

**DOKUZ EYLÜL UNIVERSITY
GRADUATE SCHOOL OF NATURAL AND APPLIED
SCIENCES**

**ROBUST SCALE ESTIMATORS IN
STATISTICAL QUALITY CONTROL:
ROBUST CONTROL CHARTS**

by
Alp Giray ÖZEN

March, 2012

İZMİR

**ROBUST SCALE ESTIMATORS IN
STATISTICAL QUALITY CONTROL:
ROBUST CONTROL CHARTS**

**A Thesis Submitted to the
Graduate School of Natural and Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Master of Science
in Statistics**

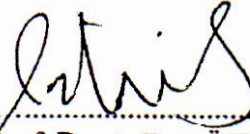
**by
Alp Giray ÖZEN**

March, 2012

İZMİR

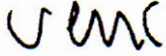
M.Sc THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**ROBUST SCALE ESTIMATORS IN STATISTICAL QUALITY CONTROL: ROBUST CONTROL CHARTS**" completed by **ALP GİRAY ÖZEN** under supervision of **ASSIST. PROF. DR. A. FIRAT ÖZDEMİR** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.



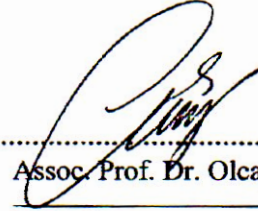
Assist. Prof. Dr. A. Firat ÖZDEMİR

Supervisor



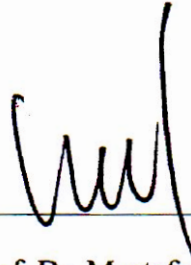
Prof. Dr. Serdar KURT

(Jury Member)



Assoc. Prof. Dr. Olcay AKAY

(Jury Member)



Prof. Dr. Mustafa Sabuncu

Director

Graduate School of Natural and Applied Sciences

ACKNOWLEDGMENTS

Undoubtedly, one of the most precious personalities I honestly want to appreciate here is the brilliant statistician: Sir Ömer GÜCELİOĞLU. He is the person that makes me “love and learn” Statistics. More importantly, he is one of the exceptional personalities for whom it is needless to consume extra words; with his kind personality, his honest desire to teach, and his legendary existence. I wish he rested in peace and were appreciating “my dealings with Statistics in order to be able to carry his flag.” I always miss you too much, Lord!

Nefise deserves so special thanks. My mother is the most beautiful woman in the world ever!

I aspire to return my sincere thanks to my dear master Sir Olcay AKAY, not only for teaching me stochastic processes, statistical estimation theory, and statistical detection theory, but also for being an excellent teacher and an excellent academician. He actually is my role model!

I also want to thank my supervisor Assist. Prof. Dr. A. Fırat ÖZDEMİR for his guidance, support, and encouragement through the course of this research.

There are two special persons in my life that I cannot pass over here, without presenting my genuine gratitude.

My legendary childhood friend Ahmet TAŞPINAR has a natural habit of helping me. Since both of us are still alive, he realized an invaluable help to me in checking the grammar errors, and in finding more efficient ways of explaining my concepts, in this research. By the way, he is an English Teacher, and he really is good in his field.

Another legendary friend of mine, Murat GÜNEŞ, is one of the most polite personalities, and one of the most creative scientists I've ever met. He is a distinguished professor of Physics, in my heart. His invaluable help to me were in preparing my presentation, and in giving the examples of my research for applied Physics.

I do love you, friends!

Finally, I want to give my very special regards to each of the people involved in the references part for their creation of the books, papers, and websites. I could be a bridge between their studies, and the findings of this research, due to our simultaneous realizations in life.

Alp Giray ÖZEN

ROBUST SCALE ESTIMATORS IN STATISTICAL QUALITY CONTROL: ROBUST CONTROL CHARTS

ABSTRACT

Control Charts are one of the most powerful tools used to detect aberrant behavior in industrial processes. A valid performance measure for a control chart is the average run length (ARL); which is the expected number of runs to get an out of control signal. The usual Shewart S Control Charts' performance in controlling the process standard deviation is based on the fundamental assumption of normality, which is a rarely consistent one in practice.

Robust estimators are of vital importance in Statistics in order to estimate population parameters independent of the data distribution. "Median Absolute Deviation" (MAD), S_n , and Q_n are such estimators for population standard deviation.

The aim of this study is to observe performance of Shewart S-Chart for heavy tailed symmetric distributions and propose alternative robust control charts that perform better. Such qualified charts are proposed, whose control limits are obtained by using bootstrap methodology. Monte Carlo simulation study is performed to simulate their performances under normal and non-normal distributions.

The findings of the study assert an equal-power design to the use of Shewart S Chart. More importantly, although the proposed design's false alarm probability (PFA) is slightly more under normal distribution, its PFA is much less than that of Shewart S Chart for heavy tailed symmetric distributions. This design employs the simultaneous use of S_n Chart and Q_n Chart.

Cauchy model is an important model in specific applications of Electrical Engineering and Physics. Shewart S chart does not work in a Cauchy model and

another design is proposed for this model. This second design makes simultaneous use of MAD and Qn Charts.

Keywords: Statistical quality control, control charts, heavy tailed distributions, Cauchy model, robust estimators, Median Absolute Deviation, Sn, Qn , average run length, bootstrap method.

**İSTATİSTİKSEL KALİTE KONTROLÜNDE DAYANIKLI
ÖLÇEK KESTİRİCİLERİ:
DAYANIKLI KONTROL GRAFİKLERİ
ÖZ**

Kontrol Grafikleri, endüstriyel süreçlerde istenmeyen sapmaların tespitinde kullanılan en güçlü araçlardandır. Kontrol grafiklerinin geçerli performans ölçütlerinden birisi, üretimin kontrol dışında olduğu sinyalinin alınması için gereken ardışık örneklem adedinin beklenen değeri olan, ortalama tekrar uzunluğudur. Klasik Shewart S Kontrol Grafiğinin kitle standart sapmasının kontrolü için performansı, temelde normallik varsayımına dayanır ki; bu varsayım, pratikte nadiren tutarlıdır.

Dayanıklı tahmin ediciler, İstatistik için, kitle parametresinin veri dağılımından bağımsız olarak tahmin edilmesinde çok önemli bir yere sahiptir. “Ortanca Mutlak Sapması” (MAD), S_n ve Q_n , kitle standart sapmasının dayanıklı tahmin edicilerinden bazılarıdır.

Bu çalışmanın amacı, Shewart S Grafiğinin performansını ağır kuyruklu dağılımlar için gözlemlenmek ve daha iyi performansa sahip olan, dayanıklı kontrol grafikleri önermektir. Kontrol limitleri bootstrap yöntemi ile belirlenen, bu özellikte grafikler önerilmiştir. Önerilen grafiklerin performansları, normal dağılan ve normal dağılmayan kitleler için, Monte Carlo benzetim çalışması yaparak karşılaştırılmıştır.

Çalışmanın bulguları, Shewart S Grafiği'nin normal dağılım altında kullanımı ile eş-güçlü olan bir tasarım öne sürer. Daha da önemlisi, önerilen tasarımın yanlış uyarı olasılığının normal dağılım için S grafiğinkinden biraz daha yüksek olsa da, ağır kuyruklu dağılımlar için bu olasılığın S grafiğinkinden çok daha düşük olmasıdır. Bu tasarım, S_n ve Q_n grafiklerinin eş zamanlı kullanılmasıyla oluşturulmuştur.

Cauchy modeli, Elektrik Mühendisliği, ve Fizikte birtakım özgül uygulamalar için önemli bir modeldir. Shewart S Grafiği Cauchy modeli için sonuç vermez ve bu

model için de yeni bir tasarım önerilmiştir. Bu yeni tasarım da, MAD ve Q_n grafiklerinin eş zamanlı kullanılmasına karşılık gelmektedir.

Anahtar Kelimeler: İstatistiksel kalite kontrolü, kontrol grafikleri, ağır kuyruklu dağılımlar, Cauchy modeli, dayanıklı tahmin ediciler, Ortanca Mutlak Sapması S_n , Q_n , ortalama tekrar uzunluğu, bootstrap yöntemi.

CONTENTS

	Page
M.Sc THESIS EXAMINATION RESULT FORM... Error! Bookmark not defined.	
ACKNOWLEDGMENTS.....	iii
ABSTRACT.....	v
ÖZ.....	vii
CHAPTER ONE - INTRODUCTION.....	1
CHAPTER TWO - STATISTICAL QUALITY CONTROL: BASIC CONCEPTS.....	8
2.1 Location and Dispersion Charts for Gaussian Data.....	9
2.2 Performance of Dispersion Charts for Non-Gaussian Data.....	31
CHAPTER THREE - ROBUST ESTIMATORS AND QUALITY APPLICATIONS.....	46
3.1 “Classical versus Robust” Estimation of Location.....	47
3.2 “Classical versus Robust” Estimation of Scale.....	55
3.3 A search for Robust Scale Control Charts.....	61
CHAPTER FOUR – CONTROL CHARTS USING ROBUST SCALE ESTIMATORS.....	72
4.1 Bootstrap Confidence Intervals.....	73
4.2 Robust Control Charts.....	77
4.2.1 Normal Distribution.....	78
4.2.1.1 Sample Variance.....	78
4.2.1.2 Median Absolute Deviation.....	78
4.2.1.3 S_n	82

4.2.1.4 Q_n	85
4.2.2 Logistic Distribution.....	89
4.2.2.1 Sample Variance.....	89
4.2.2.2 Median Absolute Deviation	90
4.2.2.3 S_n	91
4.2.2.4 Q_n	93
4.2.3 Laplace Distribution	97
4.2.3.1 Sample Variance.....	97
4.2.3.2 Median Absolute Deviation	97
4.2.3.3 S_n	99
4.2.3.4 Q_n	101
4.2.4 Cauchy Distribution.....	104
4.2.4.1 Sample Variance.....	105
4.2.4.2 Median Absolute Deviation	106
4.2.4.3 S_n	108
4.2.4.4 Q_n	110
4.3 Proposed Control Designs	113
4.3.1 Proposed Design for Finite Moment Symmetric Distributions.....	116
4.3.2 Proposed Design for Cauchy Model.....	125
CHAPTER FIVE - CONCLUSION	130
REFERENCES	133
APPENDIX – 1	136
APPENDIX – 2	137
APPENDIX – 3	138
APPENDIX – 4	142

CHAPTER ONE

INTRODUCTION

To start with, I want to make “the first aphorism of Hippocrates” remembered:

*“[The] art is long,
Life is short,
Crisis fleeting,
Experiment perilous,
Judgement difficult....”*
(Hippocrates, 400 BC)

To be accustomed to thinking judgement as a single variable function of observations and the dependence on the truth of feelings about observations make judgement easy in daily life. However, this is not the case and judgement is a difficult task as Hippocrates asserts. A better, or let’s say more reliable model for judgement may be considering it as a bivariate function of observations and assumptions.

To handle the discussion in a different manner, I replace judgement with inference, observation with data set, and function with estimator. Considering the assumptions on distribution of the data set forms the essence of my thesis’ subject. That’s what I would write:

*“Population is infinite,
Sample size is small,
Life is random,
Experience memoriless,
Inference difficult...”*

The ability in statistical thinking improves the quality of inferences made. Moreover, qualified inferences yield a general control over the future.

For that reason, I think that Control is a natural instinct rather than a technique to maximize profit. In fact, Quality makes life easier and magnificent, and Statistics is the unique tool to perform Quality Control.

Besides being a professional art of living, Statistical Quality Control has a wide range of applications in industrial processes. The mean and standard deviation of products must be controlled so as to standardize the production. By this way, the product quality is improved and production costs are minimized.

Control Charts are one of the most powerful tools used to detect aberrant behavior in industrial processes. The usual Shewart Control Charts' efficiency is based on the fundamental assumption of normality.

However, normality assumption is rarely consistent in practice. In general, we essentially want to control the process mean and the process standard deviation, independent from the data distribution. In order to monitor these parameters, it is important to advance the control charts based on robust statistics, because these statistics are expected to be more resistant to moderate changes in the underlying process distribution.

The usual performance measures for a control chart are false alarm probability; which is the probability of getting an out of control signal when the process is in control, and probability to miss; which is the probability of failure in detecting the case that process is out of control. Based on these probabilities, average run length (ARL); which is the expected number of runs to get an out of control signal, is of great importance. The aim of this thesis study is to determine this performance measure for normal and non-normal symmetric distributions and compare the

performance of usual “Shewart S Control Chart” and proposed “Robust Scale Control Charts.”

