

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

**SCHOOL OF INFORMATION SCIENCES
& TECHNOLOGY**

DEPARTMENT OF STATISTICS

POSTGRADUATE PROGRAM

**Statistical Process Control and Monitoring with Big
Data**

By

Xeni D. Kokkinopoulou

A THESIS

Submitted to the Department of Statistics
of the Athens University of Economics and Business
in partial fulfilment of the requirements for
the degree of Master of Science in Statistics

Athens, Greece
September 2018

**ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ**



ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

**ΣΧΟΛΗ ΕΠΙΣΤΗΜΩΝ & ΤΕΧΝΟΛΟΓΙΑΣ
ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ**

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ

**Στατιστικός Έλεγχος Ποιότητας σε Μεγάλα
Δεδομένα**

Ξένη Δ. Κοκκινοπούλου

ΔΙΑΤΡΙΒΗ

Που υποβλήθηκε στο Τμήμα Στατιστικής
του Οικονομικού Πανεπιστημίου Αθηνών
ως μέρος των απαιτήσεων για την απόκτηση
Μεταπτυχιακού Διπλώματος Ειδίκευσης στη Στατιστική

Αθήνα
Σεπτέμβριος 2018

DEDICATION

To my family

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor, Panagiotis Tsiamyrtzis, who was always supportive and gave me invaluable feedback. To my family and friends, who made this journey possible and enjoyable.

VITA

My name is Xeni Kokkinopoulou. I was born in Athens, Greece. I am a mathematician. Hopefully, I will be a M.Sc. Statistics graduate. I still live in Athens and I am deeply interested in any mathematical and statistical knowledge.

ABSTRACT

Xeni Kokkinopoulou

STATISTICAL PROCESS CONTROL AND MONITORING WITH BIG DATA

September 2018

In this thesis, we examine how the most common control charts behave with big data when a process is in-control as well as in out-of-control states. Additionally, some alternative monitoring schemes are examined. Such schemes include the use of Kolmogorov-Smirnov test, the use of a non-parametric Likelihood Ratio Test for stochastically ordered random variables and the use of Q-Q plots. All these methodologies are non-parametric aiming to benefit from the large volumes of data. We conclude that use of Q-Q plots constitute the most effective methodology for both in-control and out-of-control states of a process.

ΠΕΡΙΛΗΨΗ

Ξένη Κοκκινοπούλου

ΣΤΑΤΙΣΤΙΚΟΣ ΕΛΕΓΧΟΣ ΠΟΙΟΤΗΤΑΣ ΣΕ ΜΕΓΑΛΑ ΔΕΔΟΜΕΝΑ

Σεπτέμβριος 2018

Στην παρούσα διατριβή, εξετάζουμε πώς συμπεριφέρονται τα κλασσικά διαγράμματα ελέγχου με μεγάλα δεδομένα, τόσο όταν μια διαδικασία είναι εντός ελέγχου όσο και όταν αυτή είναι εκτός ελέγχου. Επιπρόσθετα εξετάζονται και κάποια εναλλακτικά σχήματα ελέγχου. Κάποια από αυτά περιλαμβάνουν τον έλεγχο Kolmogorov-Smirnov, ένα μη παραμετρικό έλεγχο λόγων πιθανοφάνειας για στοχαστικά διατεταγμένες τυχαίες μεταβλητές καθώς και την χρήση των διαγραμμάτων ποσοστιαίων σημείων (Q-Q plots). Όλες οι προαναφερθείσες μεθοδολογίες είναι μη παραμετρικές, προκειμένου να ωφεληθούμε όσο το δυνατόν περισσότερο από τους μεγάλους όγκους δεδομένων. Καταλήγουμε ότι τα Q-Q plots είναι η πιο αποτελεσματική μεθοδολογία σε καταστάσεις εντός αλλά και εκτός ελέγχου μιας διαδικασίας.

TABLE OF CONTENTS

	Page
Chapter 1: Introduction	1
Chapter 2: Control Charts.....	5
2.1 Introduction	5
2.2 I-chart, Cusum chart and EWMA chart	5
2.3 Combining an I-chart with a p-chart	12
2.4 Times between events approach	13
Chapter 3: Kolmogorov Smirnov test & alternatives	19
3.1 Introduction	19
3.2 Kolmogorov Smirnov Test	19
3.3. Other versions of Kolmogorov Smirnov test.....	22
Chapter 4: Non-parametric LRT test for stochastically ordered random variables	27
4.1. Introduction	27
4.2. Stochastic Orders of Random Variables	27
4.3 Non-parametric Likelihood Ratio Test for First Ordered Random Variables.....	29
Chapter 5: Q-Q plots in SPC.....	35
5.1. Introduction	35
5.2. Q-Q plots.....	35
5.3. Q-Q plots for SPC	37
Chapter 6: Results of Simulation Studies.....	47
6.1. Introduction	47
6.2. I-chart and X-chart.....	47
6.3 CUSUM chart	54
6.4. EWMA chart.....	59
6.5 Combinations of I-charts with other types of control charts.....	64
6.6 Kolmogorov-Smirnov and its alternatives	72
6.7. Non-parametric LRT for stochastically ordered random variables.....	82
6.8 Q-Q plots.....	85
Chapter 7: Conclusions	93
References	95

LIST OF TABLES

Table	Page
Table 2. 1. In control Average Run Length for Cusum charts.	9
Table 2. 2. In control Average Run Length for EWMA charts.	11
Table 4. 1. The vertices and strings of the observations x_1, \dots, x_7 and y_1, \dots, y_5 for the computation of $LF \geq G$	31
Table 4. 2. Maximum restricted likelihood estimation $LF \geq G$ for the observations x_1, \dots, x_7 and y_1, \dots, y_5	32
Table 4. 3. The vertices and strings of the observations x_1, \dots, x_7 and y_1, \dots, y_5 for the computation of $LG \geq F$	33
Table 4. 4. Maximum restricted likelihood estimation $LG \geq F$ for the observations x_1, \dots, x_7 and y_1, \dots, y_5	33
Table 5. 1. Values of intercept and slope for each case of Q-Q plots.	39
Table 5. 2. Fit of regression models (5.1) and (5.2) on the data of the previous examples.	43
Table 6. 1. Results of simulations for the I-chart with 3σ control limits.	48
Table 6. 2. Results of simulations for the I-chart with 4σ control limits.	48
Table 6. 3. Results of simulations for the I-chart with 4.5σ control limits. ...	48
Table 6. 4. Results of out-of-control simulations for the I-chart with 3σ control limits.	50
Table 6. 5. Results of out-of-control simulations for the I-chart with 4σ control limits.	50
Table 6. 6. Results of out-of-control simulations for the I-chart with 4.5σ control limits.	51
Table 6. 7. Results of simulations for the X -chart with 3σ control limits.	52
Table 6. 8. Results of simulations for the X -chart with 4σ control limits.	52
Table 6. 9. Results of simulations for the X -chart with 4.5σ control limits. ...	53
Table 6. 10. Results of the simulations for the CUSUM chart with $K=0.5$, $H=4.77$	55
Table 6. 11. Results of the simulations for the CUSUM chart with $K=0.5$, $H=5$	55

Table 6. 12. Results of the simulations for the CUSUM chart with $K=0.5$, $H=6$.	56
Table 6. 13. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=4.77$.	57
Table 6. 14. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=5$.	58
Table 6. 15. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=6$.	58
Table 6. 16. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=3$.	60
Table 6. 17. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=3.5$.	60
Table 6. 18. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=4$.	61
Table 6. 19. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=3$.	62
Table 6. 20. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=3.5$.	62
Table 6. 21. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=4$.	62
Table 6. 22. Results of the in-control simulations for the p-chart with 3σ control limits based on an I-chart.	65
Table 6. 23. Results of the out-of-control simulations for the p-chart with 3σ control limits based on an I-chart.	65
Table 6. 24. Results of the in-control simulations for the g-chart with 3σ control limits based on an I-chart.	67
Table 6. 25. Results of the out-of-control simulations for the g-chart with 3σ control limits based on an I-chart.	67
Table 6. 26. Results of the in-control simulations for the probabilistic g-chart based on an I-chart.	69
Table 6. 27. Results of the out-of-control simulations for the probabilistic g-chart based on an I-chart.	69
Table 6. 28. Results of the in-control simulations for a Weibull I-chart based on an I-chart.	70

Table 6. 29. Results of the out-of-control simulations for a Weibull I-chart based on an I-chart.	71
Table 6. 30. Results of simulations for Kolmogorov-Smirnov test statistic D	73
Table 6. 31. Results of simulations for Kolmogorov-Smirnov test statistic $D1$	74
Table 6. 32. Results of simulations for Kolmogorov-Smirnov test statistic $D2$	75
Table 6. 33. Results of simulations for Kolmogorov-Smirnov test statistic $D3$	76
Table 6. 34. Results of simulations for Kolmogorov-Smirnov test statistic $D4$	77
Table 6. 35. Results of simulations for Kolmogorov-Smirnov test statistic $D5$	78
Table 6. 36. Results of simulations for Kolmogorov-Smirnov test statistic $D6$	79
Table 6. 37. Results of simulations for the non-parametric LRT for stochastically ordered random variables.	83
Table 6. 38. Results of the simulations for the $T2$	86
Table 6. 39. Results for the simulations for the M	86

LIST OF FIGURES

Figure	Page
Figure 1. 1.1: The basic characteristics of a control chart	3
Figure 2. 1. An I-chart for 1,000 observations with 3 sigma control limits.	7
Figure 2. 2. An \bar{X} -chart for 1,000 observations grouped by 50 with 3 sigma control limits.....	8
Figure 2. 3. A two-sided Cusum chart with $H=4.77$ and $K=0.5$ for 1,000 observations.....	10
Figure 2. 4. An EWMA chart for 1,000 observations with $L=3$ and $\lambda=0.25$..	11
Figure 2. 5. A monitoring scheme combining an I-chart with a p-chart for 2,000 observations.....	13
Figure 2. 6. A monitoring scheme combining an I-chart and a g-chart for 2,000 observations.....	15
Figure 2. 7. A monitoring scheme combining an I-chart and a probabilistic g-type chart for 2,000 observations.	16
Figure 2. 8. A monitoring scheme combining two I-charts for 2,000 observations.....	17
Figure 3. 1. The ecdf of a sample of 1,000 observations.	20
Figure 3. 2. The empirical c.d.f.s of two samples of 1,000 observations each and Kolmogorov Smirnov test statistic.	21
Figure 3. 3. The empirical c.d.f. of two samples of 1,000 observations each and test statistic $D1$	23
Figure 3. 4. The empirical c.d.f.s of two samples of 100 observations each and test statistic $D3$	24
Figure 4. 1. The least concave majorant of the points x_1, \dots, x_7 and y_1, \dots, y_5	30
Figure 4. 2. The least concave majorants of the points x_1, \dots, x_7 and y_1, \dots, y_5	32
Figure 5. 1. A Q-Q plot for two samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$, following the same distribution.	36
Figure 5. 2. A Q-Q plot for two samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$, following linearly related distributions.....	36

Figure 5. 3. Various cases of Q-Q plots; (a) a typical Q-Q plot, (b) change of the location parameter, (c) change of the scale parameter, (d) change of both the location and scale parameters.....	38
Figure 5. 4. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ from $N(3,3)$ and $y_1, \dots, y_{1,000}$ from $\text{Gamma}(3,1)$	40
Figure 5. 5. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ from $N(0,3)$ and $y_1, \dots, y_{1,000}$ from $t(3)$	40
Figure 5. 6. The fit of the regression models (5.1) and (5.2) on samples $x_1, \dots, x_{1,000}$ from $N(3,3)$ and $y_1, \dots, y_{1,000}$ from $\text{Gamma}(3,1)$	42
Figure 5. 7. The fit of the regression models (5.1) and (5.2) on samples $x_1, \dots, x_{1,000}$ from $N(0,3)$ and $y_1, \dots, y_{1,000}$ from $t(3)$	42
Figure 5. 8. The ellipse created by the values of a and b.....	44
Figure 6. 1. Number of false alarms of an I-chart for various choices of the control limits.....	49
Figure 6. 2. Comparison between the in-control and out-of-control alarm rates of an I-chart with 3σ control limits.	52
Figure 6. 3. Comparison between the in-control and out-of-control alarm rates of an \bar{X} -chart with (a) 3σ , (b) 4σ and (c) 4.5σ control limits.....	54
Figure 6. 4. Number of false alarms of a CUSUM chart for various choices of its control limits.	56
Figure 6. 5. Comparison between the in-control and out-of-control alarm rates of a CUSUM chart with $K=0.5$ and $H=4.77$	59
Figure 6. 6. Number of false alarms of an EWMA chart for various choices of its control limits.	61
Figure 6. 7. Comparison between the in-control and out-of-control alarm rates of an EWMA chart with $\lambda=0.25$ and $L=3$	63
Figure 6. 8. Comparison between an I-chart, an \bar{X} -chart and a CUSUM chart.	64
Figure 6. 9. Comparison of in-control and out-of-control states of a process for the combination of an I-chart and a p-chart.	66
Figure 6. 10. Comparison of in-control and out-of-control states of a process for the combination of an I-chart and a g-chart.	68

Figure 6. 11. Comparison of in-control and out-of-control states of a process for the combination of an I-chart with a probabilistic g-chart.....	70
Figure 6. 12. Comparison of in-control and out-of-control states of a process for the combination of an I-chart with a Weibull I-chart.....	71
Figure 6. 13. Distribution of the test statistic D in various cases.	73
Figure 6. 14. Distribution of the test statistic $D1$ in various cases.	74
Figure 6. 15. Distribution of the test statistic $D2$ in various cases.	75
Figure 6. 16. Distribution of the test statistic $D3$ in various cases.	76
Figure 6. 17. Distribution of the test statistic $D4$ in various cases.	77
Figure 6. 18. Distribution of the test statistic $D5$ in various cases.	78
Figure 6. 19. Distribution of the test statistic $D6$ in various cases.	79
Figure 6. 20. Detection rates of all the Kolmogorov-Smirnov test statistics, when persistent shift of the location parameter occurs.....	81
Figure 6. 21. Detection rates of all the Kolmogorov-Smirnov test statistics, when persistent shift of the scale parameter occurs.	81
Figure 6. 22. Detection rates of all the Kolmogorov-Smirnov test statistics, when outliers are present.....	82
Figure 6. 23. Detection rates of the non-parametric LRT for various sizes of (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter and (c) outliers.....	84
Figure 6. 24. Distribution of the test statistic $T2$ in various cases.....	87
Figure 6. 25. Distribution of the test statistic M in various cases.....	88
Figure 6. 26. Detection rates of the $T2$ test for various lengths of (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter and (c) numbers of outliers.	89
Figure 6. 27. Detection rates of the M test for increasing number of outliers.	89
Figure 6. 28. Distribution of $T2$ via simulations and via resampling.	90
Figure 6. 29. Distribution of M via simulations and via resampling.	91
Figure 6. 30. Comparison of the $D3$ test statistic (Kolmogorov-Smirnov alternative), the non-parametric LRT and the $T2$ test statistic, when either (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter or (c) outliers are present.	92

Chapter 1

Introduction

Statistics is a science that is useful in every field of research and practically in every aspect of our everyday life. It is only natural there is a specific field of statistics whose goal is the improvement of a process. There are all kinds of processes. For instance, with the term “process” we may refer to the production line of an industry. The time needed to serve each customer at a cashier also constitutes a process. Hence, the concept of manufactured products can be either material or immaterial. This field of Statistics is called Statistical Process Control (SPC).

SPC is a method of quality control that uses statistical methods, such as control charts and designed experiments, in order to ensure the quality of the outcoming products of a company. SPC tools are used to control and monitor the process. The use of SPC aims to reduce the number of products that are defective and require scrap or rework. It is also desirable to make the production process steadier. More particularly, the aim is for the process to create products with the lowest variability achievable around a desired target value. The variability of a process may come from different sources. We divide these sources into common and assignable causes (Deming, 1986). Common causes are factors embedded into the process itself and are unavoidable. Assignable causes usually arise from an exterior source, such as operator errors. When we try to eliminate common causes, we change the process. On the contrary, when we eliminate an assignable cause we stabilize the existing process by solving a specific problem. A process is considered to be in statistical control (or simply in-control) when it is operating only under the presence of chance causes of variability. A process is considered to be out-of-control when it operates with the presence of assignable causes of variation. We aim to eliminate assignable causes as soon as possible.

In order to achieve the aforementioned goals, some statistical tools are used. For example, we can use control charts and design of experiments. We

are interested particularly in control charts. These charts are quite simple to construct and interpret. Therefore, they are widely used in industrial and not only practice.

The typical control chart (initiated from W.A Shewhart, 1924) is a graphical display of a quality characteristic measured or computed from a sample versus the sample number or time (Montgomery, 2009). Each control chart contains three horizontal lines. The center line represents the mean of the quality characteristic, measured when the process is in control. The upper line is called the upper control limit (UCL) and the lower line is the lower control limit (LCL). These lines are chosen so that when the process is in-control, almost every sample point plotted will fall between them.

A control chart is constructed in two steps. Firstly, we choose several samples to estimate the center line and the control limits of the chart. These samples constitute Phase I of the application of SPC. At this point, the control limits change with each new sample. When we have sufficient samples to estimate the desired parameters of the control chart, we finalize them and we proceed to Phase II of the controlling scheme. In Phase II, the control limits and the center line of the graph do not change (see Figure 1.1). We keep their values from Phase I. Each new sample is represented with a point on the graph. If a point falls outside the control limits, we conclude that the process is out of control. If a point falls between the control limits, we can assume the process is in control. A control chart is equivalent to a hypothesis testing that examines whether the process is in statistical control or not. By stating a process is out of control, we mean that, for some reason, the parameters of the production process have changed. If a process is determined to be out of control, corrective actions must be taken. When a point plots outside the control limits, while the process is in control, we have a false alarm. The probability of having a false alarm relates to type I error in a hypothesis testing.

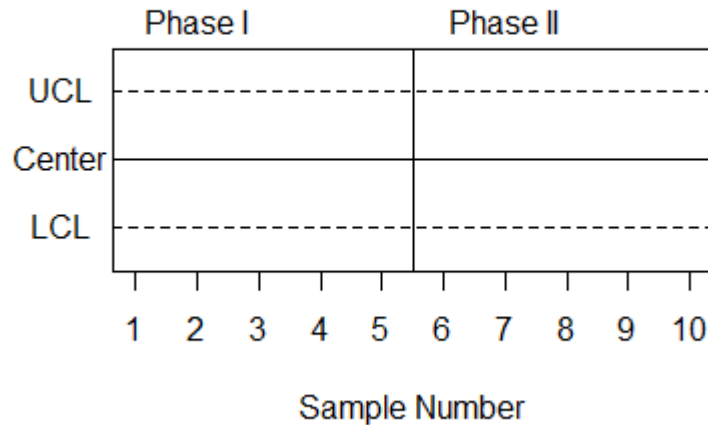


Figure 1. 1: The basic characteristics of a control chart

Big data is a concept we hear about almost every day. Most importantly, it is a concept which is present in almost every activity of our lives, sometimes without even knowing it. Due to the latest technological developments, data are produced from countless devices and applications. What makes these data so significant, is the fact they contain a lot of useful information. Therefore, we try to adapt the traditional methods of statistics to benefit from this enormous amount of information.

The characteristics of big data are volume, variety, velocity, veracity and value (Shen, Kaynak 2015). These characteristics are also known as the 5 V's of big data. Volume refers to the size of the data. Velocity is the speed of the incoming and outgoing data. Variety refers to the type of data. Data can be either of the traditional form, such as numbers, or of more complex forms, such as images, text, etc. Veracity is about the quality of the available data. Value has to do with whether the available data are indeed useful for the desired purposes or not.

In the industrial world, the constant progress of technology has made the production lines much faster. As a result, a very large number of products is manufactured in just a few hours. The produced data can be of great aid in the improvement and monitoring of the producing process, if handled appropriately. The most common forms of big data in the industries are numeric data and image data. The former are further divided into two

categories. The first category contains datasets with a very large number of observations for a small number of variables, also called “tall and thin data” (also known as $n \gg p$). In the SPC context, variables are quality characteristics of the outgoing products. The second category contains datasets with a small number of observations for a very large number of quality characteristics, known as “short and fat data” (or alternatively $n \ll p$). For instance, this case applies to motor industry. A small number of vehicles may be produced daily, but each of them has a very large number of characteristics that must be carefully monitored and controlled.

All these facts show how interesting and useful it is to study further the application of SPC on big data. In this thesis, we focus our attention to “tall and thin data”, i.e. cases with a very large number of observations for a small number of variables. We investigate how the traditional control charts behave with large volumes of incoming data. We especially examine the rate of false alarms in control charts. We also put to the test some alternative methods for controlling and monitoring a process. All methods are calibrated for a predetermined in control performance and tested for several out of control scenarios. Furthermore, there will be a new (non-parametric) proposal based on QQ plots, which can be used to efficiently monitor processes that have an excessive number of observations.

The four following chapters describe each method examined and, in the final chapter, all the methods are evaluated via simulations. More specifically, in Chapter 2 we examine several versions of control charts. In Chapter 3, we investigate an application and some variants of Kolmogorov Smirnov test. In Chapter 4, we implement a non-parametric likelihood ratio test for stochastically ordered random variables. In Chapter 5, we present the new proposed methodology, which is based on QQ plots. Lastly, in Chapter 6, we present the results of the simulations for each of the aforementioned methods. In addition, the final chapter contains comparisons between all of the tests along with conclusions and future work.

Chapter 2

Control Charts

2.1 Introduction

In this chapter, we will present the properties of some basic control charts used in SPC. These charts are the following; the chart for individual measurements (I-chart), the Cumulative Sum chart (Cusum chart) and the Exponentially Weighted Moving Average chart (EWMA chart). We also explore the behavior of some monitoring schemes based on combining two control charts. These schemes consist of an I-chart and either a p-chart or charts, which monitors the times between two consecutive alarms.

2.2 I-chart, Cusum chart and EWMA chart

The first chart examined is the individual measurements chart (I-chart). I-chart is appropriate when data come from a Normal distribution and are uncorrelated. We consider each data point as a sample of size one from the process. The formulas used to construct the control limits of the chart are the following (Montgomery,2009):

$$\begin{aligned} UCL &= \bar{x} + 3 * \frac{\overline{MR}}{d_2} \\ \text{Center Line} &= \bar{x} \\ LCL &= \bar{x} - 3 * \frac{\overline{MR}}{d_2} \end{aligned} \quad (2.1)$$

where: $\bar{x} = \frac{1}{n} * \sum_{i=1}^n x_i$ (2.2),

$$MR_i = |x_i - x_{i-1}|, i = 2, \dots, n \quad (2.3)$$

$$\overline{MR} = \frac{1}{n-1} * \sum_{i=2}^n MR_i \quad (2.4),$$

d_2 is a constant that depends on the sample size (Montgomery, 2009, Appendix VI).

Since we assume the data are normally distributed, we observe that the control limits are three standard deviations away from the mean, hence the term $\frac{\overline{MR}}{d_2}$ is multiplied by 3. In this case, we say we have a control chart with 3

sigma control limits (see Figure 2.1). A measure used to estimate the false alarm rate is Average Run Length (ARL). We can define ARL for two cases: when the process is in control (ARL_0), and when the process is out of control (ARL_1). ARL_0 expresses the expected number of observations in control before an alarm occurs, when the process is in control. We can compute ARL_0 by using the formula:

$$ARL_0 = \frac{1}{p} \quad (2.5)$$

where p denotes the probability that a data point falls outside the control limits. Assuming normal distribution, we are practically interested in computing the probability an observation is further than three standard deviations from its mean. This probability is equal to $p = 0.27\%$. Then $ARL_0 = \frac{1}{0.0027} = 370$. Some concerns regarding ARL_0 arise due to the fact its distribution is Geometric with probability of success p . Consequently, ARL_0 has a large standard deviation. More specifically, its standard deviation is equal to: $\sigma(ARL_0) = \sqrt{\frac{1-p}{p^2}} \approx 370$. Furthermore, since the Geometric distribution is quite skewed, its mean is not a representative measure. Practically, among 1,000 observations we expect 2 to 3 false alarms. However, if we combine the large standard deviation of ARL_0 and the large datasets available, we will encounter problems regarding the false alarm rate. More analytically, we expect to observe an increasing number of false alarms when the sample size increases. A first attempt to fix this problem is choosing the control limits further from the center line than 3 standard deviations. Therefore, we will implement the I-chart with the control limits given below:

$$\begin{aligned} UCL &= \bar{x} + L * \frac{\overline{MR}}{d_2} \\ \text{Center Line} &= \bar{x} \\ LCL &= \bar{x} - L * \frac{\overline{MR}}{d_2} \end{aligned} \quad (2.6)$$

where $L=4, 4.5$. For the case where $L=4$, the probability that a point is plotted outside the control limits is $p=0.063\%$, hence $ARL_0 \approx 15,787$. For the case where $L=4.5$, the probability that a point is plotted outside the control limits is $p=0.007\%$, hence $ARL_0 \approx 147,159$. These alterations will decrease the

number of false alarms. Nonetheless, this decrease is not satisfactory for two main reasons. Firstly, the number of false alarms will still increase with the increase of sample size. Secondly, the detection of an out of control state will be more difficult (decreasing the type I error leads to lower power). It is also worth noting the size of Phase I is 1,000 observations in all the aforementioned charts. With this number of observations in Phase I, we can have quite accurate estimates of the parameters of the control chart. Moreover, it is not a large number of observations compared to the total size of the datasets. Recall that Phase I is the stage where we construct the limits of a control chart and we keep their values constant in Phase II of the monitoring procedure.

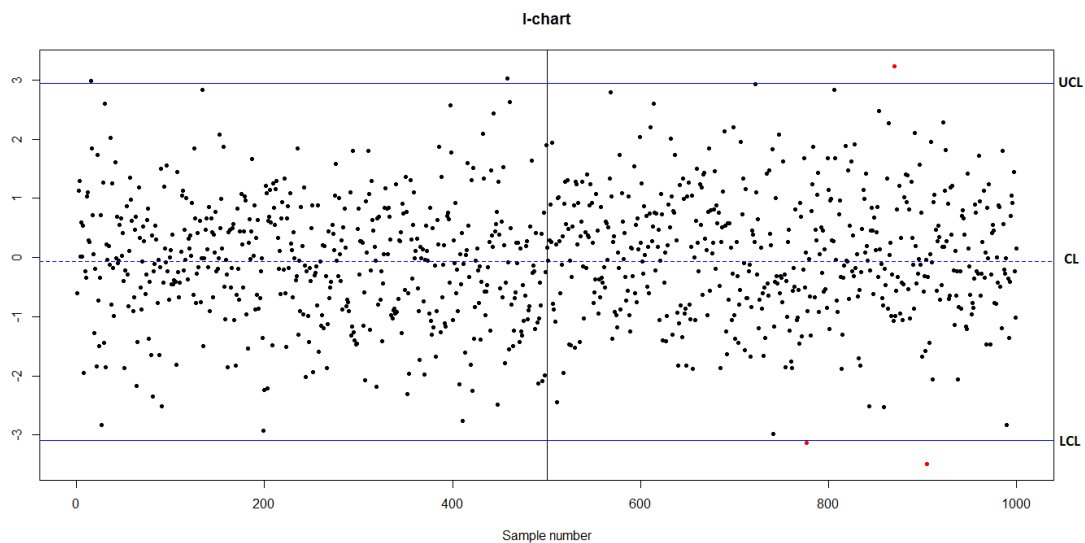


Figure 2. 1. An I-chart for 1,000 observations with 3 sigma control limits.

