DYNAMIC SAMPLING VERSIONS OF POPULAR SPC CHARTS FOR BIG DATA ANALYSIS

by

Samuel Anyaso-Samuel

A thesis
submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Mathematics
Boise State University

May 2019

BOISE STATE UNIVERSITY GRADUATE COLLEGE

DEFENSE COMMITTEE AND FINAL READING APPROVALS

of the thesis submitted by

Samuel Anyaso-Samuel

Thesis Title: Dynamic Sampling Versions of Popular SPC Charts for Big Data Analysis

Date of Final Oral Examination: 04 March 2019

The following individuals read and discussed the thesis submitted by student Samuel Anyaso-Samuel, and they evaluated his presentation and response to questions during the final oral examination. They found that the student passed the final oral examination.

Partha Mukherjee, Ph.D.	Chair, Supervisory Committee
Jaechoul Lee, Ph.D.	Member, Supervisory Committee
Jodi Mead, Ph.D.	Member, Supervisory Committee

The final reading approval of the thesis was granted by Partha Mukherjee, Ph.D., Chair of the Supervisory Committee. The thesis was approved by the Graduate College.

dedicated to my Rock of Ages.

ACKNOWLEDGMENTS

I am eternally grateful to the Faculty and Staff of the Mathematics Program at Boise State for their academic, financial and moral support throughout the course of obtaining this degree. I also express my gratitude to the Graduate College for the resources and scholarships that were awarded to me during this period. These opportunities have been instrumental in reaching this milestone in my academic career.

In addition, I wish to thank my advisor, Dr. Partha Mukherjee for his matchless commitment to my overall development. His timely counsel has set me on the right path to make advancements in my career. I am also grateful to Dr. Jaechoul Lee and Dr. Jodi Mead, both of whom have provided awesome instruction and guidance throughout this period.

Furthermore, I am also appreciative of the affection, care and understanding from my parents, my siblings, the family at UBC, and my friends. Lastly, my cohort has been amazing and I am grateful for the collaborative efforts and camaraderie enjoyed during this period.

Ultimately, to God Almighty, from whom I have graciously received all good and perfect gifts, to Him alone be all glory, praise, and honor.

ABSTRACT

The statistical process control (SPC) chart is an effective tool for the analysis, interpretation, and visualization of data from sequential processes. Commonly used SPC charts such as the Shewhart, CUSUM and EWMA charts are widely implemented in detecting distributional shifts in various processes. With recent scientific and technological advancements, massive amounts of data continue to be generated by production, medical, agricultural and many other industrial processes. Conventional SPC charts have significant drawbacks in monitoring such processes, specifically when the velocity of the data flow is greater than the run time of the monitoring procedure. In the literature, dynamic sampling control charts [15] are becoming popular due to their ability to adaptively control the next sampling time of the monitoring process. In this thesis, we incorporate similar ideas to conventional SPC charts for the real-time monitoring of big data processes.

Traditional SPC charts are designed to give a warning signal at a particular time point if a process reading plots beyond its control limit(s). This approach does not provide ample information of the likelihood of a potential shift in the process. We implement existing methods of designing control charts with p-values, which gives information about the performance of the current observations and potentially, of observations in near future. The control chart gives a signal for a mean shift if the p-value is less than some pre-specified significance level. We utilize the computed p-values of the charting statistic in designing variable sampling schemes, specifically the dynamic sampling schemes which are an increasing function of the p-value. The resulting control charts have variable sampling intervals, and hence skips several observations. Thus, their computing times are much faster than traditional charts.

This thesis provides guidance on how to incorporate dynamic sampling schemes for monitoring big data streams in other types of SPC charts. We perform extensive simulation studies to compare the performance of the dynamic sampling control charts with conventional control charts. Our results show that the dynamic sampling versions of three commonly used SPC charts can monitor big data streams efficiently.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SPC – Statistical Process Control

IC – In-Control

OC – Out-of-Control

ARL – Average Runlength

ATS – Average time to Signal

AATS – Adjusted Average Time to Signal

CUSUM – Cumulative Sum

EWMA – Exponentially Weighted Moving Average

SS – Standard Shewhart

DyS-S – Dynamic-Sampling Shewhart

DyS-CUSUM – Dynamic-Sampling Cumulative Sum

DyS-EWMA – Dynamic-Sampling Exponentially Weighted Moving Average

CHAPTER 1

BACKGROUND

1.1 Introduction

Recent developments in science and technology have given birth to the big data era in which large volumes of data are consistently being generated from several sequential processes. From health care, manufacturing and production lines, network systems, Internet services, E-commerce and so on, the proliferation of data from these sequential processes has elicited the need to develop innovative methods capable of monitoring these processes. In most cases, the observations from these processes are obtained at individual time points, and they can be described as a random sample from a parametric statistical distribution.

Furthermore, due to the high velocity of observations from these data streams, it is very likely that the parameter value(s) of the statistical distribution which describes the data changes from time to time. Thus, these observations can be partitioned in such a way where different partitions of the data correspond to different parameter values of the statistical distribution. Take for instance, in the advent of a particular disease within a specific territory, as time goes on, the prevalence of such disease may diminish and thus it becomes imperative to estimate the time point such change in prevalence occurred. More commonly, it is customary procedure for industry manufacturers to monitor the conformance and quality of products obtained from a production line. In this sense, certain quality characteristics of the product from the routine processes are checked to ensure they meet some desired requirements.

Statistical Process Control (SPC) provides statistical tools which are employed to visualize patterns, monitor, and detect shifts (changes in parameters and/or statistical distribution) in a sequential process. The variation in a sequential process can be attributed to two basic sources − common cause variation and the assignable (special) cause variation. Common cause variation results from uncontrollable and unavoidable random variation in a production process which can only be eliminated by changing the entire process, while assignable cause variations are due to malfunction of certain components of the sequential process. When the variation in the production process is due to **only** common causes, we say that the process is in-control (IC), while if the source of variation in the process results from any assignable cause, we say the process is out-of-control (OC). The main idea behind SPC is to monitor the sequential process and detect when such a system has shifted from being IC to OC.

In the SPC literature, the most efficient procedure for monitoring a sequential process is the control chart. The control chart is a plot of successive points of certain quality characteristics on a chart which is bounded by upper and/or lower control limit(s). The control limit(s) is chosen in such a way where the time to detection of a shift is reduced and false alarms are mitigated. Figure 1.1 shows a simple control chart which has the red dashed lines as control limits. The control chart is quite advantageous for the visualization of the performance of the sequential process, checking for shifts in the distribution of the sequential process and also understanding the effects of various interventions made in the process. A control chart gives an OC signal when one or more points fall beyond certain control limit(s).

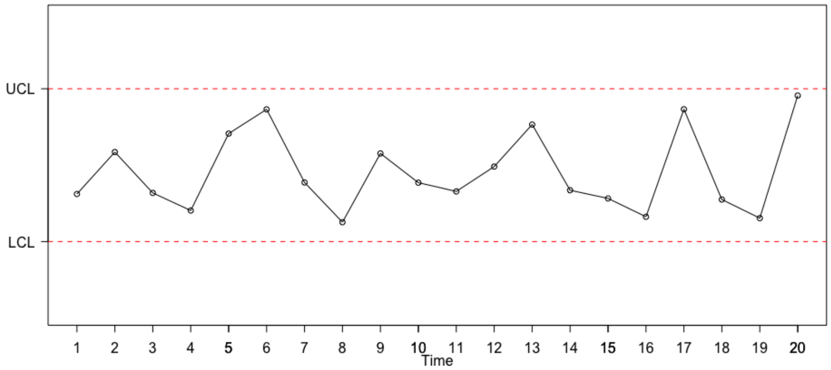

Figure 1.1: A sample control chart with upper and lower control limits depicted by the red dashed lines.

1.1.1 The Average Run Length

Given the varying underlying assumptions for the implementation of the different control charts to be presented later, the performance of each control chart varies to a significant extent. In order to evaluate the performance of the control charts, the average run length (ARL) has been used extensively in the literature. The run length of a chart can be defined as the number of process readings considered to be IC before an OC reading is observed. Thus, the average run length is the expected number of points plotted on the chart until an OC signal is obtained. An IC ARL, denoted as ARL_0, is the ARL associated with a zero-valued shift (an IC data set is being analyzed). Since there should be no shift when an IC dataset is analyzed, ARL_0 represents the number of observations until a false OC signal is given. In contrast, an OC ARL, denoted as ARL_1 is the ARL associated with a non-zero shift. It represents the number of observations from the time point at which the shift occurred to the time point at which the control chart gives a signal. Ideally, when the process is IC, we want the run length of the process to be long (as large as possible), however, when the process goes OC, the time to detection should be as short as possible. However, this desideratum is quite difficult to achieve. This is analogous to the idea behind the

Type-I and Type-II error probabilities in hypothesis testing. In the SPC literature, we usually fix the ARL_0 at a given level and try to make the ARL_1 value as small as possible. In other words, we fix the false-alarm rate and then minimize the chance of missing an actual shift.

1.1.2 Phase-I and Phase-II monitoring

Generally speaking, the monitoring of process observations can be categorized into two phases, the Phase I and Phase II monitoring. Each phase has a distinct objective. For the Phase-I SPC, data from a sequential process are collected and analyzed in a backward-looking fashion. This retrospective analysis of the process observations aims at estimating the distribution of the sequential process and also getting the control limits for the control charts to be used in subsequent analysis. This is usually done by understanding the relationship between certain controllable input variables and the quality characteristics of interest. The controllable input variables are then set at optimal values and a set of process observations is collected and analyzed with the trial control limits. If a fault is noticed while monitoring the observations, the OC observations are investigated and discarded. Then, the input variables are re-adjusted and a new set of observations is collected and analyzed. This process of fine-tuning the controllable input variables and constructing control limits is done repeatedly until all assignable causes of variation have been eliminated, thus making the process stable, then clean data are obtained from the process.

The primary objective of the Phase-II SPC is to give a signal when there is an evidence of distributional shifts during the online monitoring of the process. The Phase-II process begins with the IC process observations and control limits which were obtained from the Phase-I SPC, and are then used for online monitoring of subsequent observations obtained from the quality characteristic of interest. However, adaptive charts recalculate the control limits as more observations are collected.

In addition, a distributional shift in the sequential process can either be transient or persistent. If the shift is transient, the process goes OC but thereafter returns to being IC without any intervention. For persistent shifts, when the process leaves the IC state, it remains OC or even goes farther away from the IC state until a corrective intervention is made.

1.2 Traditional SPC charts

The SPC framework has several control charts for detecting several kinds of distributional shifts in a production process. In this section, we give a brief overview of some commonly used charts. The control charts discussed here are primarily used for the Phase-II SPC (in other chapters involving the corresponding charts, the case for the Phase-I SPC is discussed). Also, we assume that the quality characteristic of interest is univariate and numeric observations from this quality characteristic are obtained at equally spaced time points.

Consider the following independent observations obtained from the Phase-II monitoring of the sequential process

$$\begin{cases} X_1, X_2, ...X_\tau & \sim N(\mu_0, \sigma^2) \\ X_{\tau+1}, X_{\tau+2}, ... & \sim N(\mu_1, \sigma^2) \end{cases}$$

where τ is an unknown change point, μ_0 and μ_1 are the respective IC and OC means of the process ($\mu_0 \neq \mu_1$), and σ^2 is the process variance. In order to describe the SPC charts, we assume that a shift occurs only in the mean of the process. The charts to be presented can also be modified to give signals for shifts in the variance and shifts in both mean and variances of the process. Also, the IC parameters, μ_0 and σ^2, are usually unknown and should be estimated in the Phase-I SPC. With the process defined above, we begin the discussion of the charts.

1.2.1 The Shewhart Control Chart

Developed by Walter A. Shewhart in 1931, the Shewhart chart is a control chart based on the framework of hypothesis testing. The upper, center and lower control limits of the Shewhart control chart are defined as

$$U = \mu_0 + Z_{1-\alpha/2}\, \sigma; \quad C = \mu_0; \quad L = \mu_0 - Z_{1-\alpha/2}\, \sigma \tag{1.1}$$

where α is the significance level and $Z_{1-\alpha/2}$ is the (1 - $\alpha/2$)th quantile of the standard normal distribution. At time point n, the Shewhart control chart gives a signal for a mean shift if

$$X_n < L \qquad \text{or} \qquad X_n > U \tag{1.2}$$

Due to its simplicity, ease of implementation and interpretation of results, the Shewhart chart has gained a wide range of applications in industrial processes. The chart has proven to be efficient in detecting large and transient shifts in the mean of the process, this makes it appealing to the Phase-I SPC where such shifts are usually encountered.

However, since it disregards historical data when evaluating the performance of the process it performs poorly in detecting small and persistent shifts in the distribution of the process. In Chapter 2, we propose ideas to overcome this limitation. Also, the Shewhart chart works under the assumption that the process observations are normally distributed.

1.2.2 The Cumulative Sum Chart

In other to overcome the inability of the Shewhart chart to detect small and persistent in a process, Page [20] proposed the CUSUM chart which uses historical data to evaluate the performance of the sequential process at each time point. Historical data may contain vital information about the IC and OC performance of the process. The charting statistic of the CUSUM chart is based on the cumulative sum of the observation at the current time point and previous time points. This is given as

$$C_n = \sum_{i=1}^{n}(X_i - \mu_0) = C_{n-1} + (X_n - \mu_0) \quad \text{for } n \geq 1, \tag{1.3}$$

where $C_n = 0$. In order to detect upward and downward shifts in the process, (1.3) can be written as

$$\begin{cases} C_n^+ = \max\big(0, C_{n-1}^+ + (X_n - \mu_0) - k\big) \\ C_n^- = \min\big(0, C_{n-1}^- + (X_n - \mu_0) - k\big) \end{cases} \tag{1.4}$$

The CUSUM control chart gives an OC signal for a mean shift in the process if

$$C_n^+ > h \quad \text{or} \quad C_n^- < -h, \qquad \text{for } n \geq 1, \tag{1.5}$$

where $k > 0$ is a pre-specified reference parameter, and $h > 0$ is a control limit chosen to achieve a desired ARL_0 value. The $C_n^+(C_n^-)$ resets to 0 whenever there is an evidence of a shift in the process. Also, the CUSUM chart performs well when the process observations are normally distributed. In Chapter 3, we provide more discussion on the performance and implementation of the CUSUM control chart.

1.2.3 The Exponential Weighted Moving Average Chart

Another chart that circumvents the inability of the Shewhart chart to detect small and persistent shifts is the EWMA chart. This control chart which was proposed by Roberts [24] uses historical data to evaluate the performance of the process. The charting statistic of the EWMA chart is based on the weighted average of observation at the current time point and previous observations. This is given as

$$E_n = \nu X_n + (1-\nu)E_{n-1} \qquad \text{for } n \geq 1, \tag{1.6}$$

where $\nu \in (0,\ 1]$ is a pre-specified weighting parameter, and $E_0 = \mu_0$. And to detect upward and downward shifts in the mean of the process, we write (1.6) as

$$\begin{cases} E_n^+ = \max\big(0, \nu(X_n - \mu_0) + (1-\nu)E_{n-1}^+\big) \\ E_n^- = \min\big(0, \nu(X_n - \mu_0) + (1-\nu)E_{n-1}^-\big) \end{cases} \tag{1.7}$$

The chart gives an OC signal for a mean shift if

$$E_n^+ > \rho_U\sqrt{\frac{\nu}{2-\nu}} \quad \text{or} \quad E_n^- < \rho_L\sqrt{\frac{\nu}{2-\nu}}, \qquad \text{for } n \geq 1, \tag{1.8}$$

where $\rho_L, \rho_U > 0$ is a parameter chosen to achieve a desired ARL_0 value. Just like the CUSUM chart, the EWMA chart performs well for detecting small and persistent shifts in a process, this makes it suitable for the Phase-II SPC where such shifts are usually encountered. However, it performs fairly well in most applications where the process observations are not normally distributed.

1.2.4 Other SPC Control Charts

From Section 1.1, we notice that for the SPC problem, the univariate process has a common distribution before a shift occurs (IC distribution) and another distribution (OC distribution) after the shift occurs at an unknown time point. Change point detection (CPD) is a research area in the field of statistics that seeks to detect the specific position at which the distribution of a sequence of random variables changes from one to another. In the literature of CPD, the sample sizes are usually fixed and the distributions follow a parametric nature. Since the number of observations in the Phase-II SPC increases sequentially, change point detection cannot be directly applied to the SPC problem. However, on going research ([9], [10]) in SPC have modified CPD methods to handle the SPC problem and thus, CPD charts have been developed for the detection of distributional shifts in a sequential process.

The measurements from a quality characteristic of interest could be continuous, discrete or categorical. The charts presented in the previous sections focused on the case when continuous numerical observations are available. In cases when the quality characteristic is categorical or discrete but the number of different observations is small, control charts for categorical quality characteristics exist for monitoring the process. Examples of these charts include control charts for monitoring the proportion of non-conforming products of a production process and control charts for monitoring the number of defects in an inspection unit.

In addition, our description of the control charts in the preceding sections focused on monitoring of individual observations from a univariate process. In practice, multiple characteristics of a production unit may be needed to judge the quality of the product. Multivariate version of Shewhart, CUSUM, EWMA and CPD charts exist for detecting shifts in the mean and covariance matrix of the distribution of a multivariate production process.

Furthermore, for cases where the process observations are correlated, the CUSUM and the EWMA chart can be modified to handle such scenarios. Also, for monitoring a process when non-normal data are observed, several non-parametric charts have been developed. These include rank based non-parametric control chart which is based on ranking or ordering of information in the observed and non-parametric control chart by categorical data analysis which is based on observation categorization.

1.3 SPC and Big Data Analysis

A data stream can be simply defined as a constant stream of data flowing from a particular source. This includes data from a sensory machine, data from complex industrial and agricultural machines, data from web services or data from social media websites. In this case, each data is generally timestamped or geo-tagged. Furthermore, we define a stream as a possibly unbounded sequence of data items or records. These data items may be independent of each other or correlated with each other. In this setting, each data item is treated as an individual event in a synchronized sequence [14].

In the case where we have a sequential process that generates large volumes of data at a high velocity, the traditional charts presented above may not give optimal performance for monitoring such process. In this setting, the velocity of the data influx could be greater than the monitoring time of the SPC chart. Since traditional SPC charts monitor each and every observation in the process, they may not keep up with the pace at which process readings become available and therefore fail to give a signal for a distributional shift as early as possible. Considering this scenario, it becomes imperative to design and modify traditional control charts so that the complexity of monitoring large volumes of data is minimized and the run time of the monitoring process is reduced.

In order to improve the efficiency of SPC charts for monitoring big data processes, we use some existing methods in the literature to modify SPC charts. Particularly, we use p-values to design control charts and with the information obtained from the p-values, we skip observations that are IC during the monitoring procedure. The dynamic sampling scheme [15] will be used to determine how many observations are to be skipped during the monitoring procedure. Despite the goal of reducing the run time of the charts, we also intend to maintain the ability of the charts in quickly detecting distributional shifts.

1.4 Overview of Thesis

In this thesis, we primarily focus on the monitoring of independent numeric observations obtained from a univariate process at consecutive time points. In Chapter 2, we propose an adaptive Shewhart Chart for detecting small to moderate persistent mean shifts. We begin the discussion of the dynamic sampling schemes in Chapter 3, where the existing methodology is described and then implemented in the design of dynamic sampling Shewhart charts. We propose the dynamic sampling EWMA chart in Chapter 4. In Chapter 5, we restate the underlying assumptions for the implementation of the proposed charts and we provide directions for future research regarding design of SPC charts with dynamic sampling schemes.