In general, a production process is desired to perform with its specified value. However, even if the process is designed perfectly, there exists a natural variability due to unavoidable causes. Then, the specified value becomes the mean value of the produced items’ measures. Moreover, this natural variability results in a need to determine the standard deviation of the process, and is often called a “stable system of chance causes.” A process that is operating with only chance causes of variation is said to be in statistical control (Montgomery, 2009).

On the other hand, the sources of variability that are not part of the chance cause pattern are referred as “assignable causes of variation.” A process that is operating in the presence of assignable cause(s) is said to be an out of control process (Montgomery, 2009).

Control charts are statistical tools that are used to monitor the system and to detect the assignable cause when an out of control signal is observed. Basically, a control chart is a confidence interval whose limits are determined assuming that the process is in control. For this purpose, a random sample is selected from the process periodically, and the realization of the relevant statistics is used to decide between:

H_0 : The process is in control

H_A : The process is out of control

In fact since this hypothesis testing is made at the end of each period, the test statistics can be viewed as a discrete time stochastic process. Additionally, “the significance,” and “the power of the test” are of great importance especially in terms of the reduction of long term costs that occur by false alarms and misses.

Obviously, the relevant statistics for the estimation of the population mean is a location statistics and that of standard deviation is a dispersion statistics. Under the assumption that the data follows a normal distribution, from statistical theory, sample mean and sample variance are the uniformly minimum variance unbiased estimators for population mean and population variance respectively.

A pitfall of the statistics sample mean and sample variance is that their efficiency is highly dependent on the underlying assumption. To obtain more efficient estimates, robust methods are frequently used when the underlying assumption is violated. Robust methods offer operative alternatives to the traditional statistical methods which yield greater statistical power and efficiency when the underlying assumptions are not satisfied. In this study, a search for robust scale estimators to use in Statistical Quality Control will be presented.

To control the process variability, process standard deviation is mostly monitored by Shewart S-control chart or Shewart R-control chart, which use sample standard deviation and sample range, respectively. The theory under the formulation of these charts is based upon normality assumption and, hence, their performances are expected to be very good if the data fits to a normal distribution. Central Limit Theorem does not support their performances for non-normal case because most of the industrial processes do not permit large sample sizes and actual distributions of the data may have heavy tails or may be highly skewed. That is the reason to include a search for some robust estimators of scale. Specifically, the estimators that will be studied are “Median Absolute Deviation” (MAD), S_n , and Q_n . Parallel to the goal of the study, robust scale charts alternative to Shewart S-Chart will be proposed and their control limits will be constructed.

As stated previously, ARL is the expected number of samples to take an out of control signal. When the process is in control, it is desired to obtain large run lengths, since an out of control signal will be a false alarm and when a shift occurs in the process, a small value of run length is desired because we want to detect the out of control case as soon as possible. Considering the probabilities for the two cases; run

length is a geometric random variable, whose parameter is α when the process is in control and is $1 - \beta$ when the process is out of control.

My story begins with an introduction to Statistical Quality Control. This will be the theoretical background and Control Chart examples using simulated data. At first, data will follow a Gaussian (Normal) Distribution. To ensure the reliability of the study and the simulations, the simulated run length values will be compared with their theoretical expected values.

Next, I will base my research on answering the following question: “What if the data does not fit a normal distribution?” For this purpose, the run length performance of Shewart S-Chart will be simulated for some heavy tailed symmetric distributions. In particular, the Non-Gaussian distributions used in this study are: Logistic, Laplace, and Cauchy distributions. Including the Gaussian distribution, these four distributions constitute a good set in the sense that they scale from slight to strong in terms of heaviness of tail characteristics. The poor performance of Shewart S-Chart for Non-Gaussian distributions strongly supports the need for the research of alternative robust scale control charts.

Before searching control limits of robust scale control charts, a formal definition of robustness will be presented. Some location and scale estimators will be compared with respect to basic characteristics of robustness. Although the subject of the study is scale estimation, starting with location estimation will be complementary.

Influence function of an estimator is very important for understanding its robustness. Basically, it reflects the effect of an additional data to the estimator. Although being efficient for Gaussian distribution, sample mean is non-robust for its influence function is unbounded. I will represent the empirical influence function of mean, with those of some robust location estimators, which are median and trimmed mean, using a simulated data. These aim to enable a comparison so that it will be easier to express the concepts of breakdown point and gross error sensitivity.

A similar pattern will be followed for the scale estimation counterpart. The efficient estimator “sample standard deviation” has an unbounded influence function and those of MAD, S_n , and Q_n are all bounded. All these three estimators are highly robust since they have highest possible breakdown point, which is 50%. However, their efficiency and gross error sensitivity under Gaussian distribution change, Q_n being the most efficient (of these three) and MAD having the lowest possible gross error sensitivity. This is a wonderful motivation for me, to go behind.

Control limits for Shewart S-Chart are based on the standard error of sample standard deviation, which can be formulated by the help of Chi-square distribution. Alternatively, one can use “Variance Control Chart” with a direct use of Chi-square distribution in order to gain the advantage of having a constant (not a function of sample size) false alarm probability.

On the other hand, formulas for standard errors of our robust estimators do not exist. I tried to propose a MAD-Chart for a start, and applied two formulations, those of the former is similar to the S-Chart and latter to the variance chart. Since simulated run length performances are not satisfactory, the study continues with some other technique of Glorious Statistics.

Bootstrapping is a useful method to estimate the standard error of relevant statistics. Moreover, bootstrapping is a brilliant method since it somehow enables the data to talk for itself. The last part of the thesis before the Conclusion chapter is devoted to the robust control chart studies using bootstrap confidence intervals. At this part, the run length performances are compared for the Gaussian and three Non-Gaussian symmetric distributions.

The results obtained are quite satisfactory to propose control designs and to advise for future studies. Interestingly, S_n and Q_n charts present different characteristics for their run lengths, for the finite moment symmetric distributions. S_n performs very well in ARL_0 but is considerably slow in detecting shifts. On the contrary, Q_n has very low ARL_1 values, which mean that it is really good in detecting the shifts, but

also aim to give false alarms frequently. These observations yield the idea for simultaneous use of S_n and Q_n , whose features will be discussed in detail.

Moreover, the corresponding proposal for Cauchy distribution, which is a distribution that does not have finite moments, is the simultaneous use of MAD and Q_n control charts. The reasoning is exactly the same as the previous design.

CHAPTER TWO

STATISTICAL QUALITY CONTROL: BASIC CONCEPTS

Perfectness is something we create in our minds and we improve using philosophy, mathematics, or some other specific science. Our way of thinking and usage of language result in the perception of perfectness. To illustrate, when we call a leaf, we idealize “an image of a leaf” and think that leaf as a representative image for the thing called.

However, nothing is perfect in nature. Neither two things nor two moments in life exactly matches each other. To see or understand this imperfectness, some kind of a numerical measurement is needed such as weight, dimension, or volume. It will undoubtedly be observed that any measure varies from one object to another or from time to time. Therefore, a specific observation within the same class of objects -say length of a leaf- is a random variable.

Having identified the imperfectness that is dealt with, we need to develop some strategy and technique to reduce the degree of imperfectness. Here, the Glorious science Statistics takes the floor. He asks Pupil two questions that will get the story started. The former: “What is the length of the leaf you imagine?” Pupil answers, “That of my image is exactly 20 cm but my observations are around 20 cm.” And the latter: “Up to what level you will consider your observations as acceptable?”

Pupil is a fan of nature and she loves trees. She lives in a village near a forest, in which there are a lot of quassia amara (bitter-wood) trees. She takes special care of the health of the trees and observes their leaves in her daily walks through the forest.

Sometimes, the trees get ill and need to be pruned. When a tree becomes ill, its leaves show unexpected characteristics in their length. Therefore, a tree’s healthiness –say quality– can easily be understood from the length of its leaves. In order to

detect the illness of a tree and the time to prune it, Pupil decided to control the length of the leaves of a tree. She needs to develop some methodology for this purpose.

2.1 Location and Dispersion Charts for Gaussian Data

Let, (X_1, X_2, \dots, X_n) is a random sample of the size n with mean \bar{X} , range R , and standard deviation s . By Central Limit Theorem (CLT), the limiting distribution of \bar{X} is Gaussian with mean μ and standard error σ/\sqrt{n} . Furthermore, the probability is $1 - \alpha$ that any sample mean will fall within

$$\mu - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \text{ and } \mu + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad (2.1)$$

When $Z_{\alpha/2} = 3$, confidence level $1 - \alpha$ is 0.9973 and so 99.73% of the sample means fall within

$$\mu + 3 \frac{\sigma}{\sqrt{n}} \text{ and } \mu - 3 \frac{\sigma}{\sqrt{n}} \quad (2.2)$$

It is customary to use 3σ control limits. Letting the constant $A = \frac{3}{\sqrt{n}}$, the upper and lower control limits (UCL and LCL) for \bar{X} chart are obtained:

$$\text{UCL} = \mu + A\sigma \quad (2.3)$$

$$\text{LCL} = \mu - A\sigma \quad (2.4)$$

(Montgomery, 2009).

Thinking in terms of detection terminology, when the sample mean is within confidence interval, one may conclude that the population mean is NOT significantly different from $\mu = \mu_0$. Then, to control the mean of the process, it makes sense to obtain periodic samples and calculate the mean of the observations. If the sample mean is out of the control limits, the conclusion will be that the population mean is

different from μ_0 and the process is said to be out of statistical control. When this happens, the process mean is said to shift to a new mean μ_1 .

Besides the location parameter of the process random variable, its dispersion should also be controlled. The population standard deviation σ can be controlled via two estimators. The first one is the sample standard deviation, whose theoretical background is defined as follows. We know from statistical theory that when the distribution of the data is Gaussian, sample variance s^2 is Uniquely Minimum Variance Unbiased Estimator (UMVUE) for population variance σ^2 . However, s is NOT an unbiased estimator for σ since $E(s) = c_4\sigma$. Hopefully, c_4 is a constant which depend on the sample size n . Moreover, we have $Var(s) = (\sigma\sqrt{1 - c_4^2})^2$ and considering CLT by the same manner yields the following three sigma control limits:

$$UCL = c_4\sigma + 3\sigma\sqrt{1 - c_4^2} \quad (2.5)$$

$$LCL = c_4\sigma - 3\sigma\sqrt{1 - c_4^2} \quad (2.6)$$

The following two constants are defined to reduce the formulas:

$$B_6 = c_4 + 3\sqrt{1 - c_4^2} \text{ and } B_5 = c_4 - 3\sqrt{1 - c_4^2} \quad (2.7)$$

Then, the control limits of s chart becomes:

$$UCL = B_6\sigma \quad (2.8)$$

$$LCL = B_5\sigma \quad (2.9)$$

(Montgomery, 2009).

An alternative estimator of σ is the sample range R . To introduce its theoretical background, we need to consider the random variable $\xi = \frac{R}{\sigma}$ which is called the

Relative Range. The parameters of the distribution of ξ are functions of the sample size n . The expected value and standard deviation of ξ are d_2 and d_3 , respectively. Then, we have $E(R) = d_2\sigma$ and $Var(R) = (d_3\sigma)^2$ where d_2 and d_3 are functions of n . Similar to the construction of the s chart parameters, the following constants are defined:

$$D_2 = d_2 + 3d_3 \text{ and } D_1 = d_2 - 3d_3 \quad (2.10)$$

Finally, control limits of R chart are:

$$UCL = D_2\sigma \quad (2.11)$$

$$LCL = D_1\sigma \quad (2.12)$$

(Montgomery, 2009).

Before continuing, I need to put a marker here to turn back, recall, and go on further discussions. The construction of methodology is based on two important assumptions. First, control limits for sample mean are based on large sample case using CLT. Second, s is the best estimator for σ under Gaussian distribution.

Having learned some introductory theory about Quality Control from Glorious Statistics, Pupil decided to apply her knowledge to control the health of quassia amara trees in the forest. She decided to take a random sample of only $n = 5$ leaves from each tree in order to check more trees a day. Since a healthy tree has an average of 20 cm length leaves, she specified $\mu = 20$. After a research on standard deviation of the leaves, she set $\sigma = 2.5$.