We could also try grouping the observations and create an \bar{X} -chart of the formed subgroups. An \bar{X} -chart is used to monitor the mean value of a subgroup of size n from the data. An \bar{X} -chart has the following control limits (Montgomery, 2009):

$$\begin{aligned}
 UCL &= \bar{\bar{x}} + L * \frac{\bar{R}}{d_2\sqrt{n}} \\
 \text{Center Line} &= \bar{\bar{x}} \\
 LCL &= \bar{\bar{x}} - L * \frac{\bar{R}}{d_2\sqrt{n}}
 \end{aligned}
 \quad (2.7)$$

where: $\bar{\bar{x}} = \frac{1}{n*m} * \sum_{i=1}^n \sum_{j=1}^m x_j^i$ (2.8)

$$\bar{R} = \frac{1}{m} * \sum_{i=1}^m (x_{i,max} - x_{i,min}) \quad (2.9)$$

L determines the distance of the control limits from the center line,

L=3, 4 or 4.5

d_2 denotes a constant which depends on the sample size of each subgroup (Montgomery, 2009, Appendix VI).

The observations can be divided into subgroups of 50 or 100 data points (see Figure 2.2). L was given three different values resulting in control charts with 3σ , 4σ and 4.5σ control limits respectively. We expect similar behavior from the \bar{X} -chart to the I-chart. We shall implement a Phase I of 1,000 observations for this chart as well.

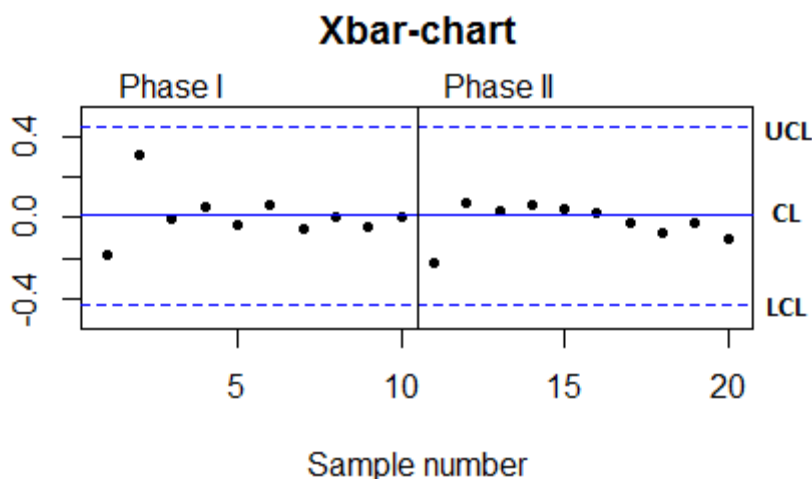


Figure 2. 2. An \bar{X} -chart for 1,000 observations grouped by 50 with 3 sigma control limits.

Our next step is to examine how the cumulative sum chart (Cusum chart), introduced by Page (1954) reacts to large datasets. In general, this chart is preferred to the I-chart when we handle observations individually and the focus is in detecting (small/moderate) step changes. We will present the basic features of a Cusum chart for the process mean. Let us denote by μ_0 the target value for the process mean and let x_i be the i th observation of the process. We assume that the data are normally distributed. When the process is in control, x_i follows a normal distribution with mean μ_0 and standard deviation σ . We also assume that σ is known or that a reliable estimate is found. Then, instead of plotting the mean of each sample, we plot the quantities defined below against the sample number i (Montgomery, 2009):

$$C_i^+ = \max[0, x_i - (\mu_0 + K) + C_{i-1}^+] \quad (2.10)$$

$$C_i^- = \max[0, (\mu_0 - K) - x_i + C_{i-1}^-] \quad (2.11)$$

where, $C_0^+ = C_0^- = 0$. C_i^+ is called one-sided upper Cusum and C_i^- is called one-sided lower Cusum. In practice, we typically use a double-sided Cusum chart. The values of C_i^+ are plotted on a graph as computed from the above formula and instead of using the values C_i^- , the values $-C_i^-$ are plotted. K is called the reference value and is related to the magnitude of the shift we aim to detect. More specifically, K is one-half of the magnitude of the shift that we aim to detect and is given by the formula $K = \frac{|\mu_1 - \mu_0|}{2} = \frac{\delta}{2} \sigma$ where μ_1 denotes the shifted mean. We often express μ_1 in standard deviation units with the formula $\mu_1 = \mu_0 + \delta * \sigma$, hence the third part of the previous formula. In order to decide whether a process is in control or not, we define a decision interval H . When a point is plotted outside the decision interval H , we may conclude the process is out of control. We choose H so that the Cusum chart has a predetermined in control average run length performance. A common choice for its value is 5 times the standard deviation of the process. We implemented to our data Cusum charts with $K=0.5$ and $H=4.77, 5$ and 6 (see Figure 2.3). The ARL_0 for each case is shown in the table below (Farouk, Mohamad, 2012).

<i>K</i>	<i>H</i>	<i>ARL₀</i>
0.5	4.77	370
0.5	5	465
0.5	6	1,318

Table 2. 1. In control Average Run Length for Cusum charts.

These charts are expected to display a similar behavior as the previous charts used, regarding the number of false alarms when the sample size increases. The size of Phase I will be 1,000 observations for these charts too.

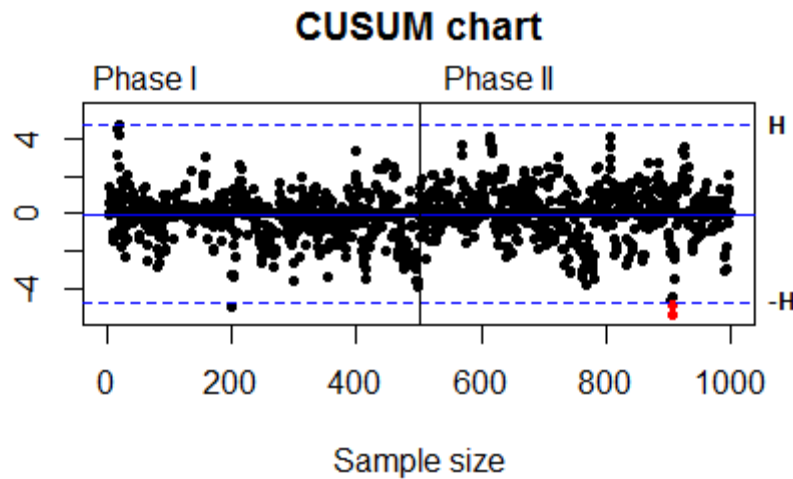


Figure 2. 3. A two-sided Cusum chart with $H=4.77$ and $K=0.5$ for 1,000 observations.

Another widely used control chart for individual observations is the Exponentially Weighted Moving Average chart (EWMA chart). It is a very useful chart for detecting small process shifts and it is known to be somewhat robust in the presence or autocorrelation (Montgomery, 2009). Let x_i denote the i th observation of the process. We define the exponentially weighted moving average as:

$$z_i = \lambda * x_i + (1 - \lambda) * z_{i-1} \quad , \quad (2.12)$$

where λ is a constant such that $0 < \lambda \leq 1$ and we also define $z_0 = \mu_0$, where μ_0 denotes the target mean of the process. So instead of plotting the observations x_i on a graph, we plot the points z_i . It is worth mentioning that EWMA is more robust to violations regarding Normality. The control limits of this chart are given by the following formulas (Montgomery,2009):

$$\begin{aligned} UCL &= \mu_0 + L\sigma \sqrt{\frac{\lambda}{(2-\lambda)} [1 - (1-\lambda)^{2i}]} \\ Center\ Line &= \mu_0 \\ LCL &= \mu_0 - L\sigma \sqrt{\frac{\lambda}{(2-\lambda)} [1 - (1-\lambda)^{2i}]} \end{aligned} \quad (2.13).$$

L is the factor, which determines the width of the control limits. If we observe closely the above formulas, we will notice the UCL and LCL depend on the

sample number i . As i increases, the term $(1 - \lambda)^{2i}$ approaches zero. This means after several samples, the control limits will stabilize to the values

$$UCL = \mu_0 + L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}} \quad \text{and} \quad LCL = \mu_0 - L\sigma\sqrt{\frac{\lambda}{(2-\lambda)}} \quad (2.14)$$

respectively. We implemented the EWMA charts with $\lambda=0.25$ and $L=3, 3.5$ and 4 (see Figure 2.4). The in control ARL for each case is shown in the following table (Crowder, 1987):

λ	L	ARL_0
0.25	3	502.90
0.25	3.5	2,640.16
0.25	4	18,069.9

Table 2. 2. In control Average Run Length for EWMA charts.

We expect similar behavior from these graphs just as with the previously examined ones, regarding the occurring number of false alarms. Phase I of each chart will consist of 1,000 observations, as was the case for the previous control charts.

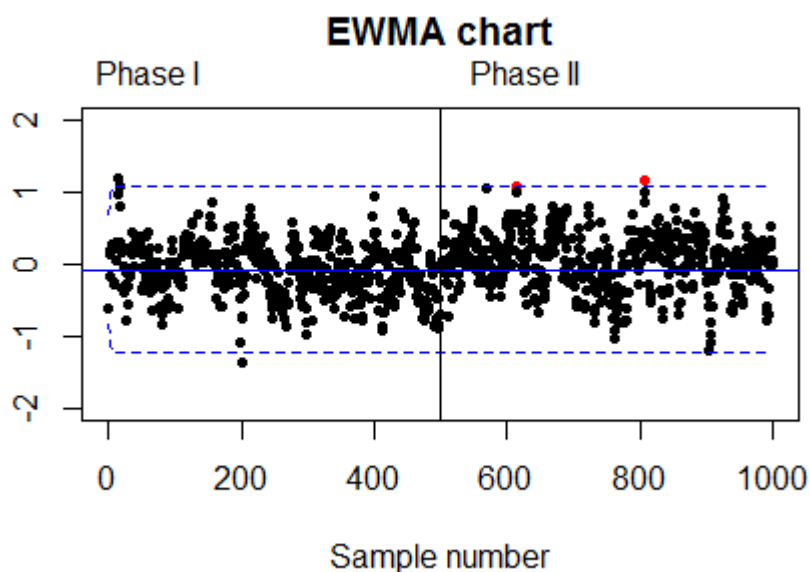


Figure 2. 4. An EWMA chart for 1,000 observations with $L=3$ and $\lambda=0.25$.

2.3 Combining an I-chart with a p-chart

The next step includes a combination of control charts. We will use an I-chart, as described in the previous section. We use a chart to monitor the number of points outside the control limits for a subgroup of observations. Each subgroup contains 100 observations. Each observation is plotted on an I-chart and we keep the number of points outside the control limits per 100 observations. Through this procedure, we compute the percentages of points beyond the control limits and use a control chart to monitor these quantities. This chart is called the control chart for the fraction nonconforming (p-chart). The statistical assumptions made are that the data follow a Binomial distribution. So, we need to know the probability of success is stable throughout the process and the observations are independent. In this case, a data point plotted outside the control limits of the I-chart is considered a success. Notice these conditions are satisfied when the process operates in an in-control state. The control limits of this chart are specified below (Montgomery, 2009):

$$\begin{aligned}
 UCL &= \bar{p} + 3 * \sqrt{\frac{\bar{p} * (1 - \bar{p})}{n}} \\
 \text{Center Line} &= \bar{p}
 \end{aligned}
 \tag{2.15}$$

$$LCL = \bar{p} - 3 * \sqrt{\frac{\bar{p} * (1 - \bar{p})}{n}}$$

where: $\hat{p}_i = \frac{D_i}{n}$, $i = 1, 2, \dots, m$ and $\bar{p} = \frac{\sum_{i=1}^m D_i}{m * n} = \frac{\sum_{i=1}^m \hat{p}_i}{m}$ (2.16)

D_i denotes the number of points in a subgroup of size n plotted outside the control limits in an I-chart.

In the previous formulas, the term $\sqrt{\frac{\bar{p} * (1 - \bar{p})}{n}}$ is multiplied by 3. This means we use a p-chart with 3 sigmas control limits. LCL may sometimes take a negative value. In these cases, we set the LCL equal to 0 (see Figure 2.5). For this scenario, we may use 2,000 observations as Phase I. This number of observations corresponds to 20 points on the p-chart for Phase I.

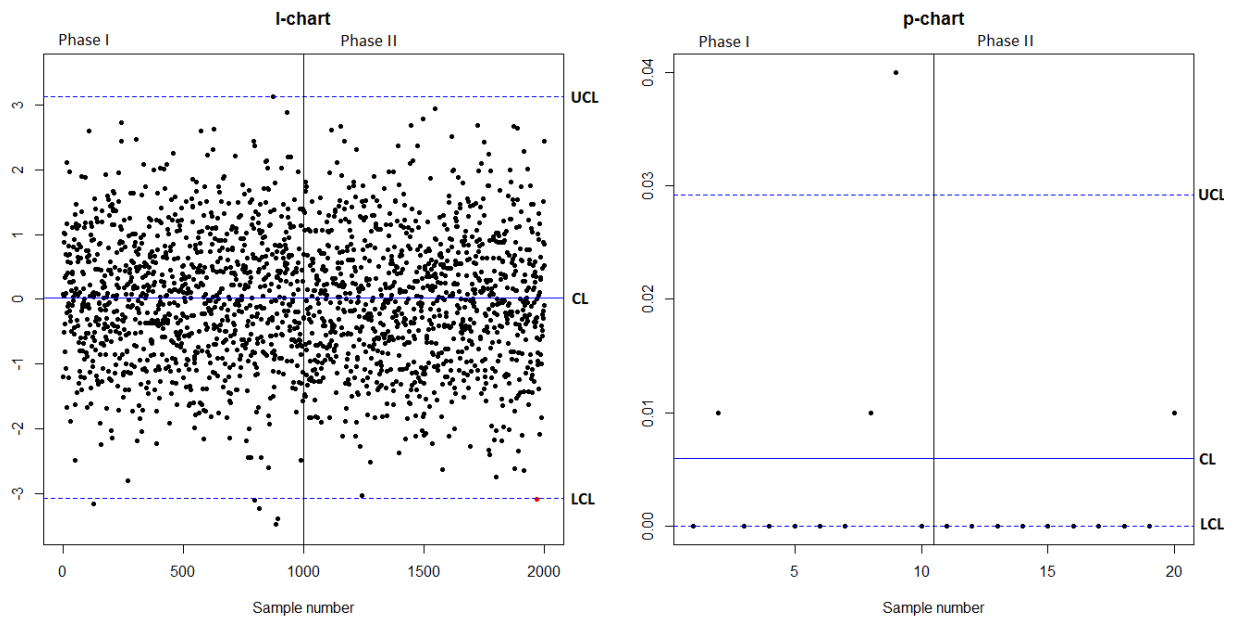


Figure 2. 5. A monitoring scheme combining an I-chart with a p-chart for 2,000 observations.

2.4 Times between events approach

Another approach, which may prove useful, is monitoring the times between consecutive alarms. The term “times” refers to the number of samples observed in between consecutive alarms (g-chart). It is essentially a control chart to monitor the number of non-events until one event occurs. Consider the following experiment implemented on the data of an I-chart. For each point, we check whether it is between the control limits or beyond them. This procedure can be modeled as a Bernoulli trial. Recall that success is observed when a data point falls outside the control limits. Using this point of view, we have consecutive independent Bernoulli trials and we are interested in the distribution of the number of trials needed until the first success. This is a Geometric distribution. However, the quantity we need for the described chart is the sum of these variables. The sum of independent Geometric variables follows a Negative Binomial distribution. This distribution will be used to compute the control limits. More particularly, the control limits of the g-chart will be the following (Benneyan, 2001):

$$\begin{aligned}
 UCL &= \bar{\bar{x}} + 3 * \sqrt{\bar{\bar{x}} * (\bar{\bar{x}} - 1)} \\
 \text{Center Line} &= \bar{\bar{x}}
 \end{aligned}
 \tag{2.17}$$

$$LCL = \bar{\bar{x}} - 3 * \sqrt{\bar{\bar{x}} * (\bar{\bar{x}} - 1)}$$

where: $\bar{\bar{x}} = \frac{1}{n*m} * \sum_{i=1}^n \sum_{j=1}^m x_j^i$.

Recall that x_i denotes the number of points until a point is plotted outside the limits of an I-chart. In this particular case, we take the observations individually. Hence, the control limits are converted to:

$$UCL = \bar{\bar{x}} + 3 * \sqrt{\bar{\bar{x}} * (\bar{\bar{x}} - 1)}$$

$$Center\ Line = \bar{\bar{x}} \quad (2.18).$$

$$LCL = \bar{\bar{x}} - 3 * \sqrt{\bar{\bar{x}} * (\bar{\bar{x}} - 1)}$$

If LCL is negative, we could set it equal to 0 (see Figure 2.6). However, it will not be useful for the purposes of the monitoring procedure. More specifically, we are mainly interested in the lower control limit. Alarms occurring more often than expected could be a sign the process is out of control, while alarms occurring less often is not a concerning phenomenon. Therefore, a positive lower control limit is of great importance. Another possible problem with this chart is the size of Phase I. This problem occurs due to plotting the number of observations between consecutive alarms versus the number of the alarms and not versus the sample number. For instance, we can use, for Phase I, the data until 25 alarms occur. Yet we cannot surely know how many original observations will be needed until this number of alarms is observed.

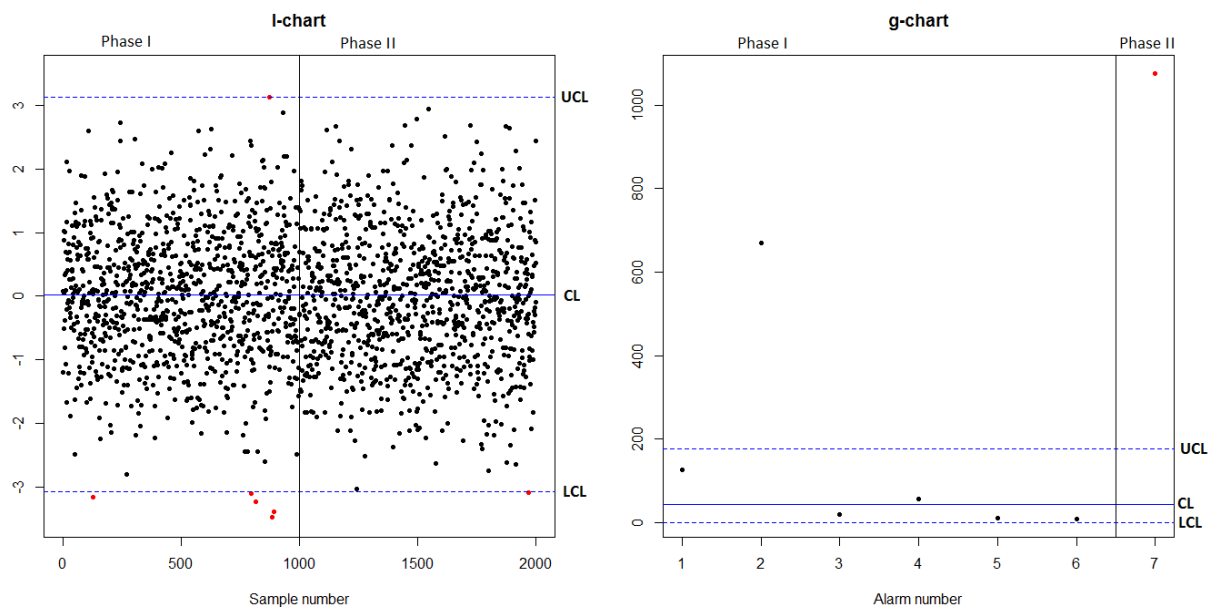


Figure 2. 6. A monitoring scheme combining an I-chart and a g-chart for 2,000 observations.

For the aforementioned reasons, we will use an additional chart based on the Geometric distribution. Its design is different from the previous control charts. This graph will contain only a lower control limit, which will be computed in a probabilistic way. More analytically, the probability mass function of a random variable $X \sim \text{Geometric}(p)$ is:

$$P(X = x) = (1 - p)^{x-1} * p, \quad x = 1, 2, 3, \dots \quad (2.19)$$

and its cumulative distribution function is

$$P(X \leq x) = 1 - (1 - p)^x, \quad x = 1, 2, 3, \dots \quad (2.20).$$

If we set $P(X \leq x) = 0.0027$, as a typical control chart, we have:

$$P(X \leq x) = 1 - (1 - p)^x \Leftrightarrow (1 - p)^x = 1 - P(X \leq x) \Leftrightarrow x = \frac{\log(1 - P(X \leq x))}{\log(1 - p)} \Leftrightarrow x = \frac{\log(1 - 0.0027)}{\log(1 - p)} \quad (2.21).$$

If the initial control chart is an I-chart with 3 sigma control limits, then $x = 1$ (see Figure 2.7). In this case, we can set the lower control limit equal to

$$LCL = 1 + \varepsilon, \quad 0 < \varepsilon \ll 1 \quad (2.22).$$

If the initial control chart is an I-chart with 4 sigma control limits, then $x \approx 43$. Lastly, if the initial control chart is an I-chart with 4.5 sigma control limits, then $x \approx 398$. Phase I is not needed for this chart. The control limit is determined in a probabilistic way, according to the control limits of the initial I-chart.

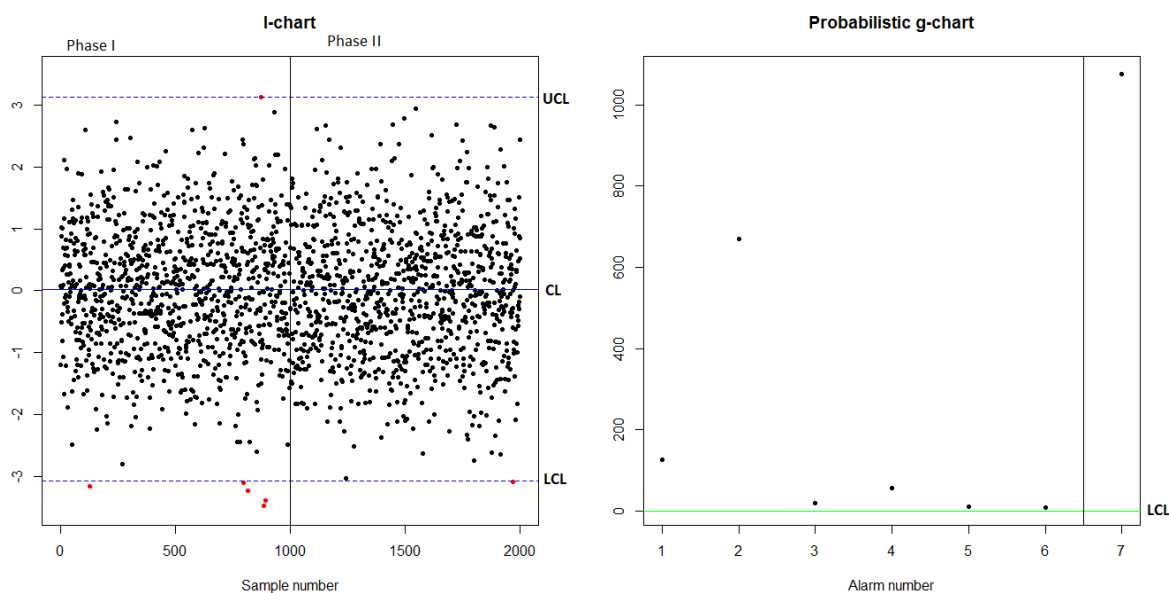


Figure 2. 7. A monitoring scheme combining an I-chart and a probabilistic g-type chart for 2,000 observations.

A more common approach to modelling the time between events is by considering these times as a continuous variable. When this assumption is made, the resulting variable follows the Exponential distribution. The underlying hypothesis is that the number of the observed alarms for a standard number of observations follows a Poisson distribution. Let x_i denote the number of observations in between consecutive alarms. Then x_i is an Exponential random variable. A control chart for these variables cannot be designed, due to the characteristics of the Exponential distribution. This distribution is highly skewed. Thus, symmetric control limits are not appropriate for the construction of a control chart. Nevertheless, a control chart with asymmetric control limits would be more difficult in terms of interpretation for a person who is not familiar with statistics. Fortunately, there is a solution to this problem. Instead of constructing a control chart for x_i , we transform their values by using the formula (Montgomery, 2009):

$$y_i = x_i^{1/3.6} = x_i^{0.2777} \quad (2.23).$$

If x is an Exponential random variable, then y is a Weibull random variable. The resulting distribution can be well approximated by a Normal distribution. Since we can assume a Normal distribution, we may use an I-chart for the transformed variables y_i . In conclusion, this monitoring scheme uses two I-charts. The first I-chart plots the actual observations of the production process. The second I-chart plots the transformed number between consecutive alarms observed on the first control chart. The size of Phase I is a potential issue for the second I-chart as well. The possible problems are the same as stated before for the g-chart.

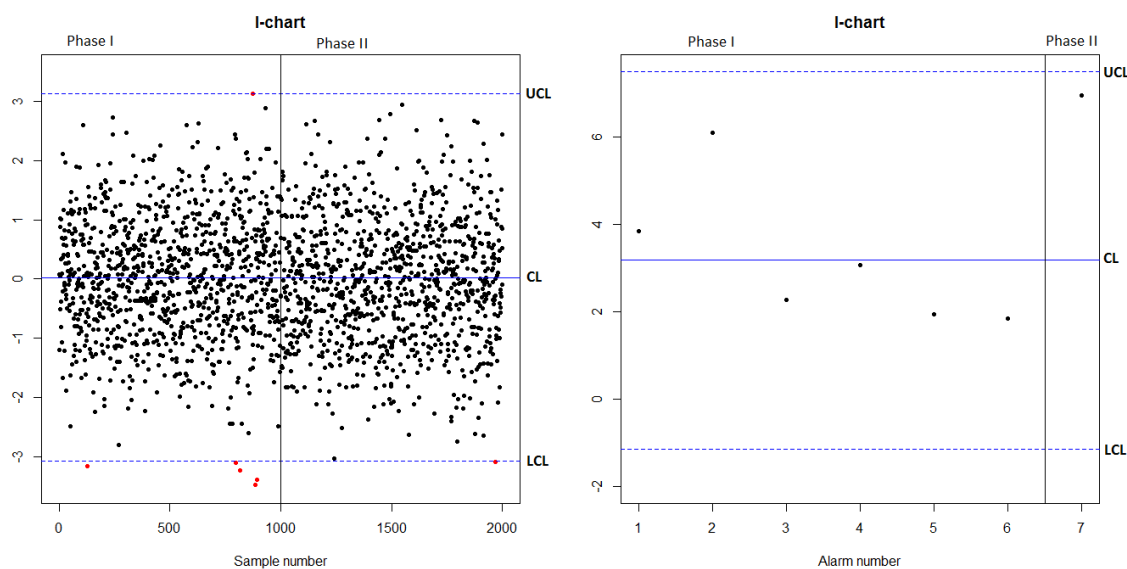


Figure 2. 8. A monitoring scheme combining two I-charts for 2,000 observations.

One may notice that two different approaches are used to model the number of samples between consecutive alarms. The first approach (g-chart), which was based on the Negative Binomial distribution, assumes time is discrete, while the second approach (I-chart), based on the transformation of Exponential random variables, assumes time is continuous. Both methods were used in order to determine which is more accurate and gives us a better understanding of the behavior of the data.

Chapter 3

Kolmogorov Smirnov test & alternatives

3.1 Introduction

In this chapter we shall study the application of Kolmogorov Smirnov test in the SPC context. We will also examine some alternative forms of its test statistic, which have the potential to ensure better detection of shifts in the process parameters.

3.2 Kolmogorov Smirnov Test

In the previous chapter, traditional methods of SPC were put to the test with big data. More analytically, we studied the behavior of basic control charts used in SPC with big data. All these charts use strict statistical assumptions. For instance, in order to implement an I-chart, we need to know the data follow a normal distribution. Since large datasets are available, we shall follow an alternative approach for their analysis.

