.61
0.3	0.62	0.62	0.63	0.63	0.63	0.64	0.64	0.64	0.65	0.65
0.4	0.66	0.66	0.66	0.67	0.67	0.67	0.68	0.68	0.68	0.69
0.5	0.69	0.69	0.70	0.70	0.71	0.71	0.71	0.72	0.72	0.72
0.6	0.73	0.73	0.73	0.74	0.74	0.74	0.75	0.75	0.75	0.75
0.7	0.76	0.76	0.76	0.77	0.77	0.77	0.78	0.78	0.78	0.79
0.8	0.79	0.79	0.79	0.80	0.80	0.80	0.81	0.81	0.81	0.81
0.9	0.82	0.82	0.82	0.82	0.83	0.83	0.83	0.83	0.84	0.84
1	0.84	0.84	0.85	0.85	0.85	0.85	0.86	0.86	0.86	0.86
1.1	0.86	0.87	0.87	0.87	0.87	0.87	0.88	0.88	0.88	0.88
1.2	0.88	0.89	0.89	0.89	0.89	0.89	0.90	0.90	0.90	0.90
1.3	0.90	0.90	0.91	0.91	0.91	0.91	0.91	0.91	0.92	0.92
1.4	0.92	0.92	0.92	0.92	0.93	0.93	0.93	0.93	0.93	0.93
1.5	0.93	0.93	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
1.6	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
1.7	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96	0.96
1.8	0.96	0.96	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.97
1.9	0.97	0.97	0.97	0.97	0.97	0.97	0.97	0.98	0.98	0.98
2	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
2.1	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.99
2.2	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.3	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.4	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
2.5	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	1.00	1.00
2.6	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.7	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.8	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2.9	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
3	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table A.1.: Z-table for Normal Distribution

A.2. R codes

A.2.1. R-code for the EWMA control chart

This is the code for the EWMA control chart for both real and estimated distribution

```
library(spc)
library(ggplot2)
library(ANOVAreplication)
library(IQCC)
library(qcc)
library(gridExtra)
library(xtable)

#function for calculating z statistic for EWMA control chart.
ewmadistributionszx <- function(ARL, l, nest, muest=0,
                              sdest=1, mu, sd=1, nrl,
                              Narl){

  # Derivation of threshold L
  L <- xewma.crit(l,ARL,sided='two')

  # Estimation of distribution
  xest <- rnorm(nest, muest, sdest)
  muestimated <- mean(xest)
  sdestimated <- sd(xest)

  # Calculation of Control Limits (steady-state)
  UCL <- +L*sqrt(1/(2-l))
  LCL <- -L*sqrt(1/(2-l))
  CL <- 0

  # Calculation of statistic z
  szx=matrix(nrow=Narl,ncol=nrl)
  RL=numeric(nrl)
  szx1=matrix(nrow=Narl,ncol=nrl)
  RL1=numeric(nrl)

  # Setting starting values of x,y and
  # szx(standardized z)
  x <- rnorm(nrl,mu,sd)
  y <- (x-muestimated)/sdestimated
  szx[1,] <- l*y
  y1 <- (x-0)/1
  szx1[1,] <- l*y1

  # for loop for each szx (with replacing)
  # for estimated distribution
  for(i in 2:Narl){
    x <- rnorm(nrl,mu,sd)
```

```

    y <- (x-muestimed)/sdestimated
    szx[i,] <- l*y+(1-l)*szx[(i-1),]
    y1 <- (x-0)/1
    szx1[i,] <- l*y1+(1-l)*szx1[(i-1),]
  }

  # Spotting values which exceed the
  # Control Limits
  count <- 0
  for(i in 1:nrl){
    RL[i] <- which(szx[,i]>UCL | szx[,i]<LCL)[1]
    if(is.na(RL[i])){
      RL[i] <- Narl
      count <- count+1
    }
  }

  # Compute the Average Run Length
  ARL <- round(mean(RL),0)

  # for real distribution
  count1 <- 0
  for(i in 1:nrl){
    RL1[i] <- which(szx1[,i]>UCL | szx1[,i]<LCL)[1]
    if(is.na(RL1[i])){
      RL1[i] <- Narl
      count1 <- count1+1
    }
  }
  ARL1 <- round(mean(RL1),0)

  # Returns a list of the statistics to plot,
  # the estimated distribution
  # the control limits and the Run Length
  outcome <- list(szx,szx1,RL,RL1,muestimed,
                 sdestimated,UCL,CL,LCL)
  return(outcome)
}