To calculate the control limits, she obtained the constant values of the charts for $n = 5$ which are as follows:

$$A = 1.342; \quad B_6 = 1.964; \quad B_5 = 0; \quad D_2 = 4.918; \quad D_1 = 0 \quad (2.13)$$

She calculated the corresponding control limits for the charts as follows:

\bar{X} chart:

$$UCL = \mu + A\sigma = 20 + 1.342 * 2.5 = 23.355 \quad (2.14)$$

$$LCL = \mu - A\sigma = 20 - 1.342 * 2.5 = 16.645 \quad (2.15)$$

s chart:

$$UCL = B_6\sigma = 1.964 * 2.5 = 4.910 \quad (2.16)$$

$$LCL = B_5\sigma = 0.000 * 2.5 = 0.000 \quad (2.17)$$

R chart:

$$UCL = D_2\sigma = 4.918 * 2.5 = 12.295 \quad (2.18)$$

$$LCL = D_1\sigma = 0.000 * 2.5 = 0.000 \quad (2.19)$$

To learn, search, and make calculations whole day made Pupil tired and it was a little later than her usual sleeping hour. To be fresh and happy with each starting day, she got accustomed to sleeping early in her childhood. While she was falling asleep, she thought how Glorious is the Statistics. It was a waste of 23 years of her life to be unaware of this lofty wisdom. However, it was still lucky to meet him in her youth. In her dream, she saw Glorious Statistics as a wisdom granddaddy, but his bread was yellow.

It was a beautiful morning and she felt the sunshine warming her heart. She took a bottle of water, a notebook, and a ruler and she went to the forest. She randomly selected 5 leaves from each of the 30 different quassia amara and collected the following data:

Table 2.1 Length of 5 randomly selected leaves from each of the 30 trees. Leaves data are generated from a Normal distribution with mean 20 and standard deviation 2.5. Their statistics mean, standard deviation, variance, and range are calculated at the right part of the table to construct corresponding control charts. Control limits are at the right bottom part of the table. Yellow shaded point is out of control limits.

Leaves Data										
TREE	LEAF					MEAN	STD_DEV	VARIANCE	RANGE	
	1	2	3	4	5					
1	18.12	18.93	17.11	20.07	19.82	18.81	1.22	1.49	2.95	
2	19.60	22.24	22.81	19.84	22.12	21.32	1.49	2.21	3.21	
3	17.13	17.15	21.27	20.72	21.11	19.48	2.14	4.59	4.14	
4	20.51	19.20	15.68	18.75	23.29	19.48	2.77	7.66	7.61	
5	20.76	19.29	15.47	22.78	17.71	19.20	2.80	7.86	7.32	
6	16.48	16.49	21.25	21.57	23.64	19.89	3.23	10.46	7.15	
7	16.17	21.42	22.72	16.49	16.45	18.65	3.16	9.97	6.55	
8	23.42	16.98	23.50	18.64	20.72	20.65	2.88	8.32	6.52	
9	22.61	19.81	22.45	17.07	18.07	20.00	2.51	6.30	5.55	
10	20.08	20.32	19.37	19.41	19.68	19.77	0.42	0.17	0.95	
11	19.22	20.45	18.56	16.20	21.64	19.21	2.06	4.23	5.44	
12	18.62	20.73	15.57	20.66	17.63	18.64	2.17	4.73	5.17	
13	20.93	20.44	17.03	22.92	15.72	19.41	2.96	8.75	7.21	
14	17.98	18.27	15.70	15.51	19.53	17.40	1.74	3.02	4.02	
15	22.33	18.19	22.34	17.34	21.32	20.30	2.37	5.63	5.00	
16	16.30	17.49	18.81	25.54	16.73	18.97	3.79	14.37	9.24	
17	24.71	22.93	18.18	19.67	18.95	20.89	2.80	7.84	6.53	
18	22.05	17.52	22.12	18.94	20.49	20.22	2.00	3.99	4.61	
19	19.91	19.20	21.44	20.40	20.37	20.26	0.81	0.66	2.23	
20	22.94	20.83	22.14	17.96	20.84	20.94	1.89	3.58	4.98	
21	18.97	17.13	18.45	20.69	15.41	18.13	1.98	3.93	5.27	
22	20.55	19.00	17.28	19.45	16.80	18.62	1.55	2.42	3.75	
23	20.91	21.24	17.69	21.80	20.75	20.48	1.61	2.59	4.11	
24	16.94	15.97	21.17	22.39	18.63	19.02	2.73	7.43	6.42	
25	19.81	25.19	20.30	19.17	14.94	19.88	3.65	13.32	10.25	
26	20.92	21.65	18.40	22.73	19.87	20.71	1.66	2.76	4.33	
27	12.81	20.12	20.27	13.08	24.20	18.10	4.98	24.80	11.39	
28	24.20	16.43	17.50	15.26	20.42	18.76	3.59	12.90	8.94	
29	19.94	17.12	19.54	14.05	16.14	17.36	2.45	5.98	5.89	
30	19.11	20.85	20.66	20.72	22.98	20.87	1.38	1.91	3.87	
AVERAGE =						19.514	2.360	6.463	5.686	
LCL =						16.645	0	0.1653	0	
UCL =						23.355	4.91	27.8098	12.295	

To obtain the first observations and to see that all the trees are healthy made Pupil happy. There seems to be a little problem for 27th tree's leaf's standard deviation since Standard Deviation Chart gave an out of control limit value. However, this value is only a little over the upper control limit and still inside the control limits for Range Chart. Just in case, she marked that tree and decided to observe it later again.

Since it is hard to observe each statistics via numbers, she decided to construct the control charts and check if there is an aberrant behavior in the data pattern. The reason is that, although all the data values are within limits, some specific patterns of the data points may be suspicious for out of quality tendency. These patterns are called "Sensitizing Rules for Shewart Control Charts." For example, two of the three consecutive points being outside the two sigma warning limits, six points in a row steadily increasing or decreasing, and a non-random pattern of the data are some of these rules (Montgomery, 2009). The corresponding charts are in the following figures:

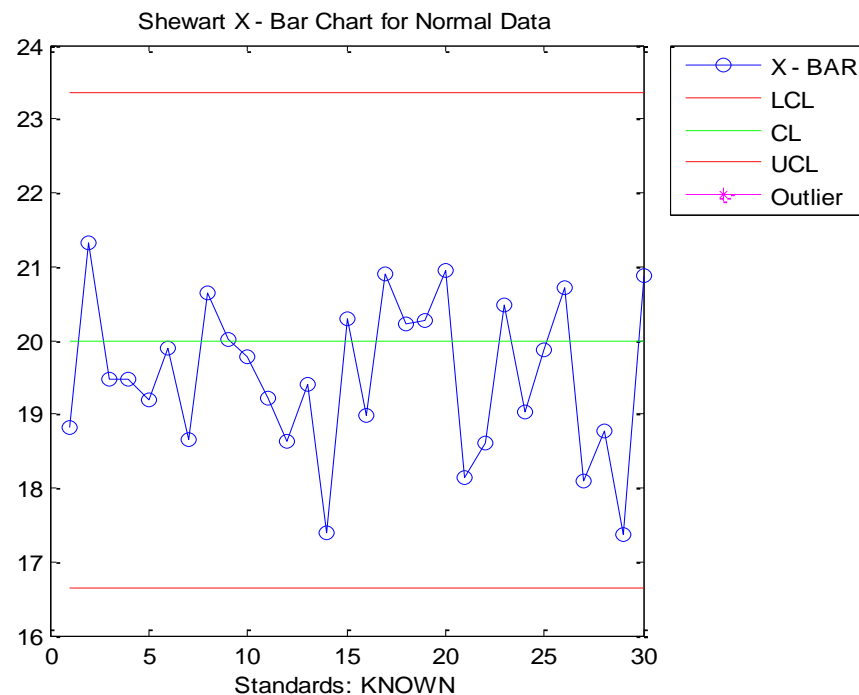


Figure 2.1 Shewart \bar{X} Control Chart for leaves data given standards $\mu = 20$ and $\sigma = 2.5$

Data points of X-bar chart are completely random and are not even close to Control Limits. Process mean is in statistical control.

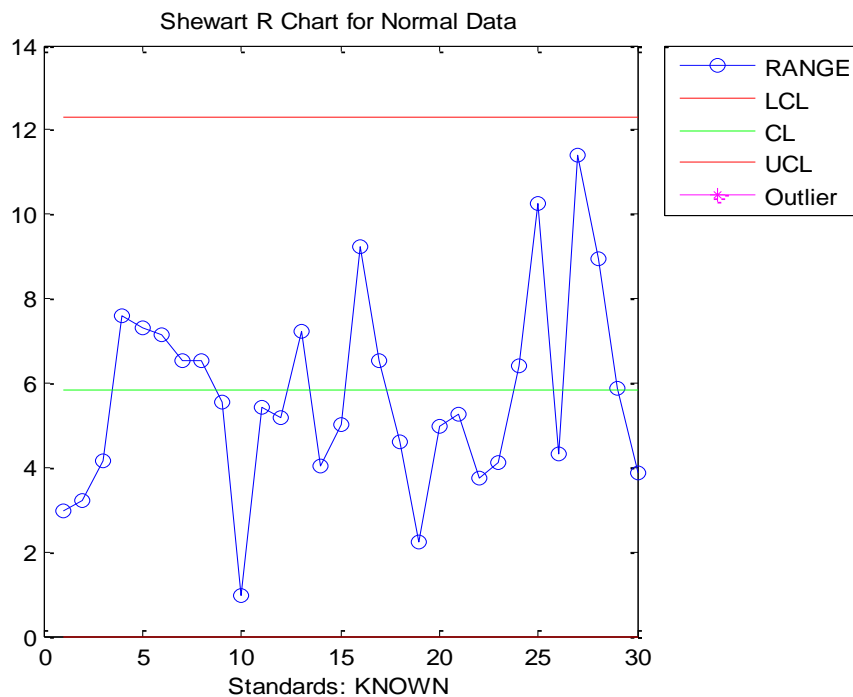


Figure 2.2 Shewart R Control Chart for leaves data given standard $\sigma = 2.5$

Data points of R chart are also completely random and are not even close to Control Limits. Process standard deviation is in statistical control.

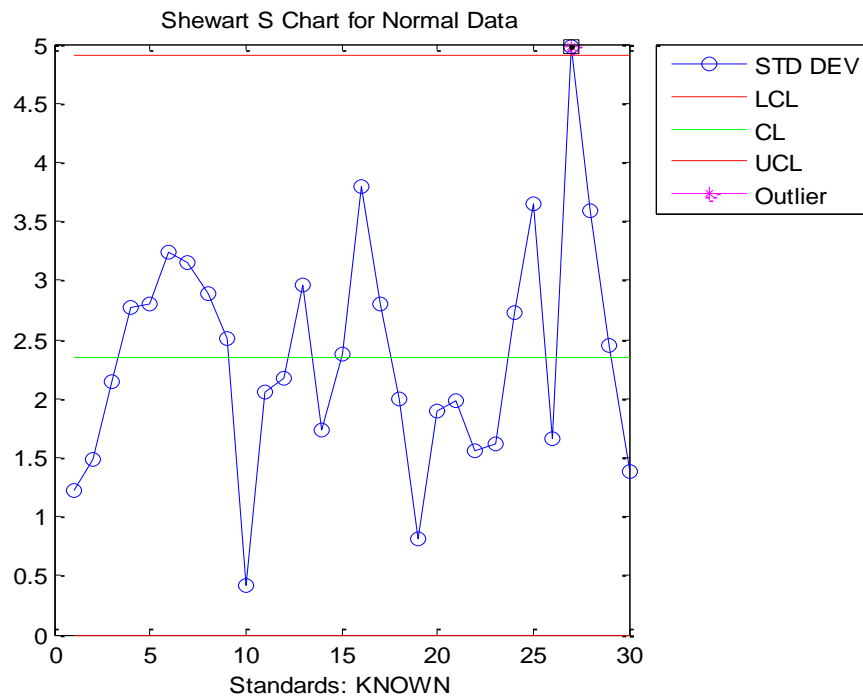


Figure 2.3 Shewart S Control Chart for leaves data given standard $\sigma = 2.5$

Contrary to the R chart, Standard Deviation Chart may indicate some small problems about the process standard deviation. Although their appearances look similar, 27th observation is out of the upper control limit and the first six points of the chart are steadily increasing.

A few days later, Pupil performed a special check to the 27th labeled observation and saw that the tree is quite healthy. This experience confused her lovely mind because this tree had given an out of control signal in the standard deviation chart. That was simply a false alarm. What is the frequency of having this experience? Moreover, she thought that the converse is also possible. Namely, it is possible to miss an ill tree since its measured statistics fall within control limits. She got the feeling that she had new things to learn from Glorious Statistics, which will be whispered to her ears soon. This whisper was going to turn into a scream in time...