CHAPTER 2

AN ADAPTIVE R-OUT-OF-M CONTROL CHART FOR DETECTING SMALL AND PERSISTENT PROCESS MEAN SHIFTS

2.1 The r-out-of-m control chart

It well-known that the Shewhart chart does not perform well in detecting small and persistent distributional shifts. This is because it does not utilize historical data which may contain useful IC and OC information during when evaluating the performance of a production process. In order to increase its sensitivity to small shifts, the Shewhart chart is usually implemented alongside with several other supplementary criteria.

One notable case discussed extensively in the literature is the accompanying rules earlier used in conjunction with the Shewhart chart by the Western Electric Company. To this effect, the Western Electric Company [8] proposed a set of decision rules for detecting nonrandom patterns on control charts. These decision rules increased the sensitivity of the Shewhart chart to small shifts, however, Champ and Woodall [5] studied the ARL performance of these rules and showed that they are usually suboptimal, in the sense that there is an increase in the number of false alarms when these rules are employed. For instance, the simultaneous use of the decision rules yields an IC ARL of 91.75 which is significantly lower than 370.4 of the standard

Shewhart chart.

To overcome the problem of the increase in the false alarm rate of the sensitivity rules, Klein [13] proposed two alternative schemes based on the standard runs rules. In the first scheme called the two-of-two scheme, the control chart gives an OC signal if two successive points plot above (below) an upper (lower) control limit. In the second scheme, called the two-of-three scheme, the control chart gives an OC signal if two of three successive points plot above (below) an upper (lower) control limit. The control limits for these schemes are symmetric and were estimated in such a way that the schemes have the same IC ARL as that of the standard Shewhart chart. His study shows that both schemes have better ARL_1 performance than the Shewhart chart for process mean shifts up to 2.6σ and they can be easily implemented.

In another study, Khoo [12] noted that obtaining the control limits for the two-of-three scheme proposed by Klein would be difficult for quality control engineers and thus proposed a more user-friendly approach. Khoo expounded on Klein's approach by using a simulated study to evaluate the performance of various schemes. In this setting, the analyst chooses a pre-specified ARL value for a zero-mean shift, and then obtain the control limits from the simulated values using less tedious steps. Khoo studied the 2-of-2, 2-of-3, 2-of-4, 3-of-3, and 3-of-4, amongst which he concluded that the 3-of-4 had the best ARL performance for small to moderate process mean shifts.

Antzoulakos and Rakitzis [1] further improved the schemes discussed above and proposed the modified r-out-of-m (M: r/m) scheme. The ARL performance of the modified r/m scheme outperforms the standard Shewhart chart and both methods presented by Klein [13] and Khoo [12] for process mean shifts up to 2.6σ. However, the standard Shewhart chart performs better for process mean shifts above 2.6σ. In this setting, for $r < m$, the control chart gives an OC signal if r points plot above (below) the upper (lower) control limit which are separated by at most $(m-r)$ points placed between the center line and the upper (lower) control limit. They suggested that the M: $r/5$ scheme were more reliable for detecting small to moderate process mean shifts.

Despite the good performance of these schemes in detecting small mean shifts, they do not perform well in detecting persistent mean shifts because of their inability to use sufficient history data during the monitoring procedure. In the case when the sequential process is IC, and additional less severe and persistent assignable causes, shift the process away from the IC mean in an intermittent but yet consistent manner, the modified schemes are not capable of detecting such irregular shifts. For instance, suppose we begin the Phase-II SPC monitoring with the M: 4/5 scheme, this scheme is primed to detect small shifts of about 0.2σ. So in the sequel considerations presented, we assume that the average shift size is not above 0.2σ. At time point t, the scheme looks back through $(t-5)+1$ observations to check for observations that are OC within this window. However, if there are less than 4 OC observations (either in the upper or lower region) within any window, the scheme will fail to give an OC signal.

Consider Figure 2.1(a), at time point $t = 5$, notice that the chart will not give an OC signal despite the fact that the observations at $t = 2, 3, 4$ are all OC. In this case, the process keeps on running and the M: 4/5 scheme fails to give any OC signal despite significant sequence of shifts around time points 2 to 4, 8 to 9, and 14 to 17.

Still consider the moving window of size 5, with intermittent shifts which occur in a persistent fashion within the process. Consider the sequential process which is visualized in Figure 2.1(b), it is easy to see that there is a consistent pattern at which the process goes OC after every two time points. Again, the M: 4/5 scheme will not give an OC signal if the process continues running in this fashion.

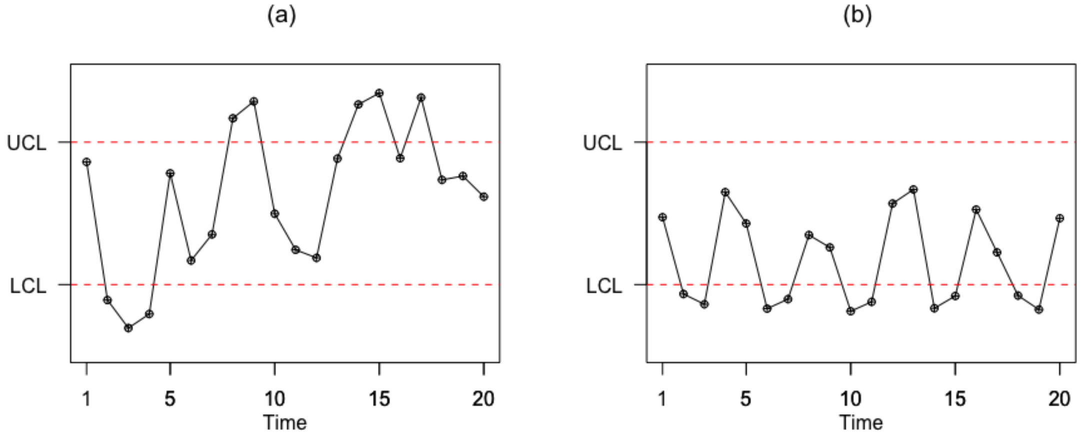

Figure 2.1: Two Shewhart control charts (a) and (b) which illustrate the inability of the M: 4/5 scheme to give signals for persistent shifts.

The illustration presented in Figure 2.1 can be generalized to other M: r/m schemes. It may be argued that a simultaneous combination of several M: r/m schemes can be used to detect such persistent shifts, even though this is plausible, this approach would be hindered by the problem of an increase of false-alarms. Furthermore, the M: r/m scheme will require prior knowledge of the shift size before any specific scheme can be employed, this may also hinder its usage since the magnitude of the shift size to be encountered in the process is usually unknown in most cases.

In this Chapter, we propose an adaptive r-out-of-m control chart, in which the values of r and m are chosen adaptively. We show that the chart will detect small to moderate persistent mean shifts and efficiently estimate the shift positions in the sequential process.

We acknowledge that our approach will be more suitable for individual observations rather than group data. This is because, since we intend to detect mean shifts for process observations where the quality characteristic of the product changes slowly over time, samples taken consecutively or very close in time would be virtually identical, apart from measurement or analytical error.

2.2 Description of the adaptive r-out-of-m control chart

Suppose

$$X_i \sim \begin{cases} N(\mu_0, \sigma^2) & \text{if } i \leq \tau \\ N(\mu_1, \sigma^2) & \text{if } i > \tau \end{cases}$$

where $\mu_0 \neq \mu_1$, τ is an unknown shift position and X_i's are independent observations. In order to detect small to moderate persistent shifts in the mean of a production process, we use an adaptive sampling procedure. In this sense, we adaptively select r-out-of-m process observations that plot beyond certain control limits. Here, the maximum value of m is set in advance and then, we use an adaptive procedure to obtain the values of r $\{r \leq m\}$ at each time point. The Shewhart control chart consists of three regions − the region above the upper control limit, the region below the lower control limit and the region between the two control limits. In this case, we separately consider points that plot in the region above the upper control limit and the points that plot in the region below the lower control limit. We consider points that plot within the control limits to be IC. At each time point t, when m $\{m \leq t\}$ observations have been obtained, let us call the number of points that plot in the region above the upper control limit r_1, and the number of points that plot in the region below the lower control limit r_2.

Given a preset maximum value of m, we begin monitoring the process in a retrospective fashion. Thus, at each time point t, for each unit increase from 1 to m, we obtain the values of $r_1(r_2)$ that plot beyond the upper(lower) control limit. Also, for each value of $r_1(r_2)$ obtained at time point t, we compute the probability of observing a more extreme value of $r_1(r_2)$ given that the process is IC. By statistical convention, this probability is the p-value.

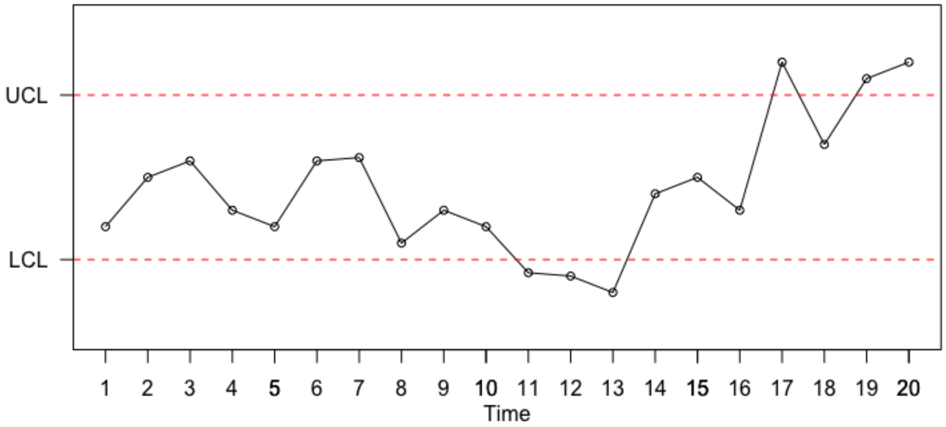

Figure 2.2: A Shewhart control chart of individual data points collected at equally spaced sampling intervals from time point $t = 1, ..., 20$. This chart is used to illustrate the mechanism of the adaptive r-out-of-m scheme at $t = 20$.

Let us define the random variables X_1 and X_2 to be the number of points that plot beyond the upper and lower control limits respectively. Following the assumption that the process observations are independent, it is easy to see that we can model the adaptive r-out-of-m scheme by a binomial probability distribution with parameters m and probability of success $\tilde{\alpha}$, where $\tilde{\alpha}$ is the probability of observing a point beyond the control limit given that the process is IC.

The adaptive r-out-of-m control chart will give an OC signal at time point t if the minimum of the set of p-values for different m obtained for $r_1(r_2)$ is less than a threshold value γ. This threshold value γ is obtained in such a way where a pre-fixed ARL$_0$ value is achieved. For instance, consider the process which is visualized in Figure 2.2. In order to illustrate the mechanism of the adaptive r-out-of-m scheme, we present the details of the scheme when it gets to time point $t = 20$. Here, we assume that $\tilde{\alpha} = 0.1$, and we set the maximum value of m to be 10. The simulated series is displayed in Table 2.1. At time point $t = 20$, we begin checking for points that plot beyond either control limits in a retrospective manner. For increasing values of m, the minimum p-values associated with the respective random variables X_1 and X_2, indicate an extreme case.

From Table 2.1, when $m = 4$, $r_1 = 3$. The p-value at such instance is computed as

$$P_1 = P(X_1 \geq r_1) = 1 - P(X_1 \leq (r_1 - 1)) = 1 - \sum_{k=0}^{r_1 - 1} \binom{m}{k} \tilde{\alpha}^k (1 - \tilde{\alpha})^{(m-k)}$$

$$\begin{aligned} P(X_1 \geq 3) = 1 - &\left\{ \binom{4}{0} (0.1)^0 (1 - 0.1)^{(4-0)} + \binom{4}{1} (0.1)^1 (1 - 0.1)^{(4-1)} \right. \\ &\left. + \binom{4}{2} (0.1)^2 (1 - 0.1)^{(4-2)} \right\} \\ &= 1 - \{0.6561 + 0.2916 + 0.0486\} = 0.004 \end{aligned}$$

From Table 2.1, it is easy to see that 3-out-of-4 and 3-out-of-10 OC observations correspond to the minimum p-values in the upper and lower regions respectively. Thus, the control chart will give an OC signal at this time point if either of the minimum p -value is less that the pre-fixed threshold value γ.

This process is repeated at each time point and the scheme gives an OC signal when the minimum p-value for either the upper or lower region is less than the pre-fixed threshold value γ. In this illustration, the maximum value of m was chosen to be 10, however, depending on the nature of the process and the extent to which persistent shifts are to be determined, m can be chosen to be smaller or larger. When the analyst aims to detect long-staying persistent shifts, m should be large because the scheme will need sufficient history to detect such shifts. Otherwise, setting $m = 10$ should be sufficient for detecting persistent shifts. Either ways, the value of m will influence the run time of the process.

Table 2.1: r_1, r_2, and m alongside their corresponding p-values for the process depicted in Figure 2.2. This illustrates the mechanism of the adaptive r-out-of-m chart at time, $t = 20$.

m	r_1	$\mathrm{P}(X_1 \geq r_1)$	r_2	$\mathrm{P}(X_2 \geq r_2)$
1	1	0.100	0	1.000
2	2	0.010	0	1.000
3	2	0.028	0	1.000
4	3	0.004	0	1.000
5	3	0.009	0	1.000
6	3	0.016	0	1.000
7	3	0.026	0	1.000
8	3	0.038	1	0.570
9	3	0.053	2	0.225
10	3	0.070	3	0.070

In the subsequent sections, we provide pseudo codes when using the adaptive r-out-of-m control chart, particularly for detecting OC signals, estimating the ARL_0 values and obtaining the threshold value γ for some certain ARL_0 values.

2.2.1 Pseudo Code for detecting an OC signal

Let $\tilde{\alpha}$ be the pre-specified probability of observing a point beyond the control limit given that the process is IC. Also, let γ be the pre-specified threshold value which is chosen to achieve a given ARL_0 value. Let n be the number of observations in the sample and m the number of most recent observations we wish to consider. In the i-th iteration, for $m \leq i \leq n$,

Step (1)

- In the j-th iteration, for $1 \leq j \leq m$, obtain the number of points, r_{1j}, that plot above the upper control limit, and the number of points, r_{2j}, that plot below the lower control limit.
- Compute the p-value for both r_{1j} and r_{2j}, which is given

$$\begin{aligned} P_{1j} &= P(X_1 \geq r_{1j}) = 1 - P(X_1 \leq (r_{1j} - 1)) \\ &= 1 - \sum_{l=0}^{r_{1j}-1} \binom{j}{l} \tilde{\alpha}^l (1 - \tilde{\alpha})^{(j-l)} \\ P_{2j} &= P(X_2 \geq r_{2j}) = 1 - P(X_2 \leq (r_{2j} - 1)) \\ &= 1 - \sum_{l=0}^{r_{2j}-1} \binom{j}{l} \tilde{\alpha}^l (1 - \tilde{\alpha})^{(j-l)} \end{aligned}$$

Step (2) If $\min(P_{1j},\ j = 1, 2, ..., m) < \gamma$ or $\min(P_{2j},\ j = 1, 2, ..., m) < \gamma$, print out the values of i together with the corresponding values of r and m, and **stop** the algorithm. Otherwise, $i = i + 1$, and return to step (1).

For the adaptive scheme, the control limits will be determined by the value of $\tilde{\alpha}$. For instance, $\tilde{\alpha} = 0.0027$ yields 3-sigma control limits. In the case of the standard Shewhart chart, the value of $\tilde{\alpha}$ determines the ARL$_0$ value, where ARL$_0 = \frac{1}{\tilde{\alpha}}$. Here, the run length of the process follows a geometric distribution with parameter, $\tilde{\alpha}$. However, since the proposed adaptive scheme follows some other criteria for giving an OC signal, the ARL$_0$ is computed differently. In the literature, the Markov chain approach has been extensively used to obtain the ARL$_0$ values for several runs rules. However, considering the fact the scheme proposed in this study follows an adaptive nature and the value of m may vary (it could be large), we resort to simulation for estimating the ARL$_0$ value. Certainly, computational complexities will arise if we compute the ARL$_0$ value using the conventional Markov chain approach. This is because there will be too many transient states in the Markov chain, thus the transient space may be totally large and out of computation ability.

In this study, Monte Carlo simulations are used to estimate the ARL$_0$ value. We notice that this procedure is more efficient and allows for numerical experimentation to understand several properties of the ARL. The algorithms given in Sections 2.2.2 and 2.2.3 closely follow the methods described by ([21], page 127 and 129). Section 2.2.2 provides a stepwise process to compute the ARL$_0$ value.

2.2.2 Pseudo Code to Compute an Estimate of the ARL$_0$ Value

Let R be the number of replicated simulations. In order to obtain stable values, this number should be a large positive integer. Specify the values of $\tilde{\alpha}$ and γ. In the g-th replicated simulation for $1 \leq g \leq R$,

Step (1) Generate n observations from $N(0,\ 1)$

Step (2) Compute the run length $RL(g)$ by the following loop; for $m \leq i \leq n$

- Compute the necessary values from Section 2.2.1.
- From step (ii) in Section 2.2.1, if $\min(P_{1j},\ j = 1, 2, ..., m) < \gamma$ or $\min(P_{2j},\ j = 1, 2, ..., m) < \gamma$, which indicates an OC signal, set $RL(g) = i$ and break out of the loop; otherwise, let $i = i + 1$ and continue the loop.

Step (3) Proceed to $g = g + 1$, and return to step (1) until R is reached.

Step (4) The ARL$_0$ is the average of R run length values. i.e ARL$_0 = \frac{\sum_{g=1}^{R} RL(g)}{R}$.

In subsequent sections, we provide some interesting properties of the ARL$_0$. Nonetheless, we see that the ARL$_0$ value depends on the threshold value γ. Thus, it becomes imperative to obtain the threshold value which will yield a certain ARL$_0$ value. We utilize the bisection method to search for the threshold value which reaches the expected ARL$_0$ to a certain accuracy. The algorithm presented in Section 2.2.3 below describes the step-wise procedure for the search.

2.2.3 Pseudo Code to Search for the Threshold Value

Let A_0 be the pre-specified ARL_0 value and let $[\gamma_L,\ \gamma_U]$ be the interval from which the threshold value, γ is searched. Let $\rho > 0$ be a small number denoting the estimation accuracy of the search. Set R to be the number of replications used in obtaining the run length of the process. Set M to be the number of required iterations for the search, and then for $1 \leq j \leq M$ perform the following steps iteratively.

Step (1) Compute $\gamma = (\gamma_L + \gamma_U)/2$. Using γ

- For $1 \leq g \leq R$, compute the run length, $RL(g)$.
- Set $\text{ARL}_0 = \text{mean}(RL)$

Step (2) If the ARL_0 value obtained from step (1) lies in the interval $[A_0 - \rho,\ A_0 + \rho]$, **stop** the algorithm. Thus, the value of γ obtained from step (1) is the searched value. Otherwise, set

$$\begin{cases} \gamma_L = \gamma_L;\ \gamma_U = (\gamma_L + \gamma_U)/2 & \text{for } \text{ARL}_0 > A_0 \\ \gamma_L = (\gamma_L + \gamma_U)/2;\ \gamma_U = \gamma_U & \text{for } \text{ARL}_0 < A_0 \end{cases}$$

continue to $j + 1$, and return to step (1).

If the algorithm does not stop before or at M-th iteration, then the value of the ARL_0 obtained still lies outside the interval $[A_0 - \rho,\ A_0 + \rho]$. Thus, the estimation accuracy specified by ρ cannot be reached.

In order to choose optimal starting values (γ_L and γ_U) for the search, we make sure that the pre-specified value, A_0, lies well in the interval of ARL_0 values obtained when $\gamma = \gamma_L$ and $\gamma = \gamma_U$, respectively. Otherwise, the computation would be expensive.

The magnitude of the estimation accuracy ρ should be small, a number in the interval [0, 1] is usually chosen.