```

A.2.2. R code for the distribution of ARL for estimated distribution

```

library(spc)
library(ggplot2)
library(ANOVAreplication)
library(IQCC)
library(qcc)
library(gridExtra)

```

```
library(xtable)
```

```
# Function for deriving a single ARL
```

```
ewmadistribution1 <- function(ARL, l, nest, muest=0,  
                             sdest=1, mu, sd=1,  
                             nrl, Narl){
```

```
# Calculation of L
```

```
L <- xewma.crit(l,ARL,sided='two')
```

```
# Estimation of the distribution
```

```
xest <- rnorm(nest, muest, sdest)
```

```
muestimated <- mean(xest)
```

```
sdestimated <- sd(xest)
```

```
# Derivation of steady-state control limits
```

```
UCL <- +L*sqrt(1/(2-l))
```

```
LCL <- -L*sqrt(1/(2-l))
```

```
CL <- 0
```

```
# Setting starting values for statistic z and RL
```

```
szx=matrix(nrow=Narl, ncol=nrl)
```

```
RL=numeric(nrl)
```

```
# Simulation x[1] and calculating y[1] and z[1]
```

```
x <- rnorm(nrl, mu, sd)
```

```
y <- (x-muestimated)/sdestimated
```

```
szx[1,] <- l*y
```

```
# For loop for obtaining Narl z with replacing
```

```
for(i in 2:Narl){
```

```
  x <- rnorm(nrl, mu, sd)
```

```
  y <- (x-muestimated)/sdestimated
```

```
  szx[i,] <- l*y+(1-l)*szx[(i-1),]
```

```
}
```

```
# Find RL
```

```
count <- 0
```

```
for(i in 1:nrl){
```

```
  RL[i] <- which(szx[,i]>UCL | szx[,i]<LCL)[1]
```

```
  if(is.na(RL[i])){
```

```
    RL[i] <- Narl
```

```
    count <- count+1
```

```
  }
```

```
}
```

```
# Calculating ARL
```

```
ARL <- round(mean(RL),0)
```

```

# Return estimated distribution , and ARL.
# Count is the number of trials that RL
# is bigger than Narl
outcome <- list(muestimated , sdestimated ,ARL,count)
return(outcome)
}

# Function for repeating the ewmadistribution1
# in order to find the ARL distribution
arlsim <- function(narl , ARL, l , nest , muest=0,
                  sdest=1, mu, sd=1, nrl ,
                  Narl){
  ARL <- replicate(narl , ewmadistribution1(ARL, l ,
                                          nest , muest , sdest ,
                                          mu, sd , nrl , Narl))

  return(ARL)
}

```

A.2.3. R code for adjusted ARL distribution

Following is the code for the ARL distribution for adjusted EWMA control chart. The code refers to both adjustments

```

library(spc)
library(boot)
library(spcadjust)

# Reads the matrix fro the Diko thresholds
Diko500 <- readRDS('DikoLarl500.RDS')
Diko001 <- readRDS('DikoLlamda001.RDS')
Diko370 <- readRDS('DikoL.RDS')

# Function for derivation of elements
# from previous matrices
findLD <- function(l ,nest ,ARL){
  if(ARL==200 & nest!=16000 & l!=0.01 & l!=0.02){
    if(l==0.1){i=1}
    if(l==0.2){i=2}
    if(l==0.5){i=3}
    if(l==1){i=4}
    if(nest==1000){j=2}
    if(nest==300){j=3}
    if(nest==100){j=4}
    L <- findL[i , j]
  }
  if(ARL==500){
    if(l==0.1){i=1}
    if(l==0.2){i=2}
    if(l==0.5){i=3}
    if(l==1){i=4}
  }
}