The major advantage of large volumes of data is that we can have quite reliable estimates of their characteristics. Those characteristics can be either the distribution moments, such as the mean and the variance, or even the cumulative distribution function. If we compute the empirical cumulative distribution function (ecdf) of a large sample, it will appear as if it is a continuous curve. Recall that if we denote by F_n the empirical cdf of a random sample X_1, X_2, \dots, X_n from distribution function (df) F , then:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^n I_{(X_i \leq x)}(x) \quad (3.1)$$

where $I_{(X_i \leq x)}$ is the indicator function with

$$I_{(X_i \leq x)}(x) = \begin{cases} 1, & \text{if } X_i \leq x \\ 0, & \text{else} \end{cases} \quad (3.2)$$

(see Figure 3.1). The empirical cumulative distribution function F_n is a consistent estimator of the cumulative distribution function F (Van der Vaart, 2000).

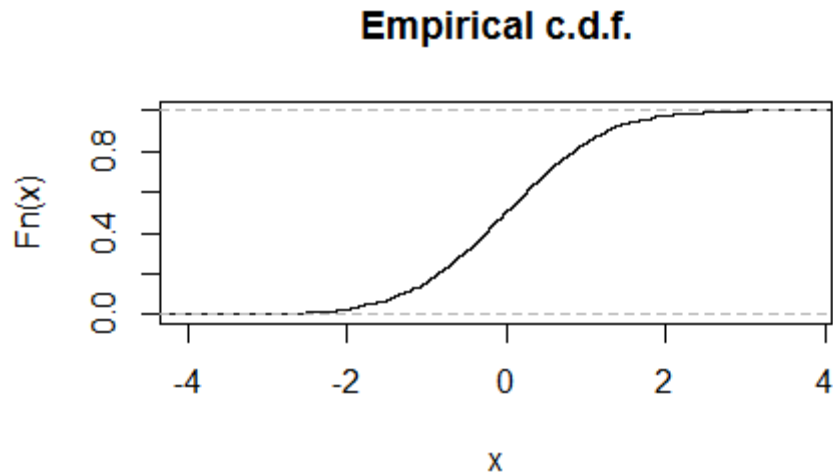


Figure 3. 1. The ecdf of a sample of 1,000 observations.

A quite popular test which uses the empirical cdf of a sample is the Kolmogorov Smirnov test. This test was introduced by Kolmogorov and Smirnov around 1930 and can be used in a twofold way; either with one sample or with two samples. Its first form is used to determine whether we can assume if a sample follows a specific theoretic distribution or not. Its second form is used to examine whether two independent samples come from the same distribution or not. Our interest is focused on Kolmogorov Smirnov test for two independent samples. More analytically, let X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m be two independent random samples from the distribution functions F and G respectively. We wish to perform the following test for F and G ; $H_0: F = G$ versus $H_1: F \neq G$. The test statistic for this hypothesis is given by the formula (Corder & Foreman, 2014):

$$D_{m,n} = \sup_x |F_m(x) - G_n(x)| \quad (3.3)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively. In practice, the quantity

$$D = \max(|F_m(x) - G_n(x)|) \quad (3.4)$$

is used instead of $D_{m,n}$, since maximum is the discrete analogue for supremum (see Figure 3.2). The distribution of D is not known. Therefore, we shall compute its distribution via simulations. The critical values of the test statistic will be chosen so that the probability of type I error is equal to α . In

other words, we will choose c_α such that $P(D \geq c_\alpha) = \alpha$. For instance, if $\alpha=0.05$, then c_α will be the 95% quantile of the distribution of D .

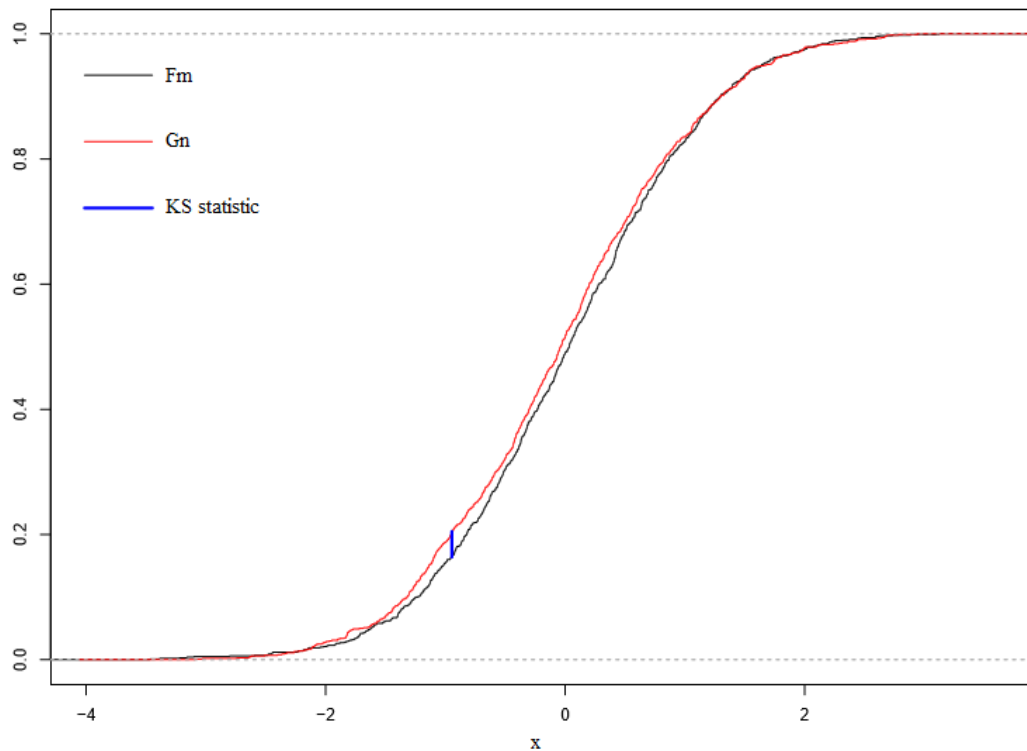


Figure 3. 2. The ecdfs of two samples of 1,000 observations each and Kolmogorov Smirnov test statistic.

Having presented the basic properties of Kolmogorov Smirnov test, we are ready to proceed to its implementation on SPC. We divide the available observations into subsamples of 1,000 observations. The first subsample is considered to be a prototype sample in the sense that we assume it is produced when the process is in-control. Every other subsample will be compared to it. We should note that each subsample needs to be sorted. Then, for every pair of subsamples the Kolmogorov Smirnov statistic D is computed. If an observed D is greater than the defined critical value c_α , then there is evidence that the process is out of control.

Essentially, in this monitoring scheme we replace the concept of control charts with another methodology. Instead of monitoring a specific characteristic of the process, such as the mean or the variance, we use a broader image of the distribution of the data; their cumulative distribution

function. Therefore, instead of comparing the first or second moment of the distribution of two different datasets we compare their cdfs. A control chart is replaced with a hypothesis testing. Since we use the actual distribution of the test statistic, reduced false alarm rates are expected.

3.3. Other versions of Kolmogorov Smirnov test

One potential drawback of the Kolmogorov Smirnov test is the absence of $x_i, i = 1, \dots, n$ for the computation of its statistic. In order to compute the test statistic D , we do not use the actual values of the observed data. This fact may cause difficulty in identifying a shift in the process, that is an out-of-control state of the process.

As previously stated, there are several out-of-control conditions of a process, which may occur due to assignable causes. We will mainly examine three types of out-of-control scenarios. The first one concerns the presence of at least one outlier in a subsample. The term outlier refers to the presence of individual observations generated from a different distribution than the in-control one. For instance, let us suppose the in-control distribution is Normal with mean μ_0 and standard deviation σ . An outlier would be an observation from a Normal distribution with mean $\mu_1 = \mu_0 + k\sigma$, for large k and standard deviation σ . Thus, we consider a large shift in the mean of the process for several individual observations. The second one refers to the persistent shift of a characteristic of the process. More particularly, a change in the mean or the variance of the process may occur persistently in a sample. For instance, let us again assume the in-control distribution is Normal with mean μ_0 and standard deviation σ and let us have a sample of 1,000 observations. Then a persistent shift is considered to occur at this sample, if say its last 100 observations are generated from a Normal distribution with mean $\mu_2 = \mu_0 + k\sigma$, for small k and standard deviation σ . We would consider a persistent shift has occurred if the last 100 observations are generated from a Normal distribution with mean μ_0 and standard deviation $\sigma_1, \sigma_1 > \sigma$. We are interested in detecting any of the aforementioned out-of-control scenarios. Hence, we present some alternative test statistics, which may prove more effective for

the desired purposes. The first suggested alternative is given by the following formula (instead of test statistic D):

$$D_1 = \sum_{i=1}^{m+n} |F_m(x_i) - G_n(x_i)| \quad (3.5)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively (Figure 3.3). Recall that X_1, X_2, \dots, X_n and Y_1, Y_2, \dots, Y_m are two independent random samples from distribution functions F and G respectively and we use the joint sorted sample to compute the test statistic. The particular alternative is expected to be more sensitive to process shifts, because it takes into account the total deviation of the two ecdfs. We operate with D_1 as we described above when using D .

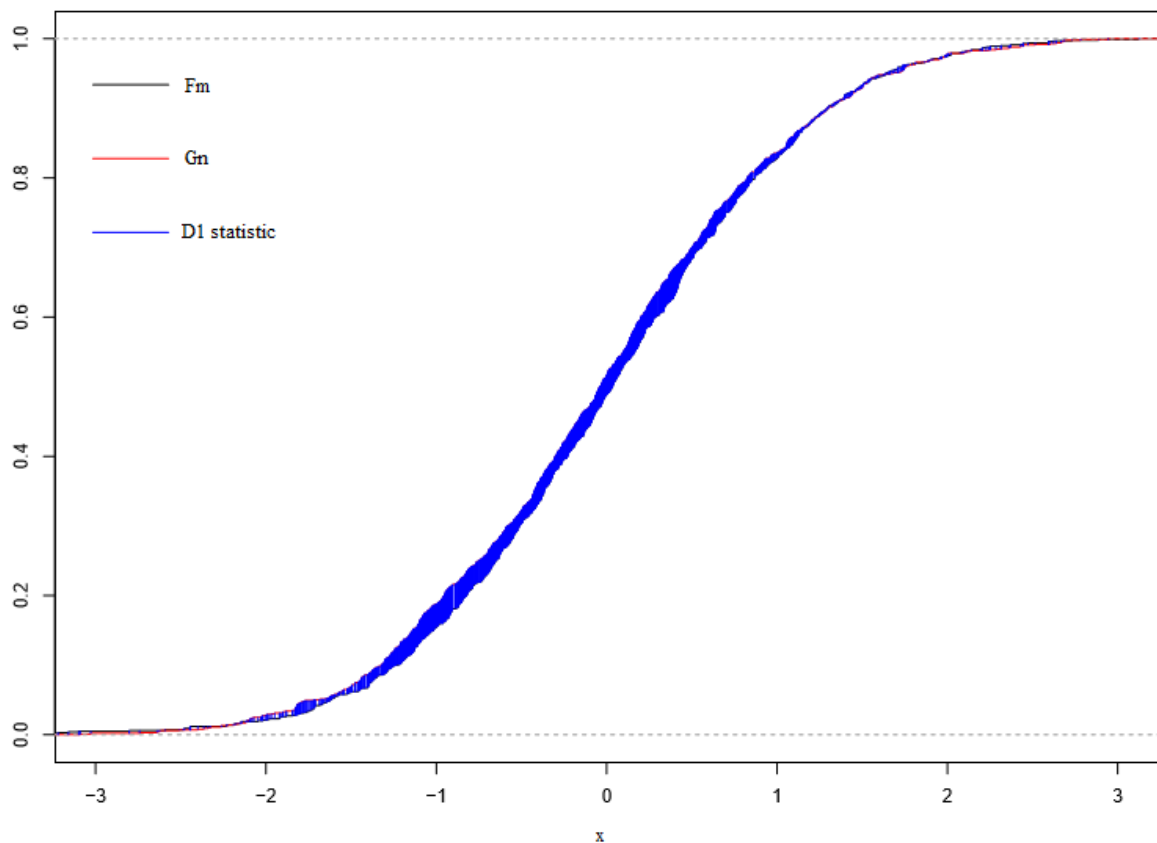


Figure 3. 3. The ecdfs of two samples of 1,000 observations each and test statistic D_1 .

The second suggested alternative is given by the following formula:

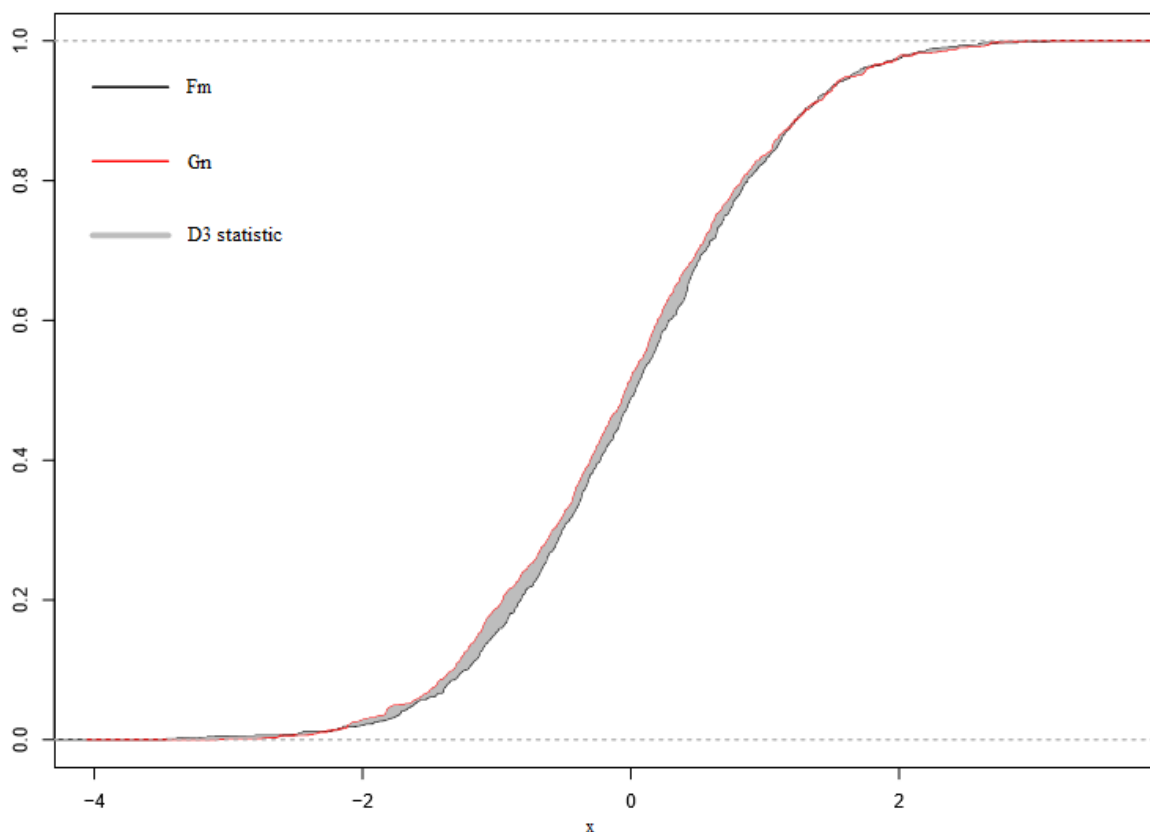
$$D_2 = \frac{1}{m+n} \sum_{i=1}^{m+n} (F_m(x_i) - G_n(x_i))^2 \quad (3.6)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively. This test statistic expresses the mean square deviation between the two curves.

The third suggested alternative is given by the following formula:

$$D_3 = \sum_{i=2}^{m+n} \frac{|F_m(x_i) - G_n(x_i)| + |F_m(x_{i-1}) - G_n(x_{i-1})|}{2} (x_i - x_{i-1}) \quad (3.7)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively (Figure 3.4). This formula computes approximately the total area between the two curves. We approximate the total area between the two curves by adding the areas created by the curves and two consecutive x_i 's.



The formula used is that of the area of a trapeze.

Figure 3. 4. The ecdfs of two samples of 100 observations each and test statistic D_3 .

The fourth suggested alternative is given by the following formula:

$$D_4 = \max(x_i - x_{i-1}) \sum_{i=1}^{m+n} |F_m(x_i) - G_n(x_i)| \quad (3.8)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively. By using the term $\max(x_i - x_{i-1})$ in formula (3.8), we attempt to increase the sensitivity of the test statistic against the detection of outliers.

The fifth suggested alternative is given by the following formula:

$$D_5 = \sum_{i=1}^{m+n} (F_m(x_i) - G_n(x_i)) \quad (3.9)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively. Test statistic D_5 is quite similar to test statistic D_1 . Their difference lies in the absence of the absolute value in D_5 . Consequently, the signs of the differences are taken into account. We wish to examine if there is an impact of these signs on the detection of a process shift.

The sixth and final suggested alternative is given by the following formula:

$$D_6 = \max(F_m(x_i) - G_n(x_i)) \quad (3.10)$$

where F_m and G_n are the empirical cumulative distribution functions for F and G respectively. This test statistic is used to perform the following one-sided hypothesis testing (Sriboonchita et al., 2009): $H_0: F \leq G$ versus $H_1: F > G$.

The performance of Kolmogorov Smirnov test and of each of its alternatives will be studied via simulations in Chapter 6. We will test the statistics' performance for in-control scenarios as well as for the aforementioned out-of-control scenarios.

Chapter 4

Non-parametric LRT test for stochastically ordered random variables

4.1. Introduction

In this chapter, the concept of stochastic ordering of random variables is introduced. We will examine how we can incorporate this idea into the monitoring procedure for a process, through a methodology proposed by Franck in 1984. This methodology uses a non-parametric likelihood ratio test to determine whether two random variables are stochastically ordered or not.

4.2. Stochastic Orders of Random Variables

Our first step is to define the stochastic ordering of random variables. Stochastic ordering is a way of quantifying the concept of one random variable X being “bigger” than another random variable Y . There are many kinds of stochastic orders, such as the usual stochastic order and the convex stochastic order. We are mainly interested in the usual stochastic ordering of random variables, which is defined as follows (Shaked & Shanthikumar, 2007).

Definition 4.1: “Let X and Y be two random variables such that

$$P\{X > x\} \leq P\{Y > x\} \text{ for all } x \in (-\infty, \infty) \quad (4.1).$$

Then X is said to be smaller than Y in the usual stochastic order, denoted by $X \leq_{st} Y$.”

An equivalent condition of (4.1) is the following; if $F(x) \geq G(x)$ for all $x \in (-\infty, \infty)$ then $X \leq_{st} Y$, where F denotes the cdf of X and G denotes the cdf of Y .

Through the concept of stochastic ordering, we can be guided to the concept of stochastic dominance, which is of great interest for the purposes of SPC. More specifically, stochastic dominance is a stochastic ordering used in decision theory. There are several orders of stochastic dominance. We are particularly interested in first order stochastic dominance, which is equivalent

to usual stochastic ordering (see Definition 4.1). In the stochastic dominance field, we use the notation $X \leq_1 Y$ instead of $X \leq_{st} Y$. In some cases, we may be also interested in second order stochastic dominance (Sriboonchita et. al, 2009).

Definition 4.2: “Let X, Y be two random variables with distribution functions F and G respectively. Then X is said to dominate Y in second-order stochastic dominance (SSD), if and only if the following condition holds:

$$\int_{-\infty}^x F(y)dy \leq \int_{-\infty}^x G(y)dy, \text{ for all } x \in \mathbb{R} \quad (4.2)$$

and is denoted by $X \leq_2 Y$.”

In practice, the first two orders of stochastic dominance are mainly important for SPC. Let us assume we have two samples of the monitored process and we wish to examine whether the process is in-control or not. Let us suppose X follows a Normal distribution with mean μ_1 and standard deviation σ and Y follows a Normal distribution with mean μ_2 and standard deviation σ , where $\mu_1 > \mu_2$. It can be easily verified that $X \geq_1 Y$ (equivalently $X \geq_{st} Y$). This property is quite useful to detect an out-of-control condition of a process. As previously mentioned, a possible out-of-control state of a process occurs when its mean value changes. We may be able to detect such changes efficiently by using methods of stochastic ordering and stochastic dominance and test the available samples. We know stochastic dominance is valid for another out-of-control state of a process as well. More analytically, we know that, if a sample contains even one outlier observation, then its distribution partially stochastically dominates the in-control distribution. Recall outliers are considered individual observations, which come from a distribution with different mean than the desired process mean. The aforementioned examples refer to first-order stochastic dominance. Let us now assume X follows a Normal distribution with mean μ and standard deviation σ_1 and Y follows a Normal distribution with mean μ and standard deviation σ_2 , where $\sigma_1 > \sigma_2$. We can easily conclude that $X \geq_2 Y$. Therefore, we wish to examine if stochastic ordering and stochastic dominance can prove helpful in identifying shifts of a process.

4.3 Non-parametric Likelihood Ratio Test for First Ordered Random Variables

It is of great importance to use the concepts of stochastic ordering and stochastic dominance in a non-parametric manner for the purposes of SPC. Recall that we wish to use a non-parametric test, since large volumes of data are available. Therefore, we wish to let the data speak for themselves. A test with these desirable properties is a likelihood ratio test suggested by Franck (1984).

Let X and Y be two random variables with cumulative distribution functions F and G respectively. Let us also consider x_1, x_2, \dots, x_m and y_1, y_2, \dots, y_n independent random samples from the cdfs F and G respectively. We also need to assume F and G are absolutely continuous. The suggested methodology performs a likelihood ratio test to test the following hypothesis:

$$H_0: F \geq G \text{ versus } H_1: F < G.$$

In order to define the test statistic for this hypothesis testing, it is necessary to describe in detail the methodology followed by the author.

Firstly, we need to estimate the cdfs \hat{F} and \hat{G} . Instead of using their empirical cdfs we will use some alternative estimators. These are the maximum likelihood estimators of their cumulative distribution functions obtained, subject to the restrictions that the supports of both F and G are contained in $\{x_1, \dots, x_m, y_1, \dots, y_n\}$ and $\hat{F} \geq \hat{G}$. The formulas used to obtain \hat{F} and \hat{G} are the following:

$$\hat{F}(x) = \sum_{\{i: x_i \leq x\}} \hat{f}(x_i) + \lambda_1(x) \quad (4.3)$$

$$\hat{G}(y) = \sum_{\{i: y_i \leq y\}} \hat{g}(y_i) + \lambda_2(y) \quad (4.4)$$

where \hat{f} , \hat{g} , λ_1 and λ_2 will be discussed in more detail later on. Let $L_{F \geq G}$ denote the maximum of the likelihood function. Then,

$$L_{F \geq G} = \prod_{i=1}^m \hat{f}(x_i) \prod_{j=1}^n \hat{g}(y_j) \quad (4.5).$$

We will use the test statistic LRT, where

$$LRT = L_{G \geq F} / L_{F \geq G} \quad (4.6).$$

The quantity $L_{G \geq F}$ is defined similarly as $L_{F \geq G}$. The particular test is essentially a rank test. We cannot characterize it as a “pure” likelihood ratio test, since the distributions of the quantities used are not known. They are

estimated in a non-parametric fashion. We shall note that we will use the test statistic

$$\log(LRT) = \log(L_{G \geq F}) - \log(L_{F \geq G}) \quad (4.7),$$

for computational efficiency.

It is necessary to explain how the quantities \hat{f} and \hat{g} are computed. We will describe their computation through an example. Let x_1, x_2, \dots, x_7 and y_1, y_2, \dots, y_5 be two independent samples of F and G respectively such that: $y_4 < x_6 < x_5 < y_3 < x_4 < y_5 < y_2 < x_2 < x_1 < y_1 < x_7 < x_3$. Let us also assume we are interested in computing $L_{F \geq G}$. We perform a random walk in the plane, going over one unit for every x and up one unit for every y . Then we draw the least concave majorant of the points visited during this random walk. The outcome of this random walk is the line shown in the following figure (Figure 4.1).

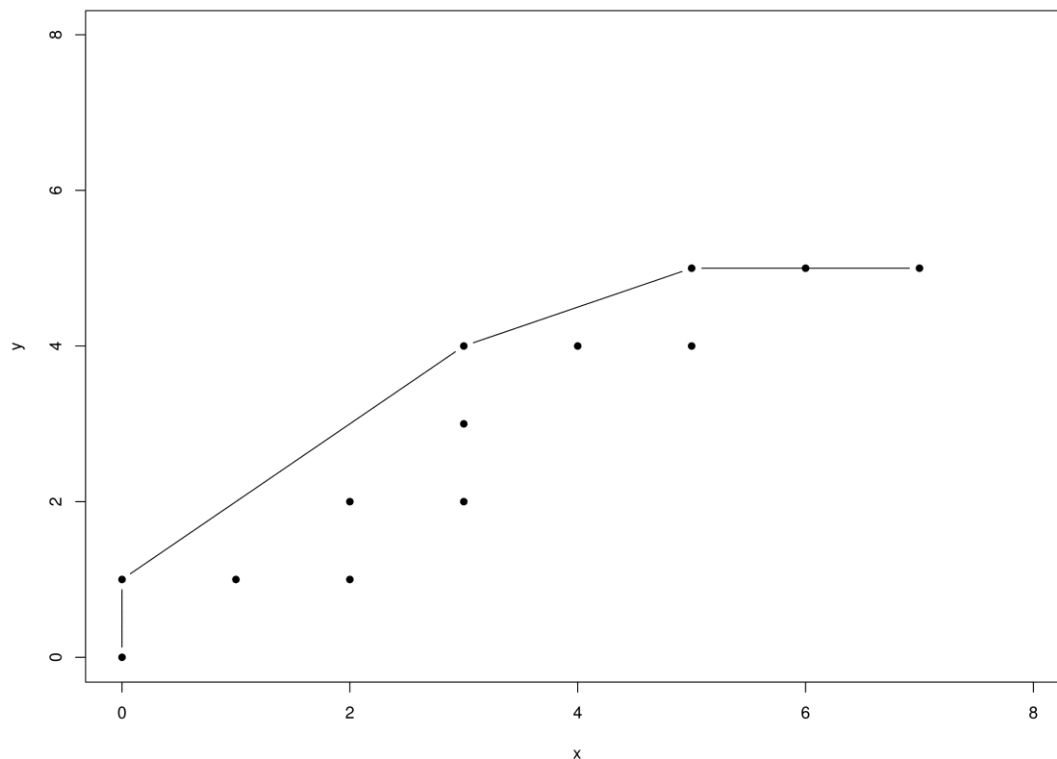


Figure 4. 1. The least concave majorant of the points x_1, \dots, x_7 and y_1, \dots, y_5 .

We need to determine which are the vertices of this curve. These vertices divide the available data into strings. Then the functions \hat{f} and \hat{g} are defined for each x_i and y_j respectively, according to the string that each data point belongs to.