2.3 Performance of the r-out-of-m scheme

Large values of $\tilde{\alpha}$ and γ will detect small and transient shifts in the process. In this setting, the control limits will be constricted and the scheme frequently yields small combinations of r-out-of-m observations that plot beyond the control limits such as 1-out-of-2, 2-out-of-2, 2-out-of-4, and 2-out-of-5. Since the resulting p-values will be small and may be often less than the threshold value, the ARL performance of the process will be poor, and thus there will be substantial false alarms. However, the adaptive scheme is advantageous in the sense that we can reduce the threshold value in order to detect persistent shifts and also reach some larger ARL values. In a similar fashion, small values of $\tilde{\alpha}$ are primed to detect large and transient shifts. Also, the value of the threshold can be set to achieve certain ARL values and detect long-staying shifts in the process.

Furthermore, from numerical experimentation shown in Table 2.2, we observe that the maximum value of m chosen for the r-out-of-m scheme does not have a substantial impact on the ARL performance. Given the values of $\tilde{\alpha}$ and γ, we see that the variation in the ARL_0 values for increasing values of m is very minimal.

Table 2.2: ARL_0 values obtained for several maximum values of m used in the adaptive r-out-of-m scheme. In this case $\tilde{\alpha} = 0.1$ and $\gamma = 0.01$.

m	5	7	10	12	15	20	25	30
ARL_0	800.9	798.6	738.2	741.0	739.6	742.0	732.0	743.3

To detect transient shifts in the process, it will be ideal to set m to be at most

5 because such shifts will only require information around the current time point. While for persistent shifts, m should be set to at least 10, because such shifts will require sufficient history data.

Next, we investigate the ARL performance of a process whose distribution is N(0, 1). In this setting, we select $\tilde{\alpha} = 0$ and γ is chosen from the interval [0, 0.1]. The obtained ARL_0 values are displayed in Figure 2.3. From this process, we observe that as γ increases, the ARL_0 value decreases. Furthermore, notice the jumps in the ARL_0 values displayed in Figure 2.3, the gaps grow farther apart when the γ is small. In this setting, we discovered that certain ARL_0 values will not be achieved when the adaptive scheme is employed. This is one limitation of the adaptive r-out-of-m scheme, in subsequent sections, we provide a detailed discussion of this limitation and a potential approach to overcome it.

In addition, we present the threshold values obtained for some commonly used ARL_0 values in Table 2.3. For this illustration, the maximum value of m is set to be 10, and we assume that the process is N(0, 1). Notice that for some slight change in γ, especially when this value is small, there seems to be significant jumps in the resulting ARL_0 values.

2.4 Limitations of the Adaptive r-out-of-m Scheme

From the description of the adaptive r-out-of-m chart provided earlier, we indicated that at each time point t, the minimum of a set of p-values is compared to a pre-fixed threshold value γ. That is, if $\min(P_{t1j},\ j = 1, .., m)$ or $\min(P_{t2j},\ j = 1, .., m) < \gamma$, then the control chart gives a signal. Once r and the maximum m are fixed, possible p-values at each time point are discrete. For instance, suppose at time

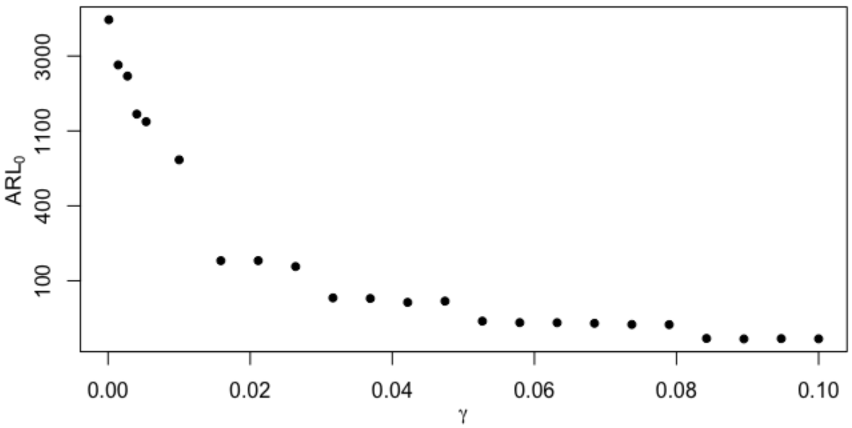

Figure 2.3: ARL$_0$ values obtained for a process whose IC distribution is N(0, 1), the threshold value γ ranges from 0 to 0.1 and $\tilde{\alpha} = 0.1$

point t, we obtained 3-out-of-5, and 3-out-of-6 when $m = 5$ and 6 respectively. There will be a jump in the resulting p-values for both cases because of the discrete nature of r and m. So, when the maximum value of m is fixed, we can only attain certain p-values for different combinations of r and m. Since we are checking if the minimum of these p-values is less than γ, there will be substantial impact of the discreteness of the charting statistic on the performance of the chart, i.e. the ARL$_0$. This limitation is presented in the graph displayed in Figure 2.3. This explains why we may not be able to obtain specific ARL$_0$.

Since specific ARL$_0$ values may be desired by practitioners using SPC charts, this limitation may pose a challenge to its usability and acceptance. If the threshold value required to reach some specific ARL$_0$ values cannot be computed, such monitoring process cannot be evaluated effectively. One possible way to avoid this issue is to use some kind of randomized comparison between the minimum p-values and γ. In this sense, we use the threshold $\gamma \pm \epsilon$, where ϵ follows some statistical distribution with zero-mean. Thus, rather than comparing the minimum p-value with a fixed number

Table 2.3: Computed threshold values, γ of the adaptive r-out-of-m scheme described in Section 2.2 for some commonly used ARL_0 and $\tilde{\alpha}$ values. * denotes that γ could not be obtained for such combination of ARL_0 and $\tilde{\alpha}$

	$\tilde{\alpha}$			
ARL_0	0.1	0.05	0.01	0.0027
50	0.1	*	*	*
100	0.0486	*	0.8999	*
200	0.0141	0.03277	*	*
370	*	0.01054	*	0.1125
500	*	0.00724	*	*
750	0.01	*	*	*
1000	0.00856	0.000110	0.00307	0.0027

γ, we compare it with a random number $\gamma \pm \epsilon$. Indeed, more research is needed in this area to address the issue of the discreteness in the model which prevents certain ARL_0 values from being achieved.

CHAPTER 3

DYNAMIC SAMPLING SCHEMES

3.1 Introduction

In Section 1.2, we introduced some commonly used SPC charts. The traditional versions of these charts are designed to monitor each and every observation during the monitoring procedure.

Suppose we aim to know whether our process is likely to yield a distributional shift in the near future. Then, a numerical measure that takes the state of the process at the current and previous time points into account would be ideal for determining the possibility of such shift. Even though the charting statistics of traditional charts provide information about the performance of the process at the current time point and at previous time points, they do not provide necessary information about the performance of the process in the near future. Furthermore, visual representations of the charting statistics may not be an ideal indicator of a possible shift in the distribution of the process, because in many cases the charting statistics rests on the assumption that the process observations are independent. Li et al. [16] proposed using p-values to design SPC charts. In this regard, at each time point during the Phase-II SPC, the p-value of the observed charting statistic is computed under the assumption that the process is IC. The control chart gives an OC signal if the computed p-value is less than a pre-specified level of significance, α. The p-values provide information about the potential of a distributional shift in the process. In this sense, the information obtained from the p-value at each time point can be used to adjust the sampling scheme of the monitoring process. That is, the sampling time and the sampling size of the next sample will be dependent on the magnitude of the p-value of the charting statistic at the current time point. This approach enables the practitioner to make informed decisions when handling future observations. For instance, if the p-value is much larger than α, this provides sufficient evidence that the process is likely to be stable at such time point.

In this context, since subsequent sampling decisions will be dependent on this numerical measure, variable sampling rates (VSR) rather than conventional fixed sampling rates (FSR) $-$ which depends on fixed time intervals or sample sizes $-$ will be incorporated in the design of traditional SPC charts. The VSR is somewhat analogous to the adaptive SPC control chart, a control chart in which either the sampling interval or the sampling size (or both) can change depending on the value of the charting statistics [18]. As discussed in the literature, one notable advantage of the VSR over the FSR is that, given the IC ARL_0 and the IC average sampling rate, the VSR has good performance in detecting small to moderate shifts.

The VSR scheme depends on several features which are changed according to state of the process at the current time point. A typical VSR scheme would depend on either the variable sampling interval (VSI), the variable sampling size (VSS) or both variable sampling interval and variable sampling size (VSSI).

During the Phase-II SPC, if a sample point plots beyond the warning limits, the control chart with the VSI is designed to wait lesser than usual before observing the next batch of observations, while for the VSS case, the next batch is set to be larger than usual. The control chart with VSSI combines both methods when a sample point plots beyond the control limits. On the contrary, if the sample point plots within the central region, the control chart with the VSI delays the next batch, and for the VSS chart, fewer observations are taken in the next batch. In this case, the VSSI combines both methods again [7].

While taking into account the potential shift size in a process, several researchers have suggested a variety of methods to adaptively control subsequent sampling times during the Phase-II SPC. The VSI schemes were introduced by Reynolds et al. [23], and were also implemented in the $\bar{X}$ chart. Reynolds et al. [22] also proposed VSI schemes for CUSUM charts. In this case, the sampling interval, $d(\cdot)$, is defined to equal either one of two values (d_1 or d_2) based on the membership of its charting statistic in a specific region defined by its control limits. More recently Li and Qiu [15] proposed the dynamic sampling interval scheme which is defined as a continuous function of the p-value of a charting statistic. These sampling interval schemes form the framework of some VSI control charts used for the detection of potential but unknown mean shifts in the distribution of a production process. The VSI schemes allow the sampling time to be changed according to the current state of the process readings.

Luo et al. [17] implemented the VSI scheme proposed by Reynolds et al. [22] in their design of a VSI adaptive CUSUM (VSI-ACUSUM) chart. Li and Qiu [15] implemented the dynamic sampling scheme in the design of the dynamic sampling CUSUM (DyS-CUSUM) chart. These charts have shown to have good performance, that is, they detect unknown shifts quicker than traditional charts. For comparison, while the VSI-ACUSUM chart uses the conventional VSI scheme which takes two possible values, the DyS-CUSUM chart employs the dynamic sampling scheme. Also, the VSI-ACUSUM chart uses the conventional control limits in its design while the

DyS-CUSUM chart uses the p-value of the CUSUM chart in its design. However, both VSI-ACUSUM and DyS-CUSUM charts use the adaptive selection of the reference value of the CUSUM chart which was developed by [26].

Numerical studies shown in [15] shows that in general, the DyS-CUSUM chart has the advantage of quickly detecting certain shift sizes when compared to the VSI-ACUSUM chart. This means that the dynamic sampling scheme has better performance than the conventional 2-interval sampling scheme. Thus, our study focuses on the incorporation of the dynamic sampling schemes in other conventional SPC charts. We emphasize this method because of its computational efficiency and optimal performance when handling different shift sizes.

Among popular SPC charts with a fixed ARL_0 value, the CUSUM chart has optimal performance $-$ the lowest ARL_1 $-$ for detection of distributional shifts in a production process that is normally distributed if the reference value k of its charting statistic is chosen properly for a particular shift size [19]. Nevertheless, if the production process follows some other distribution that is not normal, the CUSUM chart does not perform well in detecting distributional shifts. Specifically, the CUSUM chart is sensitive to the assumption that both the IC and OC distributions of the sequential process are normally distributed [6]. In most real-world applications, the distribution of the production process is usually unknown, hence it becomes imperative to derive dynamic sampling schemes for other control charts which are robust to the normality assumption.

In the next section, we provide a brief overview of the dynamic sampling scheme for the CUSUM chart proposed by Li and Qiu [15].

3.2 The Dynamic Sampling Scheme for the CUSUM chart

We begin the discussion with the design of the CUSUM control chart using p-values. For this design, let us assume that the IC process distribution is known. Li et al. [16] provide a rigorous discussion of this design which is presented below. The p-value of the charting statistic of the CUSUM chart is described as follows.

Suppose we have a sequence of independent X_i observations from a production process, where

$$\begin{cases} X_1, X_2, ..., X_\tau \sim N(\mu_0, \sigma^2), & \text{if the process is IC} \\ X_{\tau+1}, X_{\tau+2}, ... \sim N(\mu_1, \sigma^2), & \text{if the process is OC} \end{cases}$$

where τ is an unknown change point in the mean of the process, and $\mu_0 \neq \mu_1$, and $\sigma_0^2 = \sigma_1^2 = \sigma^2$. Then as shown in (1.4), the charting statistic of the conventional CUSUM chart for detecting an upward mean shift is defined by

$$\begin{cases} C_0^+ = 0 \\ C_n^+ = \max(0,\ C_{n-1}^+ + (X_n - \mu_0) - k) \end{cases} \tag{3.1}$$

If reference parameter k is chosen as $(\mu_1 - \mu_0)/2 = \delta/2$, then the chart is optimal for detecting the particular shift μ_1. The chart gives a signal of an upward mean shift when

$$C_n^+ > h \tag{3.2}$$

where $h > 0$ is a control limit chosen to achieve a given ARL_0 value. The p-value of the charting statistic of the CUSUM chart is described as follows. Let C_n^{+*} be the observed value of the charting statistic C_n^+, then the p-value at the n-th time point is defined by

$$P_{C_n^{+*}} = P(C_n^+ > C_n^{+*}) \tag{3.3}$$

We would conclude that the process has gone out-of-control at the n-th time point if

$$P_{C_n^{+*}} < \alpha \tag{3.4}$$

Otherwise, we say that the process is still IC. The analogy follows much from the classical statistical test of hypothesis, where the null hypothesis that the process is IC is rejected if the p-value is less than a significant level α. As stated in the Section 3.1, this approach has some pivotal benefits. Specifically, the p-value informs the practitioner of the likelihood of a potential shift in the distribution of the process. This information would be a beneficial tool for adjusting the next sample and also for taking subsequent actions. In the case when the p-value is much larger than the significant level α, this signifies that the process is still very much in control and thus we would want to delay the time of observing the next sample or collect less observations at the next regular time. In contrast, when the p-value is much lesser than α, this indicates that the process must have gone out-of-control, thus, the process should be stopped immediately. In the case where the p-value is only marginally less than or marginally greater than α, we would want to observe the next sample sooner than usual. Therefore, we notice that subsequent actions is dependent on the magnitude of the p-value. With this in mind, how does the practitioner decide how long to delay the process when the p-value is greater than α or how soon to observe the next sample when the p-value is only marginally greater than or less than

α? The waiting time to observe the next sample is dependent on the sampling interval function $d(\cdot)$. Logically, the $d(\cdot)$ should be an increasing function of the p-value, that is, $d(\cdot)$ increases as $P_{C_n^{+*}}$ increases. We proceed by describing the sampling interval function proposed by Li and Qiu [15]. Their sampling interval function is chosen from the Box-Cox transformation family and is defined as

$$d(P_{C_n^{+*}}) = \begin{cases} a + bP_{C_n^{+*}}^{\lambda} & \text{if } \lambda > 0 \\ a + b\log(P_{C_n^{+*}}) & \text{if } \lambda = 0, \end{cases} \tag{3.5}$$

Next, we review the methods used to estimate the parameters of the model above. But before then, we know that the ARL is commonly used to evaluate the performance of the traditional SPC charts which have fixed sampling rates. Thus, when these charts are employed, the (FSR) sampling interval is usually constant. For the variable sampling rate (VSR) control chart, the ARL would not be an idealistic measure of performance of the chart since sampling interval in this setting varies over time. In the literature, two widely used performance measures are usually employed. These are the average time to signal (ATS) and the adjusted average time to signal (AATS). The ATS is defined as the expected value of the time interval from the start of the Phase-II process monitoring to the time when a chart gives an OC signal. While the AATS is defined as the expected value of the time interval from the occurrence of a shift to the time when the chart gives an OC signal. As in the case of the FSR schemes, the chart with a larger IC ATS will have lower false alarm rate, and the chart with the smallest OC AATS will perform best for detecting a specific shift size.

3.2.1 Estimation of Parameters

In order to estimate the parameters a, b, and λ in the sampling interval $d(\cdot)$, Li and Qiu [15] carried out several simulation studies to obtain optimal values for these parameters. The authors primarily used the AATS_1 as a measure of the performance of the control chart for detecting several shift sizes. In this section, we provide a brief overview of the results provided by the authors.

For their numeric experimentation of the parameter a, the authors chose a to be in interval $[0,\ 1]$, negative values of a, could result to negative values for the sampling interval. Also, when $\lambda > 1$, $d(P_{C_n^{+*}})$ would be consistently larger than 1 if $a > 1$. This would not be ideal when a potential shift has been observed. When $\lambda = 0$, a was chosen to be 1, and when $\lambda > 0$, a was chosen to be 0. Using the information from the selection of a, it was shown that the performance of the scheme is almost identical when $\lambda \geq 2$. Thus, the authors chose $\lambda = 2$. With these chosen parameters, the sampling interval, $d(P_{C_n^{+*}})$, now becomes

$$d(P_{C_n^{+*}}) = b \cdot P^2_{C_n^{+*}} \tag{3.6}$$

The parameter, b, which can be determined to satisfy the requirement that ATS_0 = ARL_0, is selected as an integer multiple of the smallest time unit in a specific application, and thus, the sampling interval needs to be rounded when necessary.

Furthermore, the reference value k of the CUSUM chart is selected adaptively using the method proosed by Sparks [26]. At each time point, k is chosen according to the estimated shift size. A brief description of the scheme is given here. The estimator of a potential mean shift at the current time point is given as

$$\hat{\delta}_n = \max\left\{\delta_{\min}, (1-r)\hat{\delta}_{n-1} + r(X_n - \mu_0)\right\} \tag{3.7}$$

where $\delta_{\min} > 0$ is the minimum shift size of interest, $\hat{\delta}_0 = \delta_{\min}$ and $0 < r < 1$ is a weighting parameter. Define $k_n = \hat{\delta}_n/2$, and the resulting charting statistic becomes

$$\begin{cases} C_0^+ = 0, \\ C_n^+ = \max(0, C_{n-1}^+ + (X_n - \mu_0 - k_n)/h_n), \end{cases} \tag{3.8}$$

where $h_n > 0$ is a control limit. In order to approximately reach a pre-specified ARL_0 value, Shu and Jiang [25] provided the following formula to compute the control limit

$$h_n = \frac{\log(1 + 2k_n^2 \cdot ARL_0 + 2.332k_n)}{2k_n} - 1.166 \tag{3.9}$$

These authors provide some practical guidelines for choosing the parameters δ_{min} and r, and also showed that the CUSUM chart with adaptive selection scheme shown above performs well in various cases.

3.2.2 Calculating the p-values

In order to compute the p-value, $P_{C_n^{+*}}$, it is imperative to specify the distribution of the CUSUM charting statistic, C_n^+. Here, two common cases are usually considered – when the IC process distribution is either known or unknown. For the case when the IC process distribution is known, Monte Carlo simulations have been used in the literature to estimate the IC distribution of C_n^+. In this setting, random observations are generated from the known distribution, then these observations are used to estimate the IC distribution parameters of C_n^+. These parameters are then used to compute the p-value of the charting statistic as if the IC distribution is known.

When the IC process distribution is unknown, availability of an IC dataset would be handy in estimating the distribution of C_n^+. In this setting, the bootstrap approach is another alternative for estimating the IC distribution of C_n^+. Thus, resampled data are repeatedly drawn from the available IC process data, and these resampled data are then used to compute C_n^+. This process is repeated as much as B times, after which the B number of C_n^+ are used to compute the p-values of the observed charting statistic, $P_{C_n^{+*}}$, in the Phase-II SPC.