```

```

    if (nest==1000){j=2}
    if (nest==300){j=3}
    if (nest==100){j=4}
    L <- Diko500 [ i , j ]
  }
  if (ARL==370){
    if (l==0.1){i=1}
    if (l==0.2){i=2}
    if (l==0.5){i=3}
    if (l==1){i=4}
    if (nest==1000){j=2}
    if (nest==300){j=3}
    if (nest==100){j=4}
    L <- Diko500 [ i , j ]
  }
  if (ARL==200 && (nest==16000 | l==0.01 | l==0.02 | l==0.5)){
    if (l==0.01){i=1}
    if (l==0.02){i=2}
    if (l==0.5){i=3}
    if (l==1){i=4}
    if (nest==16000){j=2}
    if (nest==300){j=3}
    if (nest==100){j=4}
    L <- Diko001 [ i , j ]
  }

  return (L)
}

# EWMA simulation
ewmadistributionszxadjD <- function (ARL, l, nest, muest=0,
                                     sdest=1, mu, sd=1,
                                     nrl, Narl){
  # Derivation of estimation sample
  xest <- rnorm (nest, muest, sdest)

  # Set up chart for adjusted threshold
  EWMAchart <- new ("SPCEWMA", model=SPCModelNormal (Delta=0),
                  lambda=l)
  thresholdadj <- SPCproperty (data=xest, nrep=50,
                              property="calARL",
                              chart=EWMAchart,
                              params=list (target=ARL))@res

  # Unadjusted threshold
  L <- xewma.crit (l, ARL, sided='two')

  # Diko's adjusted threshold
  LD <- findLD (l, nest)

```



```

# Estimation of distribution
muestimated <- mean(xest)
sdestimated <- sd(xest)

# Finding of the three different control limits
UCL <- +L*sqrt(1/(2-1))
LCL <- -L*sqrt(1/(2-1))
CL <- 0
UCLadj <- thresholdadj
LCLadj <- -thresholdadj
UCLadjD <- +LD*sqrt(1/(2-1))
LCLadjD <- -LD*sqrt(1/(2-1))

# Setting up starting values for all variables
szx=matrix(nrow=Narl, ncol=nrl)
RL=numeric(nrl)
RLadj <- numeric(nrl)
RLadjD <- numeric(nrl)
szx1=matrix(nrow=Narl, ncol=nrl)
RL1=numeric(nrl)
x <- rnorm(nrl, mu, sd)
y <- (x-muestimated)/sdestimated
szx[1,] <- 1*y
y1 <- (x-0)/1
szx1[1,] <- 1*y1

# For loop for derivation of statistic, RL and ARL
# for both adjusted and unadjusted thresholds
for(i in 2:Narl){
  x <- rnorm(nrl, mu, sd)
  y <- (x-muestimated)/sdestimated
  szx[i,] <- 1*y+(1-1)*szx[(i-1),]
  y1 <- (x-0)/1
  szx1[i,] <- 1*y1+(1-1)*szx1[(i-1),]
}
count <- 0
for(i in 1:nrl){
  RL[i] <- which(szx[,i]>UCL | szx[,i]<LCL)[1]
  if(is.na(RL[i])){
    RL[i] <- Narl
    count <- count+1
  }
}
ARL <- round(mean(RL), 0)
countadj <- 0
for(i in 1:nrl){
  RLadj[i] <- which(szx[,i]>UCLadj | szx[,i]<LCLadj)[1]
  if(is.na(RLadj[i])){