The starting point of this random walk is (0,0) and we consider it to be the 0th vertex. We should note a vertex is a point where the slope of the least concave majorant changes and not just any point visited during the random walk. An observation belongs to the i^{th} string if it occurs after the $(i - 1)^{\text{th}}$ vertex but not after the i^{th} . Let m_ν denote the number of x_i 's in the ν th string and n_ν the number of y_i 's in the ν th string and let us assume the observations are divided into k strings. Then we define \hat{f} and \hat{g} as:

$$\hat{f}(x_i) = \frac{m_\nu + n_\nu}{m_\nu(m+n)} \quad \text{and} \quad \hat{g}(y_j) = \frac{m_\nu + n_\nu}{n_\nu(m+n)} \quad (4.8)$$

where observations x_i and y_j are in the ν th string. We also define the quantities $\hat{f}(y_1)$ and $\hat{g}(x_m)$ as follows:

$$\hat{f}(y_1) = \begin{cases} n_1/(m+n), & x_1 < y_1 \\ 0, & x_1 \geq y_1 \end{cases} \quad \text{and} \quad \hat{g}(x_m) = \begin{cases} m_k/(m+n), & x_m > y_n \\ 0, & x_m \leq y_n \end{cases} \quad (4.9).$$

We get the results shown in tables 4.1 and 4.2 by applying formulas (4.5) and (4.8) to the data of our example.

<i>strings</i>	<i>vertices</i>	m_i	n_i
1	(0,1)	0	1
2	(3,4)	3	3
3	(5,5)	2	1
4	(5,7)	2	0

Table 4. 1. The vertices and strings of the observations x_1, \dots, x_7 and y_1, \dots, y_5 for the computation of $L_{F \geq G}$.

x	<i>string</i>	\hat{f}	y	<i>string</i>	\hat{g}
x_1	3	0.125	y_1	3	0.25
x_2	3	0.125	y_2	2	0.167
x_3	4	0.083	y_3	2	0.167
x_4	2	0.167	y_4	1	0.083
x_5	2	0.167	y_5	2	0.167
x_6	2	0.167			
x_7	4	0.083			

$L_{F \geq G} = 4.84 \times 10^{-11}$, $\log(L_{F \geq G}) = -23.75$

Table 4. 2. Maximum restricted likelihood estimation $L_{F \geq G}$ for the observations x_1, \dots, x_7 and y_1, \dots, y_5 .

The computation of $L_{G \geq F}$ is performed in a similar manner. More analytically, we reverse the roles of x_i 's and y_j 's. We assume observations y are generated from the c.d.f. F and observations x are generated from c.d.f. G . In other words, we use x_i 's as y_j 's and y_j 's as x_i 's. Hence, we get the results shown in figure 4.2 and tables 4.3, 4.4.

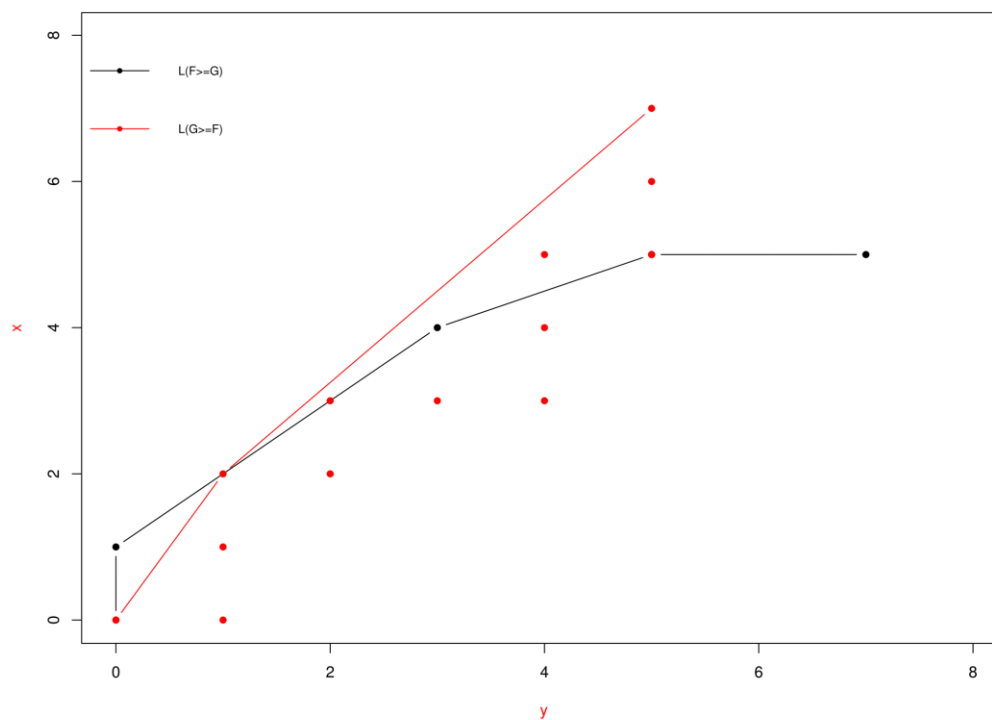


Figure 4. 2. The least concave majorants of the points x_1, \dots, x_7 and y_1, \dots, y_5 .

<i>Strings</i>	<i>vertices</i>	m_i	n_i
1	(1,2)	1	2
2	(5,7)	4	5

Table 4. 3. The vertices and strings of the observations x_1, \dots, x_7 and y_1, \dots, y_5 for the computation of $L_{G \geq F}$.

y	<i>string</i>	\hat{f}	x	<i>string</i>	\hat{g}
y_1	2	0.1875	x_1	2	0.15
y_2	2	0.1875	x_2	2	0.15
y_3	2	0.1875	x_3	2	0.15
y_4	1	0.25	x_4	2	0.15
y_5	2	0.1875	x_5	1	0.125
			x_6	1	0.125
			x_7	2	0.15
$L_{G \geq F} = 3.67 \times 10^{-10}$, $\log(L_{G \geq F}) = -21.73$					

Table 4. 4. Maximum restricted likelihood estimation $L_{G \geq F}$ for the observations x_1, \dots, x_7 and y_1, \dots, y_5 .

Consequently, the test statistic for the observations x_1, \dots, x_7 and y_1, \dots, y_5 is equal to $\log(LRT) = \log(L_{G \geq F}) - \log(L_{F \geq G}) = -21.73 - (-23.75) = 2.02$. In order to determine whether the two cdfs F and G differ, we shall compare the observed value with the appropriate threshold c_α . This threshold is determined by the desired power or the hypothesis testing and it is a quantile of the distribution of the test statistic. More analytically, c_α is the value such that $P[\log(LRT) > c_\alpha] = 1 - \alpha$. However, the distribution of the test statistic is not known. Therefore, we will determine the desired thresholds via simulations.

We shall also highlight the properties of the particular test. This test is consistent and unbiased. These properties can be proved with the use of the following theorems and lemmas.

- Lemma 4.1: Let x_1, \dots, x_m and y_1, \dots, y_n (ordered according to subscript) have least concave majorant V . Let x_1^*, \dots, x_m^* and y_1^*, \dots, y_n^* (ordered according to subscript) have least concave majorant V^* . If $V^* \leq V$, then $L_{F \leq G} \leq L_{F^* \leq G^*}$.
- Lemma 4.2: Let $F \geq G$. Suppose n and $m \rightarrow \infty$ such that $n \leq m \leq Bn$, $0 < A \leq B < 1$. Then for $\varepsilon > 0$ arbitrary and m sufficiently large, $L_{F \leq G} \geq m^{-m} n^{-n} D(m, n, \varepsilon)$, with probability 1, where $D(m, n, \varepsilon) = (m+n)^{-m\varepsilon_m - n\varepsilon_n} \times \left[\frac{m(1-\varepsilon_m) + n(1-\varepsilon_n)}{(m+n)(1-\varepsilon_m)} \right]^{m(1-\varepsilon_m)} \times \left[\frac{m(1-\varepsilon_m) + n(1-\varepsilon_n)}{(m+n)(1-\varepsilon_n)} \right]^{n(1-\varepsilon_n)}$.
- Lemma 4.3: Let m and $n \rightarrow \infty$ such that $An \leq m \leq Bn$, where $0 < A \leq B < 1$. If $F < G$, then for m sufficiently large, $L_{F \leq G} \leq m^{-m} n^{-n} R_1^m R_2^n$, with probability 1, where $0 < R_1 < 1$ and $0 < R_2 < 1$.
- Theorem 4.1: The test of $H_0: F \geq G$ versus $H_1: F < G$ given by rejecting H_0 when LRT is too large is an unbiased test.
- Theorem 4.2: The test of $H_0: F \geq G$ versus $H_1: F < G$ given by rejecting H_0 when $LRT \geq c_\alpha$ is a consistent test, where c_α is such that $\sup_{F \geq G} \Pr_{(F,G)}[LRT \geq c_\alpha] = \alpha$. Assume that $0 < \alpha < 1$ and m and $n \rightarrow \infty$ as in Lemma 4.2. Since $\Pr_{(F,G)}[LRT \geq c_\alpha]$ is a nondecreasing function of G , in finding c_α one may assume that $F = G$.

Lemma 4.1 is used for the proof of Theorem 4.1. Lemmas 4.2 and 4.3 are used for the proof of Theorem 4.2. Proofs of all the theorems and lemmas are available at Franck (1984).

Now, we shall describe how this test will be implemented to SPC with big data. Firstly, we will consider a dataset of 1,000 observations as a prototype sample. Then, each new sample of the process will be compared to this prototype dataset, using the described LRT test. Each sample will consist of 1,000 observations. The observed value of the test statistic will be compared to the threshold. If the observed value is larger than the threshold, we will conclude there exists evidence the process is out-of-control. Otherwise, we will conclude the process is in-control. Finally, we should note that the $\log(LRT)$ will be used instead of LRT , as previously mentioned due to computational efficiency.

Chapter 5

Q-Q plots in SPC

5.1. Introduction

In this chapter we will present Q-Q plots and their basic characteristics. Additionally, we will analyze a monitoring methodology using these graphs. We have designed this particular methodology, as it is non-parametric and serves best the purposes of our analysis.

5.2. Q-Q plots

Q-Q plots are graphs widely used for exploratory statistics. More analytically, they constitute a method for making detailed comparisons of the distribution of two datasets. Q-Q plot is short for Quantile-Quantile plot. Such plots are constructed by plotting the quantiles of one empirical distribution against the corresponding quantiles of the other. Estimates of these quantiles are more commonly used in designing Q-Q plots (Chambers et al., 1983).

The simplest case is when the two datasets are of equal size. In this case, the sorted observations of one sample are plotted against the sorted observations of the second sample. When the two datasets consist of different number of observations, quantiles must be estimated. If the two distributions are identical, then the points lie around the line $y = x$ (see figure 5.1). If the two distributions are linearly related, then the points lie around a straight line but not $y = x$ (see figure 5.2). The “central” point of a Q-Q plot is the point created by the two medians of the samples. A Q-Q plot is quite a useful tool, since it constitutes an overall image of the two distributions of the available samples. More specifically, it provides information about the following elements of each distribution: the shape, the location parameters, and the scale parameters.

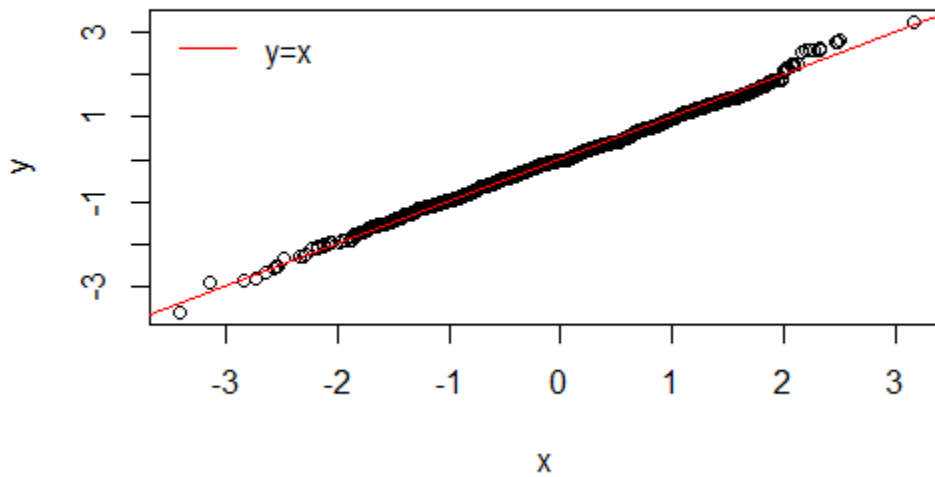


Figure 5. 1. A Q-Q plot for two samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$, following the same distribution.

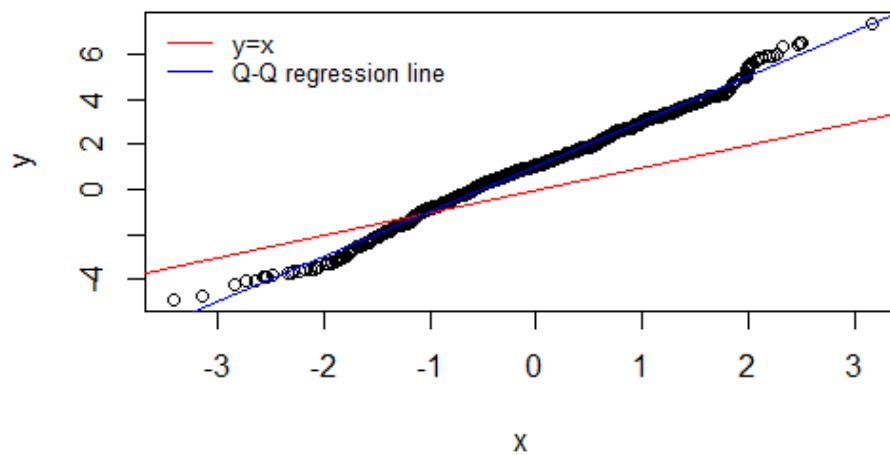


Figure 5. 2. A Q-Q plot for two samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$, following linearly related distributions.

Let us assume we are interested in samples of equal size. Let x_1, \dots, x_n and y_1, \dots, y_n be two such samples. Then, the formula used to construct a Q-Q plot is the following:

$$y_{(i)} = a + b * x_{(i)}, i = 1, \dots, n \quad (5.1)$$

where $x_{(i)}$ and $y_{(i)}$ are the sorted observations of each sample. The intercept a gives a measure of the relative location of the two samples. The slope b gives

a measure of the relative scale of the two samples. More particularly, if the median of the distribution plotted on the horizontal axis is equal to 0, then the intercept of the regression line plotted by (5.1) the intercept a is a measure of location and the slope b is a measure of scale. Consequently, this methodology is non-parametric.

5.3. Q-Q plots for SPC

Since some basic properties of Q-Q plots have been mentioned, we are ready to examine how these graphs can be useful for SPC. As a reminder, the definition of location-scale families of distributions is the following (Casella & Berger, 2002).

Definition 5.1: “Let $f(x)$ be any p.d.f. Then, for any μ , $-\infty < \mu < \infty$, and any $\sigma > 0$, the family of p.d.f.s $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$, indexed by the parameter (μ, σ) , is called the location-scale family with standard p.d.f. $f(x)$; μ is called the location parameter and σ is called the scale parameter.”

It is of great interest to investigate how a Q-Q plot reacts when there is a change in the location or scale parameter. Let us explore these changes through some examples. Firstly, let us assume we have two random samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$ following a Normal distribution with mean 0 and variance 1. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ and $y_1, \dots, y_{1,000}$ is shown in Figure 5.3(a). It is worth mentioning the regression line in Figure 5.3(a) is quite close with line $y = x$. Let us have another random sample $x_1^*, \dots, x_{1,000}^*$ from Normal distribution with mean 0.5 and variance 1. Hence, we have a shift in the location parameter. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ and $x_1^*, \dots, x_{1,000}^*$ is shown in Figure 5.3(b). By comparing the two regression lines, we also observe the slope of the two lines is approximately the same, while there is a shift in the intercept of those lines. In addition, the intercept of the regression line in Figure 5.3(b) is greater than the intercept of the regression line in Figure 5.3(a). Let us now consider a random sample $y_1^*, \dots, y_{1,000}^*$ following a Normal distribution with mean 0 and variance 1.5, where a shift in the scale parameter is observed. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ and $y_1^*, \dots, y_{1,000}^*$ is shown in Figure 5.3(c). In this case, we observe a shift in the slope of the two lines, whereas the intercept is approximately equal for both.

It is worth noting the slope of the regression line is approximately equal to the standard deviation of the distribution $N(0,1.5)$. Let us also explore the case where shifts in both location and scale parameters are observed. We assume that a random sample $w_1, \dots, w_{1,000}$ from a Normal distribution with mean 0.5 and variance 1.5 is available. As shown in Figure 5.3(d), both the intercept and the slope of the line differ from the initial values. All these results are also shown in Table 5.1.

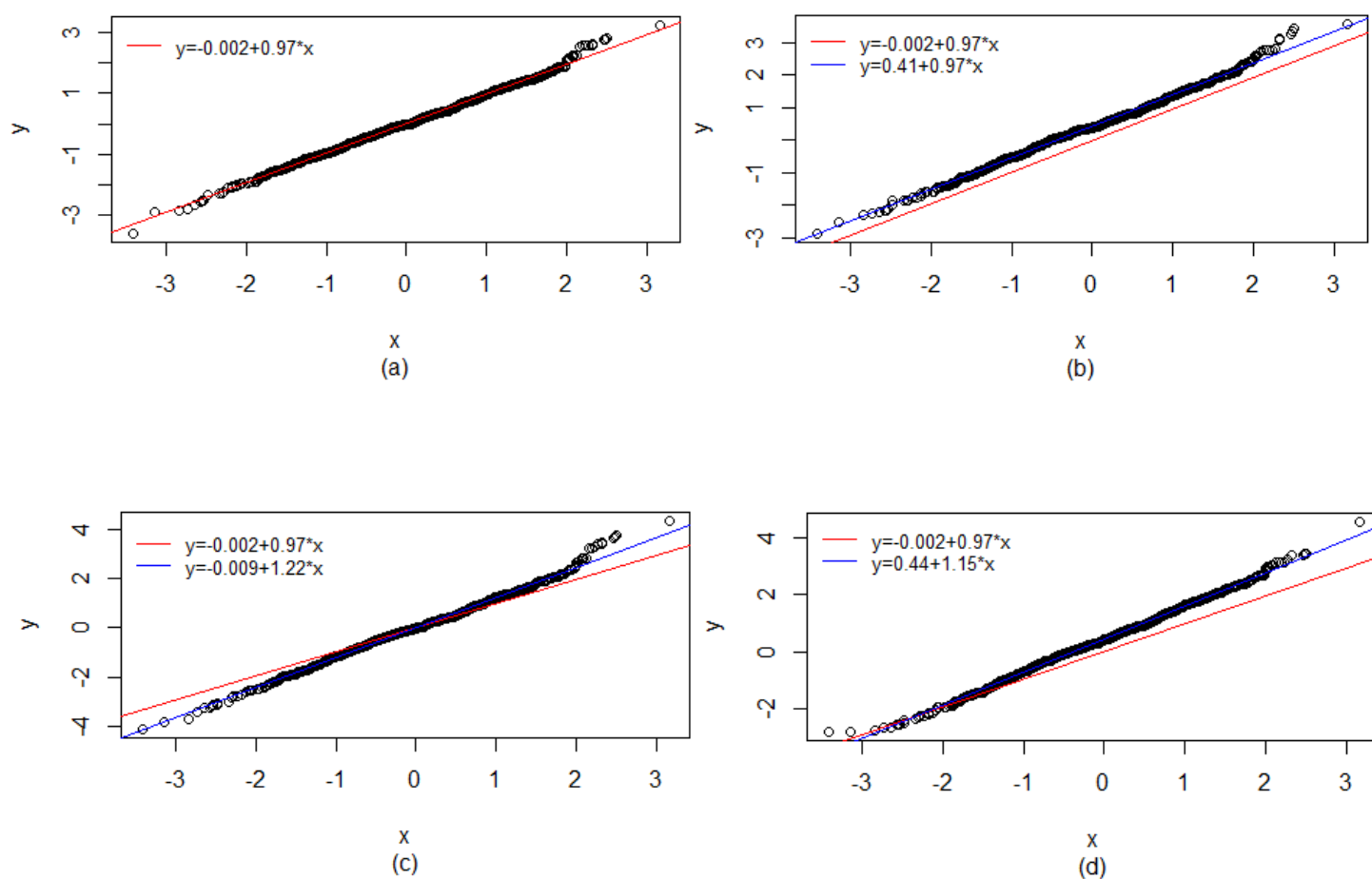


Figure 5. 3. Various cases of Q-Q plots; (a) a typical Q-Q plot, (b) change of the location parameter, (c) change of the scale parameter, (d) change of both the location and scale parameters.

<i>Distribution 1</i> (<i>x-axis</i>)	<i>Distribution 2</i> (<i>y-axis</i>)	<i>Intercept</i>	<i>Slope</i>
N(0,1)	N(0,1)	-0.002	0.970
N(0,1)	N(0.5,1)	0.410	0.968
N(0,1)	N(0,1.5)	-0.010	1.218
N(0,1)	N(0.5,1.5)	0.437	1.155

Table 5. 1. Values of intercept and slope for each case of Q-Q plots.

Taking these results into account, we may conclude the Q-Q plot is a useful tool to detect shifts in location and scale parameters within a distribution family.

This is a desirable property for the purposes of SPC. As previously stated, one of the main goals of the monitoring of a process is the detection of shifts in its operating parameters. We are usually interested in detecting such shifts in the mean and the variance of the process. These changes constitute shifts in the location and scale parameters within a location-scale family of distributions, given there has not been a change of distribution family. Therefore, Q-Q plots can prove helpful for this purpose of SPC.

It is also worth examining how a Q-Q plot can be used when there is a change in the distribution family. When such a change occurs, we cannot rely on the coefficients of the regression line to reach a conclusion about the state of the process. This becomes quite apparent in the Q-Q plot shown in Figure 5.4. More analytically, let us assume we have a random sample $x_1, \dots, x_{1,000}$ from a Normal distribution with mean 3 and variance 3, and another random sample $y_1, \dots, y_{1,000}$ from a Gamma distribution with shape parameter 3 and scale parameter 1. Notice these two distributions have equal means and equal variances. These two distributions also differ in terms of symmetry. After careful observation of the regression line, we will notice that it is very close to the line $y = x$. However, the data are plotted relatively far from the regression line. This is a characteristic we should take into account when developing another technique to detect distribution changes. One possible solution to this problem would be to fit a quadratic line, instead of a simple regression line, to our data. This solution shall be discussed further on.

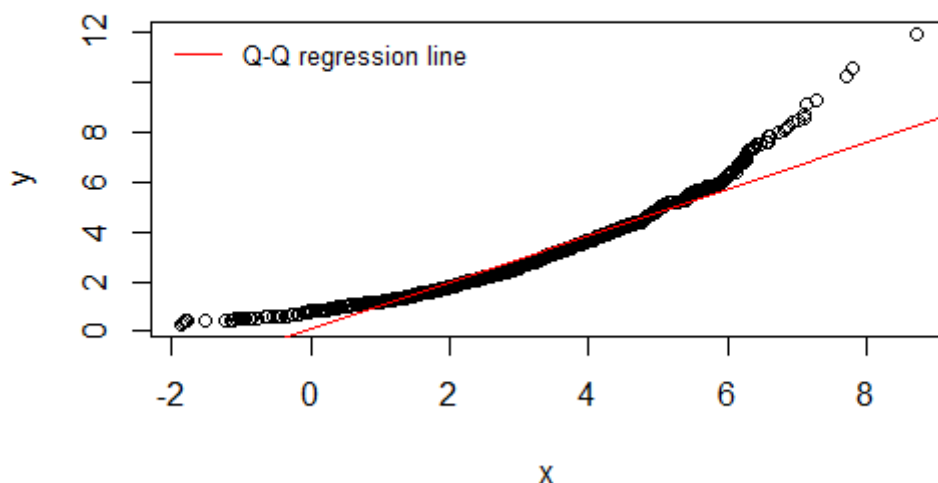


Figure 5. 4. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ from $N(3,3)$ and $y_1, \dots, y_{1,000}$ from $\text{Gamma}(3,1)$.

Moreover, we shall examine what happens to a Q-Q plot when one distribution has heavier tails than the other one. For instance, let us assume $x_1, \dots, x_{1,000}$ is a random sample from a Normal distribution with mean 0 and variance 3 and $y_1, \dots, y_{1,000}$ is a random sample from a t -Student distribution with 3 degrees of freedom. The Q-Q plot for these samples is shown in Figure 5.5. It is evident that most of the points in the center are plotted near a straight line. The points which correspond to the tails of the distributions, deviate significantly from the regression line. Thus, the simple regression line is not adequate for describing these data. A polynomial regression of third degree seems more appropriate.

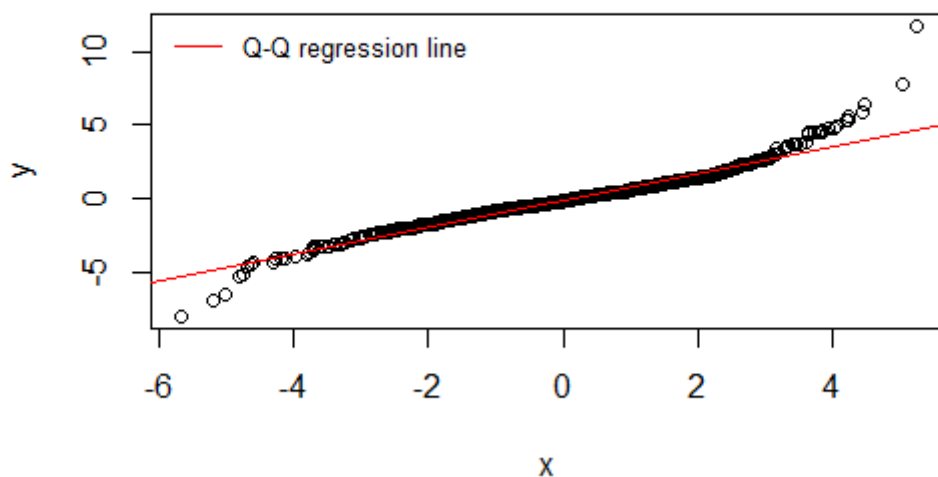


Figure 5. 5. The Q-Q plot for samples $x_1, \dots, x_{1,000}$ from $N(0,3)$ and $y_1, \dots, y_{1,000}$ from $t(3)$.

Based on these properties of the Q-Q plots, we suggest the following methodology, in order to use the Q-Q plots for the monitoring of a process. We shall fit two regression models to our data. The first one will be a simple linear regression, as given by formula (5.1). The second one will be a third-degree polynomial regression, i.e.

$$y_{(i)} = a + b_1 * x_{(i)} + b_2 * x_{(i)}^2 + b_3 * x_{(i)}^3, i = 1, \dots, n \quad (5.2).$$

Afterwards, the mean absolute error will be computed for each model. The mean absolute error for the simple regression model is given by the formula (5.3):

$$MAE_1 = \frac{1}{n} \sum_{i=1}^n |y_{(i)} - \hat{a} - \hat{b} * x_{(i)}| \quad (5.3)$$

and the mean absolute error for the polynomial regression model is given by the formula (5.4):

$$MAE_2 = \frac{1}{n} \sum_{i=1}^n |y_{(i)} - \hat{a} - \hat{b}_1 * x_{(i)} - \hat{b}_2 * x_{(i)}^2 - \hat{b}_3 * x_{(i)}^3| \quad (5.4),$$

where \hat{a} , \hat{b} , \hat{b}_1 , \hat{b}_2 and \hat{b}_3 are estimates of the corresponding coefficients of the models (5.1) and (5.2). The quantity of interest is

$$M = MAE_1 - MAE_2 \quad (5.5).$$

Models (5.1) and (5.2) are fitted to the data of the two aforementioned examples. The results are shown in Figures (5.6) and (5.7) and Table (5.2). More analytically, MAE_1 is expected to be greater than MAE_2 . Especially when there is a change in distribution family, MAE_1 will be much greater than MAE_2 leading to relatively large values of M . Hence, M seems to be a suitable quantity for detecting changes of distribution. Its effectiveness will be examined via simulations. It is also worth mentioning that it is not essential to fit a second-degree polynomial regression to the data. When the third-degree term is redundant, its coefficient will be approximately equal to 0. Consequently, the third-degree polynomial regression model is adequate for every possible case.