Using Monte Carlo simulations, Li and Qiu [15] further showed that their control chart with the adaptively selected reference value k_n has the steady-state property for $n \geq 50$. By steady-state, we mean that as the shift time τ increases, the value of AATS_1 remains quite stable. In general, the control chart given in (3.8) with the sampling interval (3.6) is called the dynamic sampling CUSUM chart (DyS-CUSUM).

As stated earlier, we adopt the dynamic sampling approach for monitoring big data processes because it shows to have the best performance in the class of VSI schemes. From this point, we begin the design of other traditional SPC charts using dynamic sampling schemes.

3.3 The Dynamic Sampling Scheme for the Shewhart Control Chart

In Section 1.2.1, we introduced the Shewhart control chart. Before we begin the discussion regarding the integration of the dynamic sampling scheme in the Shewhart chart, we further discuss its application in the Phase-I and Phase-II SPC.

3.3.1 The Phase-I SPC

In this Phase, we present an overview of the design of the control limits. Suppose we are have an independent sequence of X_i $\{i = 1, ..., n\}$ observations with unknown change point τ. Let μ_0 denote the IC mean and σ denote the IC standard deviation of the process. In this case, we assume that a shift is observed only in the mean of the process, whereas the variance remains stable. Typically, the Shewhart control chart is commonly used when batch data are observed. Nevertheless, under mild adjustments of the charting statistic, the chart can be used to monitor individual observations. In order to employ the control limits of the traditional Shewhart chart to monitor individual observations, we bin the observations into groups using a moving window technique with window size w and a total of $n - w - 1$ groups. Thus, we have

$$\begin{aligned} \text{Group 1: } & X_1, ..., X_w \\ \text{Group 2: } & X_2, ..., X_{w+1} \\ & \vdots \\ \text{Group } n-w-1\text{: } & X_{n-w+1}, ..., X_n \end{aligned}$$

Several researchers [21] advise against grouping the observed data in such a way where the first group consists of the first $\tilde{w}$ observations, the second group consists of the next $\tilde{w}$ observations, and so on; where $\tilde{w} > 1$ is the group size. In this context, it will be difficult for the practitioner to pinpoint the exact time at which the process went OC. Another limitation of this approach is that, the exact ARL_0 value which is used to evaluate the performance of the control chart becomes speculative since the exact OC timepoint is difficult to obtain.

Evaluating the performance of the process at each time point can be constructed as a test of hypothesis problem. That is, we test the following hypothesis

$$H_0 : \ \mu = \mu_0; \quad H_1 : \ \mu \neq \mu_0$$

where μ denotes the true process mean. Thus, an appropriate test statistics for this hypothesis is given as

$$Z = \frac{X_i - \mu_0}{\sigma} \sim N(0, \ 1) \tag{3.10}$$

Given the observed value of the test statistic $|Z^*|$, the null hypothesis is rejected at a pre-specified level of significance α if

$$|Z^*| > Z_{1-\alpha/2}$$

where $Z_{\alpha/2}$ is the $\alpha/2$ critical value of the standard normal distribution. Thus, in this setting, given the observed data at time point i, the process is said to OC if

$$X_i < \mu_0 - Z_{1-\alpha/2}\ \sigma \ \text{ or } \ X_i > \mu_0 + Z_{1-\alpha/2}\ \sigma \tag{3.11}$$

In practice, the IC mean μ_0 and standard deviation σ are usually unknown. Given the observed values from the process, we estimate the IC mean as

$$\hat{\mu}_0 = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{3.12}$$

In order to estimate σ, we know from statistical theory that the sample standard deviation s is a biased estimate of the population standard deviation σ. According to Kenney and Keeping [11], this bias depends on w, and thus we have

$$E\left(\frac{1}{d_3(w)}s_i\right) = \sigma \tag{3.13}$$

where s_i is the sample standard deviation of each moving window with size w, and $d_3(w)$ is a constant that corrects for the bias, and its expression is given as

$$d_3(w) = \begin{cases} \frac{2(v-1)(2^{v-2}(v-2)!)^2}{(2v-3)!}\sqrt{\frac{2}{\pi(2v-1)}} & \text{if } w = 2v \\ \frac{(2v-1)!}{2(2^{v-1}(v-1)!)^2}\sqrt{\frac{\pi}{v}} & \text{if } w = 2v+1 \end{cases}$$

notice that $3 \leq w < 170$, otherwise, $d_3(w)$ does not exist. Therefore, we have that the estimate of σ is

$$\hat{\sigma} = \frac{\bar{s}}{d_3(w)}$$

where $\bar{s} = \frac{1}{n-w+1}\sum_{i=1}^{n-w+1} s_i$. Some researchers have also used the range of each batch to estimate σ, however we prefer the sample standard deviation, because for large batch sizes, the range loses statistical efficiency when it is used to estimate σ [18]. Since we have adequate computational resources and w will be mostly large, it is natural to use to the sample standard deviation. Thus, given the estimated parameters and by rewriting (3.11) the Shewhart control chart for individual observations is given as

$$U = \bar{X} + Z_{1-\alpha/2}\ \hat{\sigma}; \quad C = \bar{X}; \quad L = \bar{X} - Z_{1-\alpha/2}\ \hat{\sigma}$$

and the control chart gives an OC signal if

$$X_i < L \quad \text{or} \quad X_i > U \tag{3.14}$$

With this modification, the process observations are usually assumed to follow a normal distribution. Borror, Montgomery, and Runger [3] studied the performance of the Shewhart control chart for individual observations when the process observations are not normally distributed. Their study showed that if the process follows some other distribution, then the control limits presented in (3.14) could be inappropriate. Specifically, suppose the IC process follows a non-normal distribution such as the t distribution, Exponential distribution or any other distribution with a long right tail, we notice that the ARL performance for these processes are poor. For $\alpha = 0.0027$, the control charts for these distributions yield ARL values that are constantly less than 370 which is the standard ARL value to achieve when this control chart is employed. Therefore, it will be necessary to check the normality assumption before the Shewhart chart for monitoring individual observations can be employed.

3.3.2 Phase-II SPC

In this Phase, we begin monitoring the process observations. After obtaining the control limits and IC dataset from the Phase-I SPC, suppose we have an incoming sequence of independent process observations Y_i $\{i = 1, 2, 3, ...\}$ with unknown change point τ. Again, we assume that random shifts occur only in the mean of the process, whereas the variance remains stable. Now, we begin the monitoring of the sequential process.

From the control limits given in (3.14), the control chart will detect the time point at which a mean shift in the process was observed, thus, at each time point, the process is either IC or OC. Since our primary goal is to design a control chart with dynamic sampling scheme, we are majorly concerned with the detection of the likelihood of possible mean shifts in the sequential process. The current set up cannot give us vital information about potential mean shifts. In order to make our control chart robust to potential shifts in the process, we proceed by using the p-value of the individual process observation to detect shifts in the mean of the process. Given that the process is IC, the p-value is a measure of the extremity of each sample observation [2]. Thus, it gives us vital information for assessing evidence of a mean shift in the process. For the p-value approach, rather than comparing the observation at each time point with the control limits, we compute the p-value corresponding to each observation and then, compare the obtained p-value at each time point with a pre-specified level of significance α. This comparison replaces the initial decision expression in (3.14) which is used to decide if the process is IC at each time point. The p-value of the observed value Y^* at the n-th point is defined as

$$P_{Y_n^*} = P\left(|Z| > \left|\frac{Y_n - \mu}{\sigma}\right|\right) \approx P\left(|Z| > \left|\frac{Y_n - \bar{X}}{\bar{s}/d_3(\tilde{w})}\right|\right) \tag{3.15}$$

where μ and σ are the unknown parameters of the IC distribution. From (3.12), $\mu_0 = E(\bar{X})$, and from (3.13), $\sigma = E\left(\frac{\bar{s}}{d_3(\tilde{w})}\right)$ which are estimated in the Phase-I SPC. The chart gives evidence of an OC mean shift if

$$P_{Y_n^*} < \alpha \tag{3.16}$$

Otherwise, the process is considered to be IC. We employ the two-sided p-value because we are interested in detecting both upward and downward mean shifts in the distribution.

Indeed, using the p-value approach has several advantages. The most paramount advantage being that it is able to inform the practitioner about the likelihood of a potential shift in the mean of the process. In this sense, if the p-value is way larger than α, which indicates that the process is likely to be stable and likely to remain stable in the near future, the practitioner can delay the time before the next sample is collected or collect fewer observations at the next regular sampling time. In contrast, if the p-value is less than α, this indicates that the process is unstable at such time and the process should be stopped. However, if the p-value is only marginally less than α, this indicates that the process is on its way to be unstable and the chart is likely to give a signal in the near future. In this case, the process may still be allowed to continue running, monitoring of the next sample should be sooner than usual and with the collection of more observations at this sampling time. In each setting presented above, the sampling time is variable and also, it is a function of the p-value. In subsequent sections, we will discuss how the sampling time will be determined.

This approach of skipping observations that are judged to be IC during the monitoring procedure will be highly instrumental for sequential processes generating large volumes of data. Rather than monitoring the observation at each time point, we reduce the complexity of the monitoring procedure by placing more emphasis on time points where potential shifts are noticed. Thus, we are able to reduce the run time of the monitoring procedure while still maintaining quick detection of distributional shifts in the process.

In addition, the p-value approach provides a clear and intuitive interpretation of the process control scheme because of the weight of information it carries. It is commonplace to report the results of most statistical analyses in terms of the p-value, where a pre-specified level of significance is used to judge if the hypothesis should be either rejected or not rejected. Using the decision criteria of (3.16), at each time point, the practitioner will be able to clearly report the status of the process. Also, this approach allows the practitioner to make more informed decisions and take insightful actions in cases when the process is still IC or when a shift has been detected.

In order to compute the p-value at each time point, first, we need to indicate the parameters of the IC process distribution. Given that the IC process distribution family is known, we can estimate the mean and the standard deviation of the IC process distribution from IC observations obtained from this distribution family using the estimation approach described in Section 3.3.1, in which μ is estimated by $\bar{X}$ and σ is estimated by $\frac{\bar{s}}{d(w)}$. As an alternative to the estimation approach described in Section 3.3.1, we can estimate the parameters of the IC process distribution using the bootstrap approach when IC process observations are available. In this sense, resampled data are obtained by repeatedly drawing observations of size $\tilde{w}$ with replacement from the IC data set. This process is carried out B times, then, the B number of samples are used to estimate the parameters of the IC process distribution. In the same vein, the resampled data can also be used to design the control limits of the Shewhart chart. For approximately large B, and $w > 1$, the bootstrap method gives a good approximation of the parameters of the IC process distribution.

Next, we investigate the behavior of the distribution of the p-value for different values of the moving window size w, used for estimation of the IC parameters in the Phase-I SPC. Suppose the actual distribution of the IC process is N(0,1), and we have IC dataset from this distribution. From Figure 3.1, we see that the distribution of the p-value is almost identical for several values of w used in estimating the parameters of the IC distribution. Also, notice that as the observation values depart from the IC

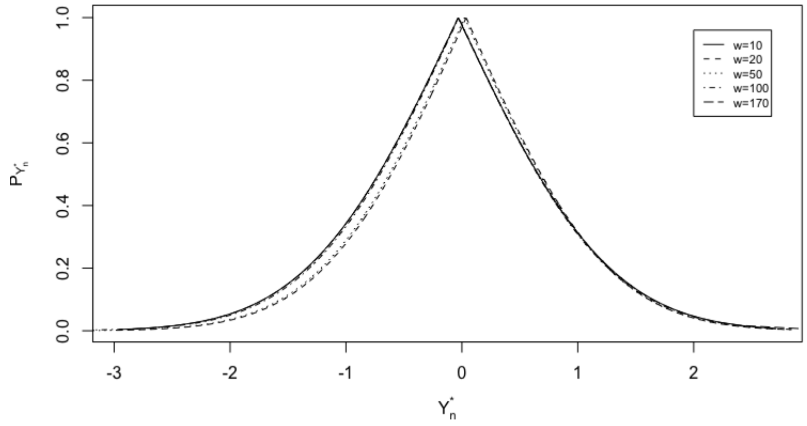

Figure 3.1: Distribution of the p-values of a $N(0,\ 1)$ process where different values of w were used in the computation of $\hat{\sigma} = \frac{\bar{s}}{d_3(w)}$ during the Phase-I SPC.

mean (in this case $\mu = 0$) in either direction, the corresponding p-value decreases, but the observations clustered about the IC mean have the largest p-values. Therefore using the expression in (3.16), the control chart is more likely to give a signal for a mean shift when observations drift away from their IC mean. Also, since the dynamic sampling scheme is an increasing function of the p-value, we would delay the sampling time of the next observation when we notice that a sequence of process readings are consistently clustered around the IC mean, that is, these sequence of observations have large p-values. However, if the p-value begins to get closer to the significant level, the practitioner is alerted to be become more cautious of the process.

3.3.3 Estimation of the Sampling interval

In this section, we present the selection procedure of the size of the sampling interval. We adopt the dynamic sampling scheme proposed by Li and Qiu [15] to estimate the sampling interval. Section 3.2 provides a brief overview of this scheme when applied to the CUSUM control chart. In that section, we saw that the sampling interval function $d(\cdot)$ was chosen from the Box-Cox transformation family. The expression given in (3.5) which is restated below shows the sampling interval function with parameters a, b and λ.

$$d(P_{C_n^{+*}}) = \begin{cases} a + bP^{\lambda}_{C_n^{+*}} & \text{if } \lambda > 0 \\ a + b\log(P_{C_n^{+*}}) & \text{if } \lambda = 0, \end{cases}$$

Section 3.2 also briefly describes the estimation of these parameters for the CUSUM control chart. Subsequently, we discuss the estimation of the parameters of the dynamic sampling scheme for the Shewhart control chart.

The parameters, a, b and λ will be evaluated using the ATS and the AATS of the Shewhart control chart. For a pre-specified ATS$_0$ value and for a specific shift size, the optimal chart will be the chart with the least AATS$_1$ value. By convention, we want to achieve the situation where ATS$_0$ = ARL$_0$. This allows us to estimate the a and λ and then set b to reach this requirement.

First, we begin by estimating the parameter a. As advised by Li and Qiu [15], let a be chosen from the interval [0, 1]. Let us consider the case when the IC process distribution is N(0,1) with a mean shift at the initial time point of size $\{0, 0.1, 0.5, 0.75, 1.0, 1.5, 2, 2.5, 3\}$. For investigative purpose, let us also consider the cases when $\lambda = 0$ and $\lambda = 0.5$. Figure 3.2 shows the AATS values of the chart (3.15)-(3.16) when this process is monitored. For the case when $\lambda = 0$, the values of AATS$_1$ decrease as the values of a increase, and the chart has the best performance when $a = 1$. However, when $\lambda = 0.5$, the AATS$_1$ decreases when a decreases, and the chart performs best when $a = 0$. Therefore, we set $a = 1$, when $\lambda = 0$ and $a = 0$ when $\lambda > 0$.

In order to investigate the effect of λ, we choose its values from [0, 0.5, 1, 1.5, 2,

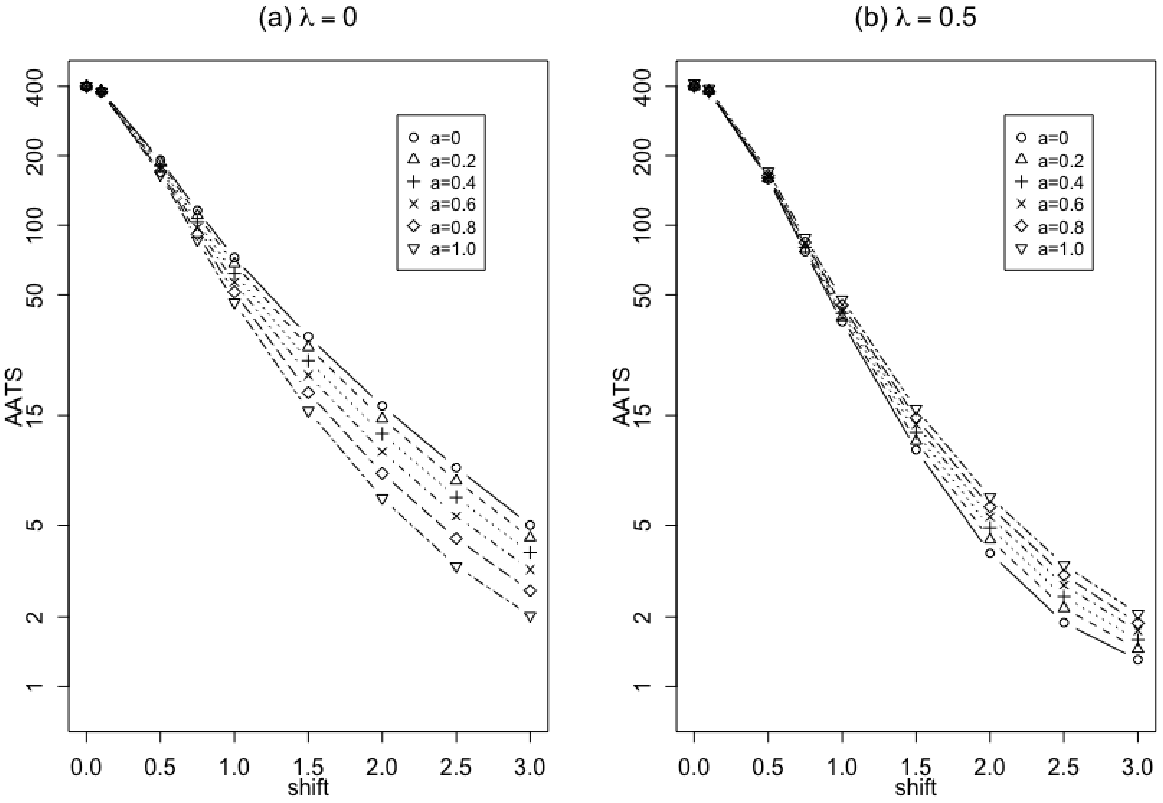

Figure 3.2: AATS values of the control chart (3.15)-(3.16) with the dynamic sampling interval (3.5) for monitoring a process whose IC distribution is N(0,1) with mean shift of size $\{0, 0.1, 0.5, 0.75, 1.0, 1.5, 2, 2.5, 3\}$ occurring at the initial observation time. For the dynamic scheme, two cases are cosidered $-$ (a) $\lambda = 0$ and (b) $\lambda = 0.5$. In both cases, the value of a is cosidered to be $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and b is chosen to achieve $\text{ARL}_0 = \text{ATS}_0 = 400$.

2.5, 3, 6, 10]. From Figure 3.2, we set $a = 1$, when $\lambda = 0$ and $a = 0$ when $\lambda > 0$. Other settings provided in the preceding paragraph are maintained. Figure 3.3 shows the AATS values for the monitoring process. Apparently, the AATS values decrease when λ increases. Furthermore, the AATS performance becomes identical when $\lambda > 2$. Figure 3.3(b) shows the AATS values for the case when $\lambda \in [2, 10]$, the AATS values obtained are virtually the same in this setting. Li and Qiu [15] recommends that $\lambda = 2$.