When a data point –let’s say in \bar{X} control chart– gives an out of control signal, Pupil decides that the mean length of the leaves in the tree is different from 20 cm

and that the tree is ill. This decision has a false alarm probability of $\alpha = 0.0027$. In fact, each observation is a statistical hypothesis test:

$$H_0: \mu = 20$$

$$H_A: \mu \neq 20$$

In a statistical hypothesis test, two hypotheses and two decisions construct a cross product of 4 cases:

Table 2.2 Terminology and notation of hypothesis testing cases and the corresponding probabilities.

Hypothesis Testing Cases		
	H_0 is true	H_A is true
Do NOT Reject H_0	Confidence Level: $1 - \alpha$	Miss: β
Reject H_0	False Alarm: α	Detection: $1 - \beta$

There is a threshold between false alarm and miss probabilities. Namely, as α decreases, β increases or vice versa. Since 3σ limits are used, mean control chart has a constant probability of false alarm. Reducing probability of miss (and therefore increasing detection probability) can be achieved by two different ways. First one is increasing the sample size and the second is decreasing the standard deviation of the process.

Let's assume, a tree got ill and its mean length of leaves became $\mu_1 = 22$. What is the probability that this illness is detected? The calculations follow:

$$\begin{aligned} \beta &= P(LCL < \bar{X} < UCL; \mu = \mu_1) = P(16.645 < \bar{X} < 23.355; \mu = 22) \\ &= P\left(\frac{16.645 - 22}{2.5/\sqrt{5}} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{23.355 - 22}{2.5/\sqrt{5}}\right) = P(-4.7897 < Z < 1.2119) \end{aligned}$$

$$= \Phi(1.2119) - \Phi(-4.7897) = 0.8871 \quad (2.20)$$

where $\Phi(z)$ is the cumulative standard normal distribution.

When the true mean shifts to $\mu = 22$, the probability to miss that illness is 0.8872. A shift is often measured in standard deviation units. For example, this shift is 0.8σ shift since $\frac{22-20}{2.5}=0.8$. Then, the detection probability of such a shift is:

$$1 - \beta = 1 - 0.8871 = 0.1129 \quad (2.22)$$

The number of samples to get an out of control signal is a random variable, which is called *Run Length*. Given the constant mean value of μ , $R = \text{Run Length}$ is a Geometric Random Variable with parameter $P(\text{Out of control signal})$. Identifying the true processes' "in control" and "out of control" cases with corresponding subscripts, we have:

$$R_0 \sim \text{Geometric}(\alpha)$$

$$E(R_0) = \frac{1}{\alpha} \text{ and } \text{Var}(R_0) = \frac{1-\alpha}{\alpha^2} \quad (2.23)$$

$$R_1 \sim \text{Geometric}(1 - \beta)$$

$$E(R_1) = \frac{1}{1-\beta} \text{ and } \text{Var}(R_1) = \frac{\beta}{(1-\beta)^2} \quad (2.24)$$

(Montgomery, 2009).

The expected value of the Run length is called Average Run Length (ARL). When the process of the mean is in control, we have:

$$ARL_0 = E(R_0) = \frac{1}{0.0027} = 370.37 \quad (2.24)$$

$$\text{Var}(R_0) = \frac{1-0.0027}{0.0027^2} = 136803.84 \quad (2.25)$$

The average run length and variance of run length for a 0.8σ shift when the sample size $n = 5$ is used are:

$$ARL_1 = E(R_1) = \frac{1}{0.1129} = 8.86 \quad (2.26)$$

$$\text{Var}(R_1) = \frac{1-0.1129}{0.1129^2} = 69.73 \quad (2.27)$$

Since α is a constant value, ARL_0 of \bar{X} chart does not depend on n . However, β is a decreasing function of n which results in decreasing values of ARL_1 with increasing n . This means that the more sample is collected, the more accurate information is gained, and in return, the quicker the shift is detected.

Similar calculations show that the detection probability for sample size $n = 20$ increases to 0.7183 and ARL_1 reduces to 1.392. Sample size $n = 50$ has corresponding values of 0.9961 and 1.004 respectively.

Pupil had stormed her brain and improved her statistical ability. She now knows the concept of hypothesis testing, Type-I and Type-II Errors, Random variable, mean, and variance. She was also satisfied with her question: “How frequently can I expect an out of control signal?” She thought that it may be too late to detect an ill tree for her current sample size and she decided to increase her sample size to $n = 20$.

Glorious Statistics taught her how to make a simulation and wanted her to see applicable results of the theory she learnt. She decided to check the mean and standard deviation of the random variable R .

For many applications, using standardized scores improves the computational efficiency. For a realization of a random variable Y_i , its standard score $T_i = \frac{Y_i - \mu_Y}{\sigma_Y}$ shows how far the observed value Y_i is away from its mean μ_Y in standard deviation units. Therefore, T_i has mean zero and standard deviation 1. It is customary to show standardized score with Z_i , but more generally Z is a standard normal random variable. Due to the fact that T does not necessarily follow a normal distribution, I decided this notation to be more appropriate.

Pupil generated $r = 1000$ replications of R_0 for the \bar{X} chart, designed for standard normal T with different sample sizes and calculated the mean for each of the simulated runs for 10 independent streams, each of which is \bar{R}_0 . The final mean is represented as $\overline{\bar{R}_0} = ARL_0$.

Table 2.3 Simulated run lengths with different sample sizes for mean control chart of Normal data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards are at the bottom part of the table.

ARL ₀ for mean given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL ₀	sims	ARL ₀	sims	ARL ₀
388.9910	375.5783	391.2840	372.1857	381.1700	373.3449
372.1630		384.7030		376.7660	
376.4210		347.7520		373.0490	
359.7050		371.3350		374.2000	
353.9880		373.5040		380.9160	
388.4730		345.4720		366.4210	
364.6110		371.6030		366.8730	
400.1050		384.3830		371.8470	
375.6640		381.0000		381.1330	
375.6620		370.8210		361.0740	
14.1363	=stdev	15.1408	=stdev	6.9472	=stdev
		$\mu =$	0.0000		
		$\sigma =$	1.0000		
UCL =	1.3416	UCL =	0.6708	UCL =	0.4243
LCL =	-1.3416	LCL =	-0.6708	LCL =	-0.4243

The left column sims shows the mean run length for each n obtained from each of the $r = 1000$ runs and ARL_0 is the mean of these 10 values. The simulated ARL_0 estimates are not significantly different from the theoretical mean 370.37. However, sims column values lie in a wide range since standard deviation of the R_0 is $\sqrt{136803.84} = 369.87$, and standard deviation for mean of $r = 1000$ runs is $369.87/\sqrt{1000} = 11.70$. It is important to mention that all these calculations are valid for \bar{X} chart and under Gaussian case.

Next, she generated $r = 1000$ replication of R_1 again for the control charts designed for standard normal T , but this time she used a normal random number generator with mean 0.8 and standard deviation 1. The results of the simulation are given in the following table. The simulated ARL_1 values are much closer to the theoretical values this time. Moreover, sims values lie in a narrower range since increase in parameter of R reduces its variance.

Table 2.4 Simulated run lengths with different sample sizes for mean control chart of Normal data and the average run length, when the process is out of control with a 0.8σ shift. Standard deviation of the mean run lengths and the control limits given standards are at the bottom part of the table.

ARL₁ for mean with a shift of 0.8σ given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₁	sims	ARL₁	sims	ARL₁
8.8270	8.8773	1.3970	1.3977	1.0060	1.0053
9.2400		1.4280		1.0060	
8.9610		1.3770		1.0050	
8.6320		1.3740		1.0030	
8.6490		1.3960		1.0010	
8.7120		1.3910		1.0100	
9.0120		1.3890		1.0020	
8.8310		1.3990		1.0070	
8.9330		1.3930		1.0030	
8.9760		1.4330		1.0100	
0.1867	=stdev	0.0192	=stdev	0.0031	=stdev
		$\mu =$	0.0000		
		$\sigma =$	1.0000		
UCL =	-1.3416	UCL =	0.6708	UCL =	0.4243
LCL =	1.3416	LCL =	-0.6708	LCL =	-0.4243
$1 - \beta$	0.1128	$1 - \beta$	0.7183	$1 - \beta$	0.9961

Calculation of ARL for s chart in the same manner will not be true since s does not follow a Gaussian distribution. Hopefully, we can calculate the probability of getting an out of control limit signal using Chi-Square distribution. The random variable $W = \frac{(n-1)s^2}{\sigma^2}$ follows a Chi-Square distribution with degrees of freedom $n - 1$, where s^2 is the sample variance of a Gaussian data with variance σ^2 . If standardized score T is used, the control limits of s chart for $n = 5$ will be:

$$UCL = B_6\sigma = 1.964 * 1 = 1.964 \quad (2.29)$$

$$LCL = B_5\sigma = 0.000 * 1 = 0.000 \quad (2.30)$$

False alarm probability for s chart is calculated as follows:

$$\begin{aligned}
1 - \alpha &= P(LCL < s < UCL; \sigma = 1) = P(0.000^2 < s^2 < 1.964^2; \sigma = 1) \\
&= P\left(\frac{(5 - 1) * 0.000^2}{1^2} < \frac{(n - 1) * s^2}{\sigma^2} < \frac{(5 - 1) * 1.964^2}{1^2}\right) \\
&= P(0.000 < \chi^2 < 15.423) = F(15.423) - F(0.000) = 0.9961 \quad (2.31)
\end{aligned}$$

$$\alpha = 1 - 0.9961 = 0.0039 \quad (2.32)$$

Therefore, Run Length for standard deviation chart has distribution:

$$R_0 \sim \text{Geometric}(\alpha = 0.0039) \quad (2.33)$$

Finally, in control average run length is:

$$ARL_0 = E(R_0) = \frac{1}{0.0039} = 256.42 \quad (2.33)$$

Unlike the distribution of Z used to calculate average run length of \bar{X} chart, the distribution of W used to calculate that of s chart is a function of n . For that reason, average run length of s chart depends on the sample size n . The corresponding values of ARL_0 for $n = 20$ and $n = 50$ are 357.14 and 367.06, respectively. The calculations are similar.

Following table shows the simulation results of ARL_0 . Simulation parameters are the same as the previous one and Run Lengths are calculated for s chart. The results are similar to that of \bar{X} chart in that simulated ARL_0 values are close to theoretical ones and there exist a variation within sims.

Table 2.5 Simulated run lengths with different sample sizes for standard deviation control chart of Normal Data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards are at the bottom part of the table.

ARL ₀ for standard deviation given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL ₀	sims	ARL ₀	sims	ARL ₀
252.1050	258.2923	372.5630	361.4899	381.1980	365.4255
258.6720		364.5900		344.9180	
252.1900		370.1940		367.1300	
260.7010		345.9530		360.7810	
262.0200		382.1870		361.1380	
254.5820		352.7760		380.3800	
248.5270		348.2580		378.1370	
256.9360		360.7160		347.4870	
258.8570		340.1970		365.6360	
278.3330		377.4650		367.4500	
8.2211	=stdev	14.2905	=stdev	12.5757	=stdev
		σ =	1.0000		
UCL =	1.9636	UCL =	1.4703	UCL =	1.2972
LCL =	0.0000	LCL =	0.5036	LCL =	0.6926

The change in ARL₀ values of *s chart* with respect to sample size may cause some practical problems in interpreting the chart results. It is a good idea to develop a chart that has the same ARL₀ value with \bar{X} *chart*, which is the constant 370.37. It is easy to develop such a chart using χ^2 statistics that follows the same logic of \bar{X} *chart* development.

Considering the statement “The probability is $1 - \alpha$ that χ^2 lies within the interval $(\chi_{1-\alpha/2}^2; \chi_{\alpha/2}^2)$ ” will follow that:

$$\chi_{1-\alpha/2}^2 < \chi^2 < \chi_{\alpha/2}^2$$

$$\chi_{1-\alpha/2}^2 < \frac{(n-1) * s^2}{\sigma^2} < \chi_{\alpha/2}^2$$

$$\frac{\chi_{1-\alpha/2}^2 * \sigma^2}{n-1} < s^2 < \frac{\chi_{\alpha/2}^2 * \sigma^2}{n-1} \quad (2.35)$$

Therefore, the control limits of the s^2 chart are:

$$UCL = \frac{\chi_{\alpha/2}^2}{n-1} \sigma^2 \quad (2.36)$$

$$LCL = \frac{\chi_{1-\alpha/2}^2}{n-1} \sigma^2 \quad (2.37)$$

(Montgomery, 2009).