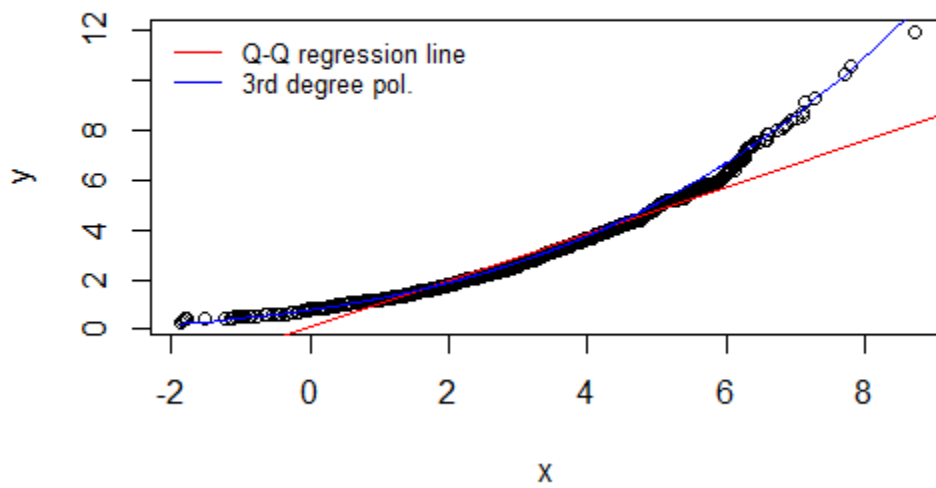


Figure 5. 6. The fit of the regression models (5.1) and (5.2) on samples $x_1, \dots, x_{1,000}$ from $N(3,3)$ and $y_1, \dots, y_{1,000}$ from $\text{Gamma}(3,1)$.

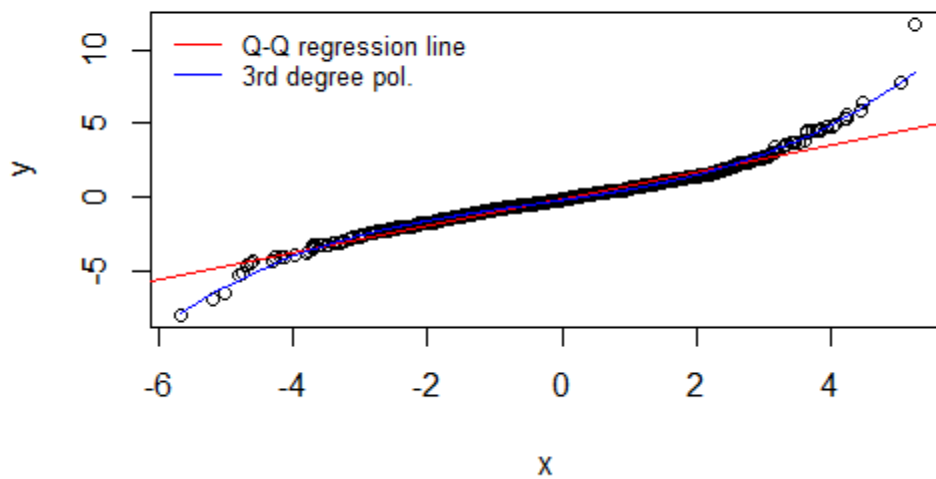


Figure 5. 7. The fit of the regression models (5.1) and (5.2) on samples $x_1, \dots, x_{1,000}$ from $N(0,3)$ and $y_1, \dots, y_{1,000}$ from $t(3)$.

<i>Distribution 1</i> (<i>x-axis</i>)	<i>Distribution 2</i> (<i>y-axis</i>)	<i>Regression models</i>
N(3,3)	Gamma (3,1)	$y = 0.12 + 0.94 * x$ $y = 0.79 + 0.40 * x + 0.06 * x^2 + 0.01 * x^3$
$MAE_1 = 0.33$	$MAE_2 = 0.05$	$M = MAE_1 - MAE_2 = 0.28$
N(0,3)	$t(3)$	$y = -0.07 + 0.92 * x$ $y = -0.17 + 0.66 * x + 0.04 * x^2 + 0.03 * x^3$
$MAE_1 = 0.21$	$MAE_2 = 0.09$	$M = MAE_1 - MAE_2 = 0.12$

Table 5. 2. Fit of regression models (5.1) and (5.2) on the data of the previous examples.

Let us now describe the proposed methodology for SPC with big data, using Q-Q plots, in more detail. The examined data will be divided into subsamples of 1,000 observations. There will be a prototype dataset of 1,000 observations, with which every other sample will be compared. This prototype dataset will be from the in-control process. Our first step shall be checking whether the distribution of the process has changed or not. We will compute M from (5.5) for each subsample (compared to the prototype). If the value of M exceeds a predefined threshold c_α , then we may conclude the distribution of the process has changed and, as a result, the process will be considered to be out-of-control. Note that this threshold c_α will be such that $P(M > c_\alpha) = 1 - \alpha$. If the process is out-of-control, the production will stop and we will seek possible assignable causes for this shift. If the process is in-control, we will examine whether the location and scale parameters have the desired values or not. Consequently, we shall examine if there has been a location or scale shift of the process. To achieve this, we will use the values of the coefficients, as shown in model (5.1). However, the coefficients of a regression model are correlated. Therefore, we cannot examine each coefficient individually. We shall use a test statistic which is widely used for multivariate control charts. This statistic is called the *Hotelling's* T^2 and is given by the formula (5.6):

$$T^2 = (X - \bar{X})' * S^{-1} * (X - \bar{X}) \quad (5.6),$$

where $X = (X_1, \dots, X_p)'$ is a random vector, $\bar{X} = (\bar{X}_1, \dots, \bar{X}_p)'$ is the vector of

sample means and $S = \begin{bmatrix} S_{11} & \dots & S_{1p} \\ \vdots & \ddots & \vdots \\ S_{p1} & \dots & S_{pp} \end{bmatrix}$ is the sample covariance matrix.

In our case, *Hotelling's* T^2 takes the following form:

$$T^2 = \left(\begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} \right)' * S^{-1} * \left(\begin{bmatrix} a \\ b \end{bmatrix} - \begin{bmatrix} \bar{a} \\ \bar{b} \end{bmatrix} \right) \quad (5.7),$$

where $S = \begin{bmatrix} s_a^2 & s_{ab} \\ s_{ab} & s_b^2 \end{bmatrix}$. We will operate in a similar manner as with M . For

each subsample, T^2 is computed and is compared to a threshold c'_α . c'_α is such that $P(T^2 > c'_\alpha) = 1 - \alpha$. If the observed value of T^2 exceeds c'_α , we may conclude the process is out-of-control. It would also be very helpful to plot the values of the coefficients to determine the kind of the occurred shift in a graph as shown in Figure 5.8.

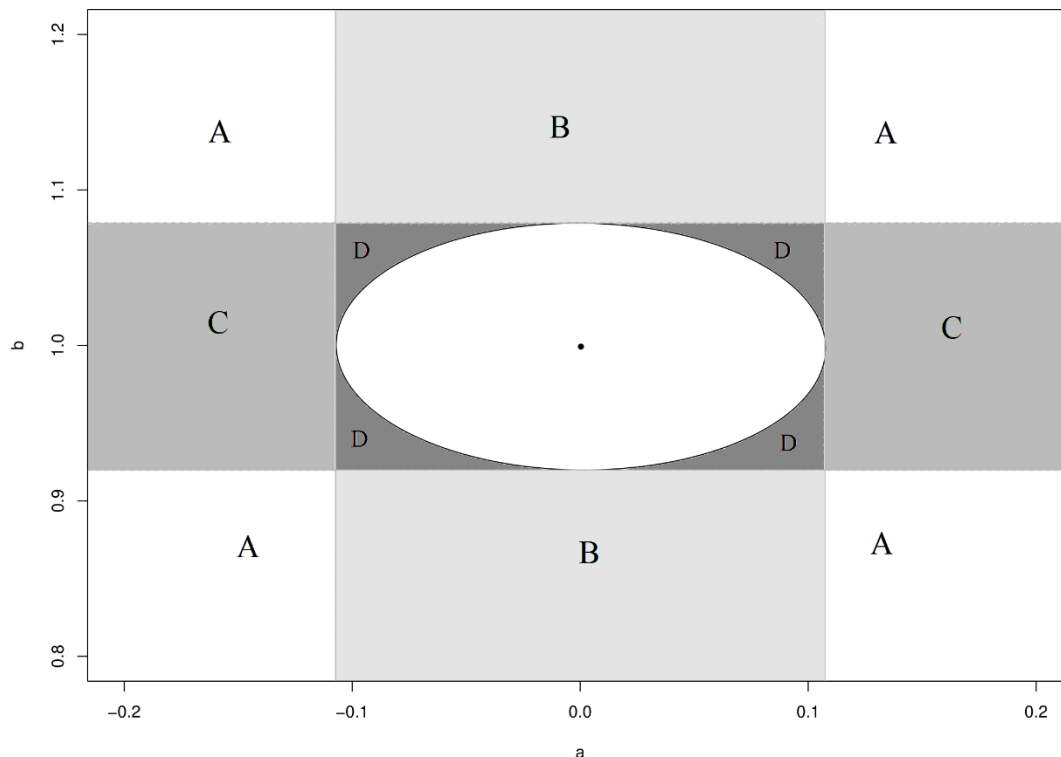


Figure 5. 8. The ellipse created by the values of a and b.

It is worth analyzing Figure 5.8 in more detail. In the center of the graph, the ellipsis created by the T^2 statistic is drawn. When a pair of a and b values plots inside this ellipsis, we may conclude there has neither been a shift of location nor a shift of scale parameter. We shall note the pairwise

confidence region is of 95% level. When a point (a, b) plots in one of the regions marked by A, then we may conclude both location and scale parameters are out-of-control. When a point (a, b) plots in one of the regions marked by B, then the scale parameter might have changed, hence there is a scale shift of the process. When a point (a, b) plots in one of the regions marked by C, then we may conclude the location parameter is out-of-control. Unfortunately, when a point (a, b) plots in one of the regions marked by D, we do not have a quite clear understanding of the problem occurred in the process. One possible scenario is there has occurred a shift in the family of the distribution of the process. Finally, as previously mentioned, this methodology is non-parametric. Therefore, it is suitable for the desired purposes. Its properties and performance will be thoroughly examined via simulations in the following chapter.

Chapter 6

Results of Simulation Studies

6.1. Introduction

In this final chapter, we present and discuss all the results of the conducted simulation studies. We performed several simulation scenarios for each method discussed in the previous chapters. All the scenarios are described thoroughly. This chapter concludes with some final remarks and some recommendations for future research.

6.2. I-chart and \bar{X} -chart

These control charts were described in detail in Chapter 2. We ran 1,000 iterations of simulating observations. These observations came from a Normal distribution with mean 0 and variance 1, which is considered to be the in-control distribution. The sample sizes were 2,000, 10,000, 100,000, 500,000 and 1,000,000 observations for each iteration. Phase I consisted of 1,000 observations in each case. Our initial goal is to explore how the rate of false alarms of these charts behaves when the sample size increases. We should point out we are only interested in the alarms occurring during Phase II of the monitoring process. Recall that the alarms occurring in Phase I are not taken into account in order to construct the limits of a control chart. Therefore, all analyses from this point forward are about the number of alarms in Phase II. We should also note we ran simulations for various choices of control limits. More specifically, simulations were performed for 3σ , 4σ and 4.5σ control limits. Firstly, let us examine the results for the I-charts.

- I-chart with 3σ control limits

The results for this chart are shown in Table 6.1.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	2.758
10,000	9,000	24.688
100,000	99,000	281.432
500,000	499,000	1,395.788
1,000,000	999,000	2,838.800

Table 6. 1. Results of simulations for the I-chart with 3σ control limits.

- I-chart with 4σ control limits

The results for this chart are shown in Table 6.2.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	0.086
10,000	9,000	0.615
100,000	99,000	7.005
500,000	499,000	35.812
1,000,000	999,000	74.478

Table 6. 2. Results of simulations for the I-chart with 4σ control limits.

- I-chart with 4.5σ control limits

The results for this chart are shown in Table 6.3.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	0.008
10,000	9,000	0.073
100,000	99,000	0.770
500,000	499,000	1.540
1,000,000	999,000	8.160

Table 6. 3. Results of simulations for the I-chart with 4.5σ control limits.

Let us discuss these results. We observe that the number of false alarms increases when the sample size increases, as expected. It is worth mentioning the sample size and the number of false alarms appear to behave as proportional quantities. We also observe that, when we use wider control limits, the number of false alarms decreases. However, it is still affected by sample size. This information is also shown in Figure 6.1. Notably, the lines which correspond to the I-chart with 4σ limits and to the I-chart with 4.5σ limits, appear to coincide due to the scale of the vertical axis.

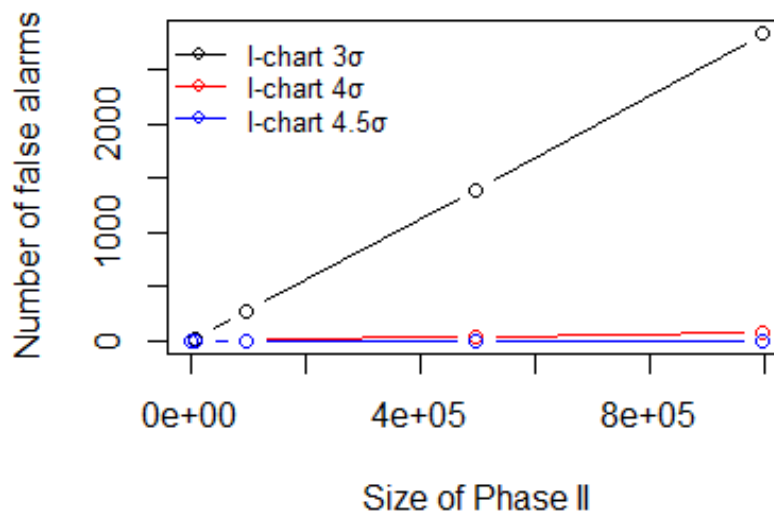


Figure 6. 1. Number of false alarms of an I-chart for various choices of the control limits.

One interesting aspect of this study is to investigate how these I-charts behave in out-of-control states of the process. In order to achieve this, we ran another set of simulations for two out-of-control scenarios. The number of iterations and the choices of sample sizes were the same for the in-control state. The first out-of-control case is a persistent shift in the location parameter. Namely, the last 500 observations of each sample were generated from a Normal distribution with mean 0.5 and variance 1. Hence, we had a shift of 0.5 standard deviations. Another aspect of interest is the presence of outliers. In particular, we insert an outlier per 1,000 observations in Phase II of the process. The outliers are generated from a Normal distribution with mean 3

and variance 1. Thus, we have a shift of 3 standard deviations, which is considered to be a large shift. We used various choices of control limits again. The results are presented in the following tables (Tables 6.4 to 6.6).

- I-chart with 3σ control limits

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	4.718	3.277
10,000	9,000	27.338	29.322
100,000	99,000	280.029	325.304
500,000	499,000	1,394.294	1,661.936
1,000,000	999,000	2,859.827	3,320.159

Table 6. 4. Results of out-of-control simulations for the I-chart with 3σ control limits.

- I-chart with 4σ control limits

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	0.173	0.258
10,000	9,000	0.711	2.105
100,000	99,000	7.041	22.672
500,000	499,000	35.510	114.250
1,000,000	999,000	71.085	233.694

Table 6. 5. Results of out-of-control simulations for the I-chart with 4σ control limits.

- I-chart with 4.5σ control limits

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	0.024	0.079
10,000	9,000	0.076	0.687
100,000	99,000	0.770	7.423
500,000	499,000	4.102	38.093
1,000,000	999,000	8.144	76.912

Table 6. 6. Results of out-of-control simulations for the I-chart with 4.5σ control limits.

As far as persistent shift is concerned, we observe the I-chart cannot successfully detect the shift. Nevertheless, this result is expected, since we know that I-charts are mostly capable of detecting large shifts and not small ones. When it comes to the detection of outliers, the findings are different. The 3σ control limits present more alarms than we would desire. Most of the times the outliers are detected, but false alarms occur as well. The 4σ and 4.5σ control limits present less alarms than the correct number. The advantage of using these limits is they have quite decreased false alarm rates. Nonetheless, their use is not satisfactory, since the outliers are not always detected. Figure 6.2 shows graphically the results described for the I-chart with 3σ control limits.

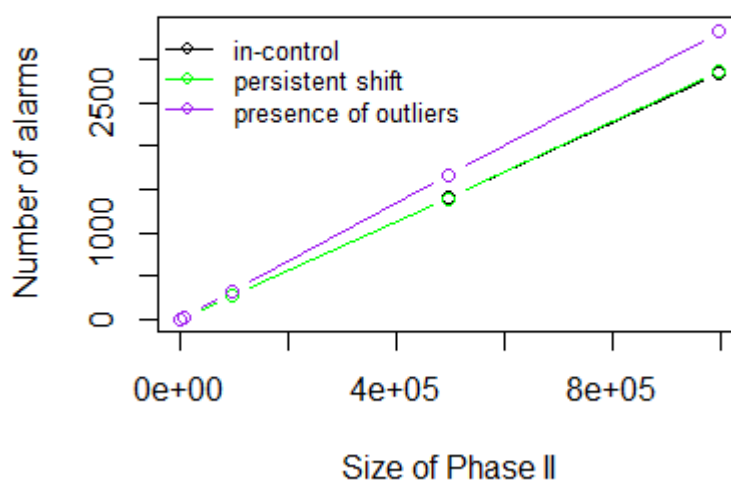


Figure 6. 2. Comparison between the in-control and out-of-control alarm rates of an I-chart with 3σ control limits.

Let us now examine the analogous results for the \bar{X} -chart. The observations were grouped into subsamples of 100 observations. The exact same simulation scenarios were performed. The size of Phase I was 1,000 observations, which corresponds to 10 points plotted on a \bar{X} -chart. Simulations for sample size of 2,000 were not performed, as very few points would be plotted.

- \bar{X} -chart with 3σ control limits

The results for this chart are shown in Table 6.7. The first column shows the size of Phase II. Keep in mind that Phase I consists of 1,000 additional observations. The second column contains the mean number of alarms, when the process is in-control, hence it contains the number of false alarms per iteration. The last two columns contain the number of alarms for each out-of-control scenario.

<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>		
	<i>In-control state</i>	<i>Persistent shift</i>	<i>Outliers</i>
9,000	0.387	5.244	0.420
99,000	4.341	8.974	4.517
499,000	21.658	26.741	21.807
999,000	44.661	47.943	44.949

Table 6. 7. Results of simulations for the \bar{X} -chart with 3σ control limits.

- \bar{X} -chart with 4σ control limits

The results of the simulations for the \bar{X} -chart with 4σ control limits are presented in Table 6.8. This table has the same structure as Table 6.7.

<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>		
	<i>In-control state</i>	<i>Persistent shift</i>	<i>Outliers</i>
9,000	0.017	4.152	0.015
99,000	0.148	4.286	0.180
499,000	0.752	4.826	0.816
999,000	1.521	5.436	1.619

Table 6. 8. Results of simulations for the \bar{X} -chart with 4σ control limits.

- \bar{X} -chart with 4.5σ control limits

The results of the simulations for the \bar{X} -chart with 4.5σ control limits are presented in Table 6.9. This table has the same structure as Tables 6.7 and 6.8.

<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>		
	<i>In-control state</i>	<i>Persistent shift</i>	<i>Outliers</i>
9,000	0.003	3.343	0.002
99,000	0.020	3.471	0.016
499,000	0.087	3.499	0.102
999,000	0.207	3.611	0.198

Table 6. 9. Results of simulations for the \bar{X} -chart with 4.5σ control limits.

Based on these results, we reach different conclusions. The \bar{X} -chart with 3σ control limits behaves exactly like the I-chart, but with reduced false alarm rates. This particular chart does not detect process shifts quite successfully. The mean number of alarms, when there is a persistent shift, is slightly higher than the number of false alarms. The mean number of alarms under the presence of outliers is equal to the number of false alarms. Therefore, the detection of outliers is completely unsuccessful. This fact was expected, since the \bar{X} -chart plots the mean value of each subgroup. Consequently, the influence of an extreme value vanishes. As far as detection of outliers is concerned, identical conclusions can be also drawn for \bar{X} -chart with 4σ and 4.5σ control limits. An unexpected observation is made when a persistent shift of the process has occurred. Recall that the last 500 observations are generated from a $N(0.5,1)$, instead of a $N(0,1)$. These observations correspond to 5 points plotted on the \bar{X} -chart. Hence, we wish to observe the last 5 points of each iteration beyond the control limits. It appears the \bar{X} -chart with 4σ or with 4.5σ control limits achieves quite closely this goal along with very low false alarm rates, despite the fact the shift is of a small magnitude. More particularly, the \bar{X} -chart with 4σ control limits achieves almost a perfect detection rate. Yet, as previously stated, all the \bar{X} -charts fail to detect the presence of outliers. Finally, we shall observe the false alarm rates are quite

low when wider control limits are used, as was expected. These results are also shown in Figures 6.3(a), (b) and (c).

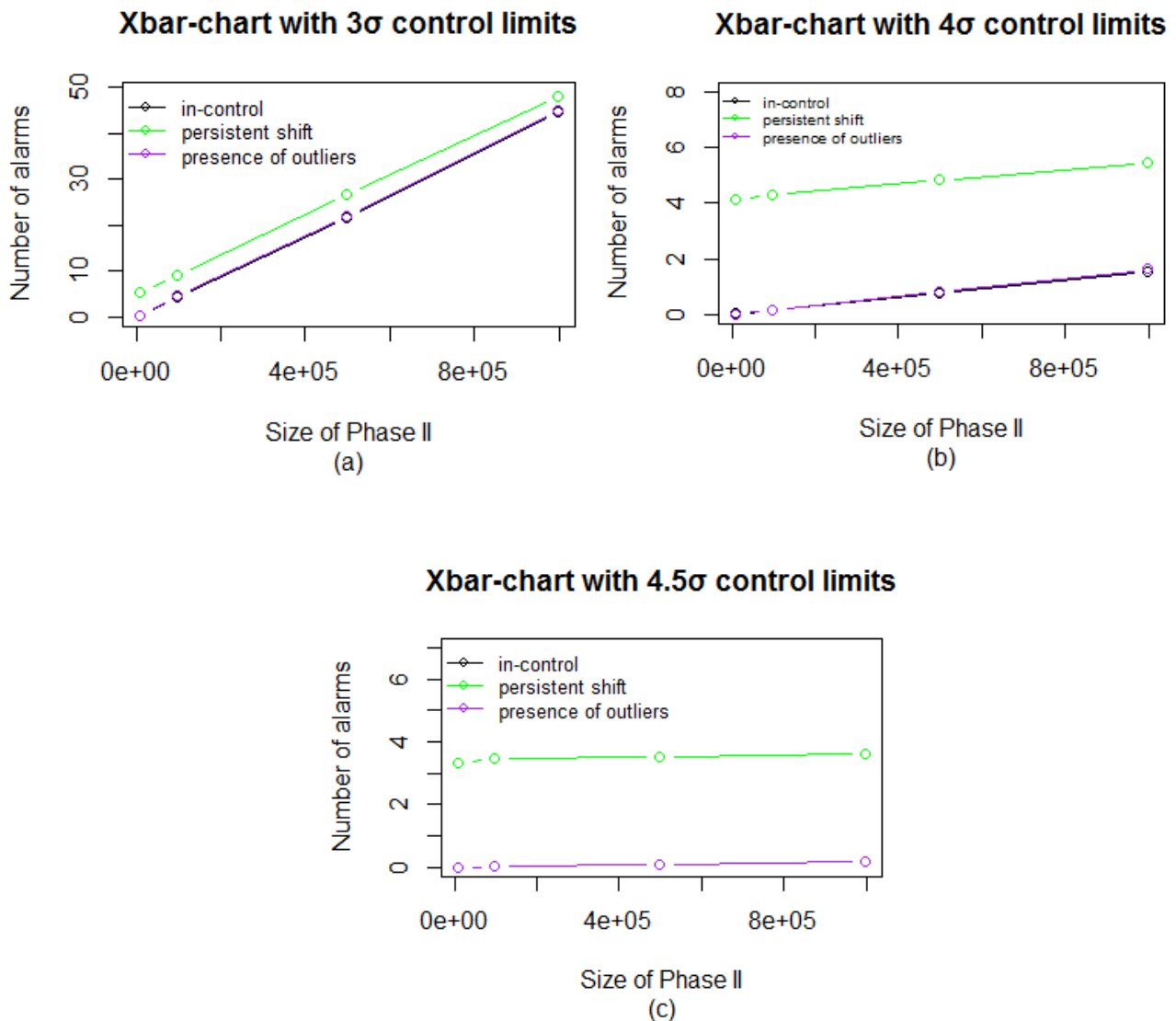


Figure 6. 3. Comparison between the in-control and out-of-control alarm rates of an \bar{X} -chart with (a) 3σ , (b) 4σ and (c) 4.5σ control limits.

6.3 CUSUM chart

After closely examining the results of the simulations for the I-chart and the \bar{X} -chart, we shall now delve into the analogous results for the CUSUM chart. The basic characteristics of this chart are also presented in Chapter 2. Simulation studies were performed for various sample sizes. More specifically, the total sample size was equal to 2,000, 10,000, 100,000, 500,000 and 1,000,000 observations.

For each sample, the first 1,000 observations were used for Phase I, i.e. the construction of the control limits. Different choices were also made concerning the width of the control limits. Namely, the parameters K and H , which define the control limits of a CUSUM chart, took the following values; $K=0.5$ for every case and $H=4.77, 5$ and 6 . The in-control ARLs for these combinations are given in Table 2.1. The in-control distribution is again the $N(0,1)$. Let us now present the results for each of the aforementioned simulation scenarios.

- CUSUM chart with $K=0.5$ and $H=4.77$

The results for the false alarm rate of a CUSUM chart with $K=0.5$ and $H=4.77$ are shown in the following table (Table 6.10).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	10.364
10,000	9,000	95.274
100,000	99,000	1,021.282
500,000	499,000	5,324.956
1,000,000	999,000	10,401.130

Table 6. 10. Results of the simulations for the CUSUM chart with $K=0.5$, $H=4.77$.

- CUSUM chart with $K=0.5$ and $H=5$

The results for the false alarm rate of a CUSUM chart with $K=0.5$ and $H=5$ are shown in the following table (Table 6.11).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	8.475
10,000	9,000	74.530
100,000	99,000	829.008
500,000	499,000	4,151.584
1,000,000	999,000	8,321.600

Table 6. 11. Results of the simulations for the CUSUM chart with $K=0.5$, $H=5$.

- CUSUM chart with $K=0.5$ and $H=6$

The results for the false alarm rate of a CUSUM chart with $K=0.5$ and $H=6$ are shown in the following table (Table 6.12).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	3.360
10,000	9,000	28.955
100,000	99,000	314.406
500,000	499,000	1,568.354
1,000,000	999,000	3,259.926

Table 6. 12. Results of the simulations for the CUSUM chart with $K=0.5$, $H=6$.

The previously shown results, are also displayed in Figure 6.4.