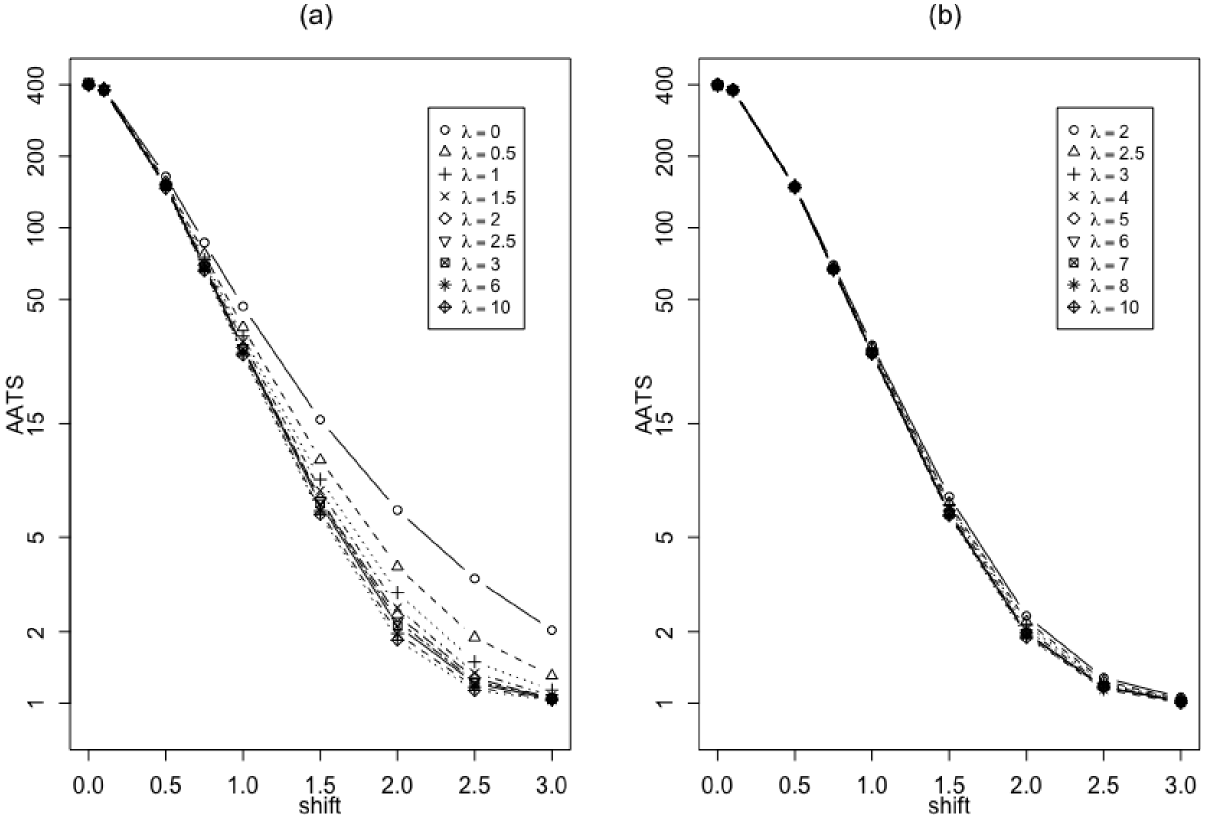

Figure 3.3: AATS values of the control chart (3.15)-(3.16) with the dynamic sampling interval (3.5) for monitoring a process whose IC distribution is N(0,1) with mean shift of size $\{0, 0.1, 0.5, 0.75, 1.0, 1.5, 2, 2.5, 3\}$ occurring at the initial observation time. For the dynamic scheme, two cases are cosidered $-$ (a) $\lambda \in [0, 10]$ and (b) $\lambda \in [2, 10]$. In both cases, $a = 0$ when $\lambda > 0$, $a = 1$ when $\lambda = 0$ and b is chosen to achieve $\text{ARL}_0 = \text{ATS}_0 = 400$.

The estimated parameters obtained from our numerical studies are largely consistent with the parameters obtained by Li and Qiu [15]. Therefore, we write the dynamic sampling interval as

$$d(P_{Y_n^*}) = b \cdot P_{Y_n^*}^2 \tag{3.17}$$

where b can be obtained to achieve $\text{ARL}_0 = \text{ATS}_0$.

If (3.16) is true, the process should be stopped, otherwise, (3.17) provides the sampling interval before the next process reading is monitored. In general, we incorporate the dynamic sampling interval which is expressed in (3.17) into the Shewhart control chart which uses the charting statistic defined in (3.15)-(3.16). This chart is called the dynamic-sampling Shewhart (DyS-S) chart .

3.3.4 Simulation Study

In this section, we discuss and compare the performance of the standard Shewhart chart (SS) and dynamic-sampling Shewhart chart. As earlier stated, the traditional Shewhart chart has a fixed sampling rate while DyS-S chart has a variable sampling rate.

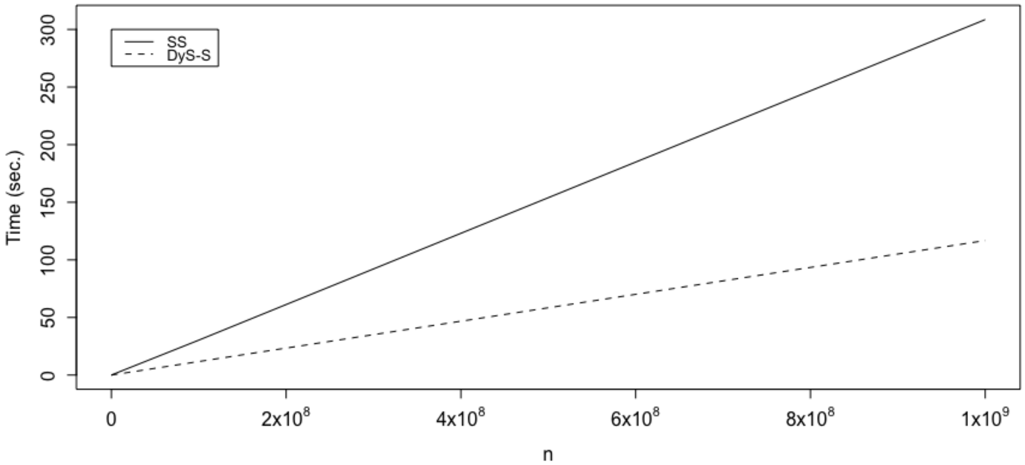

Figure 3.4: Phase-II monitoring times (in seconds) of the traditional Shewhart chart and the Shewhart chart with a dynamic sampling scheme for an IC process distribution of $N(0,\ 1)$ of several sizes n.

First, we begin by investigating the Phase-II monitoring times for both charts. Let us consider the case where we have an IC process which follows a N(0, 1) distribution and we choose different sizes of n, say, $10^3, 10^5, 10^7, 10^8, 10^9$. For both charts we use the control chart defined in (3.15)-(3.16). However, we define the sampling interval for SS as $d(P_{Y_n^*}) = 1$, while the sampling interval for DyS-S is the expression given in (3.17) in which $b = 3.0262$ achieves $\mathrm{ARL}_0 = \mathrm{ATS}_0 = 370$. Figure 3.4 displays the monitoring times (in seconds) for both control charts when n observations from the IC process distribution of N(0, 1) is monitored. The DyS-S chart consistently performs better than the SS chart, because the former does not monitor the observation at each time point whereas the SS chart monitors all observations. From Figure 3.4, the difference in the monitoring times of both methods is almost negligible when n is small, but as n increases the difference becomes more substantial. It is imperative to note that a little difference in the monitoring time of the process is capable of greatly increasing the efficiency of the chart. This advantage of DyS-S will greatly reduce the complexity of monitoring a large sequence of observations.

3.4 Performance of the Shewhart Control Chart with a Dynamic Sampling Scheme

3.4.1 Comparing the $\mathbf{AATS_1}$ values for both control charts

In this section, we further evaluate the differences in the performance of the DyS-S chart and the SS chart. For our investigative purpose, we compute the AATS values achieved for several shift sizes for both charts. Since a better chart will achieve a smaller AATS value for a non-zero mean shift, we employ this conventional method to compare both charts. Furthermore, for a zero-mean shift in the process, we have that $\mathrm{ATS}_0 = \mathrm{AATS}_0$. Also, we know that the SS chart is equivalent to the DyS-S chart when $d(P_{Y_n^*}) = 1$, thus, we can compute its AATS_1 values for several shift sizes.

From the description of both charts, we note that the major differences between the SS chart and the DyS-S Chart are (i) the SS chart uses the conventional control limits to detect shifts in the process whereas the DyS-S chart poses the monitoring of the process as a classical test of hypothesis problem, where the p-value is compared with a pre-specified α value for shift detection, and (ii) the SS chart uses a fixed sampling rate, while the sampling interval defined in (3.17) is used for DyS-S.

For the purpose of evaluation and since we aim to monitor a big data process, we set the value of the ATS_0, which is the expected number of observations from the beginning of the Phase-II process to the time when the chart gives a signal, to be quite large. In order to avoid false alarms when monitoring these large volumes of data, sticking with the conventional $\text{ATS}_0 = \text{ARL}_0 = 370$ will not be ideal. Since $\text{ARL} = \frac{1}{\alpha}$, we can obtain larger ARL values by decreasing α. Then, for the dynamic scheme, we obtain the value of b which achieves $\text{ATS}_0 = \text{ARL}_0$.

For our numerical study, we aim to achieve $\text{ATS}_0 = 1000$, thus, we set $\alpha = 0.001$. Also, we assume that the IC distribution of the process is N(0, 1), and the mean of the process shifts from 0 to $0 + \delta$, where $\delta = 0.5$, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0, 2.25, 2.5, 2.75, and 3.0 are the shift sizes we aim detect. For the sampling interval defined in (3.17), $b = 2.994215$ achieves $\text{ATS}_0 = \text{ARL}_0 = 1{,}000$. Table 3.1 shows the AATS values for the DyS-S chart and the SS chart, respectively. For all shift sizes, δ_j, the DyS-S chart consistently shows to have better AATS performance than the SS chart.

Furthermore, we investigate the performance of both charts when the parameters of IC process distribution are either known or unknown. As stated in Section 3.3.2, we can use either the distribution estimation method (IC distribution is known) or the bootstrap method (when an IC dataset is available) to estimate these parameters. Figure 3.5 displays the performance of both DyS-S and SS charts in the case where

Table 3.1: AATS values for the dynamic-sampling Shewhart Chart and the traditional Shewhart chart when both charts are used to detect several mean shifts δ_j. We assume the process is N(0, 1) and $\text{ATS}_0 = \text{ARL}_0 = 1000$ for both charts.

δ_j	DyS-S	SS
0.00	1000.000	1000.000
0.50	333.006	367.960
0.75	144.225	179.973
1.00	60.955	89.921
1.25	26.616	48.171
1.50	12.205	27.638
1.75	5.771	16.307
2.00	3.043	10.091
2.25	1.892	6.673
2.50	1.376	4.609
2.75	1.165	3.406
3.00	1.071	2.591

the bootstrap approach and the distribution estimation approach are used to estimate the parameters of the IC process distribution. Both estimation techniques perform similarly, however, the overall performance of the DyS-S chart remains better than the SS chart because of lower AATS values of the former.

3.4.2 Abrupt Shifts

From the previous two sections, it has been shown that the DyS-S chart proves to be more efficient and quicker in detecting drifts in the mean of the process. However, there may be certain limitations to its application. In the Phase-II SPC, there is a tendency of the DyS-S chart skipping an observation that is actually OC. If this happens, then it must be that there was an abrupt shift in the process mean. Consider this scenario, during the Phase-II monitoring of a process whose IC process distribution is N(0,1), we obtain four consecutive observations in the interval [-0.0001, 0.0001] which is followed by a single observation in the interval [3.0, 4.0] and then

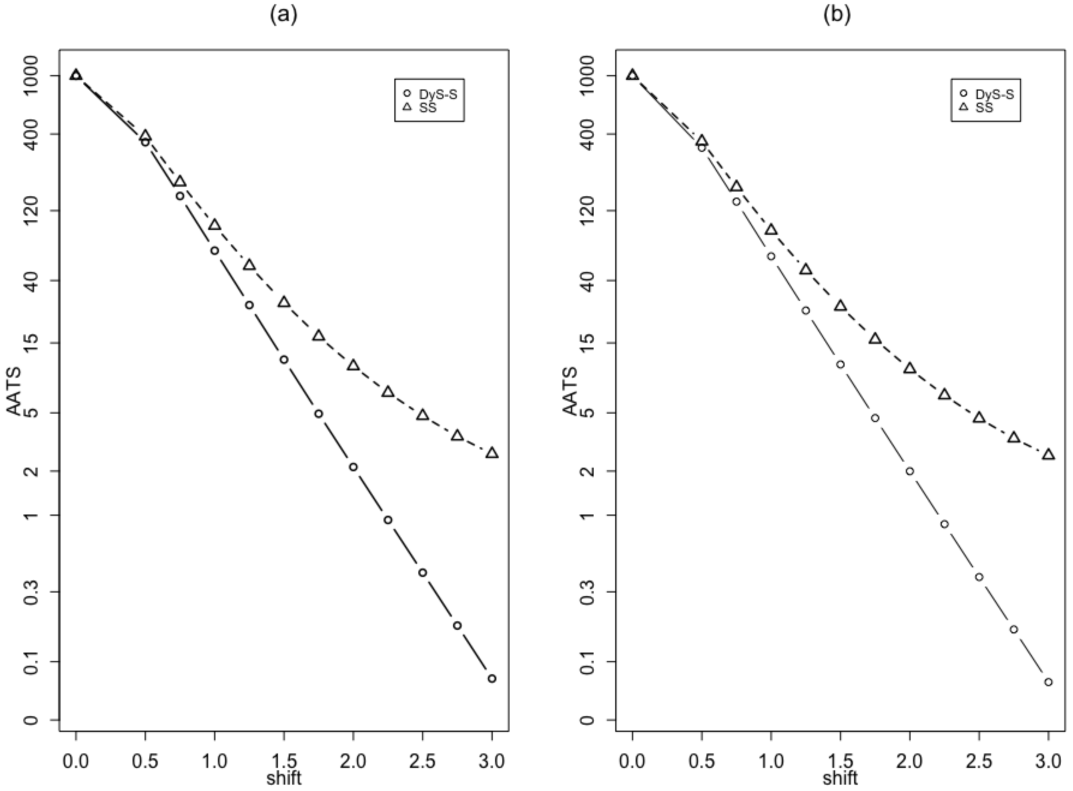

Figure 3.5: AATS values of the Shewhart chart with a dynamic scheme and the Traditional Shewhart chart, where ARL_0 = ATS_0 = 1,000. In this example, the parameters of the IC distribution are unknown. In plot (a), the parameters are estimated using the bootstrap approach. In plot (b), we compute the parameters using the distribution estimation approach.

more observations from the initial interval are obtained. The DyS-S chart may not detect that the single observation from the interval [3.0, 4.0] is actually OC. In this setting, the dynamic scheme is taken unawares and thus fails to detect such shift. However, since SS chart monitors all observations, it will not miss the shift as such time point. In theory, missing such observation may impair the efficiency of the DyS-S chart as this will likely increase the run time of the monitoring process and also increase the AATS_1 of the control chart with a dynamic scheme. In practice, the cost of letting one defective item slip away may also be substantial.

It could be argued that the parameters of the dynamic scheme may be adjusted to handle such shifts. Tweaking the parameters may affect the overall effectiveness of the scheme in detecting other gradual shifts in the process. Since the dynamic scheme will almost never skip more than 3 consecutive observations at any time point, if such shift has a long-staying shift, the dynamic scheme will detect it.

In a real world setting, say for instance in the surveillance and epidemiology of diseases, it is usually unlikely to encounter such unreasonable discrepancies between two observations. However, when the process to be monitored is known to have sudden shifts (with large discrepancies between consecutive observations), then it may be reasonable to adjust the parameters to avoid the scheme skipping more than 2 observations at any given time point. However, this may elongate the monitoring time. In general, since we are concerned with monitoring big data streams, the delay in run time caused by the abrupt change in the distribution of the process will be negligible and the computation is tractable. Nonetheless, the DyS-S chart is even more efficient in detecting gradual shifts in the process.

3.4.3 Simulated Data Example

In this section, we implement the proposed DyS-S chart. Here, we use the control chart to monitor a univariate process with simulated random numbers, and further explain the mechanism of the scheme while giving sufficient interpretation of the monitoring process.

Suppose that IC numerical measurements from a certain quality characteristic of a production process follows a normal distribution where the parameters of this distribution are unknown. An individual numerical measurement from this sequential process is observed every 0.0001 second, and thus, the process generates 36,000,000 observations per hour. Let us assume a mean shift occurs after the 10,000,000th measurement is observed. Given this scenario, we aim to use our proposed DyS-S chart to detect moderate to large shifts in the mean of the process observations.

In order to estimate the IC parameters of the distribution, we assume that we have 1,000,000 IC observations which are individual measurements from the quality characteristic of interest. We then proceed to use the bootstrap approach to estimate these parameters. Using the bootstrap approach with bootstrap sample size $B = 1,000,000$, we estimate the IC mean and IC standard deviation. Furthermore, let $\alpha = 0.001$, and we obtain $b = 2.97204$ which achieves $\text{ATS}_0 = 1{,}000$. Here, α is chosen to detect large shifts in the mean of the process.

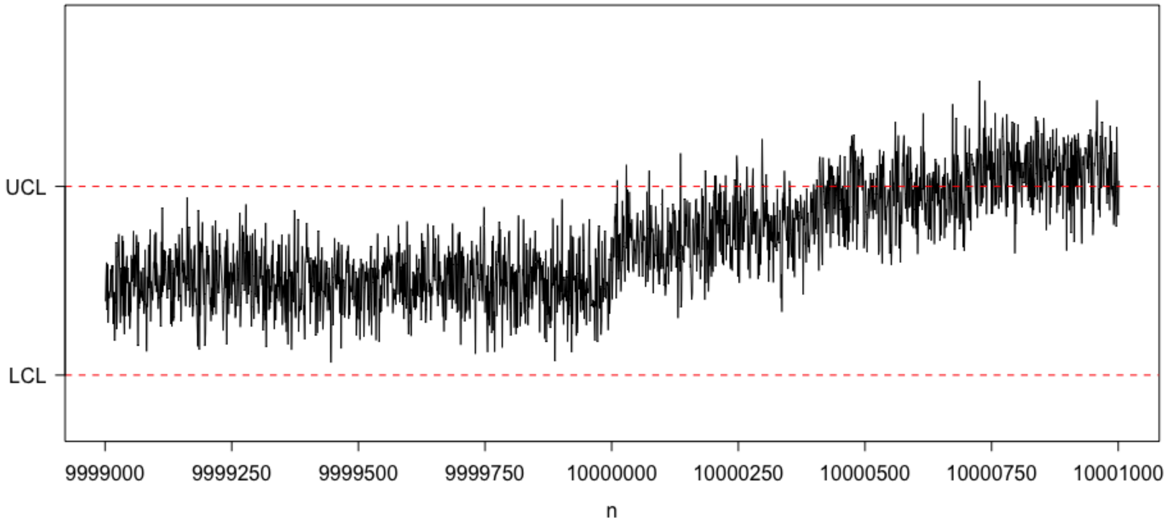

Figure 3.6: Control chart for monitoring the simulated univariate process observations. The warning lines are the control limits for the chart obtained at a significance level of $\alpha = 0.001$.

Figure 3.6 visualizes a subset of the monitoring process, from the 9,999,000th sample to the 10,001,000th sample. Notice that the process becomes unstable after the 10,000,000th sample is observed. Furthermore, the process observations from the 9,999,971th sample to the 10,000,030th sample along with their p-values and dynamic sampling intervals are presented in Table 3.2. At time point $n = 9999981$, the p-value is computed to be 0.905, which indicates that the process is likely to be stable at this time point. The resulting sampling interval is computed to be 2.436, which specifies the time unit before the next sample is monitored. In contrast, at time point, $n = 10000005$, the p-value is computed to be 0.025 and the resulting sampling interval is 0.002. This indicates that the process is on its way to being unstable and future observations are likely to be OC. From Table 3.2, we notice that the process runs stably up till the observation of the 10,000,010th sample, where $P_{Y^*_{10000010}} < \alpha$. The chart gives an OC signal at this time point, and the process should be stopped.