Using confidence level of 0.9973 and leaves data of Table (2.1), we have the following confidence limits:

$$UCL = \frac{17.8004}{4} * 2.5^2 = 27.810 \quad (2.38)$$

$$LCL = \frac{0.1058}{4} * 2.5^2 = 0.165 \quad (2.39)$$

Following figure is the control chart for variance. Its appearance is exactly the same as the standard deviation chart but this time, no data points are out of control limits. The reason is that, false alarm probability of variance chart is lower due to the design for a higher ARL_0 value.

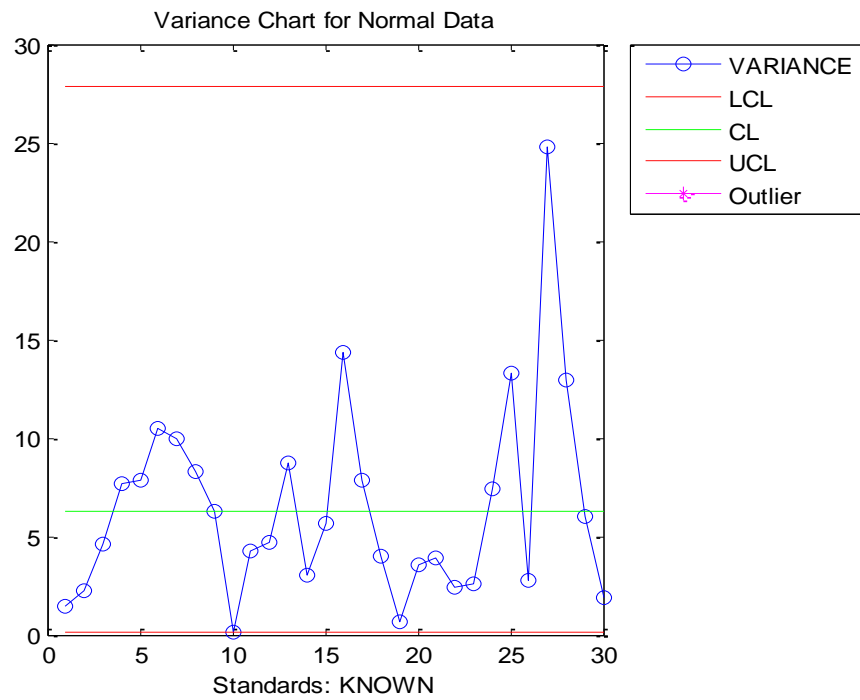


Figure 2.4 Sample Variance Control Chart for leaves data given standard $\sigma = 2.5$

Following table is the ARL_0 simulation for variance chart. Simulation parameters are the same as the previous ones. Results are very similar to that of \bar{X} chart because they have the same parameter α for the run length random variable R_0 .

Table 2.6 Simulated run lengths with different sample sizes for variance control chart of Normal Data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards are at the bottom part of the table.

ARL₀ for variance given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
367.8530	372.7727	396.8810	376.2946	388.8490	372.2155
349.4350		388.7400		352.5830	
374.3600		366.0620		373.1640	
376.8830		376.4790		364.4410	
395.1820		392.1050		358.9310	
388.9330		364.4060		392.8820	
376.9460		365.0390		384.8150	
358.0790		366.2990		352.8140	
345.6640		354.7590		371.5570	
394.3920		392.1760		382.1190	
17.5980	=stdev	14.9716	=stdev	14.7600	=stdev
		$\sigma^2 =$	1.0000		
UCL =	4.4501	UCL =	2.2564	UCL =	1.7158
LCL =	0.0264	LCL =	0.2969	LCL =	0.5007

Pupil's introductory education on Statistical Quality Control was almost completed. The thing she wonders was how reliable the mean and the standard deviation parameters of the leaves of quassia amara are. The grand mean of the data $\bar{\bar{X}} = 19.514$ and mean of the sample standard deviations $\bar{s} = 2.36$ were quite close to the standard values of mean $\mu = 20$ and $\sigma = 2.5$. However, she wanted both to be sure about the accuracy and to complete her basic knowledge.

Glorious Statistics was so generous that any kind of information improves the inference with an honest study. Consequently, he contains the scope for those who has not standardized values.

Since $E(\bar{X}) = \mu$ and $E(s) = c_4\sigma$, their mean counterpart $E(\bar{\bar{X}}) = \mu$ and $E(\bar{s}) = c_4\sigma$ are also true. This fact make $\hat{\mu} = \bar{\bar{X}}$ and $\hat{\sigma} = \frac{\bar{s}}{c_4}$ unbiased estimators of μ and σ respectively. Moreover, \bar{X} and s are complete statistics for the data set when the data

has Gaussian distribution. Then, these estimators can be replaced with the parameters in the previous interval:

$$\left(\bar{X} - 3\frac{\bar{s}}{c_4\sqrt{n}} ; \bar{X} + 3\frac{\bar{s}}{c_4\sqrt{n}} \right) \quad (2.40)$$

Letting the constant $A_3 = \frac{3}{c_4\sqrt{n}}$, the upper and lower control limits for \bar{X} chart are obtained:

$$UCL = \bar{X} + A_3\bar{s} \quad (2.41)$$

$$LCL = \bar{X} - A_3\bar{s} \quad (2.42)$$

(Montgomery, 2009).

Control limit calculations for s chart are similarly as follows. If σ is replaced with $\frac{\bar{s}}{c_4}$, the corresponding interval is:

$$\left(c_4\frac{\bar{s}}{c_4} + 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2} ; c_4\frac{\bar{s}}{c_4} - 3\frac{\bar{s}}{c_4}\sqrt{1-c_4^2} \right) \quad (2.43)$$

The following constants are defined to reduce the formulas:

$$B_4 = 1 + \frac{3}{c_4}\sqrt{1-c_4^2} \text{ and } B_3 = 1 - \frac{3}{c_4}\sqrt{1-c_4^2} \quad (2.44)$$

Finally, the control limits of s chart becomes:

$$UCL = B_4\bar{s} \quad (2.45)$$

$$LCL = B_3\bar{s} \quad (2.46)$$

(Montgomery, 2009).

Control chart constants for $n = 5$ are:

$$A_3 = 1.427 \quad ; \quad B_4 = 2.089 \quad ; \quad B_3 = 0 \quad (2.47)$$

The corresponding control limits for \bar{X} and s chart are as follows:

\bar{X} chart:

$$UCL = \bar{\bar{X}} + A_3\bar{s} = 19.514 + 1.427 * 2.360 = 22.882 \quad (2.48)$$

$$LCL = \bar{\bar{X}} - A_3\bar{s} = 19.514 - 1.427 * 2.360 = 16.146 \quad (2.49)$$

s chart:

$$UCL = B_4\sigma = 2.089 * 2.360 = 4.930 \quad (2.50)$$

$$LCL = B_3\sigma = 0.000 * 2.360 = 0.000 \quad (2.51)$$

The statistics of leaves data are exactly the same for “Standards: KNOWN” and “Standards: UNKNOWN” cases, and only the limits change a little, hence the figures for the latter is not required. However, it is necessary to simulate ARL_0 values because change in the control limits will cause a change in false alarm probabilities.

The following tables are obtained by a two stage procedure, as is the case in practice, when quality standards are not known. In the first stage, $n = 100$ standard normal numbers are generated and control limits are calculated. In the second stage, $r = 1000$ runs of ARL_0 are simulated (using the same random stream as in previous simulations) 10 times and their means are calculated as done previously.

It is absolutely obvious that ARL_0 values decreases significantly (except for the case of standard deviation for $n = 5$). The performances of the control charts are

quite sensitive to the control limits. If standards are not known in a process, they should be estimated in great care.

Table 2.7 Simulated run lengths with different sample sizes for mean control chart of Normal Data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits for the case no standards given, are at the bottom part of the table.

ARL₀ for mean without standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
343.1790	331.3637	275.6170	265.8428	304.8720	306.5488
320.6540		272.4350		299.5170	
327.2340		248.4530		302.8230	
328.0760		265.3820		311.8030	
311.9450		266.1180		309.7280	
341.1650		252.7180		304.2810	
328.6130		263.5020		303.0720	
338.4370		261.0120		311.7310	
337.8700		279.0680		312.5540	
336.4640		274.1230		305.1070	
9.8953	=stdev	9.9232	=stdev	4.5470	=stdev
$\bar{\bar{X}} =$	-0.0936	$\bar{\bar{X}} =$	-0.0491	$\bar{\bar{X}} =$	-0.0231
$\bar{s} =$	0.9482	$\bar{s} =$	0.9734	$\bar{s} =$	0.9886
UCL =	1.2598	UCL =	0.6125	UCL =	0.3985
LCL =	-1.4469	LCL =	-0.7108	LCL =	-0.4447

Table 2.8 Simulated run lengths with different sample sizes for standard deviation control chart of Normal Data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits for the case no standards given, are at the bottom part of the table.

ARL₀ for standard deviation without standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
280.5900	290.1759	289.3430	273.8020	321.5970	316.3736
288.4670		275.8650		304.4700	
273.7430		277.4160		310.4810	
297.0480		271.8470		319.3990	
297.8720		281.9100		323.6640	
291.8810		260.8090		328.1490	
283.8940		260.8450		321.3160	
291.9060		268.5070		297.4140	
289.6740		265.1460		321.1430	
306.6840		286.3320		316.1030	
9.3656	=stdev	10.1383	=stdev	9.4949	=stdev
$\bar{s} =$	0.9377	$\bar{s} =$	0.9734	$\bar{s} =$	0.9886
UCL =	1.9808	UCL =	1.4501	UCL =	1.2890
LCL =	0.0000	LCL =	0.4966	LCL =	0.6883

2.2 Performance of Dispersion Charts for Non-Gaussian Data

Pupil was suffering from false alarms for control charts, especially about the standard deviation chart. It was the case parallel to the 27th observation of her first data. She learnt that dispersion control is more important than location control. For example, if the mean is out of control for a production process, this may mean that the machines set are wrong and should be corrected. However, if the standard deviation of the process is out of control, the reason may be that the machines are old, or cannot produce within the specified limits. Another example goes with human nature. If an housewife is unhappy for a period of time, relatively simple acts of her husband can turn her back to a usual life of productivity. However, if her mind goes back and front between happiness and sadness frequently, a clinical depression can be suspected.

Pupil was feeling that she should learn some new concepts but she couldn't get a start for a period of time. Then, she suspected from the heart of the assumptions: Gaussian distribution. It was the heart because all of the control formulas for charts had developed assuming that data has a Gaussian distribution. What if it is not? Are there any alternative formulas, statistics, or methods for control?

Glorious Statistics cooled her down, recalling the fundamentals of wisdom. There are surely many estimators, distributions and patterns, but inference is a difficult art with its lower stairs and slower steps. He told Pupil to understand the logic that underlies the false alarm signals and execute the performance of her relevant chart trying some other distributions.

Pupil was relieved and satisfied. She understood that a calm mind is more likely to produce creative ideas. There should be some unexpectedly high or low values of the data that increases the standard deviation of the data and yield false alarm signals. There should be some characteristics of other distributions that make this more possible than that of Gaussian.

She finally met the definition of heavy tail. A heavy-tailed distribution has higher probabilities than Gaussian distribution to observe values from the part that is far away from its median. A measure for "far away" can be outside the middle 50% of the distribution.

To make the results comparable with Gaussian case, she decided to study some symmetric heavy tailed distributions. Logistic Distribution, Laplace (Double Exponential) Distribution, and Cauchy Distribution are three such distributions which form a good set to study because their heaviness of tail are different from each other.

A proxy for heaviness of a distribution's tail can be its kurtosis, which is a measure of its "peakedness." The relationship is that the sharper the distribution has

peak, the narrower its middle 50% interval is, and in return the more a “far away” value is probable. Therefore, a distribution having a high kurtosis has a sharper peak and longer, fatter tails or vice versa. Moreover, higher kurtosis means that outlier values contribute to its variance more than modestly sized observations.

If μ_4 is the fourth moment about the mean of the distribution and its standard deviation is σ , kurtosis is defined as:

$$\beta_2 = \frac{\mu_4}{\sigma^4} \quad (2.52)$$

(De Carlo, 1997).

Gaussian distribution has kurtosis 3 and it is customary to measure the kurtosis of a distribution (and in parallel, heaviness of tail) with reference to that of Gaussian. Subtracting 3 from the kurtosis give a parameter value, which is called “Excess Kurtosis”:

$$\gamma_2 = \frac{\mu_4}{\sigma^4} - 3 \quad (2.53)$$

(De Carlo, 1997).