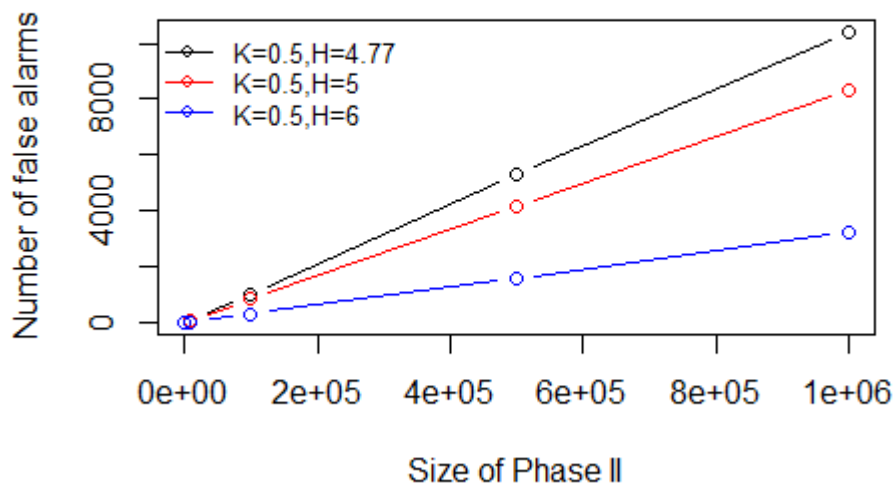


Figure 6. 4. Number of false alarms of a CUSUM chart for various choices of its control limits.

Our main observation is that the number of false alarms of a CUSUM chart increases, when the sample size increases. Accordingly, we draw similar conclusions as for the I-chart and \bar{X} -chart. It is also important to examine the

detection ability of a CUSUM chart, in order to have a more comprehensive view of its behavior with big data. To achieve this goal, we will perform additional simulation scenarios. We shall use the same procedures as with the control charts discussed in section 6.2. In effect, we will simulate persistent shifts of the process and the presence of outliers in fixed positions. For the first case, we will simulate the last 500 observations of each sample from a $N(0.5,1)$. For the second case, we will insert an outlier per 1,000 observations. This outlying observation will follow a $N(3,1)$. Remember we shall study the case of a location parameter shift with both of these scenarios. The sample sizes will be the same as for the in-control simulations. The results for these out-of-control cases are displayed in the following tables (Tables 6.13 to 6.15).

- CUSUM chart with $K=0.5$ and $H=4.77$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	328.367	11.964
10,000	9,000	411.572	99.353
100,000	99,000	1,346.895	1,094.042
500,000	499,000	5,592.760	5,580.325
1,000,000	999,000	10,714.870	11,142.190

Table 6. 13. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=4.77$.

- CUSUM chart with $K=0.5$ and $H=5$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	328.101	9.408
10,000	9,000	402.131	80.846
100,000	99,000	1,145.808	888.968
500,000	499,000	4,504.421	4,454.602
1,000,000	999,000	8,600.061	8.975.962

Table 6. 14. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=5$.

- CUSUM chart with $K=0.5$ and $H=6$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	304.572	3.760
10,000	9,000	335.124	31.579
100,000	99,000	609.621	344.444
500,000	499,000	1,906.039	1,735.226
1,000,000	999,000	3,518.190	3,480.601

Table 6. 15. Results of out-of-control simulations for the CUSUM chart with $K=0.5$ and $H=6$.

Taking a closer look at the results displayed above, we are lead to similar conclusions as for the previously discussed control charts. The CUSUM chart seems to detect efficiently the persistent shift in the data. This particular outcome was expected, due to the construction of this chart. CUSUM charts are very useful for detecting small persistent shifts, as is the case of the simulated scenario which assumes a persistent shift of the process. However, they are not quite effective at detecting large transient shifts of the process. This fact becomes apparent with the second simulated scenario, which assumes the presence of outliers in fixed positions within the sample. Thus, the results we get for the CUSUM chart are expected, if we take into account its basic principles. Additionally, we observe an increase in the number of alarms as the sample size increases, which was also the case for the in-control simulations. The results for the CUSUM chart with $K=0.5$ and $H=4.77$ are graphically displayed in Figure 6.5. In this figure, we compare the number of alarms when the process is in-control and when the process is out-of-control.

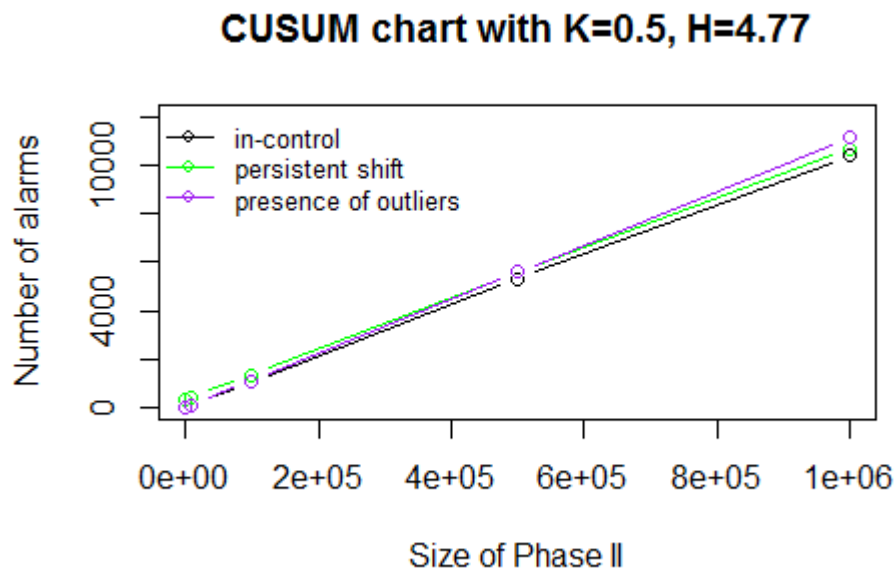


Figure 6. 5. Comparison between the in-control and out-of-control alarm rates of a CUSUM chart with K=0.5 and H=4.77.

6.4. EWMA chart

The next control chart of interest is the EWMA chart. Its basic characteristics were presented in Chapter 2. Simulations were conducted to investigate the behavior of the particular chart, when big data are encountered. The simulation scenarios were exactly the same as the ones used for the I-chart, the \bar{X} -chart and the CUSUM chart. The parameters used for constructing the control limits took the following values; $\lambda=0.25$ for every case and $L=3, 3.5$ and 4 . The in-control ARLs for these combinations are given in Table 2.2. Let us firstly examine the results for these in-control simulations.

- EWMA chart with $\lambda=0.25$ and $L=3$

The results for the false alarm rate of an EWMA chart with $\lambda=0.25$ and $L=3$ are shown in the following table (Table 6.16).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	2.912
10,000	9,000	25.981
100,000	99,000	288.260
500,000	499,000	1,438.978
1,000,000	999,000	2,901.473

Table 6. 16. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=3$.

- EWMA chart with $\lambda=0.25$ and $L=3.5$

The results for the false alarm rate of an EWMA chart with $\lambda=0.25$ and $L=3.5$ are shown in the following table (Table 6.17).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	0.495
10,000	9,000	4.793
100,000	99,000	50.701
500,000	499,000	254.914
1,000,000	999,000	518.365

Table 6. 17. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=3.5$.

- EWMA chart with $\lambda=0.25$ and $L=4$

The results for the false alarm rate of an EWMA chart with $\lambda=0.25$ and $L=4$ are shown in the following table (Table 6.18).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
2,000	1,000	0.064
10,000	9,000	0.635
100,000	99,000	7.430
500,000	499,000	36.666
1,000,000	999,000	72.722

Table 6. 18. Results of the simulations for the EWMA chart with $\lambda=0.25$ and $L=4$.

The previously described results are also displayed in Figure 6.6.

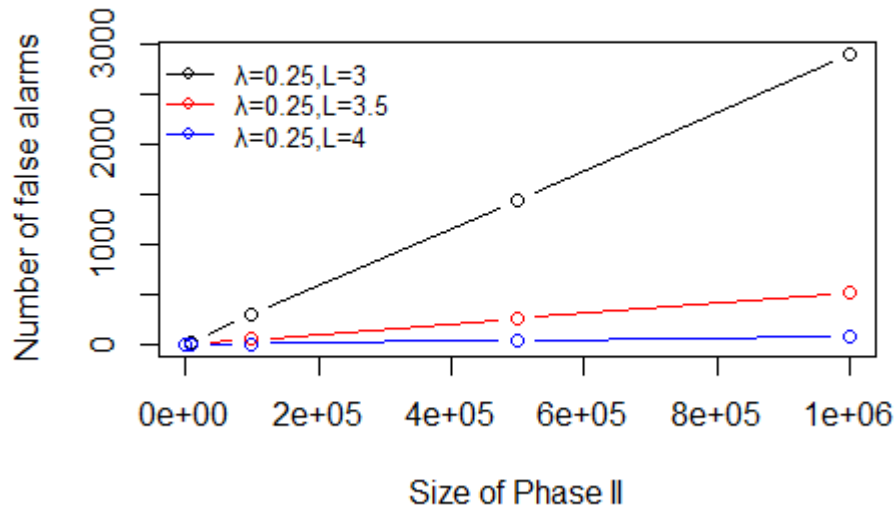


Figure 6. 6. Number of false alarms of an EWMA chart for various choices of its control limits.

Our conclusions about this chart are no different than those for the previous control charts. It is clear the number of false alarms increases when the total sample size increases. Our next step is to check the detection strength of this chart. Therefore, we will use exactly the same simulation scenarios as before, which investigate the cases of persistent shifts of the process and the presence of outliers in fixed positions. The simulations performed for these cases are exactly the same as for the I-chart, the \bar{X} -chart and the CUSUM chart. These results are displayed in the following tables (Tables 6.19 to 6.21).

- EWMA chart with $\lambda=0.25$ and $L=3$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	24.886	3.238
10,000	9,000	48.342	28.488
100,000	99,000	312.050	312.278
500,000	499,000	1,483.618	1,571.522
1,000,000	999,000	2,885.897	3,185.656

Table 6. 19. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=3$.

- EWMA chart with $\lambda=0.25$ and $L=3.5$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	7.885	0.643
10,000	9,000	11.780	5.602
100,000	99,000	59.091	60.945
500,000	499,000	266.860	310.792
1,000,000	999,000	525.149	611.047

Table 6. 20. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=3.5$.

- EWMA chart with $\lambda=0.25$ and $L=4$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
2,000	1,000	2.028	0.120
10,000	9,000	2.601	0.973
100,000	99,000	9.329	10.668
500,000	499,000	39.383	55.176
1,000,000	999,000	75.626	105.568

Table 6. 21. Results of out-of-control simulations for the EWMA chart with $\lambda=0.25$ and $L=4$.

After careful consideration of these tables, we conclude the EWMA chart has a similar behavior to the previously analyzed control charts in the out-of-control cases as well. This chart detects a persistent shift but cannot detect adequately the presence of an outlier. Again, this result was expected, since EWMA charts are quite useful for detecting small persistent shifts rather than large transient shifts. These conclusions can be seen graphically in Figure 6.7 (when $\lambda=0.25$ and $L=3$).

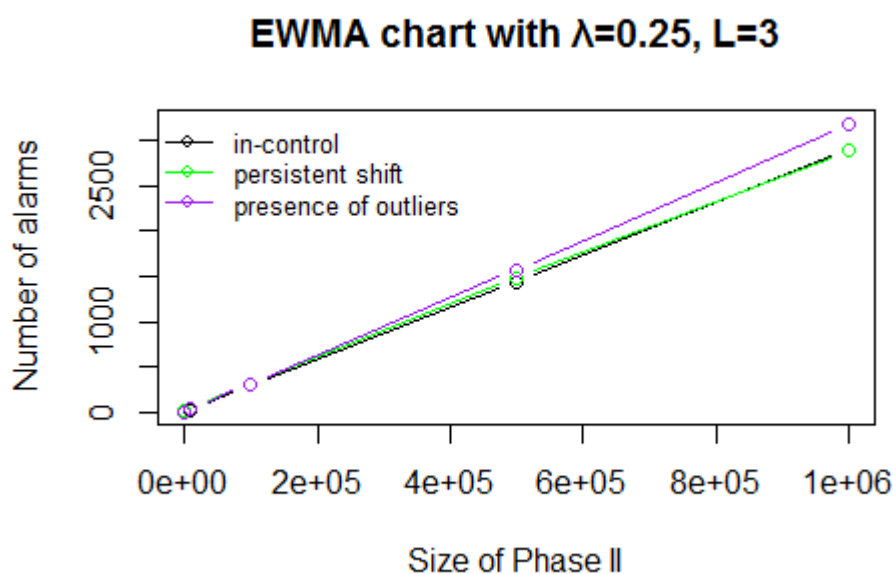


Figure 6. 7. Comparison between the in-control and out-of-control alarm rates of an EWMA chart with $\lambda=0.25$ and $L=3$.

It would also be of great interest to make a comparison between the graphs we have examined so far. We can compare control charts, which have the same in-control ARL. We can only compare the I-chart with 3σ control limits, the \bar{X} -chart with 3σ control limits and the CUSUM chart with $K=0.5$ and $H=4.77$. All these control charts have $ARL_0 = 370$. The chosen combinations for the EWMA chart do not yield an ARL_0 of this value, so a comparison is not valid. In order to make this comparison easier, the numbers of false alarms are displayed in Figure 6.8.

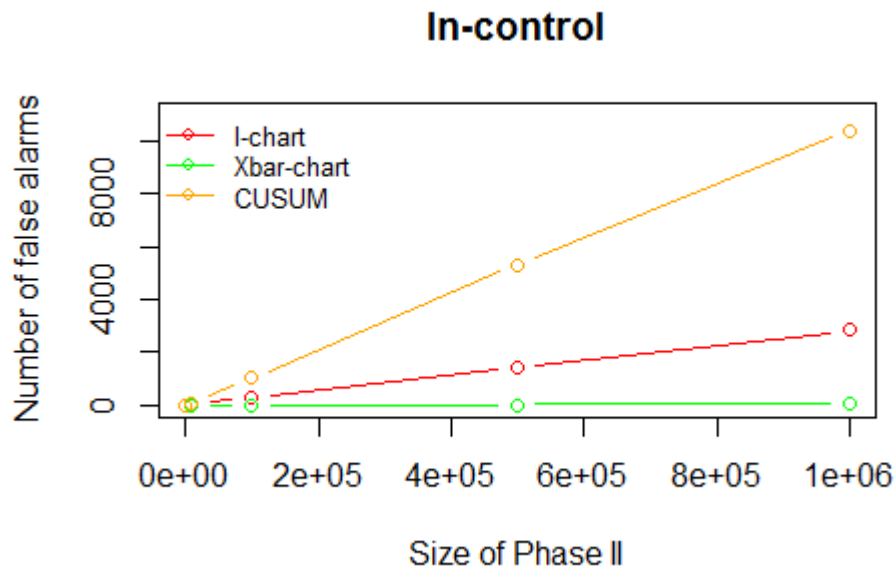


Figure 6. 8. Comparison between an I-chart, an \bar{X} -chart and a CUSUM chart.

It is apparent the best possible choice is an \bar{X} -chart. Recall data points are grouped into subsamples of 100 observations. The \bar{X} -chart presents reduced false alarms rates compared to the other two charts. Furthermore, it appears that the rate of increase in the number of false alarms for this chart is smaller than the other charts. As a consequence, it is preferable to the other control charts so far. However, this chart is not completely satisfactory, since it cannot detect outlying observations adequately. We should also point out the CUSUM chart appears to have the highest false alarm rates.

6.5 Combinations of I-charts with other types of control charts

Our next approach, discussed in detail in Chapter 2, involved the combination of an I-chart either with a p-chart or with a chart monitoring the times between alarms. We shall firstly analyze the former case.

In order to reduce the false alarm rates of a monitoring scheme, we suggested a combination of control charts. More particularly, the first control chart to be used is an I-chart. Then, an additional control chart will be used to monitor the percentage of points plotted beyond the control limits of the initial I-chart. This chart is a p-chart. To construct this particular p-chart, observations are divided into subgroups of size 100. The p-chart is built as described in Chapter 2. We ignore the alarms of the I-chart, i.e. we do not

take any corrective action. Instead, we plot the points on the p-chart and if the p-chart signals the process is out-of-control, then we take corrective actions. The simulations were performed for in-control and out-of-control cases. As previously stated, the data points are put into groups of 100. The sizes of the total sample were equal to 10,000, 100,000, 500,000 and 1,000,000 observations. We did not simulate samples of 2,000 observations, as the number of points plotted on the graph would be quite small. The initial I-chart is built with 3σ control limits. 3σ control limits are also used for the p-chart. The size of Phase I is again 2,000 observations, which corresponds to 20 plotted points. This combination of control charts was tested using in-control and out-of-control scenarios. Two cases of out-of-control states were considered; the persistent shift of the process and the presence of outliers at fixed positions. The exact same scenarios were used as with the previously discussed control charts. The following tables contain the results for the in-control as well as for the out-of-control states of the process (Table 6.22 and 6.23).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
10,000	9,000	2.243
100,000	99,000	26.579
500,000	499,000	135.972
1,000,000	999,000	267.694

Table 6. 22. Results of the in-control simulations for the p-chart with 3σ control limits based on an I-chart.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
10,000	9,000	2.175	3.084
100,000	99,000	26.992	27.019
500,000	499,000	138.496	151.261
1,000,000	999,000	272.035	287.735

Table 6. 23. Results of the out-of-control simulations for the p-chart with 3σ control limits based on an I-chart.

Based on these results, we conclude the false alarm rate of the p-chart is decreased compared to the false alarm rate when the I-chart is examined in isolation. Yet this chart is no different when it comes to the increase of false alarms with the increase of the sample size. It is evident the number of false alarms rises when large datasets are handled. We also observe that its detection ability is not very satisfactory. More particularly, the alarm rates, when a persistent shift has occurred in the process, are approximately equal to the false alarm rates. The alarm rates with the presence of outliers are slightly higher than the false alarm rates. Nevertheless, their differences are quite small, and the detection rate appears to be small as well. Hence, we may conclude the detection ability of this combination of charts is not satisfactory. These results are also displayed in Figure 6.9.

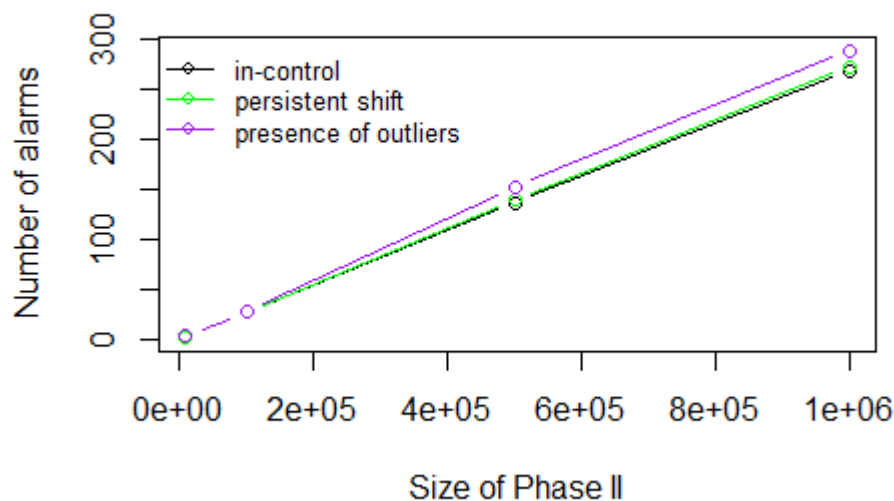


Figure 6. 9. Comparison of in-control and out-of-control states of a process for the combination of an I-chart and a p-chart.

Another suggested combination of control charts was that of an I-chart with a g-chart. Recall that a g-chart is a control chart based on the Negative Binomial distribution and monitors the times between two events (see also Chapter 2). Simulation studies were performed for this combination as well. Both in-control and out-of-control scenarios were examined. However, in the case of the I-chart and g-chart combination, there is a problem with the size of Phase I. The g-chart used essentially changes the scale of the horizontal

axis. Instead of plotting the points against the sample number, the points are plotted against the number of the occurring alarm. We cannot predict when an alarm will occur. As a consequence, we cannot determine the size of Phase I for the g-chart. To resolve this issue, we may determine Phase I consists of 25 points. Then we let the process run until we have these 25 points and construct the g-chart in the traditional fashion. An issue occurred when for some choices of sample sizes, 25 alarms did not occur throughout the whole process. In these cases, half of the alarms were used for Phase I and the rest of them were used for Phase II. The control limits of both the I-chart and the g-chart were of 3σ . The simulation scenarios were exactly the same as the ones used for the combination of an I-chart with a p-chart. The results of these simulations are displayed in the following tables (Tables 6.24 and 6.25).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
10,000	9,000	0.684
100,000	99,000	7.332
500,000	499,000	34.520
1,000,000	999,000	70.194

Table 6. 24. Results of the in-control simulations for the g-chart with 3σ control limits based on an I-chart.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
10,000	9,000	0.680	0.550
100,000	99,000	7.008	6.497
500,000	499,000	35.231	33.948
1,000,000	999,000	69.076	68.797

Table 6. 25. Results of the out-of-control simulations for the g-chart with 3σ control limits based on an I-chart.

This combination of control charts presents a completely identical image, when it comes to the increase of false alarms with the increase of sample size. It also cannot detect any change of the process. These results are shown in Figure 2.10 as well. This fact was expected due to the construction of the g-

chart. The lower control limit is almost always equal to 0. Consequently, alarms are considered to be the points plotted beyond the upper control limit. However, we are not interested in these points, since they do not provide us with useful information about the state of the process. The informative points would be those, which plot below the lower control limit.

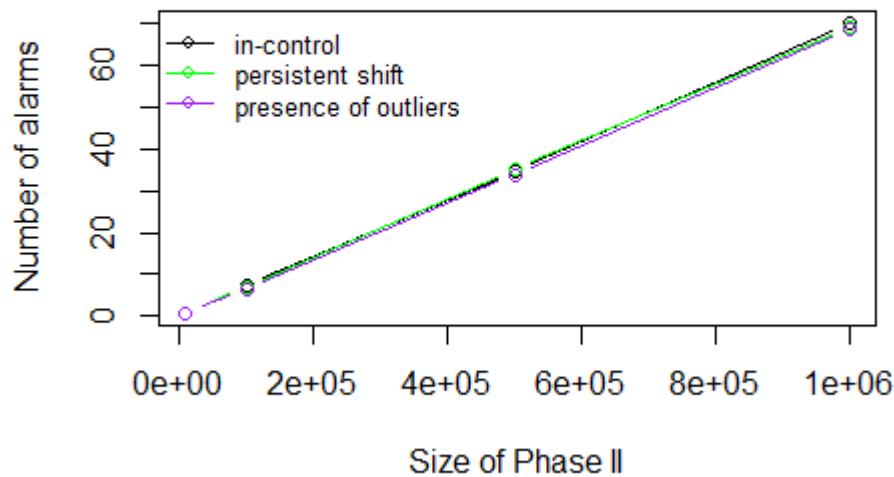


Figure 6. 10. Comparison of in-control and out-of-control states of a process for the combination of an I-chart and a g-chart.

The next combination of control charts involves an I-chart and an alternative g-chart. This g-chart is constructed in a probabilistic manner and only contains a lower control limit. The times between two consecutive alarms of the I-chart are plotted on the g-chart. The advantage of this chart is that a Phase I is not needed for its construction. Its control limit depends only on the width of the control limits of the initial I-chart (see Chapter 2). The same in-control and out-of-control simulation scenarios were performed in order to evaluate the performance of this combination. The I-chart used has 3σ control limits. These results are shown in the following tables (Tables 6.26 and 6.27).

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
10,000	9,000	0.068
100,000	99,000	0.911
500,000	499,000	4.355
1,000,000	999,000	8.420

Table 6. 26. Results of the in-control simulations for the probabilistic g-chart based on an I-chart.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
10,000	9,000	0.096	0.098
100,000	99,000	0.886	1.112
500,000	499,000	4.427	5.612
1,000,000	999,000	8.319	11.072

Table 6. 27. Results of the out-of-control simulations for the probabilistic g-chart based on an I-chart.

We observe the false alarm rates are low for this combination of control charts. The pattern of the increase of false alarms with the increase of sample size we have observed so far is true for this combination as well. In addition, its detection ability is not at the desired levels. In particular, persistent shifts in the process are only detected by chance. Furthermore, detection of outliers seems to be a little higher, but it does not reach the desirable level. These results are also shown in Figure 6.11.

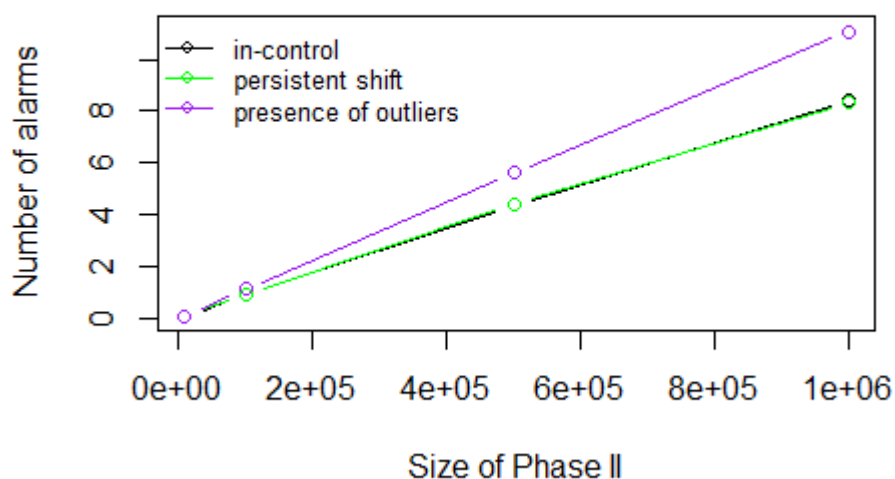


Figure 6. 11. Comparison of in-control and out-of-control states of a process for the combination of an I-chart with a probabilistic g-chart.

Last but not least, we suggested using a different chart for monitoring the times between two consecutive alarms of an I-chart. This monitoring can be achieved by firstly using a Weibull transformation on these times, approximating with the Normal distribution and therefore using another I-chart. This I-chart is quite different than the initial one. The issues we encountered with the size of Phase I remain unresolved when using this approach. Nevertheless, we are interested in gaining insight of the behavior of such a combination. The same simulation scenarios were used for in-control states of the process, as well as for out-of-control states of the process. The results of these simulations are shown in Tables 6.28 and 6.29 and in Figure 6.12.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of false alarms per iteration</i>
10,000	9,000	0.221
100,000	99,000	0.229
500,000	499,000	0.950
1,000,000	999,000	1.963

Table 6. 28. Results of the in-control simulations for a Weibull I-chart based on an I-chart.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Mean number of alarms per iteration</i>	
		<i>Persistent shift</i>	<i>Outliers</i>
10,000	9,000	0.212	0.184
100,000	99,000	0.300	0.205
500,000	499,000	1.055	0.790
1,000,000	999,000	1.906	1.576

Table 6. 29. Results of the out-of-control simulations for a Weibull I-chart based on an I-chart.