In order to show the relationship between the sampling interval $d(P_{Y^*_n})$ and the p-value in this application, we monitored all observations obtained from this process. In reality, certain observations will be skipped depending on the magnitude of the preceding sampling interval, we present this result in Table A.1 found in Appendix A. Furthermore, before the control chart gives a signal at the 10,000,010th time point, only 7,286,065 observations were monitored when the dynamic scheme is employed. Thus, the control chart skipped a total of 2,713,945 observations during the monitoring procedure. This further confirms the efficiency of the scheme.

3.5 Conclusion

Due to the ability of the DyS-S chart to skip several observations when the process is poised to be IC, the DyS-S shows to have better OC AATS performance than the SS chart, that is, the DyS-S yields quicker detection of distributional shifts. This distinct ability of the DyS-S chart lessens the run time and the complexity of monitoring procedure, thereby making the chart applicable for monitoring observations from big data applications. The underlying assumptions for optimal performance of this control chart include, the observations are independent and are normally distributed, and the process is not characterized by abrupt disturbances.

By incorporating the dynamic scheme in the design of the Shewhart chart, the DyS-S chart takes lesser time (in seconds) in detecting a shift, however, the SS chart takes lesser epoch time for detection of such shift. Furthermore, since the DyS-S chart is designed with the p-values which give information about observations in the near future, the dynamic scheme improves the performance of the Shewhart chart in detecting persistent and gradual shifts in a production process. Also, since it is common practice to report the results of most experimental studies using p-values, the practitioner employing the DyS-S will be able to understand the procedure and clearly report its results.

The DyS-S chart carries an inherent limitation of the SS chart, which is its inability to detect small shifts in the mean of a production process. Thus, it becomes imperative to design SPC charts with dynamic schemes which will detect such small distributional shifts. The EWMA chart performs well in this regard, therefore, in the next chapter, we present discussions on the design of the EWMA control chart with a dynamic scheme.

Table 3.2: The observed value Y_n^*, the p-value of the charting statistic $P_{Y_n^*}$, and dynamic sampling interval $d(P_{Y_n^*})$ at time point n for the dynamic-sampling Shewhart Chart. The values are shown for a subset of the entire process, namely from the 9,999,951th sample to the 10,000,010th sample.

n	Y_n^*	$P_{Y_n^*}$	$d(P_{Y_n^*})$	n	Y_n^*	$P_{Y_n^*}$	$d(P_{Y_n^*})$
9999971	2.893	0.035	0.004	10000001	5.290	0.773	1.775
9999972	5.855	0.393	0.459	10000002	6.489	0.137	0.056
9999973	6.643	0.101	0.030	10000003	5.840	0.402	0.479
9999974	6.921	0.055	0.009	10000004	5.823	0.411	0.502
9999975	4.065	0.349	0.362	10000005	7.243	0.025	0.002
9999976	4.835	0.868	2.240	10000006	7.054	0.040	0.005
9999977	4.727	0.784	1.827	10000007	7.677	0.007	0.000
9999978	3.139	0.063	0.012	10000008	4.394	0.544	0.879
9999979	3.632	0.171	0.087	10000009	5.960	0.338	0.339
9999980	6.375	0.169	0.085	10000010	8.489	0.000	*
9999981	4.882	0.905	2.436	10000011	5.829	0.408	*
9999982	4.282	0.472	0.662	10000012	5.071	0.945	*
9999983	5.080	0.937	2.608	10000013	6.610	0.108	*
9999984	4.448	0.580	1.001	10000014	5.325	0.746	*
9999985	6.083	0.279	0.231	10000015	5.928	0.354	*
9999986	6.737	0.083	0.020	10000016	6.642	0.101	*
9999987	4.722	0.780	1.810	10000017	4.643	0.720	*
9999988	4.387	0.539	0.864	10000018	5.220	0.827	*
9999989	4.965	0.971	2.801	10000019	5.689	0.491	*
9999990	4.948	0.958	2.727	10000020	7.777	0.005	*
9999991	4.858	0.886	2.334	10000021	5.708	0.480	*
9999992	3.306	0.090	0.024	10000022	7.257	0.024	*
9999993	4.655	0.729	1.580	10000023	7.215	0.027	*
9999994	5.767	0.444	0.585	10000024	6.177	0.240	*
9999995	3.876	0.261	0.202	10000025	4.989	0.990	*
9999996	4.369	0.527	0.826	10000026	6.027	0.305	*
9999997	5.956	0.339	0.342	10000027	6.327	0.185	*
9999998	5.691	0.490	0.714	10000028	9.023	0.000	*
9999999	6.517	0.130	0.050	10000029	6.134	0.257	*
10000000	4.375	0.531	0.839	10000030	8.166	0.002	*

CHAPTER 4

EWMA CONTROL CHART WITH A DYNAMIC SAMPLING SCHEME

4.1 Introduction

As stated in Section 1.2.3, the EWMA chart which makes use of history data for evaluating the performance of the process, is based on a weighted average of all observed data available at the current time point. This design makes the control chart effective in detecting small and persistent mean shifts. In this respect, the EWMA chart is often used in the Phase-II SPC where such shift are common. In Section 1.2.3, we presented the charting statistic and control limits of the EWMA chart. From (1.6), we have that

$$\begin{aligned} E_n &= \nu X_n + \nu(1-\nu)X_{n-1} + ... + \nu(1-\nu)^{n-1}X_{n-1} + (1-\nu)^n\mu_0 \\ &= \nu\sum_{i=1}^{n}(1-\nu)^{n-i}X_i + (1-\nu)^n\mu_0 \end{aligned} \tag{4.1}$$

Below, we indicate some interesting properties of the EWMA charting statistic, E_n,

1. From (4.1), the E_n is the weighted average of IC mean, μ_0 and all available observations at the current time point, n.

2. The control chart based on E_n is called an Exponential Weighted Moving

Average chart because at time point, n, the weight $\nu(1-\nu)^{n-i}$ received by the i-th observation decays exponentially when i moves away from n.

3. Notice that when $\nu = 1$, the observations $\{X_1, X_2, ...\}$ receive no weight, that is, the charting statistics does not consider any history data. Thus, the chart based on E_n becomes equivalent to the Shewhart control chart.

4. From (1.6), obviously, more weight will be given to the current observation, and less weight will be given to the previous observations when ν is large (say, $\nu > 0.5$). On the contrary, when ν is small, less weight will be given to current observation while more weight will be given to the previous observations.

5. It can shown that with increasing value of n, the charting statistic E_n has stable variance.

Furthermore, suppose that the IC process distribution is normally distributed with mean μ_0 and variance σ^2, and if the process is IC up to a particular time point n, the distribution of the charting statistic is given as

$$E_n \sim N\left(\mu_0,\ \frac{\nu}{2-\nu}[1-(1-\nu)^{2n}]\sigma^2\right) \tag{4.2}$$

In the case when the mean of the process drifts from μ_0 to μ_1 at time point τ, the distribution of the charting statistic is now defined by

$$E_{n,\tau} \sim N\left(\mu_0 + [1-(1-\nu)^{n-\tau+1}](\mu_1-\mu_0),\ \frac{\nu}{2-\nu}\sigma^2\right) \tag{4.3}$$

In Appendix B, we provide justifications of (4.2) and (4.3). It is easy to see that the last property of the charting statistic stated above is true.

4.2 EWMA Chart using p-values

Let $X_1, X_2, ..., X_\tau, X_{\tau+1}, X_{\tau+2}, ...$ be a sequence of independent random variables from the same distribution. The observations, $X_i,\ 1 \leq i \leq \tau$ come from an IC process which has mean μ_0, while the observations $X_i, \tau+1 \leq i \leq n$ are OC process observations with mean μ_1, where $\mu_0 \neq \mu_1$, and $\tau,\ 1 \leq \tau \leq n$, is an unknown change point. The charting statistic of the EWMA control chart for detecting an upward shift in the mean of the process is defined by

$$E_n^+ = \max\left(0, \nu(X_n - \mu_0) + (1-\nu)E_{n-1}^+\right), \tag{4.4}$$

where $E_0^+ = 0$, and the chart gives an OC signal for an upward mean shift when

$$E_n^+ > \rho_U \sqrt{\frac{\nu}{2-\nu}}\sigma \tag{4.5}$$

where ρ_U is a pre-specified parameter chosen to reach a desired ARL_0 value. The EWMA chart for detecting downward mean shifts can be determined designed similarly.

The weighting parameter ν in (1.6) and (4.4) is usually pre-specified. The value of ν is chosen in such a way where the ARL_1 value for detecting a specific shift is minimized. Since we aim to monitor a process generating large volumes of data, we ought to select combinations of ν and ρ_U values that reach a large ARL_0 value. Also, since the magnitude of ν will determine the weight which current and previous process observations receives, it is imperative to select this value with care. A small value of ν can affect the overall performance of the EWMA chart, this is because it gives more weight to previous data than the current observation. Thus, if there is a mean shift at the current time point, the EWMA chart may take longer to detect such shift. On the other hand, a very large value of ν will impede the efficiency of the EWMA chart in detecting persistent mean shifts in the process due to the reason that it gives less weights to history data. In Section 4.3.1, we provide guidelines for selecting ν.

Similar to the case of the traditional Shewhart chart and CUSUM chart, the typical EWMA chart monitors each observation during the monitoring procedure and only previous data to evaluate the performance of the process. This set up does not provide information about the performance of the process in the near future. To circumvent this, we use the p-value of its charting statistic (similar to the approach introduced in Chapter 3) in the design of the control chart. We define the p-value of the charting statistic as follows. Let E_n^{+*} be the observed value of the charting statistic E_n^+ at time point n, then, the p-value at the n-th time point is given as

$$P_{E_n^{+*}} = P(E_n^+ > E_n^{+*}) \tag{4.6}$$

We reject the null hypothesis that the process is IC at the n-th time point if

$$P_{E_n^{+*}} < \alpha \tag{4.7}$$

where α is a pre-specified level of significance. The setup above is a p-value design of the EWMA chart for detecting upward mean shift in the process, the design for detecting downward shifts in the process follows an analogous pattern.

If $P_{E_n^{+*}}$ is much larger than α, the next sampling time is delayed and fewer samples will be collected at such time. On the contrary, the process gives a signal of a mean shift if $P_{E_n^{+*}} < \alpha$, and thus the process should be stopped. Nonetheless, if the $P_{E_n^{+*}}$ is only slightly larger or less than α, the process may still be allowed to continue running. In this case, the next sample is taken quicker than usual and more observations are collected at this time point. Since this set up is based on the convention of "reject/do not reject hypothesis" which is popular among industry practitioners, the control chart designed with p-values will be easier to interpret than the standard EWMA chart. Implicitly, our chart will have variable sampling intervals. In subsequent sections, we present discussions on estimation of parameters which determines the interval size at the current time point.

Now, we discuss the computation of the p-value of the random variable E_n^+. We present two different scenarios $-$ when the IC distribution of the process is known, and when this distribution is unknown. In the case when the IC distribution of the process observations $\{X_1, X_2, ..., X_\tau\}$ is known, then the IC distribution of the charting statistic, E_n^+ can be obtained by Monte Carlo simulation. Then, the p-value of the charting statistic is computed as if the IC distribution is known. Given values of ν, the variance of the charting statistic becomes stable as n increases, that is, the charting statistic has a steady-state distribution when n is large [21].

For values of n from 1 up to 100, Figure 4.1 shows the variance of the charting statistic for several values of ν. We notice that for some values of ν, say, $\nu \geq 0.05$, $\sigma^2_{E_n}$ becomes stable when $n \geq 30$. However, larger values of n will be needed to achieve stability when $\nu < 0.05$. Here, let us assume that we choose $\nu = 0.05$ (later, we present guidelines for choosing ν when monitoring the big data process) and since a process generating large volumes of data is of interest, it is reasonable to assume that a shift can only occur after $n \geq 30$. Thus, we use the steady state distribution of E_n^+ to obtain the p-values. For $n = 10$, 20, 30, 50, 100, 200 and 500, Figure 4.2 shows the p-values defined by (4.6) of the empirical distribution of E_n^+ defined by (4.4). We

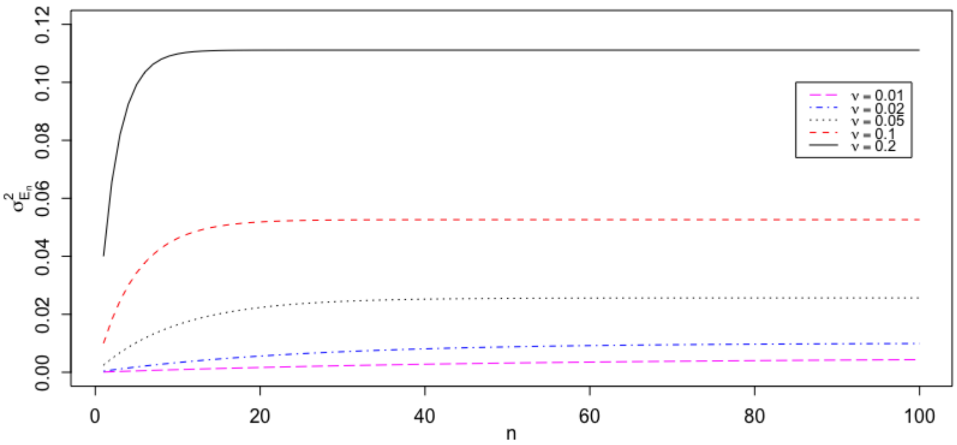

Figure 4.1: Values of $\sigma^2_{E_n}$ as given in the expression of (4.2) for $n = 1, ..., 100$, in cases when $\sigma^2 = 1$, and $\nu = 0.02,\ 0.05,\ 0.1,\ \text{or } 0.2$.

notice that for $n \geq 30$, the distribution of E_n^+ displays the the steady-state property.

Consider the second scenario, when the IC distribution of the process observations is unknown. In this case, if we have available IC dataset, we can use the bootstrap approach which is analogous to the method discussed by Chatterjee and Qiu [6] to determine the IC distribution of E_n^+. In this setting, we repeatedly obtain resampled data from the available IC data, and these resampled data are then used to compute the values of E_n^+ in the Phase-II monitoring of the process. This method is repeated B times and resulting values of E_n^+ are used to estimate the distribution of E_n^+. With sufficient IC process observations, the result obtained using this approach is very similar to the result obtained when the IC distribution is known. Likewise, we obtain the steady-state distribution when $n \geq 30$.

4.3 EWMA chart with a dynamic scheme

As stated in Section 4.2, the magnitude of $P_{E_n^{+*}}$ at the current time point n will determine the next sampling time. That is, the time interval between successive

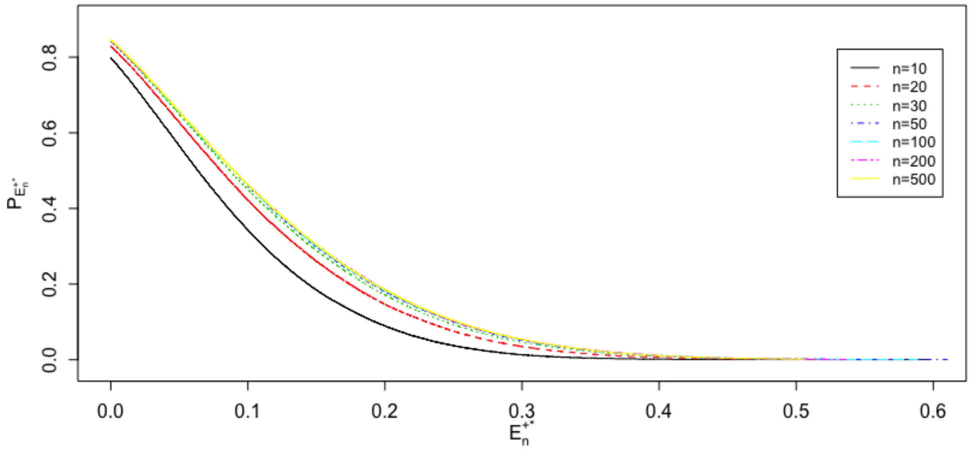

Figure 4.2: p-values of the empirical distribution for the charting statistic (4.4) when n = 10, 20, 30, 50, 100, 200 and 500.

samples follows a variable nature which is dependent on the p-value of the charting statistic at the current time point. Based on the approach proposed by Li and Qiu [15], in Section 3.2 and 3.3.3, we described the sampling interval $d(\cdot)$ as an increasing function of the p-value of the charting statistic of interest. Thus, the sampling interval $d(P_{E_n^{+*}})$ follows (3.5), that is

$$d(P_{E_n^{+*}}) = \begin{cases} a + bP_{E_n^{+*}}^{\lambda} & \text{if } \lambda > 0 \\ a + b\log(P_{E_n^{+*}}) & \text{if } \lambda = 0, \end{cases}$$

Next, using the guidelines presented by Li and Qiu, we discuss the estimation of the parameters of the a, and λ. And as seen previously, b is chosen to satisfy ATS$_0$ = ARL$_0$.

Here, we provide discussion on the selection of a. Before now, we constrained a to be chosen from the interval [0, 1], otherwise, $d(P_{E_n^{+*}})$ will be adversely impacted. Let us assume that the IC distribution of the process is N(0, 1), ATS$_0$ = 200, $\nu = 0.05$, and λ is chosen from the interval [0, 10]. Let us further assume that a mean shift occurs at the initial observation, Figure 4.3 shows the AATS_1 when shift sizes 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5 and 2.0 are considered.

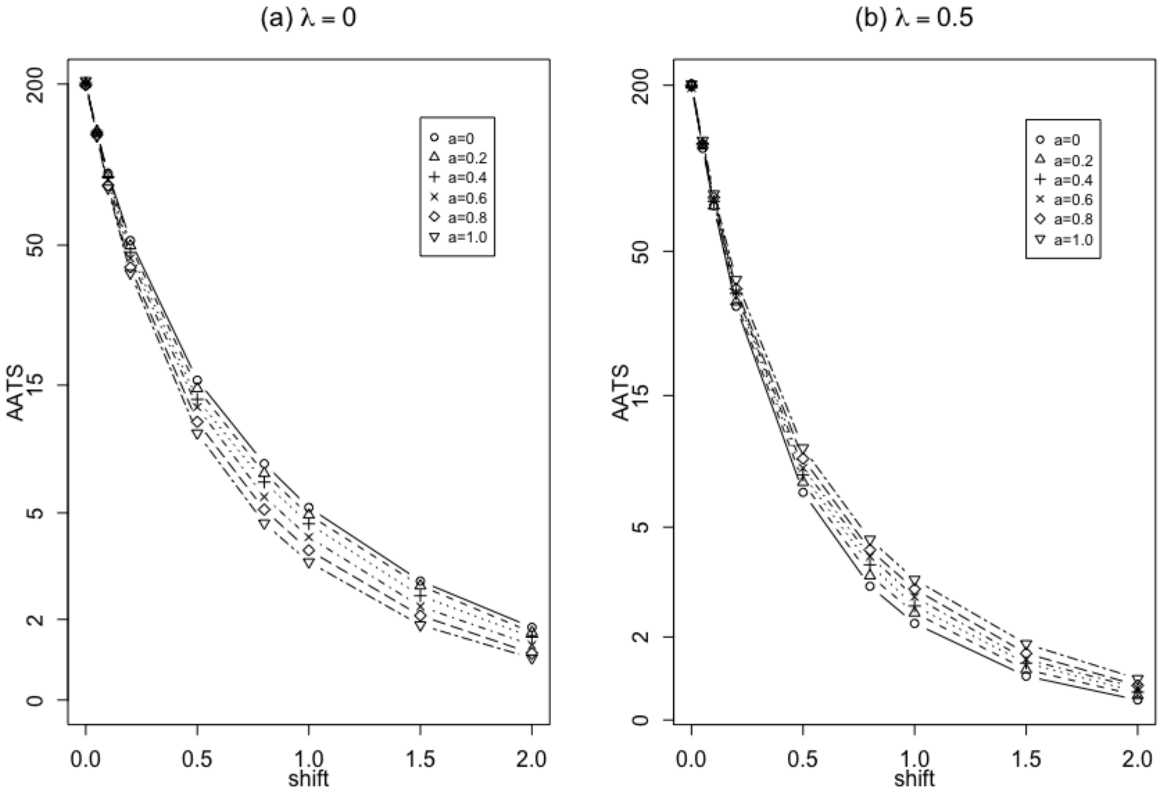

Figure 4.3: AATS values of the control chart (4.4) with the dynamic sampling interval (3.5) for monitoring a process whose IC distribution is N(0,1) with mean shift of size $\{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2\}$ occurring at the initial observation time. For the dynamic scheme, two cases are cosidered - (a) $\lambda = 0$ and (b) $\lambda = 0.5$. In both cases, the value of a is cosidered to be $\{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and b is chosen to achieve $\text{ARL}_0 = \text{ATS}_0 = 200$ and $\nu = 0.05$.