Obviously, positive excess kurtosis shows that the distribution has a more acute peak and fatter tails than Gaussian distribution and these distributions are called “leptokurtic” (lepto means slim). Likewise, distributions having negative excess kurtosis aim to have a lower and wider peak around their mean and they are called “platykurtic” (platy means wide).

The distributions Pupil will study are all leptokurtic and Logistic Distribution has excess kurtosis 1.2 whereas Laplace has that of 3. Namely, Laplace has heavier tails than Logistic. Cauchy distribution has the heaviest tail among them, but since its moments are undefined, it has no kurtosis value (De Carlo, 1997).

The study must begin with analyzing false alarm probabilities because decrease in false alarm probability will naturally increase the detection probability and this will be misleading. Secondly, it is also possible to obtain analytical calculations for this purpose, but only ARL_0 simulations will be presented since analytical results will be unobtainable for future parts of the study.

Logistic distribution has the probability density function (pdf):

$$f(x; \mu, s) = \frac{e^{-(x-\mu)/s}}{s(1+e^{-(x-\mu)/s})^2}, \quad -\infty < x < \infty \quad (2.54)$$

where μ is location parameter and $s > 0$ is scale parameter.

Mean and variance of logistic distribution is:

$$E(X) = \mu \quad ; \quad Var(X) = \frac{\pi^2}{3} s^2 \quad (2.55)$$

The cumulative distribution function (cdf) is:

$$F(x; \mu, s) = \int_{-\infty}^x f(w; \mu, s) dw = \frac{1}{(1+e^{-(x-\mu)/s})} \quad (2.56)$$

(Walck, 2007).

The graphs of Logistic pdf for some values of μ and s are shown in Figure 2.5.

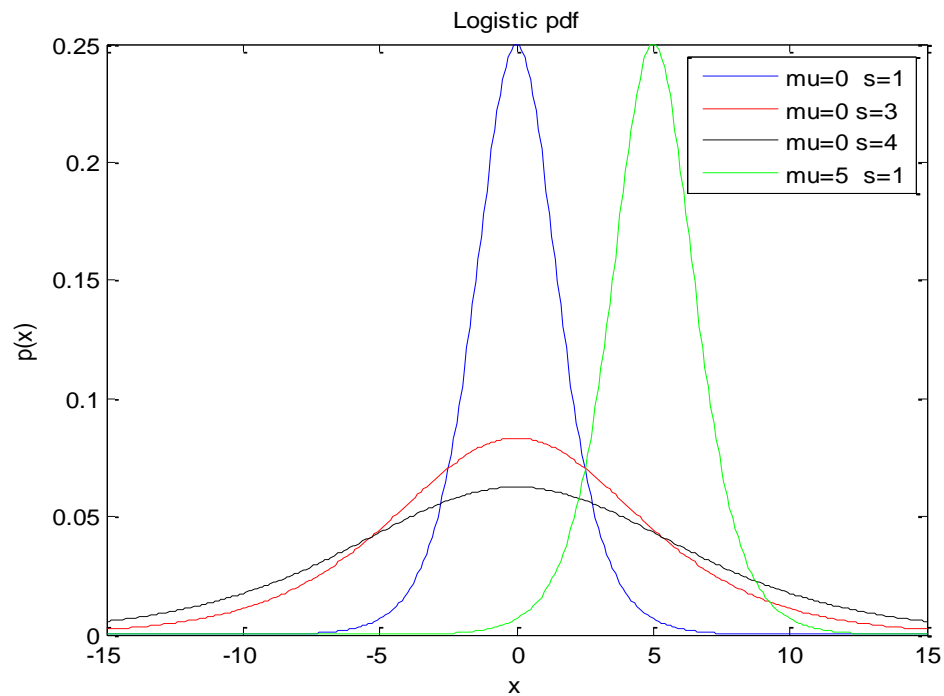


Figure 2.5 Graphs of Logistic pdfs with some specified location and scale parameter values

In the simulations that will be run, standardized random variables will be used as before. In order to run a simulation with Logistic Random Variable, its “Random Number Generator” is required. In general, a random number generator is a mapping that transforms a random number $U_i \sim Uniform(0; 1)$ to a random number of the specified distribution.

Let, $Y_i \sim Logistic(\mu = 0; s = 1)$. Then, Y has cdf:

$$F(y) = \frac{1}{(1+e^{-y})} \quad (2.57)$$

Since $F_Y(u) = u$, the inverse function $Y = F^{-1}(U)$ will map the uniform random number U_i to the Logistic random number Y_i . This is called “Inverse Transformation Technique” (Banks, Carson II, Nelson, & Nicol, 2005). The calculations are as follows:

$$U = F(Y) = \frac{1}{(1 + e^{-Y})}$$

$$1 - U = Ue^{-Y}$$

$$-Y = \ln\left(\frac{1-U}{U}\right)$$

$$Y = F^{-1}(U) = \ln\left(\frac{U}{1-U}\right) \quad (2.58)$$

But we have $E(Y) = 0$ and $Var(Y) = \frac{\pi^2}{3}$. To obtain a standardized random variable, $T = \frac{\sqrt{3}}{\pi}Y$ is defined and finally T is a standard Logistic random variable with generator:

$$T = \frac{\sqrt{3}}{\pi} \ln\left(\frac{U}{1-U}\right) \quad (2.59)$$

The code for MATLAB function “generator.m” that performs random number mappings to Logistic, Laplace, and Cauchy distributions is shown in Appendix-4. The simulations are done by the function “runlength_intro.m.” Moreover, all of the MATLAB functions that generate the tables and figures of the thesis are also given in Appendix-4.

The following table shows the ARL_0 simulation for the variance control chart of Table 2.6, but this time simulation random variable T follows a standard Logistic distribution, not a standard Gaussian distribution. There are two important facts to mention for the variance control chart.

First of all, there is a dramatic decrease in ARL_0 performances when the data is Logistic. The variance chart is quite sensitive, in other words, non-robust to deviations in the distribution of the data. Secondly, ARL_0 does not converge to its nominal value of 370.37 calculated for Gaussian case. Conversely, values decrease

as n gets higher. It might be interpreted that as a result of higher n values, there exist more extreme values in the data and variance increases.

Table 2.9 Simulated run lengths with different sample sizes for variance control chart of Logistic Data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards, are at the bottom part of the table.

ARL₀ for variance given standards (Logistic Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
106.3150	104.6312	68.1160	68.4798	62.7940	61.4027
102.4110		65.5090		59.7870	
107.3710		69.4650		59.6100	
101.7360		68.0230		61.2240	
101.2060		66.9130		65.0210	
105.6810		66.6580		64.6460	
102.8520		74.1380		60.7360	
108.1370		68.8120		63.5340	
107.0870		70.1360		55.7440	
103.5160		67.0280		60.9310	
2.5651	=stdev	2.4241	=stdev	2.7615	=stdev
UCL =	4.4501	$\sigma^2 =$	1.0000	UCL =	1.7158
LCL =	0.0264	UCL =	2.2564	LCL =	0.5007
		LCL =	0.2969		

Having seen the disappointing results for ARL₀ performances of Logistic case, one may expect that things will go worse for Laplace and the worst for Cauchy distributions because their tails are heavier. As mentioned before, Laplace distribution has excess kurtosis 3.0 which is much higher than 1.2 of Logistic.

Laplace (Double Exponential) distribution has the probability density function (pdf):

$$f(x; \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right), \quad -\infty < x < \infty \quad (2.60)$$

where μ is location parameter and $b > 0$ is scale parameter. For the special case $\mu = 0$ and $b = 1$, the positive half-line is exactly an exponential distribution scaled by 0.5, and negative one is its symmetric. That's why; "Laplace distribution" is also called as "Double Exponential distribution."

Mean and variance of Laplace distribution is:

$$E(X) = \mu \quad ; \quad Var(X) = 2b^2 \quad (2.61)$$

The cumulative distribution function (cdf) is:

$$F(x; \mu, s) = \int_{-\infty}^x f(w; \mu, s)dw = \begin{cases} \frac{1}{2} \exp\left(-\frac{\mu-x}{b}\right) & \text{if } x < \mu \\ 1 - \frac{1}{2} \exp\left(-\frac{x-\mu}{b}\right) & \text{if } x \geq \mu \end{cases} \quad (2.62)$$

(Walck, 2007).

The graphs of Laplace pdf for some values of μ and b are shown in the following figure:

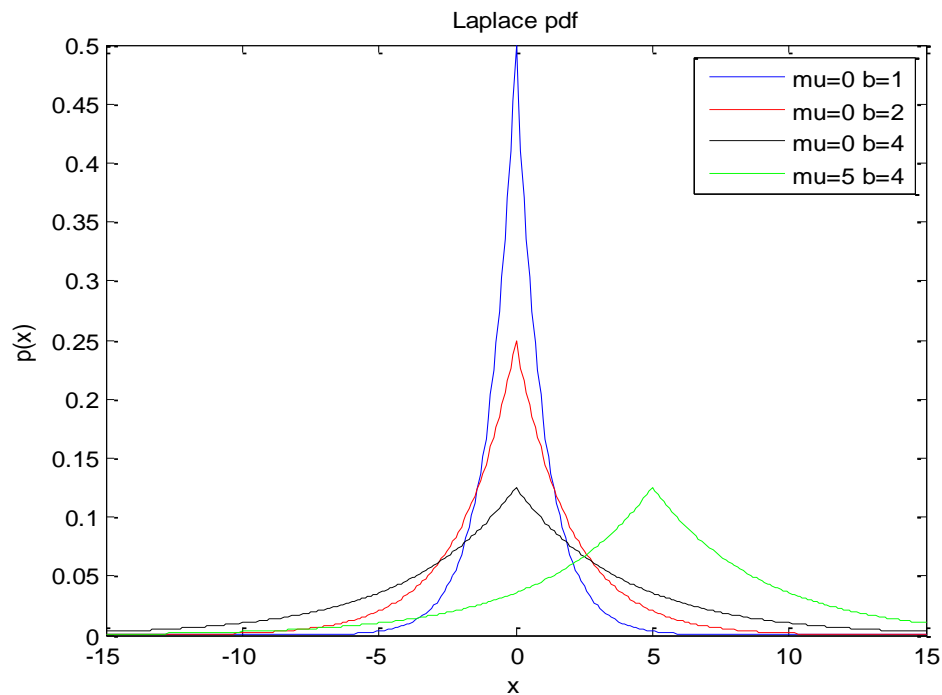


Figure 2.6 Graphs of Laplace pdfs with some specified location and scale parameter values

To generate standardized Laplace random variable T , exponential random variable should be introduced first. An exponential random variable Y with rate λ has the pdf and cdf:

$$f_Y(y) = \lambda e^{-\lambda y} \quad \text{and} \quad F_Y(y) = 1 - e^{-\lambda y} \quad (2.63)$$

Letting $1 - U \sim \text{Uniform}(0; 1)$, the inverse function $Y = F^{-1}(U)$ is:

$$Y = -\frac{1}{\lambda} \ln(U) \quad (2.64)$$

Now, if we consider two independent exponential random variables Y_1, Y_2 with $\lambda = 1/2$, their joint pdf is:

$$f(y_1, y_2) = f_1(y_1)f_2(y_2) = \frac{1}{4} e^{-\frac{y_1+y_2}{2}} \quad (2.65)$$

Let, $X_1 = \frac{Y_1 - Y_2}{2}$ and $X_2 = Y_2$. Then, $Y_1 = w_1(X_1, X_2) = 2X_1 + X_2$ and $Y_2 = w_2(X_1, X_2) = X_2$. The Jacobian of the transformation is:

$$J = \begin{vmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_1}{\partial X_2} \\ \frac{\partial Y_2}{\partial X_1} & \frac{\partial Y_2}{\partial X_2} \end{vmatrix} = \begin{vmatrix} 2 & 1 \\ 0 & 1 \end{vmatrix} = 2 \quad (2.66)$$

The joint pdf of X_1 and X_2 is:

$$g(x_1, x_2) = f(w_1(x_1, x_2), w_2(x_1, x_2))|J| = \frac{1}{2} e^{-x_1 - x_2} \quad (2.67)$$

Thus, pdf of x_1 is given by:

$$g_1(x_1) = \int g(x_1, x_2) dx_2 = \frac{1}{2} e^{-|x_1|} \quad (2.68)$$

Then, $X_1 \sim \text{Laplace}(\mu = 0; b = 1)$. Since $\text{Var}(X_1) = 2$, Random number generator for standard Laplace random variable is:

$$T = \frac{1}{\sqrt{2}} X_1 = \frac{Y_1 - Y_2}{2\sqrt{2}} = \frac{1}{2\sqrt{2}} (-2 \ln(U_1) + 2 \ln(U_2)) = \frac{1}{\sqrt{2}} \ln\left(\frac{U_2}{U_1}\right) \quad (2.69)$$

The following table shows the ARL_0 simulation for the variance control chart where T follows a standard Laplace distribution. Just like for the logistic case, ARL_0 values get smaller for increasing sample size values. Moreover, ARL_0 values are about one third of Logistic case since Laplace distribution has heavier tails.