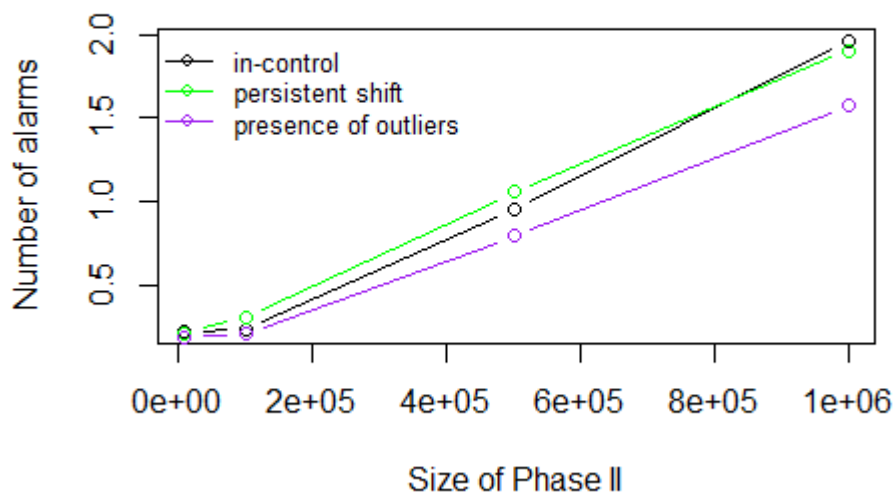


Figure 6. 12. Comparison of in-control and out-of-control states of a process for the combination of an I-chart with a Weibull I-chart.

This combination of control charts yields somehow different results. The number of false alarms increases when the sample size increases, but this increase happens differently from the previous cases. If we examine Figure 6.12 closely, we notice that the increase of alarms when sample points increase from 10,000 to 100,000 is very small. In fact, the two numbers are almost equal. We also notice the detection ability of this monitoring scheme is clearly inadequate.

Overall, all the aforementioned combinations of control charts reduced the false alarm rates. Nonetheless, they shall not be preferred for monitoring purposes, since their detection ability is not satisfactory.

6.6 Kolmogorov-Smirnov and its alternatives

In Chapter 3, we discussed how the Kolmogorov-Smirnov test could be implemented to the monitoring procedure of a process. We also suggested some alternatives for it, in order to make it more effective. We shall now describe in detail the procedure followed for the simulations. For each test statistic the following simulations were performed:

1. a set of in-control simulations to determine the appropriate thresholds of the distribution of each test statistic
2. in-control simulations to investigate the false alarm rates. The in-control distribution was the $N(0,1)$. The sample sizes used were 10,000, 100,000, 500,000 and 1,000,000. In each simulation the first 1,000 observations were considered to be the prototype dataset (i.e. data drawn from the in control distribution), to which every other subsample was compared.
3. simulations for various out-of-control states of the process. We investigated persistent shifts and the presence of outliers in fixed positions. We considered persistent shifts of the location parameter, as well as persistent shifts of the scale parameter. For the former case, the last observations were generated from a $N(0.5,1)$. For the latter case, the last observations were generated from a $N(0,1.5)$. The length of the shift was active for 50, 100, 200, 300, 400 and 500 observations. Outlying observations were generated from a $N(3,1)$. Firstly, one outlier per 1,000 observations was inserted in Phase II. We also tried increasing the number of outliers (per 1,000 observations) to 2,3,4,5,10,25,50,75 and 100.

We will now present the results for each test statistic in the following tables and figures (Tables 6.30 to 6.36, Figures 6.13 to 6.19). All the thresholds used were selected at 95% significance level. Furthermore, all the following tables display the results, either when a persistent shift of 500 observations has occurred or when one outlier is present in a subsample of 1,000 observations. Lastly, we perform one-sided hypothesis testing for each test statistic except for test statistic D_5 . The hypothesis testing for this test statistic is two-sided (see also Chapter 3 for more details).

- Kolmogorov-Smirnov test statistic D

Recall that this test statistic is equal to

$$D = \max(|F_m(x) - G_n(x)|) \text{ (formula (3.4)).}$$

Sample size	Size of Phase II	Percent of alarms			
		In-control	Persistent shift (location)	Persistent shift (scale)	Outliers
2,000	1,000	0.048	0.999	0.779	0.050
10,000	9,000	0.055	1	0.768	0.052
100,000	99,000	0.047	0.998	0.785	0.045
500,000	499,000	0.050	0.997	0.791	0.045
1,000,000	999,000	0.048	0.995	0.770	0.049

Table 6. 30. Results of simulations for Kolmogorov-Smirnov test statistic D .

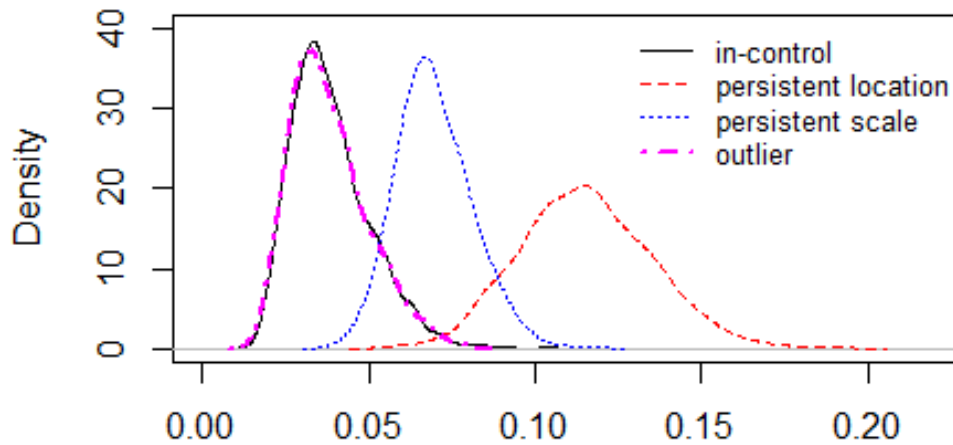


Figure 6. 13. Distribution of the test statistic D in various cases.

- Kolmogorov-Smirnov test statistic D_1

Recall that this test statistic is equal to

$$D_1 = \sum_{i=1}^{m+n} |F_m(x_i) - G_n(x_i)| \text{ (formula (3.5)).}$$

Sample size	Size of Phase II	Percent of alarms			
		In-control	Persistent shift (location)	Persistent shift (scale)	Outliers
2,000	1,000	0.060	1	0.907	0.049
10,000	9,000	0.058	1	0.912	0.047
100,000	99,000	0.049	0.999	0.907	0.053
500,000	499,000	0.050	0.999	0.912	0.054
1,000,000	999,000	0.051	1	0.921	0.058

Table 6. 31. Results of simulations for Kolmogorov-Smirnov test statistic D_1 .

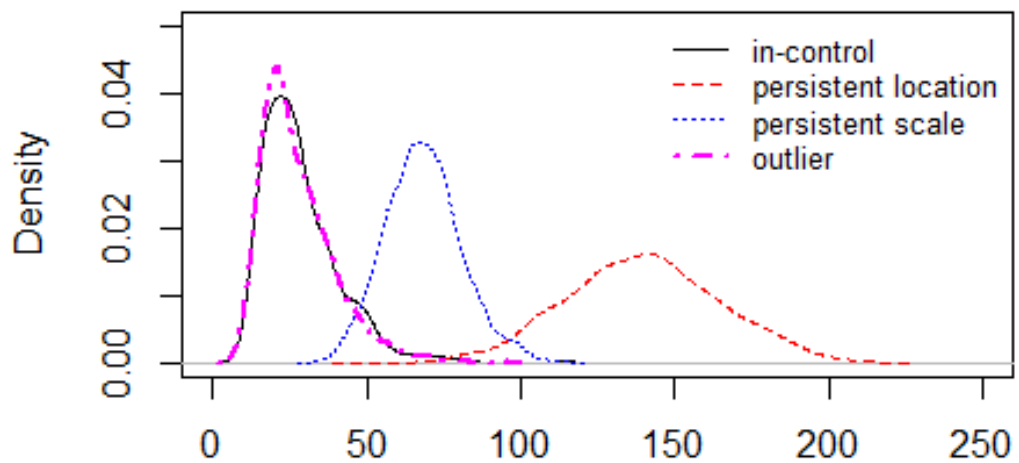


Figure 6. 14. Distribution of the test statistic D_1 in various cases.

- Kolmogorov-Smirnov test statistic D_2

Recall that this test statistic is equal to

$$D_2 = \frac{1}{m+n} \sum_{i=1}^{m+n} (F_m(x_i) - G_n(x_i))^2 \text{ (formula (3.6)).}$$

Sample size	Size of Phase II	Percent of alarms			
		In-control	Persistent shift (location)	Persistent shift (scale)	Outliers
2,000	1,000	0.061	0.999	0.900	0.049
10,000	9,000	0.053	1	0.928	0.050
100,000	99,000	0.051	0.999	0.896	0.052
500,000	499,000	0.055	1	0.900	0.054
1,000,000	999,000	0.055	1	0.907	0.052

Table 6. 32. Results of simulations for Kolmogorov-Smirnov test statistic D_2 .

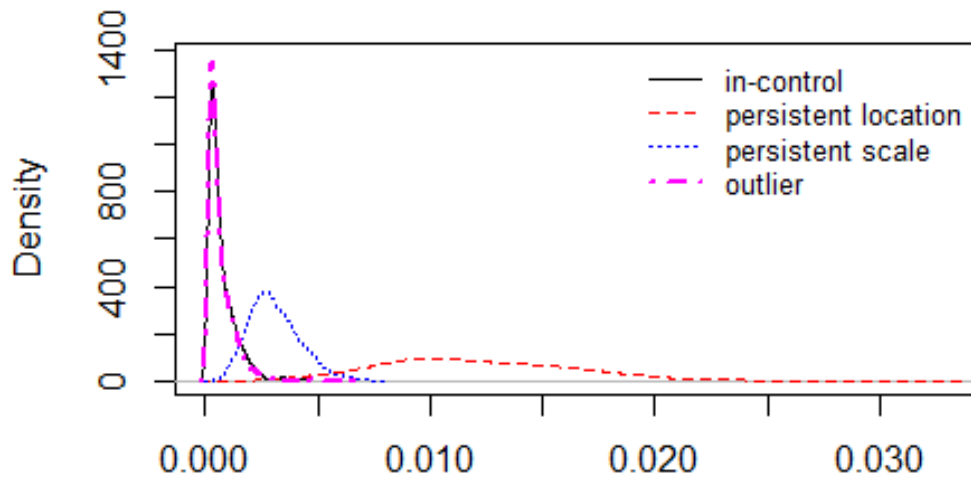


Figure 6. 15. Distribution of the test statistic D_2 in various cases.

- Kolmogorov-Smirnov test statistic D_3

Recall that this test statistic is equal to

$$D_3 = \sum_{i=2}^{m+n} \frac{|F_m(x_i) - G_n(x_i)| + |F_m(x_{i-1}) - G_n(x_{i-1})|}{2} (x_i - x_{i-1}) \text{ (formula (3.7)).}$$

Sample size	Size of Phase II	Percent of alarms			
		In-control	Persistent shift (location)	Persistent shift (scale)	Outliers
2,000	1,000	0.052	1	1	0.050
10,000	9,000	0.055	0.999	1	0.062
100,000	99,000	0.053	1	1	0.06
500,000	499,000	0.052	1	1	0.058
1,000,000	999,000	0.050	1	1	0.056

Table 6. 33. Results of simulations for Kolmogorov-Smirnov test statistic D_3 .

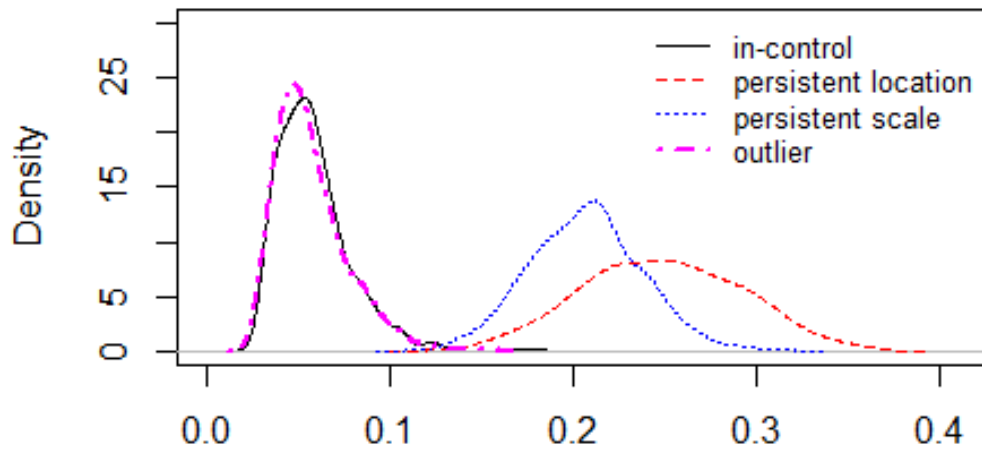


Figure 6. 16. Distribution of the test statistic D_3 in various cases.

- Kolmogorov-Smirnov test statistic D_4

Recall that this test statistic is equal to

$$D_4 = \max(x_i - x_{i-1}) \sum_{i=1}^{m+n} |F_m(x_i) - G_n(x_i)| \text{ (formula (3.8)).}$$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Persistent shift (location)</i>	<i>Persistent shift (scale)</i>	<i>Outliers</i>
2,000	1,000	0.054	0.894	0.781	0.593
10,000	9,000	0.049	0.879	0.747	0.597
100,000	99,000	0.053	0.853	0.749	0.603
500,000	499,000	0.052	0.859	0.754	0.603
1,000,000	999,000	0.055	0.879	0.747	0.599

Table 6. 34. Results of simulations for Kolmogorov-Smirnov test statistic D_4 .

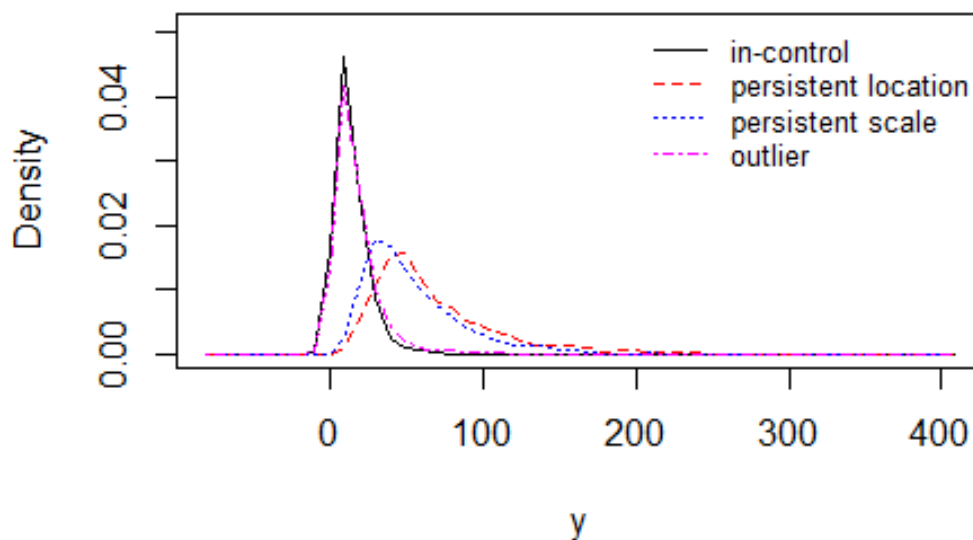


Figure 6. 17. Distribution of the test statistic D_4 in various cases.

- Kolmogorov-Smirnov test statistic D_5

Recall that this test statistic is equal to

$$D_5 = \sum_{i=1}^{m+n} (F_m(x_i) - G_n(x_i)) \text{ (formula (3.9)).}$$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Persistent shift (location)</i>	<i>Persistent shift (scale)</i>	<i>Outliers</i>
2,000	1,000	0.052	1	0.059	0.049
10,000	9,000	0.056	1	0.048	0.046
100,000	99,000	0.058	1	0.047	0.049
500,000	499,000	0.058	1	0.043	0.056
1,000,000	999,000	0.063	0.999	0.061	0.055

Table 6. 35. Results of simulations for Kolmogorov-Smirnov test statistic D_5 .

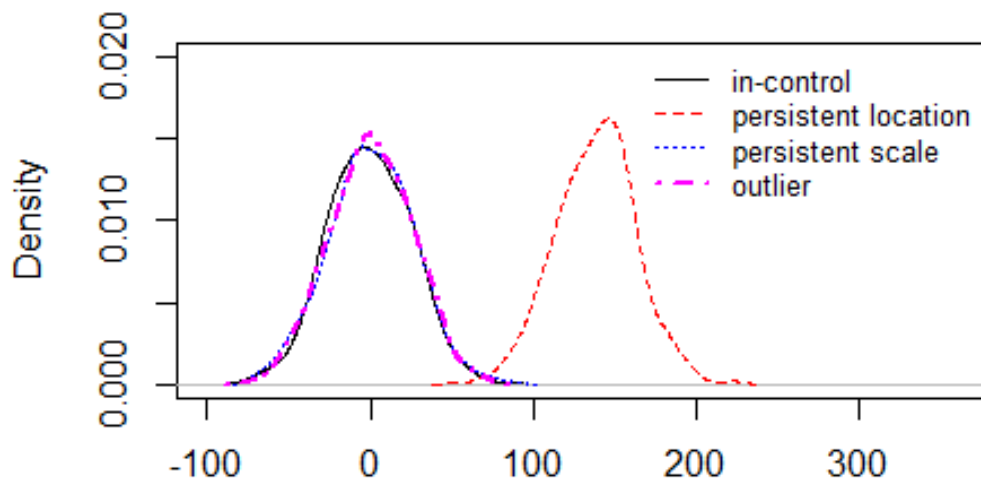


Figure 6. 18. Distribution of the test statistic D_5 in various cases.

- Kolmogorov-Smirnov test statistic D_6

Recall that this test statistic is equal to

$$D_6 = \max(F_m(x_i) - G_n(x_i)) \text{ (formula (3.10)).}$$

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Persistent shift (location)</i>	<i>Persistent shift (scale)</i>	<i>Outliers</i>
2,000	1,000	0.056	1	0.635	0.045
10,000	9,000	0.045	1	0.663	0.050
100,000	99,000	0.040	1	0.675	0.044
500,000	499,000	0.046	1	0.661	0.051
1,000,000	999,000	0.044	0.998	0.658	0.051

Table 6. 36. Results of simulations for Kolmogorov-Smirnov test statistic D_6 .

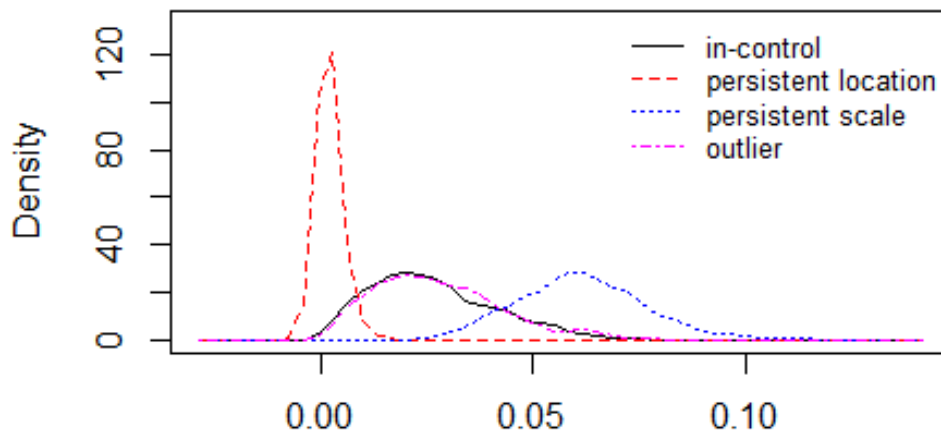


Figure 6. 19. Distribution of the test statistic D_6 in various cases.

Let us now comment on these results. The traditional test statistic of Kolmogorov-Smirnov test D does not display increased false alarm rates when the sample size increases, detects almost perfectly a persistent shift of the location parameter, detects adequately a persistent shift of the scale parameter but cannot detect the presence of one outlier in 1,000 observations. These results are also validated by the distribution of the test statistic for each case. For instance, the in-control distribution and the distribution have almost no overlap when a persistent shift of the location parameter occurs. Thus, it is expected this test statistic to be capable of detecting such shifts. The results for all the other test statistics are quite similar as far as false alarm rates and persistent shifts of the location parameter are concerned. The only test statistic which can detect at some degree the presence of one outlier is D_4 . However, this conclusion is not fully supported by its distributions. Therefore, we shall not rely on this result. All the other test statistics cannot detect the presence of one outlier. The alarm rates are equal to the false alarm rates in each case of test statistic. We also observe that the test statistics D_1 , D_2 and D_3 detect more efficiently a persistent shift of the scale parameter than the test statistics D , D_4 and D_6 . The exception is test statistic D_5 . This statistic can detect persistent shifts of the scale parameter only by chance. The alarm rates of this shift are almost equal to the false alarm rates. In order to decide which test statistic performs best, it would be useful to examine how the detection rates behave when the size of the persistent shifts decreases and when the number of outliers in 1,000 observations increases. These scenarios were also examined and the results are displayed in the following figures (Figure 6.20 to 6.22).

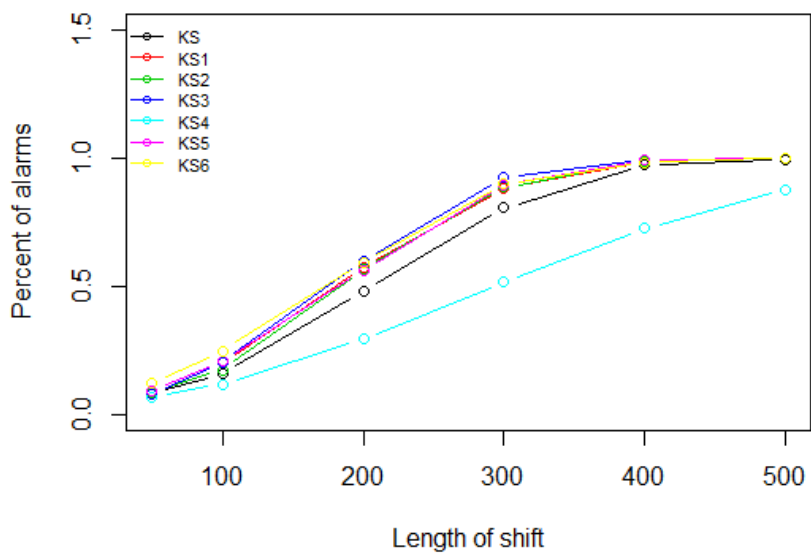


Figure 6. 20. Detection rates of all the Kolmogorov-Smirnov test statistics, when persistent shift of the location parameter occurs.

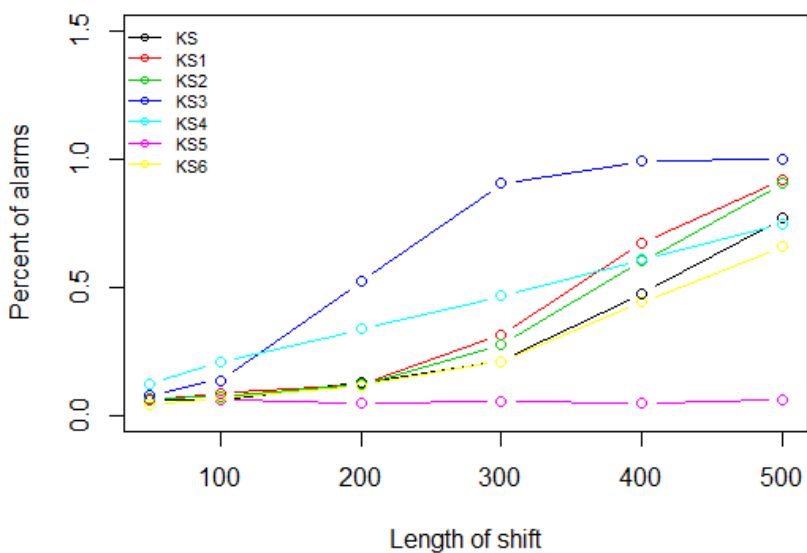


Figure 6. 21. Detection rates of all the Kolmogorov-Smirnov test statistics, when persistent shift of the scale parameter occurs.

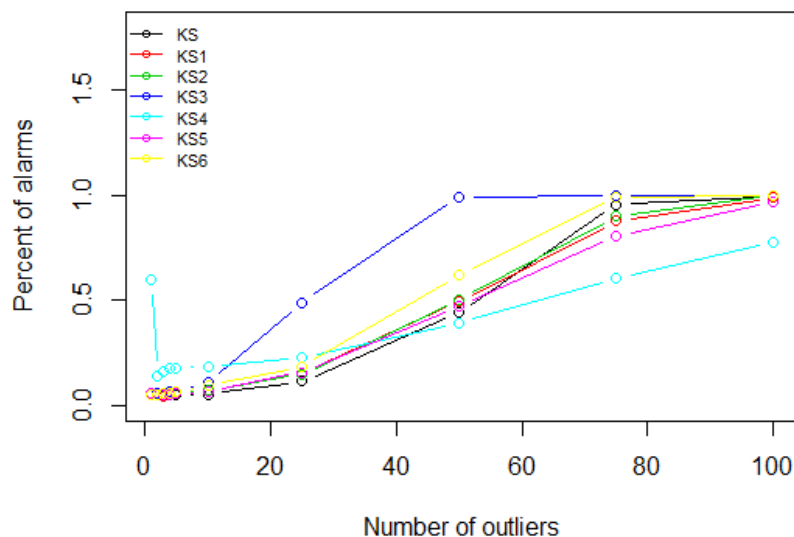


Figure 6. 22. Detection rates of all the Kolmogorov-Smirnov test statistics, when outliers are present.

The test statistic D_3 seems to perform best in every case. Consequently, if we wish to use some version of the Kolmogorov-Smirnov test for monitoring purposes, we shall use this particular test statistic.

6.7. Non-parametric LRT for stochastically ordered random variables

The next methodology, discussed in Chapter 4, is non-parametric LRT for stochastically ordered random variables. Recall that this methodology is analyzed in Franck (1984). It is important to note this test is designed for detecting usual stochastic ordering or equivalently first order stochastic dominance. We discussed how we can incorporate this methodology in the SPC context. In order to check its performance, we performed several simulation scenarios. The initial case investigated the in-control state of the process. The in-control distribution was the $N(0,1)$. We simulated samples of 10,000, 100,000, 500,000 and 1,000,000 observations and we examined the false alarm rates. Three out-of-control states of the process were also examined. The first assumed a persistent shift of the location parameter of the process. That being the case, the last 500 observations of each sample were generated from a $N(0.5,1)$. The sizes of the total sample were the same as for

the in-control case. The second assumed a persistent shift of the scale parameter of the process. The last 500 observations of each sample were generated from a $N(0,1.5)$. The third out-of-control scenario assumed the presence of one outlier per 1,000 observations. These outliers were generated from a $N(3,1)$. Each value of the test statistic was compared to the 95% quantile of its in-control distribution. The results of these simulations are shown in Table 6.37.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Persistent shift (location)</i>	<i>Persistent shift (scale)</i>	<i>Outliers</i>
10,000	9,000	0.041	0.999	0.175	0.056
100,000	99,000	0.045	0.997	0.180	0.062
500,000	499,000	0.048	0.999	0.167	0.057
1,000,000	999,000	0.046	0.998	0.152	0.063

Table 6. 37. Results of simulations for the non-parametric LRT for stochastically ordered random variables.

The false alarm rates of this methodology are close to the nominal value, as was expected. Moreover, this test is suitable for detecting a persistent shift of the location parameter, but it is not adequate for detecting a persistent shift of the scale parameter. These results were expected, since the particular test is designed to check for usual stochastic ordering of random variables (equivalently first order stochastic dominance). The scale shift constitutes stochastic dominance of second order. On that account, the test cannot detect well such shifts. Lastly, we notice that this test cannot detect the presence of one outlier in 1,000 observations. The alarm rates are almost equal to its false alarm rates.