As shown in Figure 4.3, we present the case for two scenarios, when $\lambda = 0$ and when $\lambda = 0.5$. For the case when $\lambda = 0$, we notice that as the value of a increases from 0 to 1 the chart performs better in detecting the specified shifts. In contrast, for the case when $\lambda = 0.5$, we notice that as the value of a decreases the chart performs better. Thus, we choose $a = 1$ when $\lambda = 0$ and $a = 0$ when $\lambda > 0$.

In order to investigate the selection of the λ, Figure 4.4 shows the AATS$_1$ values of 4.4 when values of λ in the interval [0, 10] are chosen to compute the sampling interval. In plot (a), where the λ values are 0, 0.5, 1, 1.5, 2, 2.5, 3, 6, 10, we notice that as the value of λ increases, the chart has better performance. We also notice that the AATS$_1$ does not change much for $\lambda \geq 2$. Plot (b) buttresses this point, it shows the AATS$_1$ values when λ values 2, 2.5, 3, 4, 5, 6, 7, 8, 10 are considered. Indeed, the AATS$_1$ values become stable when $\lambda \geq 2$. Again, the result obtained here is similar to the result obtained by Li and Qiu, and thus we choose $\lambda = 2$.

Therefore, based on the investigation of a and λ, and from Figure 4.3 and Figure 4.4, we suggest that sampling interval for the charting statistic 4.4 should be

$$d(P_{E_n^{+*}}) = b \cdot P^2_{E_n^{+*}} \tag{4.8}$$

where b is chosen to reach a pre-specified such that $\text{ARL}_0 = \text{ATS}_0$.

Therefore the EWMA chart proposed here, which we will call the dynamic sampling EWMA chart (DyS-EWMA), uses the charting statistic described in (4.4), where the chart gives an OC signal for a mean shift in the process if (4.7) is true and the sampling interval function is defined in (4.8). Also, as shown before now, the charting statistic converges to a steady-state distribution when n is reasonably large.

4.3.1 Guidelines for selecting ν

Earlier on, we stated that the shift size to be detected will determine the value of the weighting parameter ν. Also in Section 4.3, we saw that for $\nu \geq 0.05$ with moderate to large values of n, the variance of the EWMA charting statistic becomes stable. However, larger values of n will be needed when $\nu < 0.05$. These two considerations

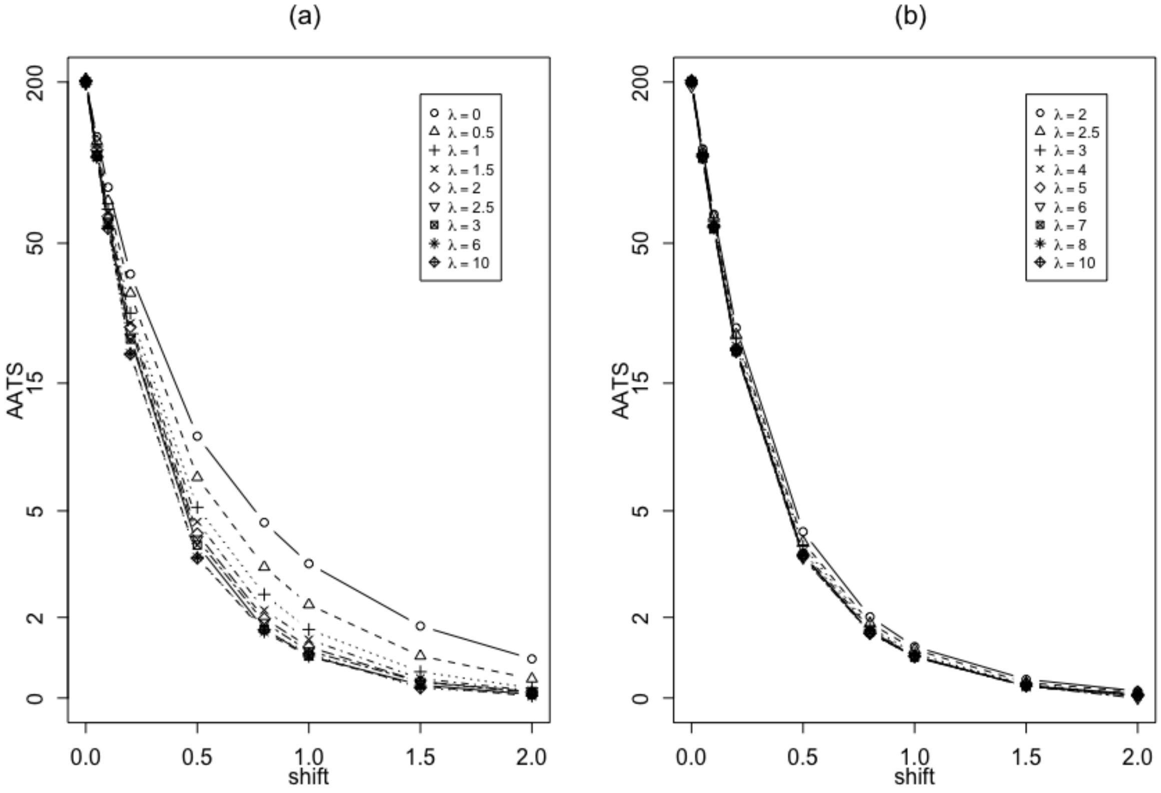

Figure 4.4: AATS values of the control chart (4.4) with the dynamic sampling interval (3.5) for monitoring a process whose IC distribution is N(0,1) with mean shift of size $\{0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1, 1.5, 2\}$ occurring at the initial observation time. For the dynamic scheme, two cases are cosidered − (a) $\lambda \in [0, 10]$ and (b) $\lambda \in [2, 10]$. In both cases, $a = 0$ when $\lambda > 0$, $a = 1$ when $\lambda = 0$ and b is chosen to achieve $\text{ARL}_0 = \text{ATS}_0 = 200$ and $\nu = 0.05$.

will be necessary in selecting ν. But since we are monitoring a big data process, we can relax the later consideration. For the former consideration, since the size of shift to be detected is usually unknown, it may be worthwhile to adaptively select ν. Capizzi and Masarotto [4] proposed an algorithm for choosing ν adaptively under different situations. Even though the method proposed by Capizzi and Masarotto may be effective in adaptively selecting ν, we are hesitant in incorporating the algorithm in the design of the EWMA chart with a dynamic scheme. We note that this algorithm is tedious and requires ample statistical knowledge on the part of the analyst. Since we aim to monitor observations which are generated at a high velocity, performing the adaptive selection of ν at every sampling time point will impede the efficiency of the chart to swiftly detect mean shifts in the process.

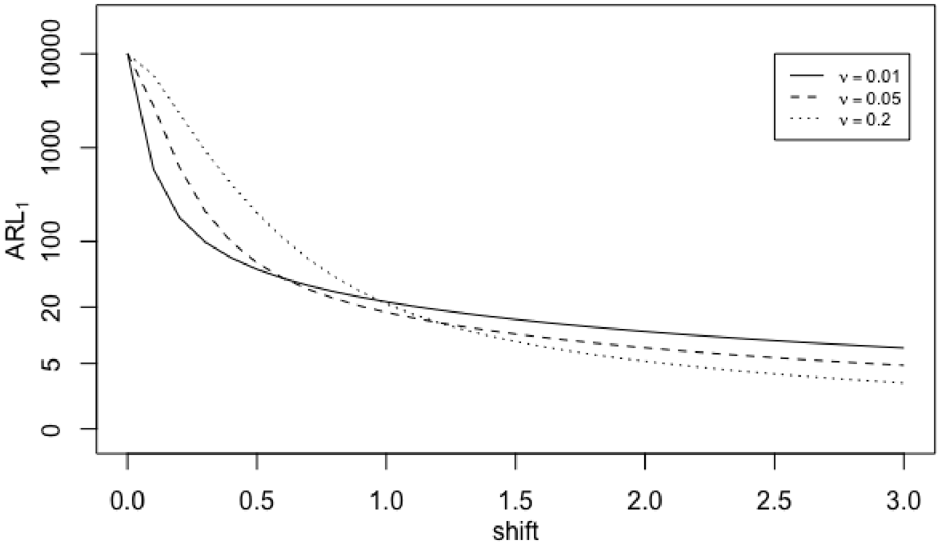

Figure 4.5: ARL_1 values of the EWMA chart when ARL_0 = 10,000, ν = 0.01, 0.05, 0.2 and the shift size changes from 0 to 3 with a step of 0.1

Other practical guidelines for selecting ν have been discussed by several researchers ([18], [21]). From Figure 4.5, we see that small values of ν will effectively detect small shifts in the process, while large values of ν will detect large shifts. In fact, $\nu = 0.01$ will detect shift sizes in the interval [0, 0.7], $\nu = 0.05$ will effectively detect shift sizes in the interval (0.7, 1.2], while $\nu = 0.2$ will effectively detect shift sizes greater than 1.2. The practitioner using this scheme is advised to perform some preliminary analysis to estimate the magnitude of the shift prevalent in the process.

4.4 Simulation Study

In this section, we provide discussions on the numerical performance of the dynamic sampling EWMA chart. For this discussion, we compare the performance of the DyS-EWMA chart to the standard EWMA (S-EWMA) chart. Both charts employ the charting statistic defined in (4.4) and use the p-value of the charting statistic defined in (4.6)-(4.7). However, for the S-EWMA chart, the sampling interval is defined as $d(P_{E_n^{+*}}) = 1$ while the sampling interval of DyS-EWMA is defined by the expression in (4.8).

In order to compare the AATS performance of DyS-EWMA to the S-EWMA chart, we set the weighting parameter to be $\nu = 0.05$. Also, $\alpha = 0.00391$, and $b = 9.0501$ are set to reach $\text{ATS}_0 = \text{ARL}_0 = 400$. Furthermore, we assume that the IC process distribution is N(0, 1), and the mean shift size changes from 0 to 2. Table 4.1 shows the AATS values for both charts when they are used for detecting shift sizes {0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0}.

Table 4.1: AATS values for the DyS-EWMA and the conventional EWMA control charts when are used for detecting mean shifts of size δ_j for a process whose IC distribution is N(0, 1). It is assumed that $\text{ATS}_0 = 400$ and $\nu = 0.05$ for both charts.

δ_j	DyS-EWMA	S-EWMA
0.00	400.000	400.000
0.05	201.749	233.092
0.10	106.018	141.649
0.20	33.724	60.760
0.40	7.017	18.949
0.60	3.156	9.268
0.80	2.015	5.598
1.00	1.565	3.876
1.50	1.158	2.117
2.00	1.049	1.522

From Table 4.1, the AATS_1 values for the DyS-EWMA control chart are smaller than that of S-EWMA when detecting the mean shifts considered. Thus, the DyS-EWMA performs better than the conventional EWMA chart. This advantage of quicker detection of mean shifts in a process makes the DyS-EWMA a better control chart for monitoring a process generating large volumes of data.

Furthermore, we evaluate the performance of both charts when either distribution estimation approach or the bootstrap approach is used to estimate the empirical distribution of the charting statistics. Also, the discussion evaluates the performance of the dynamic scheme used in DyS-EWMA. Using the control charts (4.6)-(4.7) to dectect shift of sizes ranging from 0 to 2, we employ both DyS-EWMA and S-EWMA for the monitoring process.

Suppose the IC process distribution is unknown, but there are available IC observations, we can utilize the bootstrap approach to estimate the distribution of the IC charting statistic. Figure 4.6(a) displays the AATS values for both charts when this approach is employed. In this case, we have 2,000 IC observations from the N(0, 1) distribution. Resampled data are obtained from the available observations and used to compute E_n^{+*}. By repeating this process several times, we use the obtained E_n^{+*} for computing the p-value. From the plot, we see that the DyS-EWMA control chart has better performance than the S-EWMA. Indeed, the dynamic scheme employed in DyS-EWMA improves the EWMA chart.

Suppose the distribution family of the process is known, the distribution estimation method provides an alternative for computing the p-values of the charting statistic. In this approach, the IC distribution parameters are estimated by observations gotten from the known distribution family, and they are used to compute the p-value. Figure 4.6(b) shows the AATS values for both DyS-EWMA and S-EWMA when

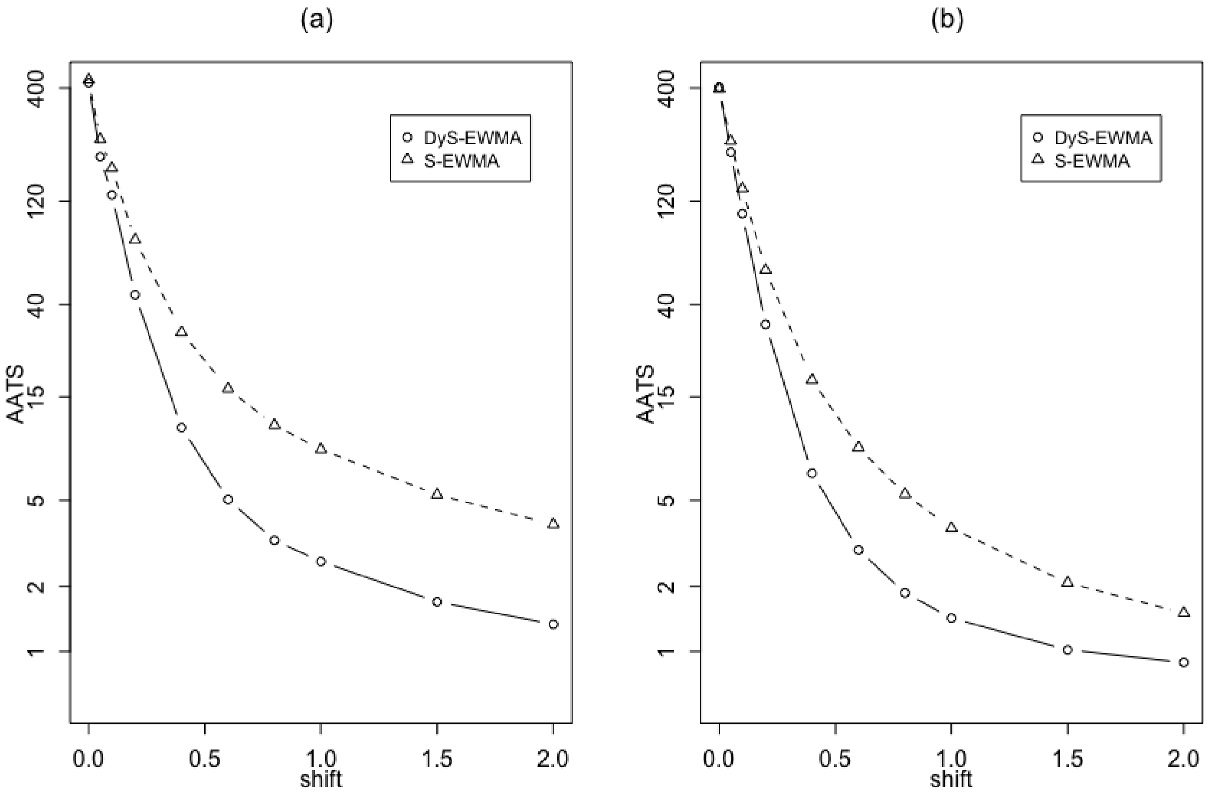

Figure 4.6: AATS_1 values of the control chart (4.6)-(4.7). The sampling interval of S-EWMA is given by $d(\cdot) = 1$ while the sampling interval for DyS-EWMA, is given by the expression in (4.8). In plot (a), the p-value of the charting statistic is computed using the bootstrap approach. In plot (b), the p-value is computed by the distribution estimation approach. Here, $\nu = 0.05$, and $\text{ATS}_0 = 400$

the given distribution family is N(0, 1), and the parameters of the IC distribution parameters are estimated from 2,000 IC process observations. Using this known distribution family, the p-values are computed based on these estimated parameters. The results obtained remain consistent with the results shown in Table 4.1 and Figure 4.6(a). In this sense, it is obvious that the dynamic scheme employed in the design of DyS-EWMA has a substantial effect on the performance of the chart. Indeed, the distribution estimation method and the bootstrap method perform similarly for estimation of the IC distribution of the charting statistic, however, IC process distribution family must be known in order to use the distribution-estimation approach, but knowledge of the distribution family is not required for the bootstrap method.

4.5 Simulated Data Example

In this section, we present discussions on the implementation and the mechanism of the dynamic-sampling EWMA chart. In addition, we utilize the control chart in monitoring a big data process.

Suppose we aim to monitor a univariate quality characteristic for which 60,000,000 process readings are available every hour. These readings are independent of each other and they are obtained at equally spaced time intervals. We use this data instance to demonstrate the mechanism of the DyS-EWMA control chart described in (4.6)-(4.7) with the sampling interval given by the expression in (4.8), to detect small to moderate upward mean shifts in the process.

Furthermore, let us assume that the IC distribution parameters of the charting statistic are unknown, but 1,000,000 readings are available from the distribution family of process which is known to be N(0, 1). In order to compute the p-values, from (4.2) we use the distribution estimation approach to estimate the distribution of the charting statistic E_n^{+*}. Then, $P_{E_n^{+*}}$ is computed as if the IC distribution of E_n^{+*} is known. In this case, we set $\nu = 0.05$. Also, let $\text{ATS}_0 = 1{,}000$. Then $\alpha = 0.00159$, and $b = 8.7380$ achieves this ATS_0.

To further highlight the mechanism, let us assume that an upward mean shift occurred after the 10,000,000th observation is collected. Figure 4.7 shows the obtained p-values of a subset of the process, specifically between the 9,999,951th and the

10,000,020th observation. The DyS-EWMA control chart shown in Figure 4.7 still maintains the properties of the standard EWMA chart. Specifically, at every time point, the chart makes use of history data in the monitoring of the current observation. The control chart gives a signal for an upward mean shift after the 1,000,0011th sample is observed. Table 4.2 shows the computed charting statistic, p-values and sampling interval from the 9,999,961th sample to the 10,000,020th sample. From the desciption of the simulated data, a shift occurred after the 10,000,000th observation was obtained, however, the control chart did not detect this shift at this time until 12 additional observations were monitored. This delay could be attributed to the choice of the weighting parameter ν. Recall, that the shift size to be detected should influence the choice of ν.