Table 2.10 Simulated run lengths with different sample sizes for variance control chart of Laplace data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards, are at the bottom part of the table.

ARL₀ for variance given standards (Laplace Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
49.6360	48.0299	25.2540	23.9981	20.6820	20.4096
48.5240		24.0990		20.1790	
50.6610		23.7670		21.1100	
47.2660		24.0490		20.0790	
47.6370		23.0730		20.0550	
46.5740		23.3520		21.0220	
47.5790		24.1170		19.6310	
48.1810		24.3020		20.4900	
47.7220		23.9610		20.1620	
46.5190		24.0070		20.6860	
1.2999	=stdev	0.5797	=stdev	0.4689	=stdev
		σ² =	1.0000		
UCL =	4.4501	UCL =	2.2564	UCL =	1.7158
LCL =	0.0264	LCL =	0.2969	LCL =	0.5007

The final distribution that is going to be studied is *Cauchy distribution*, with pdf:

$$f(x; \mu, \gamma) = \frac{1}{\pi} \left[\frac{\gamma}{(x-\mu)^2 + \gamma^2} \right], \quad -\infty < x < \infty \quad (2.70)$$

where μ is location parameter and $\gamma > 0$ is scale parameter.

The cumulative distribution is:

$$F(x; \mu, \gamma) = \int_{-\infty}^x f(w; \mu, \gamma) dw = \frac{1}{\pi} \arctan \left(\frac{x-\mu}{\gamma} \right) + \frac{1}{2} \quad (2.71)$$

(Walck, 2007).

The graphs of Cauchy pdf for some values of μ and γ are shown in the following figure:

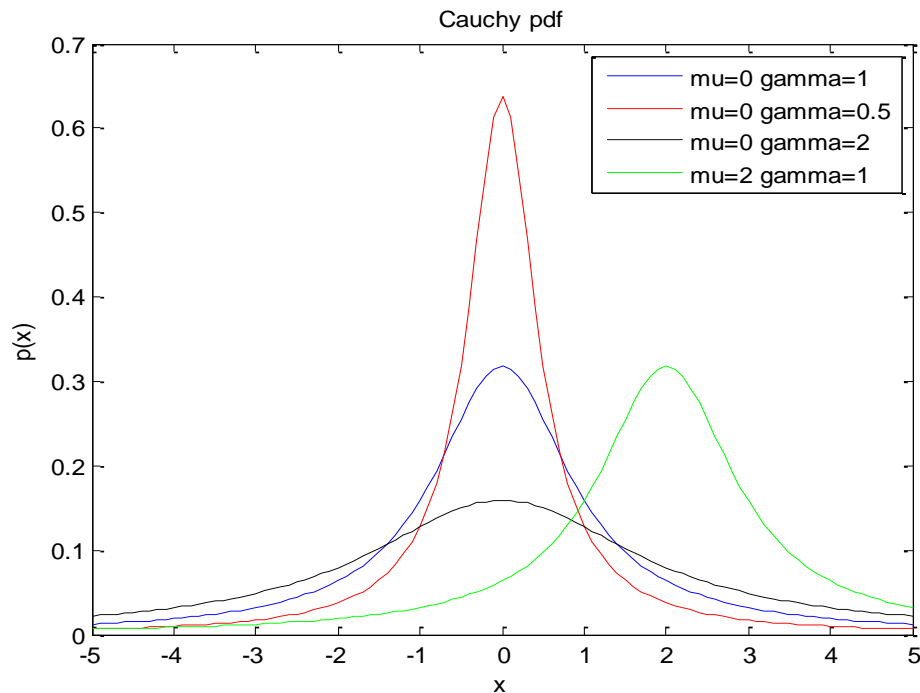


Figure 2.7 Graphs of Cauchy pdfs with some specified location and scale parameter values

In order to simulate Cauchy random variables, it is necessary to show that Cauchy random variable is the ratio of independent Gaussian random variables. Let, random variables Y_1, Y_2 are standard Gaussian. We want to find the distribution of $X_1 = Y_1/Y_2$ and we let $X_2 = Y_2$. The joint pdf of Y_1 and Y_2 is:

$$f(y_1, y_2) = f_1(y_1)f_2(y_2) = \frac{1}{2\pi} e^{-\frac{y_1^2 + y_2^2}{2}} \quad (2.72)$$

We have, $Y_1 = w_1(X_1, X_2) = X_1 X_2$ and $Y_2 = w_2(X_1, X_2) = X_2$ and the Jacobian of the transformation is:

$$J = \begin{vmatrix} \frac{\partial Y_1}{\partial X_1} & \frac{\partial Y_1}{\partial X_2} \\ \frac{\partial Y_2}{\partial X_1} & \frac{\partial Y_2}{\partial X_2} \end{vmatrix} = \begin{vmatrix} X_2 & X_1 \\ 0 & 1 \end{vmatrix} = X_2 \quad (2.73)$$

The joint pdf of X_1 and X_2 is:

$$g(x_1, x_2) = f(w_1(x_1, x_2), w_2(x_1, x_2))|J| = \frac{1}{2\pi} e^{-\frac{(1+y_1^2)+y_2^2}{2}} |y_2| \quad (2.74)$$

Thus, pdf of x_1 is given by:

$$g_1(x_1) = \int g(x_1, x_2) dx_2 = \frac{1}{\pi(1+y_1^2)} \quad (2.75)$$

Then, $X_1 \sim \text{Cauchy}(\mu = 0; \gamma = 1)$. Random number generator for standard Cauchy random variable is:

$$T = X_1 = \frac{Y_1}{Y_2} \quad (2.76)$$

Given uniform random variables U_1 and U_2 , Standard Normal random variables can be generated by the following transformation, which is suggested by Box and Muller:

$$Y_1 = \text{Cos}(2\pi U_1) \sqrt{-2 \ln U_2} \quad (2.77)$$

$$Y_2 = \text{Sin}(2\pi U_1) \sqrt{-2 \ln U_2} \quad (2.78)$$

(Hogg & Craig, 1995)

since this transformation results in the joint pdf of independent standard Gaussian random variables $f(y_1, y_2)$.

The following table shows the ARL_0 simulation for the variance control chart where T follows a standard Cauchy distribution. ARL_0 values are too small to be acceptable for any practical study.

Besides having very heavy tails, such low values can be explained considering the denominator term Y_2 of the Cauchy random variable T . For a few values of Y_2 in the data that are close to zero, corresponding T values attain a very high (absolute) value, and in return, sample variance will be out of control limits too frequently.

Table 2.11 Simulated run lengths with different sample sizes for variance control chart of Cauchy Data and the average run length, when the process is in control. Standard deviation of the run lengths and the control limits given standards, are at the bottom part of the table.

ARL₀ for variance given standards (Cauchy Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
1.7110	1.7421	1.0320	1.0287	1.0001	1.0001
1.7690		1.0330		1.0000	
1.6980		1.0280		1.0000	
1.7340		1.0240		1.0000	
1.7060		1.0350		1.0000	
1.7780		1.0240		1.0000	
1.7480		1.0210		1.0000	
1.7690		1.0350		1.0010	
1.7280		1.0290		1.0000	
1.7800		1.0260		1.0000	
0.0311	=stdev	0.0049	=stdev	0.0003	=stdev
UCL =	4,4501	UCL =	1.0000	UCL =	1,7158
LCL =	0,02644	LCL =	2,2564	LCL =	0.5007
		LCL =	0.2969		

Pupil was so upset for the results of her analysis. There was no way to be satisfied with the validity of the Gaussian assumption and when the data is not Gaussian, dispersion control chart suffers too much from false alarms and this caused too much time to be lost in control. She has to find some new ideas to develop a control chart which was resistant to changes in distribution of the data.

Pupil had graduated from Botanic Department and Statistics was a new area for her. For a plenty of time, she couldn't discover new aspects of Glorious Statistics to improve her studies. During this period, she lost considerable time and effort due to false alarm out of control signals.

During one of her morning walks, she saw a young boy who was examining a quassia amara tree carefully. He said a calm hello to her and introduced himself. Rookie was an Agricultural Technician and he was a fan of nature like Pupil. Pupil told him about her studies on the health of quassia amara in detail.

Rookie had studied "Estimation of Agricultural Productivity" for his undergraduate thesis and used many statistical techniques there. It was not quite apparent whether her study itself or her charming beauty got him interested, but they decided to continue searching together. In fact, who cares?

CHAPTER THREE

ROBUST ESTIMATORS AND QUALITY APPLICATIONS

Whether directly or indirectly, every statement is based on an “Assumption,” that is assumed to be true unquestionably. A scientific study does so by defining a “Model” with some specified parameters or characteristics, and by continuing the search using its own terminology of discipline.

How can we rely on the truth of assumptions, or more specifically, on the presupposed Model? Pure mathematical studies can give intuitive reasoning for this required reliance, but when it comes to an area in Applied Science, it may be more difficult to find such intuitions that are more than a belief.

Glorious Statistics enables us to “Check Assumptions” and provides a bridge to go back and forth between the presupposed data models and the real life data. Moreover -with some acceptable costs paid for desired properties- the concept of “Robust Statistics” enables the scientist to feel comfortable because these statistics are resistant to the changes in data distributions.

To open the discussion in detail, consider the population parameter θ we want to estimate using the estimator $\hat{\theta}$. θ can be a location parameter such as population mean μ , or a scale parameter such as population variance σ^2 .

In general, an estimator (or statistics) $\hat{\theta}$ is any mapping from the sample data to the real line, but some estimators are better than the others in some sense. Now, the question is, “What is a good estimator?” or “How can we understand its goodness?” Such a terminological distinction as “Classical Estimation” and “Robust Estimation” may be helpful to introduce the concept since the criteria of being good will refer to different intuition for each case.

This part of the story will begin with some properties of an estimator in classical sense, such as unbiasedness and efficiency. Next, it comes to the important concepts of robust estimators, which are relative efficiency, breakdown point, influence function, and gross error sensitivity. Having understood the basics of background, young enthusiasts of Glorious Statistics will search for “Robust Control Charts”...

3.1 “Classical versus Robust” Estimation of Location

In a part of his undergraduate thesis study, Rookie had studied food demand estimation and got the logic. It was a good starting point to train Pupil. Before anything else, we want an estimator to estimate the true population parameter on the average. Then, we want its values to show a small variation between observed samples. The former is the unbiasedness property $E(\hat{\theta}) = \theta$, and the latter is the efficiency, which refers to a small variance.

The term “small variance” needs a reference point here. Hopefully, smart statisticians Cramer and Rao brothers found the minimum variance that an unbiased estimator can take using Fisher Information. Fisher is another smart one. If there exists an estimator $\hat{\theta}$ of θ whose variance $Var(\hat{\theta})$ is equal to “Cramer Rao Lower Bound,” it is called “Efficient Estimator.” Another unbiased estimator $\hat{\theta}^*$ is said to have relative efficiency (RE) measured by:

$$RE(\hat{\theta}^*) = 100 * \frac{var(\hat{\theta})}{var(\hat{\theta}^*)} \quad (3.1)$$

(Hogg & Craig, 1995)

As in our traditional path, we first assume that data follows a Gaussian distribution and then we will analyze the case “What if it is not?” So, at first, let the random data of size n is: $X_i \sim Normal(\mu; \sigma^2)$. The Fisher Information of a parameter θ obtained from a single observation $X = x$ is defined as:

$$I(\theta) = - \int_{-\infty}^{\infty} \frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} f(x; \theta) dx = -E \left[\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} \right] \quad (3.2)$$

(Hogg & Craig, 1995)

Then, Fisher information of Gaussian distribution's mean $\theta = \mu$ is

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\theta)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

$$\ln f(x; \theta) = -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{(x-\theta)^2}{2\sigma^2}$$

$$\frac{\partial \ln f(x; \theta)}{\partial \theta} = \frac{x-\theta}{\sigma^2}$$

$$\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} = -\frac{1}{\sigma^2}$$

$$I(\theta) = -E \left[\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} \right] = -E \left[-\frac{1}{\sigma^2} \right] = \frac{1}{\sigma^2} \quad (3.3)$$

The joint probability distribution of the random vector $\mathbf{X} = (X_1, X_2, \dots, X_n)$ is called the likelihood function, which is due to independence:

$$L(\mathbf{x}; \theta) = f(x_1; \theta) f(x_2; \theta) \dots f(x_n; \theta) \quad (3.4)$$

Replacing the pdf $f(x; \theta)$ with the likelihood function $L(\mathbf{x}; \theta)$, we have the Fisher Information obtained from the sample data, which is shown by $I_n(\theta)$, and it is easy to show that $I_n(\theta) = nI(\theta)$. Cramer Rao Inequality, which serves a lower bound for an unbiased estimator's variance is:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)} \quad (3.5)$$

(Hogg & Craig, 1995)

The minimum variance of an unbiased estimator $\theta = \mu$ is $\frac{1}{I_n(\theta)} = \frac{\sigma^2}{n}$, which is the variance of the sample mean: \bar{X} . Therefore, $\hat{\theta} = \bar{X}$ is the “Uniformly Minimum Variance Unbiased Estimator” (UMVUE) of $\theta = \mu$.