Let us now examine how this test behaves when the length of the persistent shift decreases and when the number of outliers increases. We performed simulations for the persistent shift of the location parameter, where the last 50, 100, 200, 300 and 400 observations were generated from a $N(0.5,1)$. The same lengths of shift were used for the case of the persistent

shift of the scale parameter as well. We also inserted 2,3,4,5,10,25,50,75 and 100 outliers per 1,000 observations to examine the alarm rates. These results are shown graphically in Figure 6.23.

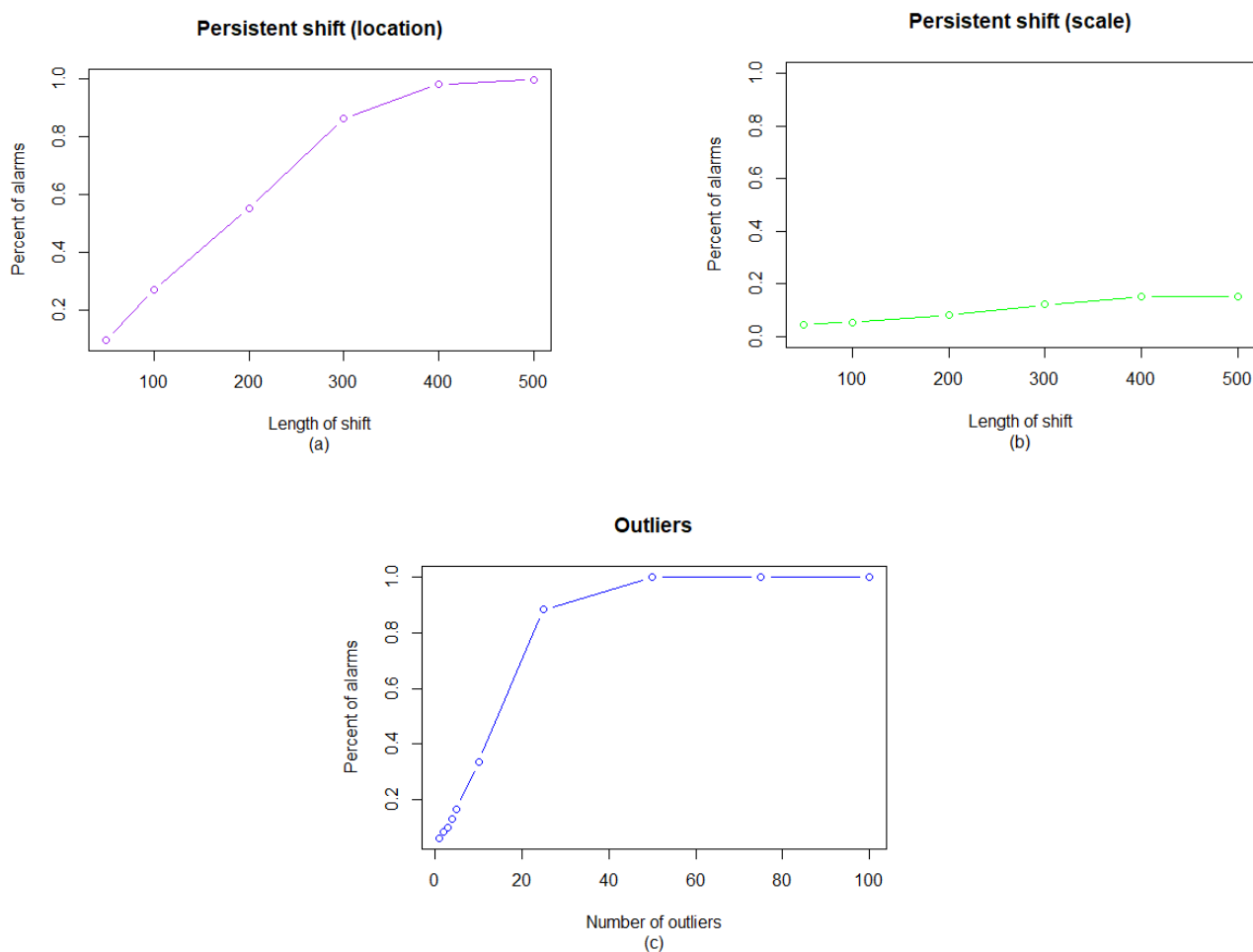


Figure 6. 23. Detection rates of the non-parametric LRT for various sizes of (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter and (c) outliers.

The detection rates increase when the size of a persistent shift in the location parameter has occurred. The same behavior is observed when the number of outliers increases. Yet this is not the case when a persistent shift in the scale parameter has occurred. The alarm rates remain in low levels when the size of the shift grows. Thus this methodology appears to be more suitable for detecting persistent shifts of the location parameter and the presence of a relatively large number of outliers (25 outliers or more).

6.8 Q-Q plots

The last methodology presented is the monitoring scheme with use of Q-Q plots (see also Chapter 5). Several simulations were performed to evaluate its performance. The following scenarios were examined:

1. in-control state of the process. We performed simulations to examine the behavior of false alarm rates, when the sample size increases. Therefore, we simulated samples from a $N(0,1)$ consisting of 10,000, 100,000, 500,000 and 1,000,000 observations.
2. persistent shift in the location parameter of the process. We generated samples from a $N(0,1)$ again, but the last 500 observations of each sample were generated from a $N(0.5,1)$. The total sample sizes were again 10,000, 100,000, 500,000 and 1,000,000 observations. We also generated shorter shifts. These shifts consisted of 50, 100, 200, 300 and 400 observations.
3. persistent shift in the scale parameter of the process. The in-control samples were generated from a $N(0,1)$, consisting of 10,000, 100,000, 500,000 and 1,000,000 data points. The last 500 observations of each sample were generated from a $N(0,1.5)$. Other choices for the length of the shift were the 50, 100, 200, 300 and 400 last observations.
4. presence of outliers in fixed positions. We inserted one outlier per 1,000 observations. These observations were generated from a $N(3,1)$. We also increased the number of outliers to 2, 3, 4, 5, 10, 25, 50, 75 and 100 per 1,000 observations to investigate how this methodology responds.
5. change in distribution family. We simulated two cases of distributional changes. In the first case, the in-control distribution was $N(3,3)$ and the out-of-control distribution was $\text{Gamma}(3,1)$. A sample of 1,000 observations was generated from each distribution. In the second case, the in-control distribution was $N(0,3)$ and the out-of-control distribution was $t(3)$. Again, a sample of 1,000 observations was generated from each one. We shall note the distributions used for each case have equal first and second moments. Moreover, Gamma distribution is more skewed than the Normal distribution and

Student's t has heavier tails than the Normal distribution. Hence, we have examined both possible problems, as mentioned in Chapter 5.

T^2 test statistic was used to detect persistent shifts, while M was used to detect changes in the distribution family of the process. Both test statistics were used to detect the presence of outliers. The hypotheses testing were performed at 5% level of significance. The appropriate thresholds were determined via separate simulations. The results of these simulations are displayed in Tables 6.38 and 6.39.

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Persistent shift (location)</i>	<i>Persistent shift (scale)</i>	<i>Outliers</i>
10,000	9,000	0.055	1	1	0.054
100,000	99,000	0.052	0.999	1	0.057
500,000	499,000	0.052	0.999	1	0.053
1,000,000	999,000	0.051	1	1	0.055

Table 6. 38. Results of the simulations for the T^2 .

<i>Sample size</i>	<i>Size of Phase II</i>	<i>Percent of alarms</i>			
		<i>In-control</i>	<i>Change in distribution</i>		<i>Outliers</i>
			<i>Gamma(3,1)</i>	<i>t(3)</i>	
10,000	9,000	0.051	1	0.980	0.050
100,000	99,000	0.048	1	0.981	0.049
500,000	499,000	0.047	1	0.981	0.046
1,000,000	999,000	0.047	1	0.981	0.047

Table 6. 39. Results for the simulations for the M .

Let us firstly discuss the results for the T^2 test statistic. The false alarm rates are approximately equal to the nominal value. In addition, this test statistic can perfectly detect persistent shifts in both the location and scale parameters. However, it cannot detect effectively the presence of one outlier

in 1,000 observations. The detection rate for this case is approximately equal to the false alarm rate.

As far as M is concerned, its false alarm rate is approximately equal to the nominal value. This test statistic can also detect changes in the distribution family quite effectively. More specifically, M can detect perfectly differences in the shape of the two distributions and can detect effectively differences in the tails of the two distributions. On the other hand, it cannot detect the presence of one outlier in 1,000 observations. These results are also validated by the distributions of the test statistics (see Figure 6.24 and 6.25).

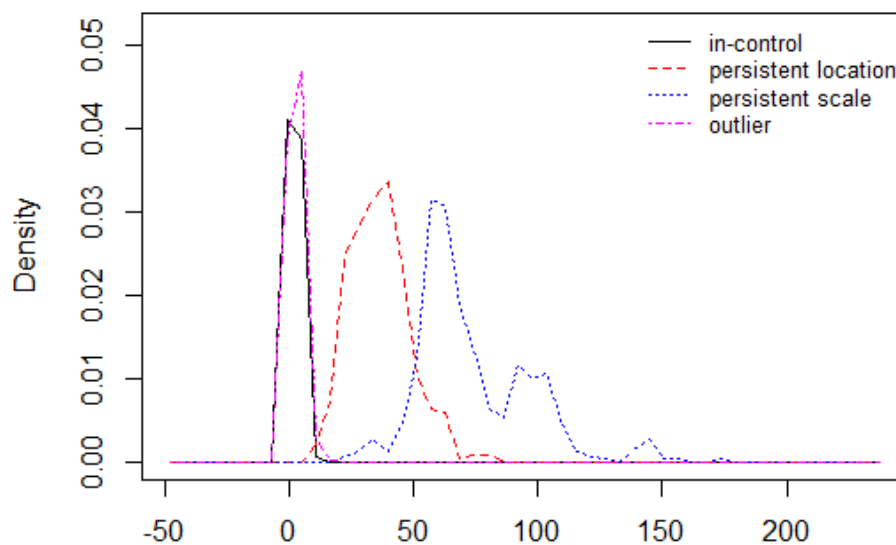


Figure 6. 24. Distribution of the test statistic T^2 in various cases.

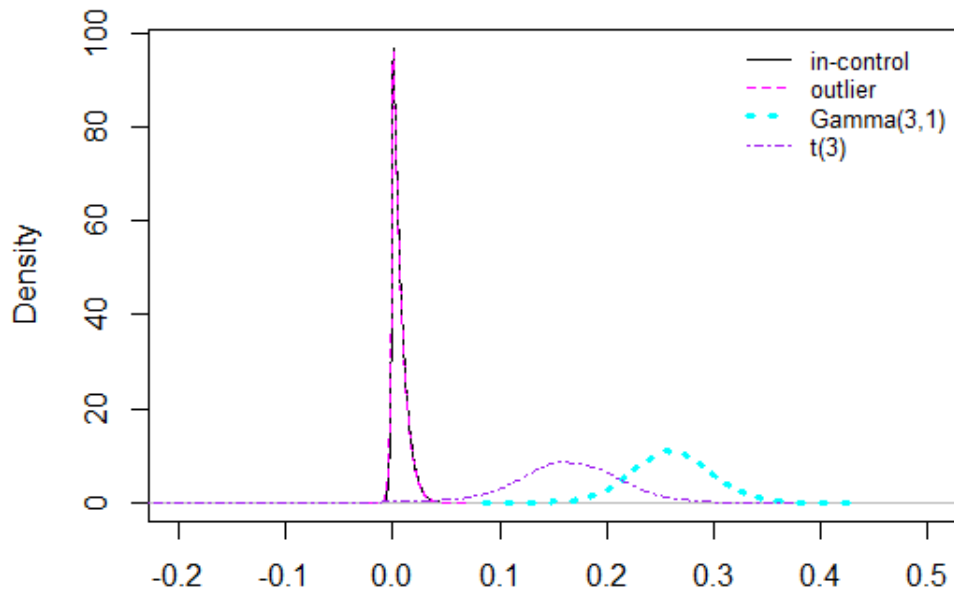
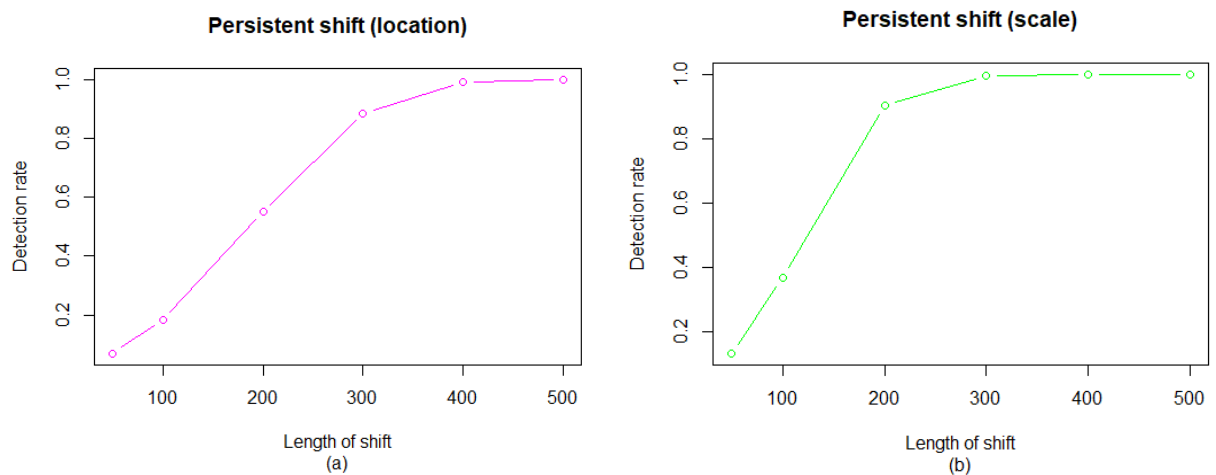


Figure 6. 25. Distribution of the test statistic M in various cases.

Let us now examine how the detection rates are affected, when the size of a shift changes. We wish to examine how the detection rate of a persistent shift reacts, when the length of the shift decreases. Furthermore, we wish to examine how the detection rate of outliers changes, when the number of outliers per 1,000 observations increases. These results are displayed in the following figures (Figure 6.26 and 6.27).



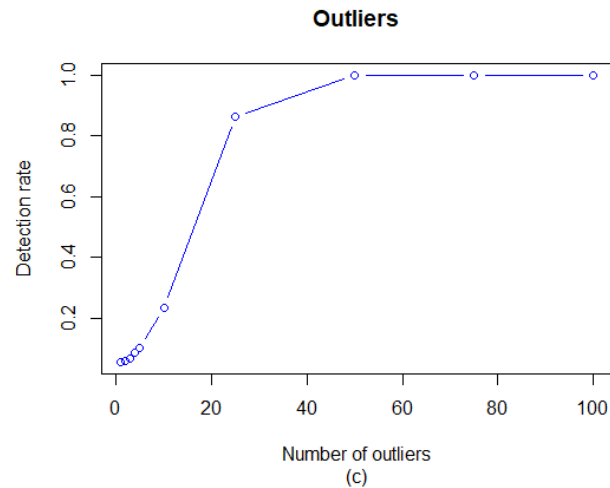


Figure 6. 26. Detection rates of the T^2 test for various lengths of (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter and (c) numbers of outliers.

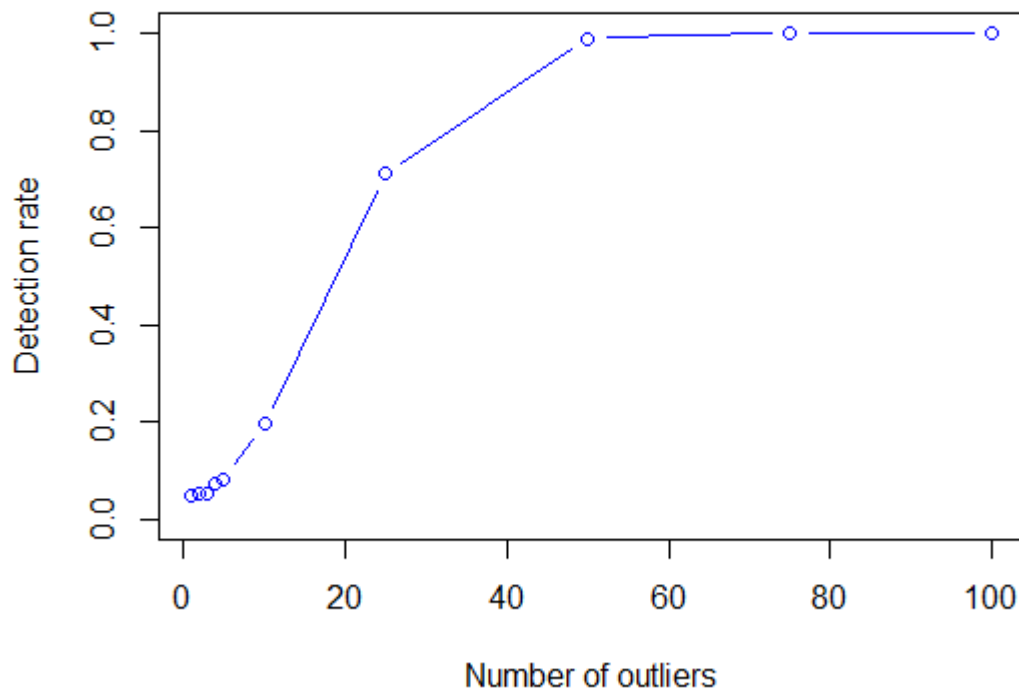


Figure 6. 27. Detection rates of the M test for increasing number of outliers. The detection rates of both test statistics increase when the size of a shift increases, as was expected. We also observe T^2 can detect effectively the presence of a relatively large number of outliers, i.e. 25 or more outliers.

As previously discussed, the thresholds used were determined via simulations. In practice, performing such simulations is a difficult task. Therefore, we also computed the necessary thresholds via resampling. We generated a sample used as the prototype consisting of 1,000 observations. Then, 999,000 other samples were generated via resampling. We computed the T^2 and the MSE for each one. The derived distributions are shown in the following figures (Figure 6.28 and 6.29).

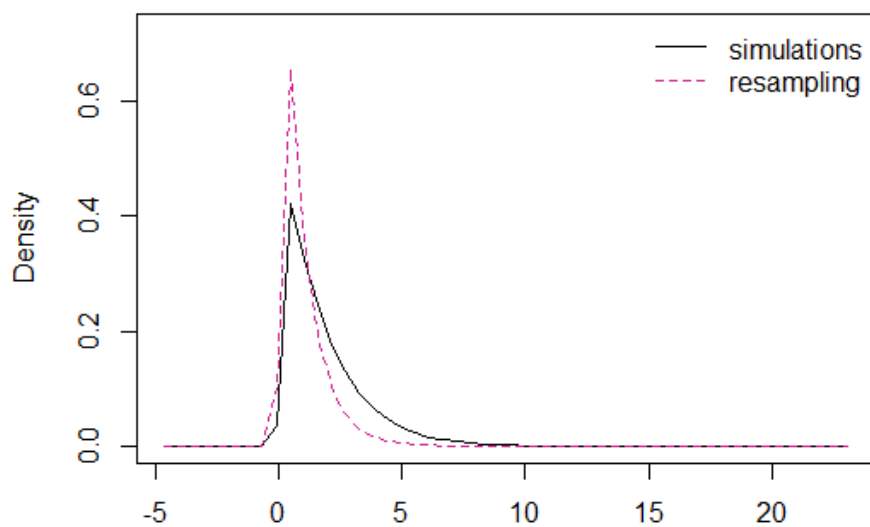


Figure 6. 28. Distribution of T^2 via simulations and via resampling.

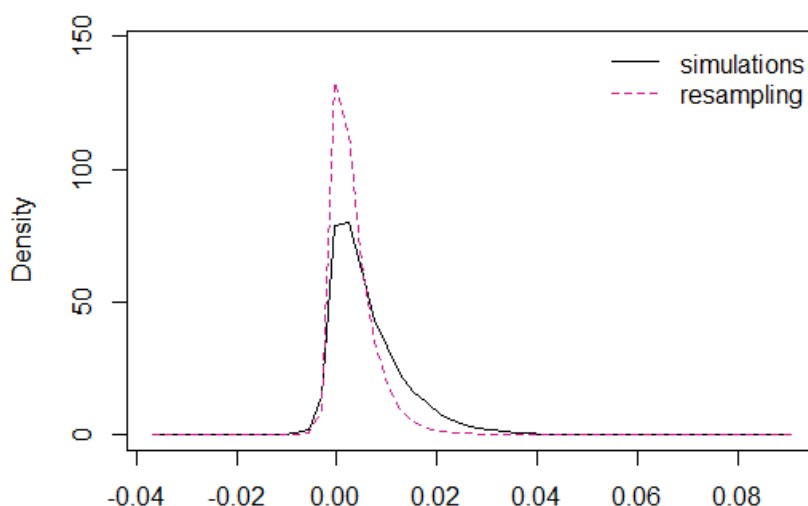


Figure 6. 29. Distribution of M via simulations and via resampling.

Unfortunately, there are slight differences in the distributions of both test statistics. One possible way to resolve this issue would be performing the hypotheses testing in different significance levels. When the resampling values are considered, the hypotheses testing would be performed in lower significance levels.

We shall now compare the performance of the proposed monitoring scheme to the ones examined in the previous sections. As shown in Figure 6.30, the T^2 test statistic performs equally well or even better than the test statistic D_3 (used for Kolmogorov-Smirnov test) and the non-parametric LRT test for stochastically ordered random variables. Furthermore, M is capable of detecting changes in distribution family almost perfectly. The results for the Kolmogorov-Smirnov alternatives were not as satisfying. LRT requires a lot of theoretical calculations to examine whether it can be used for detection of such changes or not. Overall, the proposed Q-Q plot based methodology seems to perform better.

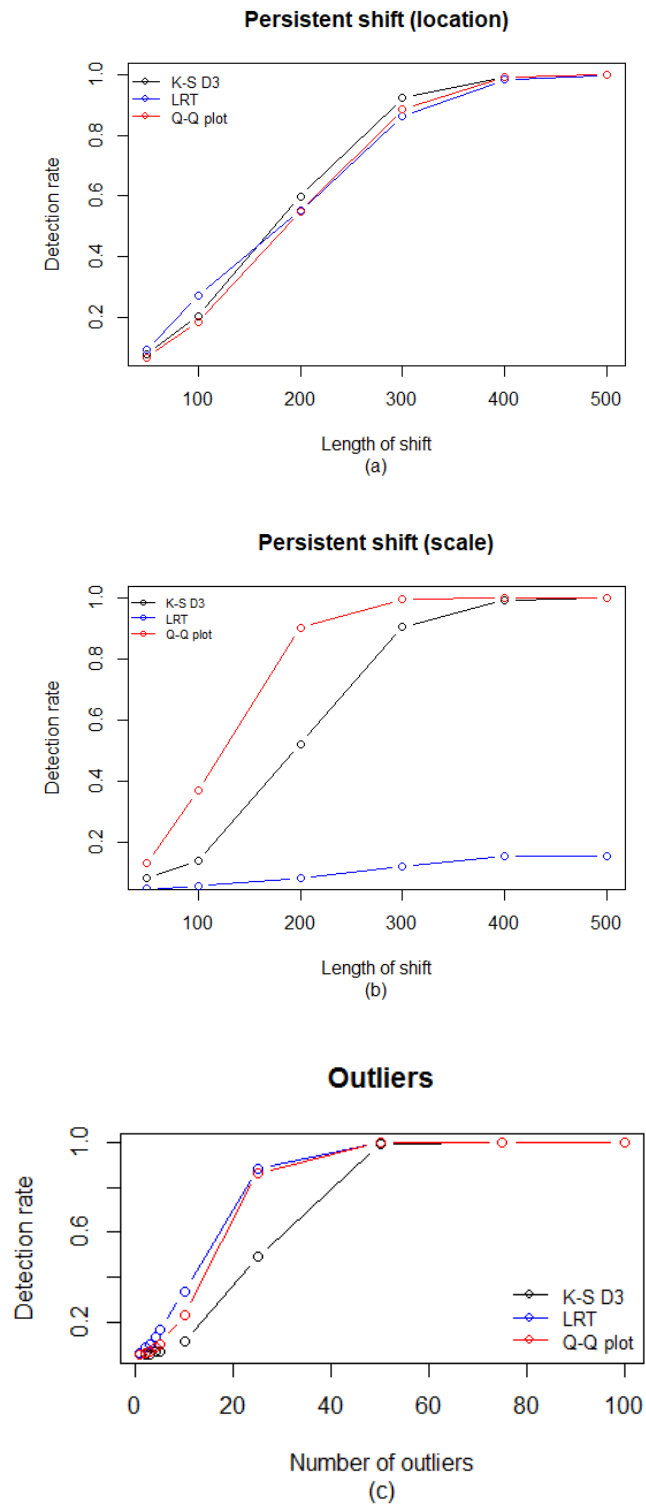


Figure 6.30. Comparison of the D_3 test statistic (Kolmogorov-Smirnov alternative), the non-parametric LRT and the T^2 test statistic, when either (a) persistent shift of the location parameter, (b) persistent shift of the scale parameter or (c) outliers are present.

Chapter 7

Conclusions

In conclusion, we examined the performance of some traditional control charts with big data. We also presented some other monitoring schemes, such as the Kolmogorov-Smirnov test, a non-parametric LRT test for stochastically ordered random variables, and the use of Q-Q plots in SPC. We concluded the use of Q-Q plots is the most effective methodology. It is capable of detecting multiple out-of-control scenarios, while the false alarm rate remains low. Future research could be focused in improving this methodology, as far as detection of outliers is concerned. For instance, we may use the residuals or some regression diagnostics to detect effectively the presence of a small number of outliers. It would also be useful to examine how the test statistics react, when the size of the subsamples used changes. It is quite important to examine the case when resampling is used to determine the appropriate thresholds for the test statistics. Finally, other cases of out-of-control states of a process could be examined, such as a time-dependent shift of the process mean.

References

- Benneyan, J. C. (2001).** Number-between g-type statistical quality control charts for monitoring adverse events. *Health care management science*, 4(4), 305-318.
- Casella, G., & Berger, R. L. (2002).** *Statistical inference* (Vol. 2). Pacific Grove, CA: Duxbury.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A (1983).** *Graphical methods for data analysis*, Wadsworth.
- Corder, G. W., & Foreman, D. I. (2014).** *Nonparametric statistics: A step-by-step approach*. John Wiley & Sons.
- Crowder, S. V. (1987).** A simple method for studying run-length distributions of exponentially weighted moving average charts. *Technometrics*, 29(4), 401-407.
- Deming, W. E. (1986).** Out of the crisis. Massachusetts Institute of Technology. *Center for advanced engineering study, Cambridge, MA, 510*, 419-425.
- Farouk, A. U., & Mohamad, I. B. (2012).** Average Run Length Efficiency of CUSUM Control Charts with Normal Distribution. *Archives Des Sciences*, 65(12).
- Franck, W. E. (1984).** A likelihood ratio test for stochastic ordering. *Journal of the American Statistical Association*, 79(387), 686-691.
- Mason, R. L., & Young, J. C. (2002).** *Multivariate statistical process control with industrial applications* (Vol. 9). Siam.
- Montgomery, D. C. (2009).** *Introduction to statistical quality control*. John Wiley & Sons (New York).
- Shaked, M., & Shanthikumar, J. G. (2007).** *Stochastic orders*. Springer Science & Business Media.
- Shen, Y., & Kaynak, O. K. Y. A. Y. (2015).** Big Data for Modern Industry: Challenges and Trends [Point of View]. *Proceedings of the IEEE*, 103(2), 143-146.

Sriboonchita, S., Nguyen, H. T., Wong, W. K., & Dhompongsa, S. (2009). *Stochastic dominance and applications to finance, risk and economics*. Chapman and Hall/CRC.

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge university press.