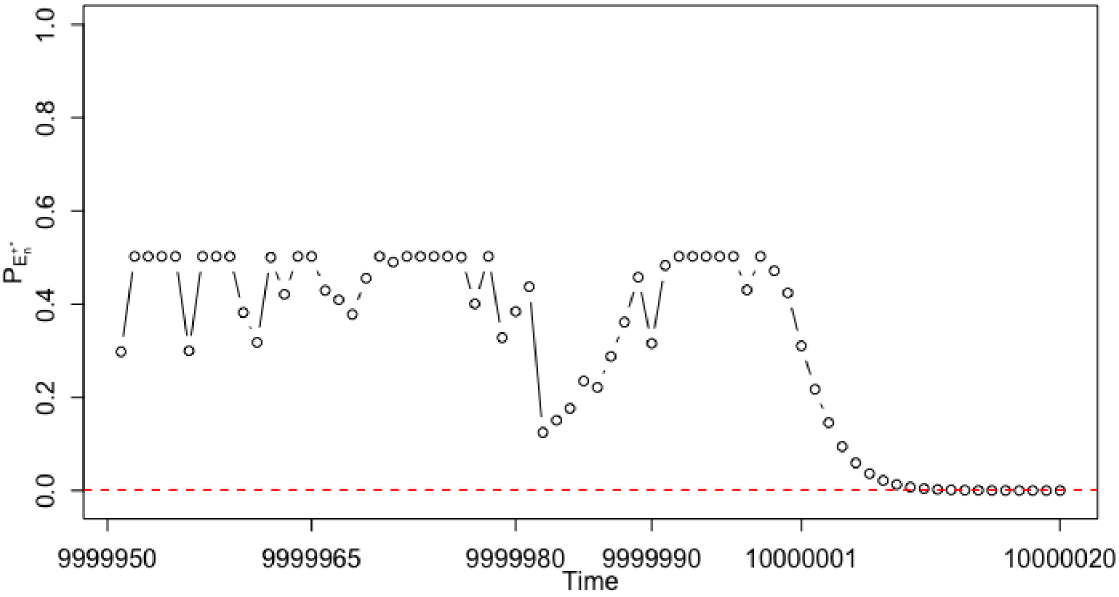

Figure 4.7: The DyS-EWMA control chart for monitoring the simulated process data described in Section 4.5. The red warning line represents the significance level, $\alpha = 0.00159$ which achieves $\text{ATS}_0 = 1{,}000$.

Furthermore, in Table 4.2, no observations were skipped during the Phase-II monitoring. We intend to show the magnitude of the sampling interval at each time point. For instance, when $n = 9,999,965$, the p-value of the charting statistic is reported as $P_{E_n^{+*}} = 0.502$ which suggests the process is stable. The resulting sampling interval, $d(P_{E_n^{+*}}) = 2.203$ indicates that we collect the next observation after 2.203 time units. In general, we see that $d(P_{E_n^{+*}})$ is an increasing function of $P_{E_n^{+*}}$, and notice that $d(P_{E_n^{+*}})$ decays as the process drifts away from the IC mean, which should make the practitioner more cautious of the process.

Prior to the DyS-EWMA conrol chart giving a signal at the 10,000,008-th time point, a total of 7,775,017 observations were monitored. If the S-EWMA chart were employed for the monitoring of this process, additional 2,224,991 observations would have been monitored before the signal is given. The reduction in the number of observations monitored connotes reduction in the total run time of the monitoring process while the ability of the chart in detecting the shift at the known time point is preserved. Thus, the DyS-EWMA chart should be preferred for the monitoring of big data processes.

4.6 Conclusion

Since the DyS-EWMA chart is designed to skip observations that are IC during its monitoring procedure, it will be efficient for monitoring observations from big data applications. Furthermore, by incorporating the dynamic schemes in the design of the standard EWMA chart, the DyS-EWMA chart yields better AATS performance than its traditional counterpart. This advantage becomes even more essential when the pace at which measurement readings from a quality characteristic of interest is greater than the run time of the monitoring scheme.

It is assumed that the observations from the process are independent. Also, the distribution of the process may not be necessarily normal. Since the charting statistics of the EWMA chart is a weighted average of the current observation and previous observations, it will be robust to the normality assumptions in some cases due to the central limit theorem. Particularly, if the distribution is non-normal, then the weighting parameter ν should be selected in such a way where sufficient history data ($n \geq 30$ independent observations) will be involved in the evaluation of process performance. Setting ν to be small will cause more observations to be involved, however, this set-up will be only efficient in detecting small mean shifts. Thus, the robustness of the DyS-EWMA chart to the normality assumption should be used with discretion when moderate mean shifts are to be detected.

Furthermore, since the shift size to be detected by the control chart is usually unknown, DyS-EWMA chart may not give its best performance when an improper value of ν is used in its charting statistic. We suggest a preliminary analysis of the sequential process to determine a proper value of ν. However, more studies are needed to develop schemes which will select ν adaptively.

Table 4.2: The observed charting statistic E_n^{+*}, the p-value $P_{E_n^{+*}}$, and dynamic sampling interval $d(P_{E_n^{+*}})$ at time point n for the dynamic-sampling EWMA chart. The values are shown for a subset of the entire process, namely from the 9,999,961th sample to the 10,000,020th sample.

n	E_n^{+*}	$P_{E_n^{+*}}$	$d(P_{E_n^{+*}})$	n	E_n^{+*}	$P_{E_n^{+*}}$	$d(P_{E_n^{+*}})$
9999961	0.077	0.318	0.882	9999991	0.008	0.483	2.038
9999962	0.001	0.500	2.187	9999992	0.000	0.502	2.203
9999963	0.033	0.421	1.549	9999993	0.000	0.502	2.203
9999964	0.000	0.502	2.203	9999994	0.000	0.502	2.203
9999965	0.000	0.502	2.203	9999995	0.000	0.502	2.203
9999966	0.029	0.430	1.614	9999996	0.000	0.502	2.203
9999967	0.037	0.410	1.466	9999997	0.029	0.431	1.620
9999968	0.050	0.378	1.249	9999998	0.000	0.502	2.203
9999969	0.019	0.456	1.813	9999999	0.012	0.472	1.943
9999970	0.000	0.502	2.203	10000000	0.031	0.424	1.572
9999971	0.005	0.490	2.100	10000001	0.080	0.311	0.843
9999972	0.000	0.502	2.203	10000002	0.126	0.217	0.412
9999973	0.000	0.502	2.203	10000003	0.169	0.146	0.185
9999974	0.000	0.502	2.203	10000004	0.211	0.094	0.078
9999975	0.000	0.502	2.203	10000005	0.250	0.059	0.031
9999976	0.000	0.501	2.193	10000006	0.288	0.036	0.011
9999977	0.041	0.401	1.403	10000007	0.323	0.022	0.004
9999978	0.000	0.502	2.203	10000008	0.357	0.013	0.001
9999979	0.072	0.328	0.940	10000009	0.389	0.008	0.000
9999980	0.048	0.384	1.289	10000010	0.420	0.004	0.000
9999981	0.026	0.437	1.671	10000011	0.449	0.003	0.000
9999982	0.185	0.125	0.136	10000012	0.476	0.001	*
9999983	0.166	0.151	0.198	10000013	0.502	0.001	*
9999984	0.149	0.176	0.272	10000014	0.527	0.000	*
9999985	0.116	0.235	0.483	10000015	0.551	0.000	*
9999986	0.123	0.221	0.428	10000016	0.573	0.000	*
9999987	0.090	0.288	0.724	10000017	0.595	0.000	*
9999988	0.057	0.362	1.142	10000018	0.615	0.000	*
9999989	0.018	0.458	1.832	10000019	0.634	0.000	*
9999990	0.078	0.316	0.870	10000020	0.652	0.000	*

CHAPTER 5

CONCLUSION

In this thesis, first, we proposed an adaptive Shewhart chart for detecting small to moderate persistent shifts in the distribution of a sequential process. The inability of this chart to reach certain ARL_0 values due to the discreteness of the scheme may impede its usage among industry practitioners.

Chapter 3 and Chapter 4 focus primarily on the integration of dynamic sampling schemes in the design of commonly used SPC charts for the efficient monitoring of big data (sequential) processes. Since the dynamic sampling versions of SPC are designed to skip certain observations, thereby reducing the run time of the monitoring procedure, they will be more applicable for monitoring observations from big data applications than the traditional SPC charts. The skipping of observations does not follow an arbitrary nature, rather information about the likelihood of potential shifts in the near future which is obtained from the p-value of the charting statistic of interest provides an ideal criteria for skipping IC observations. In addition, with the information obtained from the p-value during the monitoring procedure, the practitioner can make educated decisions while clearly interpreting the outcomes of the procedure.

The dynamic sampling charts introduced in this thesis focused on detecting mean shifts in a large sequence of independent observations obtained from a univariate quality characteristic. Furthermore, the efficiency of the charts rest on the assumptions that the process is normally distributed and is not characterized by frequent abrupt disturbances.

5.1 Future Studies

From evaluating the performance of the adaptive r-out-of-m charts in Chapter 2, we noticed that certain ARL_0 values cannot be achieved when this chart is employed. Thus, future work is needed to develop methods that will overcome this limitation.

In Section 1.2, we briefly described other SPC charts which have profound applications. Future studies in this area will be focused on the design of the dynamic sampling versions of these charts. In particular, we hope to design dynamic sampling control charts for correlated data, non-normal data and data from multivariate quality characteristics. The dynamic sampling control charts will monitor large sequence of observations from these cases.

Furthermore, in Chapter 4, we indicated that the weighting parameter, ν of the EWMA charting statistic is chosen based on the shift size to detected. But the shift size to be encountered while monitoring the process is usually unknown, therefore, future work will also be needed to incorporate schemes that select ν adaptively in the design of the dynamic sampling EWMA charts.

Also, since the DyS-S chart performs well in detecting large mean shifts while the DyS-EWMA chart performs well in detecting small to moderate mean shifts, it will be worthwhile to design dynamic sampling Shewhart-EWMA control charts for detecting small to large distributional shifts in sequential processes. Ultimately, we intend to apply the methods developed in this thesis to a real data applications.

REFERENCES

[1] D. L. Antzoulakos and A. C. Rakitzis. The modified r out of m control chart. *Communications in Statistics - Simulation and Computation*, 37(2):396–408, 2008.

[2] Y. Benjamini and Y. Kling. A look at statistical process control through the p-values. *Research Paper: RP-SOR-99-08*, Tel Aviv University, School of Mathematical Science, Israel, 1999.

[3] C. M. Borror, D. C. Montgomery, and G. C. Runger. Robustness of the EWMA control chart to non-normality. *Journal of Quality Technology*, 31(3):309–316, 7 1999.

[4] G. Capizzi and G. Masarotto. An adaptive exponentially weighted moving average control chart. *Technometrics*, 45(3):199–207, 2003.

[5] C. W. Champ and W. H. Woodall. Exact results for Shewhart control charts with supplementary runs rules. *Technometrics*, 29(4):393–399, 1987.

[6] S. Chatterjee and P. Qiu. Distribution-free Cumulative Sum control charts using bootstrap-based control limits. *Annals of Applied Statistics*, 3(1):349–369, 03 2009.

[7] A. F. Costa. $\bar{X}$ chart with variable sample size and sampling intervals. *Journal of Quality Technology*, 29(2):197–204, 1997.

[8] WESTERN ELECTRIC. *Statistical Quality Control Handbook*. Western Electric Company, Indianapolis, IN, 1956.

[9] D.M. Hawkins, P. Qiu, and C.W. Kang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

[10] D.M. Hawkins and K.D. Zamba. Statistical process control for shifts in mean or variance using a change point formulation. *Technometrics*, 47:164–173, 2005.

[11] E. F. Kenney and E. S. Keeping. *Mathematics of Statistics, Part Two.* Van Nostrand Company Inc., Princeton, NJ, USA, 2nd edition, 1951.

[12] M. B. Khoo. Design of runs rules schemes. *Quality Engineering*, 16(1):27–43, 2003.

[13] M. Klein. Two alternatives to the Shewhart $\bar{X}$ control chart. *Journal of Quality Technology*, 32(4):427–431, 2000.

[14] J. Lee, T. Wei, and S. K. Mukhiya. *Hands-On Big Data Modeling: Effective Database Design Techniques for Data Architects and Business Intelligence Professionals*. Packt Publishing, Limited, 2018.

[15] Z. Li and P. Qiu. Statistical Process Control using a Dynamic Sampling Scheme. *Technometrics*, 56(3):325–335, 2014.

[16] Z. Li, P. Qiu, S. Chatterjee, and Z. Wang. Using p-values to design statistical process control charts. *Statistical Papers*, 54(2):523–539, May 2013.

[17] Y. Luo, Z. Li, and Z. Wang. Adaptive CUSUM control chart with variable sampling intervals. *Computational Statistics & Data Analysis*, 53(7):2693 – 2701, 2009.

[18] D. C. Montgomery. *Introduction to Statistical Quality Control.* John Wiley & Sons, 5th edition, 2007.

[19] G. V. Moustakides. Optimal stopping times for detecting changes in distributions. *Annals of Statistics*, 14(4):1379–1387, 12 1986.

[20] E. S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 06 1954.

[21] P. Qiu. *Introduction to Statistical Process Control.* Chapman and Hall/CRC, 2014.

[22] M.R. Reynolds, R.W. Amin, and J.C. Arnold. Cusum charts with variable sampling intervals. *Technometrics*, 32(4):371–384, 1990.

[23] M.R. Reynolds, R.W. Amin, J.C. Arnold, and J.A. Nachlas. $\bar{X}$ charts with variable sampling intervals. *Technometrics*, 30(2):181–192, 1988.

[24] S. W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

[25] L. Shu and W. Jiang. A Markov chain model for the adaptive CUSUM control chart. *Journal of Quality Technology*, 38(2):135–147, 2006.

[26] R. S. Sparks. CUSUM charts for signalling varying location shifts. *Journal of Quality Technology*, 32(2):157–171, 2000.

APPENDIX A

SIMULATED DATA EXAMPLE - DYNAMIC SAMPLING SHEWHART CHART

Table A.1: The observed value Y_n^*, the p-value $P_{Y_n^*}$, and dynamic sampling interval $d(P_{Y_n^*})$ at time point n for the dynamic-sampling Shewhart Chart. The values are shown for a subset of the entire process, namely from the 9,999,951th sample to the 10,000,010th sample.

n	Y_n^*	$P_{Y_n^*}$	$d(P_{Y_n^*})$	n	Y_n^*	$P_{Y_n^*}$	$d(P_{Y_n^*})$
9999971	2.893	0.035	1	10000003	5.840	0.402	1
9999972	5.855	0.393	1	10000004	5.823	0.411	1
9999973	6.643	0.101	1	10000005	7.243	0.025	1
9999974	6.921	0.055	1	10000006	7.054	0.040	1
9999975	4.065	0.349	1	10000007	7.677	0.007	1
9999976	4.835	0.868	2	10000008	4.394	0.544	1
9999978	3.139	0.063	1	10000009	5.960	0.338	1
9999979	3.632	0.171	1	10000010	8.489	0.000	*
9999980	6.375	0.169	1	10000011	5.829	0.408	*
9999981	4.882	0.905	2	10000012	5.071	0.945	*
9999983	5.080	0.937	3	10000015	5.928	0.354	*
9999986	6.737	0.083	1	10000016	6.642	0.101	*
9999987	4.722	0.780	2	10000017	4.643	0.720	*
9999989	4.965	0.971	3	10000019	5.689	0.491	*
9999992	3.306	0.090	1	10000020	7.777	0.005	*
9999993	4.655	0.729	2	10000021	5.708	0.480	*
9999995	3.876	0.261	1	10000022	7.257	0.024	*
9999996	4.369	0.527	1	10000023	7.215	0.027	*
9999997	5.956	0.339	1	10000024	6.177	0.240	*
9999998	5.691	0.490	1	10000025	4.989	0.990	*
9999999	6.517	0.130	1	10000028	9.023	0.000	*
10000000	4.375	0.531	1	10000029	6.134	0.257	*
10000001	5.290	0.773	2	10000030	8.166	0.002	*

APPENDIX B

DISTRIBUTION OF THE EWMA CHARTING STATISTIC

Here, we provide justifications for (4.2) and (4.3). From (4.1),

$$\begin{aligned}
E_n &= \nu X_n + (1-\nu)E_{n-1} \\
&= \nu X_n + \nu(1-\nu)X_{n-1} + ... + \nu(1-\nu)^{n-1}X_1 + (1-\nu)^n\mu_0 \\
&= \nu\sum_{i=1}^{n}(1-\nu)^{n-i}X_i + (1-\nu)^n\mu_0
\end{aligned}$$

The expectation of E_n is given as

$$\begin{aligned}
\mu_{E_n} &= \nu E(X_n) + \nu(1-\nu)E(X_{n-1}) + ... + \nu(1-\nu)^{n-1}E(X_1) + (1-\nu)^n\mu_0 \\
&= \nu\mu_0 + \nu(1-\nu)\mu_0 + ... + \nu(1-\nu)^{n-1}\mu_0 + (1-\nu)^n\mu_0 \\
&= \mu_0
\end{aligned}$$

Since

$$\nu\sum_{i=1}^{n}(1-\nu)^{n-i} + (1-\nu)^n = 1$$

Also, since $Var(aX+b) = a^2Var(X)$, then the variance of E_n is given as

$$\begin{aligned}
Var(E_n) = \sigma^2_{E_n} &= Var[\nu\sum_{i=1}^{n}(1-\nu)^{n-i}X_i + (1-\nu)^n\mu_0] = \left(\nu\sum_{i=1}^{n}(1-\nu)^{n-i}\right)^2 Var(X_i) \\
&= \nu^2\sum_{i=1}^{n}(1-\nu)^{2n-2i}\ \sigma^2 \\
&= \nu^2\sum_{i=1}^{n}\frac{(1-\nu)^{2n}}{(1-\nu)^{2i}}\ \sigma^2 = \nu^2(1-\nu)^{2n}\sum_{i=1}^{n}\left[\left(\frac{1}{1-\nu}\right)^2\right]^i\ \sigma^2 \\
&= \nu^2(1-\nu)^{2n}\frac{\left(\frac{1}{1-\nu}\right)^2\left[1-\left(\frac{1}{1-\nu}\right)^{2n}\right]}{1-\left(\frac{1}{1-\nu}\right)^2}\ \sigma^2 \\
&= \frac{\nu}{2-\nu}\left[1-(1-\nu)^{2n}\right]\ \sigma^2
\end{aligned} \tag{B.1}$$

Therefore, we obtain (4.2),

$$E_n \sim N\left(\mu_0,\ \frac{\nu}{2-\nu}\left[1-(1-\nu)^{2n}\right]\ \sigma^2\right)$$

In the case when the mean of the process drifts from μ_0 to μ_1 at time point τ, $1 \leq \tau \leq n$, the variance of E_n is still (B.1), but the mean is given as

$$\begin{aligned}
E_{n,\tau} &= \nu \sum_{i=1}^{n-\tau+1} (1-\nu)^{n-\tau+1-i} X_i + (1-\nu)^{n-\tau+1}\mu_0 \\
&= \nu(1-\nu)^{n-\tau+1} \sum_{i=1}^{n-\tau+1} \left(\frac{1}{1-\nu}\right)^i X_i + (1-\nu)^{n-\tau+1}\mu_0
\end{aligned}$$

Taking Expectation of both sides we have

$$\begin{aligned}
\mu_{E_{n,\tau}} &= \nu(1-\nu)^{n-\tau+1}\left\{\frac{\left(\frac{1}{1-\nu}\right)\left[1-\left(\frac{1}{1-\nu}\right)^{n-\tau+1}\right]}{1-\left(\frac{1}{1-\nu}\right)}\right\}\mu_1 + (1-\nu)^{n-\tau+1}\mu_0 \\
&= (1-\nu)^{n-\tau+1}\mu_0 + \left[1-(1-\nu)^{n-\tau+1}\right]\mu_1
\end{aligned}$$

From (B.1), we notice that, for a given ν value, as n gets larger the distribution of $E_{n,\tau}$ converges to

$$E_{n,\tau} \sim N\left(\mu_1,\ \frac{\nu}{2-\nu}\sigma^2\right)$$