```

```

        RLadj[i] <- Narl
        countadj <- countadj+1
    }
}
ARLadj <- round(mean(RLadj),0)
countadjD <- 0
for(i in 1:nrl){
    RLadjD[i] <- which(szx[,i]>UCLadjD | sxz[,i]<LCLadjD)[1]
    if(is.na(RLadjD[i])){
        RLadjD[i] <- Narl
        countadjD <- countadjD+1
    }
}
ARLadjD <- round(mean(RLadjD),0)
##real distribution
count1 <- 0
for(i in 1:nrl){
    RL1[i] <- which(szx1[,i]>UCL | sxz1[,i]<LCL)[1]
    if(is.na(RL1[i])){
        RL1[i] <- Narl
        count1 <- count1+1
    }
}
ARL1 <- round(mean(RL1),0)
outcome <- list(szx, sxz1, RL, RL1, RLadj, RLadjD,
               mueestimated, sdestimated, UCL, CL, LCL,
               UCLadj, LCLadj, UCLadjD, LCLadjD)
return(outcome)
}

```

A.2.4. R code for regression model

```

# Finding the regression model
lmall <- lm(BIRTH_WEIGHT ~ PREG_COMP + INFECTION +
           CS_indication + PREECLAMP
           + BLEEDING + GEST_AGE,
           data = hbbdataphaseI)
summary(lmall)

# Deriving the EWMA control chart
# for predicted residuals
EWMAchartreg <- new("SPCEWMA", model = SPCModellm(Delta = 0,
           formula = "BIRTH_WEIGHT_~_PREG_COMP_+
           .....INFECTION_+_CS_indication_+_PREECLAMP
           .....+_BLEEDING_+_GEST_AGE"), lambda = 1)
calEWMA <- SPCproperty(data = hbbdataphaseI,
           nrep = 50, property = "calARL",
           chart = EWMAchartreg, params = list(target = ARL),
           quiet = TRUE)

```

```
xihat <- xiofdata(EWMAchartreg, hbdataphaseI)
S <- runchart(EWMAchartreg, newdata = hbdataphaseII,
              xi=xihat)
```

Bibliography

- [1] Subhabrata Chakraborti. *Nonparametric statistical process control*. In collab. with Marien Alet Graham. Hoboken, NJ: Wiley, 2019. ISBN: 978-1-118-89067-7.
- [2] Mandla D. Diko, Subha Chakraborti, and Ronald J.M.M. Does. “Guaranteed in-control performance of the EWMA chart for monitoring the mean”. In: *Qual Reliab Engng Int* 35.4 (June 2019), pp. 1144–1160.
- [3] Axel Gandy and Jan Kvaløy Terje. “spcadjust: An R Package for Adjusting for Estimation Error in Control Charts”. In: *The R Journal* 9.1 (2017), p. 458.
- [4] Axel Gandy and Jan Terje Kvaløy. “Guaranteed Conditional Performance of Control Charts via Bootstrap Methods: Guaranteed control charts”. In: *Scand J Statist* 40.4 (Dec. 2013), pp. 647–668.
- [5] L. Allison Jones, Charles W Champ, and Steven E Rigdon. “The Performance of Exponentially Weighted Moving Average Charts With Estimated Parameters”. In: *Technometrics* 43.2 (May 2001), pp. 156–167.
- [6] Sven Knoth. *spc: Statistical Process Control – Calculation of ARL and Other Control Chart Performance Measures*. Version 0.6.3. Oct. 27, 2019.
- [7] Douglas C. Montgomery. *Introduction to statistical quality control*. 6th ed. Hoboken, N.J: Wiley, 2009. 734 pp.
- [8] Nesma A. Saleh et al. “Another Look at the EWMA Control Chart with Estimated Parameters”. In: *Journal of Quality Technology* 47.4 (2015), pp. 363–382.
- [9] Luca Scrucca. “qcc: An R package for quality control charting and statistical process control”. In: *R news* 4/1 (2004), pp. 11–17.