A second estimator can be used for the Gaussian mean $\theta = \mu$ as, $\hat{\theta}^* = med_i(x_i)$, which is the sample median. Due to the symmetry of Gaussian distribution, $\hat{\theta}^*$ is also an unbiased estimator but its variance is more than that of sample mean. In particular, their variance ratio is about $\pi/2$ for large values of sample size.

The limit of Relative Efficiency as $n \rightarrow \infty$ is called “Asymptotic Relative Efficiency (ARE)” and if the limit of variance for an estimator is equal to Cramer Rao lower bound, the estimator is called “Asymptotically Efficient.” The asymptotic relative efficiency of sample median is:

$$ARE(\hat{\theta}^*) = 100 * \frac{Var(\hat{\theta})}{Var(\hat{\theta}^*)} = \frac{2}{\pi} = 63.7\% \quad (3.6)$$

(Martin & Zamar, 1991)

Although being less efficient than mean, median has desirable properties that mean does not have. A simple example may be helpful to explain the case. Let, a very rich man -say with a wealth of 50 billion TL- moves to a village in which people have moderate income. If we use mean to estimate location parameter, this movement will result in all people but one having income less than the mean. However, there will be almost no change in median income and this measure will make much more sense in terms of location parameter since many of the people still will have income around median.

The reason for our control charts not performing well was similar. Since the distributions other than Gaussian have heavy tails, estimations were subject to false alarm signals due to outlier values. If we can find resistant estimators against

outliers, it might be expected that they will outperform the current ones. Such estimators are called Robust Estimators and they are not unduly affected by a few outliers, or say moderate departures from model assumptions.

As our simple example indicates, a useful way to qualify robustness can be by looking at the response of the estimator created by an additional unit in the sample data. This response is observed by “Influence Function,” which is explained as follows. Before giving theoretical background, “Empirical Influence Function (EIF)” will be defined with an example.

Let the random sample (x_1, x_2, \dots, x_n) and the estimator (or functional) $T_n(x_1, x_2, \dots, x_n)$ be given. The empirical influence function $EIF(x)$ is given by:

$$EIF(x) = T_{n+1}(x_1, x_2, \dots, x_n, x) \quad , \quad -\infty < x < \infty \quad (3.7)$$

(Klawonn, 2009).

Besides the estimators mean and median, it is useful to define $\bar{X}_\alpha = \alpha - \text{trimmed mean}$, which is the mean of the sample after removal of lowest and highest $100 \cdot \alpha\%$ values. Consider the $n = 10$ (ordered) sample data:

[0.13 1.27 1.44 1.52 1.75 2.09 2.96 3.80 3.83 4.22]

The statistics are calculated as:

$$\bar{X} = 2.301 \quad ; \quad med_i(x_i) = \tilde{X} = 1.920 \quad ; \quad \bar{X}_{10\%} = 2.333 \quad (3.8)$$

The following table shows the Empirical Influence Functions of these three estimators based on the given sample data. Influence function of mean is unbounded, which refers to non-robustness of an estimator. This means that any additional value $x \in R$ influences the estimator sample mean. Trimmed mean is not affected by

extreme values since they are removed before calculating the mean and its influence function has a positive slope within middle 80% of the sample data. Median is almost a step function here because an additional data unit makes the median 1.75 if the unit is less than this value and 2.09 if the unit is more than this value. Additional unit is the median itself between these two values.

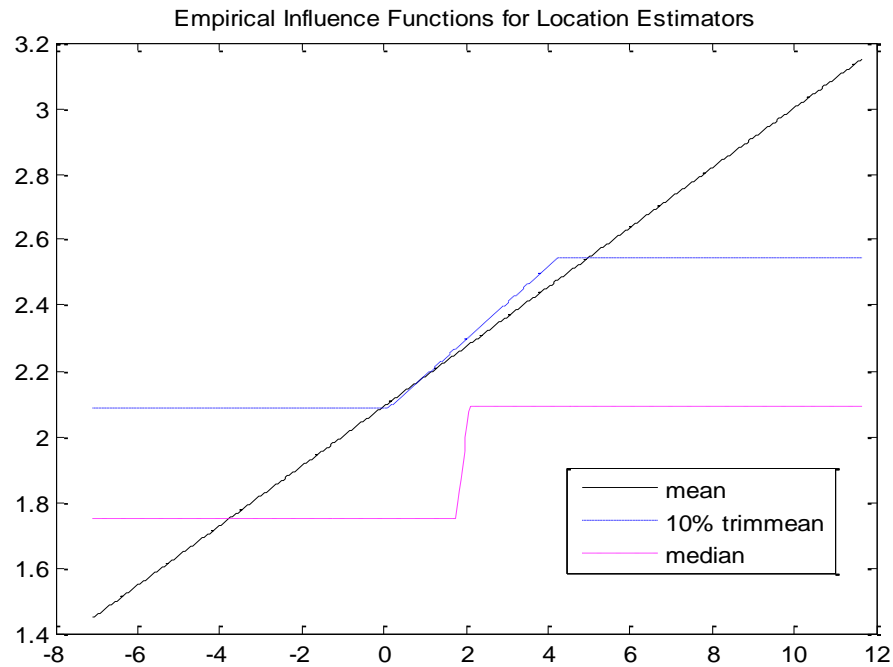


Figure 3.1 Empirical Influence Functions for the location estimators: mean, 10% trimmed mean and median using the simulated $n = 10$ sample data set.

The normalized version of EIF which is centered on zero and scaled with respect to sample size is called the Empirical Sensitivity Curve. The idea for normalization resembles that of obtaining standardized scores.

The Empirical Sensitivity Curve (ESC) is given by the following equation:

$$ESC(x) = \frac{T_{n+1}(x_1, x_2, \dots, x_n, x) - T_n(x_1, x_2, \dots, x_n)}{1/(n+1)} \quad (3.9)$$

(Klawonn, 2009).

The following figure is the ESC of the previous example:

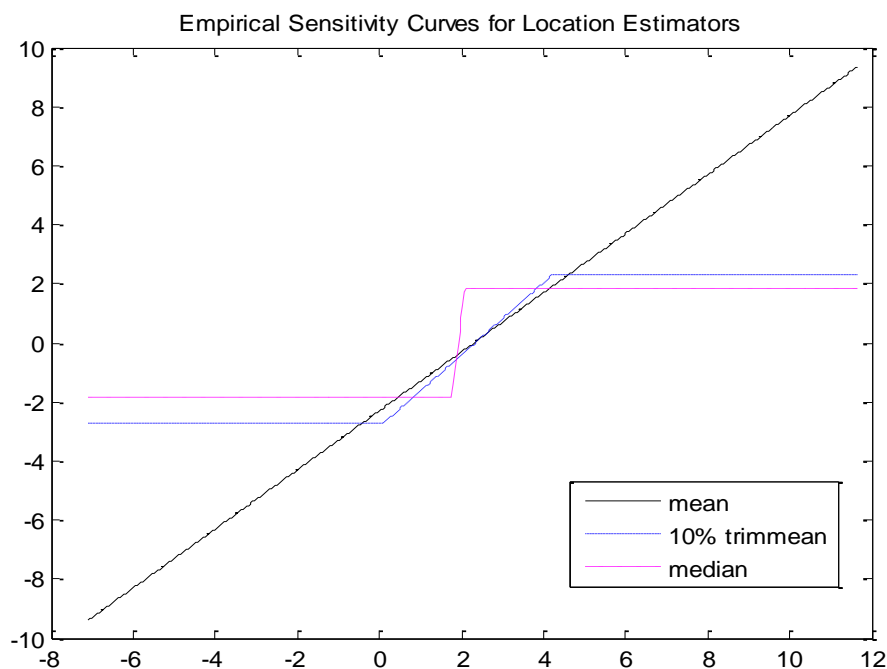


Figure 3.2 Empirical Sensitivity Curves for the location estimators: mean, 10% trimmed mean and median using the simulated $n = 10$ sample data set.

Sensitivity Curve is a bridge between EIF and its theoretical counterpart of Influence Function (IF). The influence function is defined as the limit (only if the limit exists) of sensitivity curve as sample size goes to infinity.

$$IF(x, T, F) = \lim_{n \rightarrow \infty} \frac{T\left(\left(1 - \frac{1}{n}\right)F - \frac{1}{n}\delta_x\right) - T(F)}{1/n} \quad (3.10)$$

(Klawonn, 2009).

The interpretation of IF is similar to that of EIF but it enables us to make general inferences for estimators T under an assumed distribution with cdf F . δ_x stands for a cdf resulting in the value of x with probability 1. In other words, it measures the effect of an infinitesimal contamination to the estimator T . It is clear that having a bounded Influence Function is a necessary condition for T to be a Robust Estimator.

The unbounded influence function of sample mean is given by:

$$IF(x, \bar{X}, F) = x - \mu_F \quad (3.11)$$

(Wilcox, 2005).

The bounded influence function of the sample median (a step function) is given by:

$$IF(x, \tilde{X}, F) = \frac{\text{sign}(x - \tilde{X})}{2f(\tilde{X})} \quad (3.12)$$

(Wilcox, 2005).

The maximum absolute value of the Influence Function is called ‘‘Gross Error Sensitivity’’ (GES) and in terms of the outlier value x , it shows the worst case that will happen to the estimator T . GES is defined as:

$$\gamma^*(T; F) = \sup_x |IF(u; T; F)| \quad (3.13)$$

(Klawonn, 2009).

Since the IF of mean is unbounded, its GES is infinite. On the other hand, median has a finite GES and its value for standard Gaussian distribution is $0.5 * \sqrt{2\pi} = 1.2533$, which is the minimum value an estimator for mean of a standard Gaussian distribution can have (Martin & Zamar, 1991).

Imagine that Estimators are Kings of their data lands and a black hearted witch is able to make a conversion charm that disturbs data members. What percentage of the ranked observations should the witch convert in order to reach and disturb the King?

The King Mean can be upset by converting any member of the sample data. Therefore, mean is said to have a breakdown point of 0%. A more robust estimator α – trimmed mean has a breakdown point of $100\alpha\%$ and it is more resistant to outliers. One of the most powerful kings of the “Glorious Statistics World” is the median with maximum breakdown point of 50%. Therefore, it will be too hard to defeat median for the witch.

Walking through the wonders of “Glorious Statistics World” made Pupil so excited. She was also feeling safe for experiencing these wonders with Rookie. Their tenderness on the nature was a strong tie between their souls. She was so eager to learn about robust scale estimators, hoping to find well performed control charts...

Rookie was also happy, but he was a bit confused. He was afraid of suffering from pangs of love. He thought that he should keep it slow, but it may be riskier than the current situation. He remembered the famous aphorism of Nietzsche on beauty, which can give an explanation to the case:

“The slow arrow of beauty... The noblest kind of beauty is not that which suddenly transports us, which makes a violent and intoxicating assault upon us (such beauty can easily excite disgust), but that which slowly infiltrates us, which we bear away with us almost without noticing and encounter again in dreams, but which finally, after having for long lain modestly in our heart, takes total possession of us, filling our eyes with tears and our heart with longing. –What is it we long for at the sight of beauty? To be beautiful ourselves: we imagine we would be very happy if we were beautiful. – But that is an error.” (Hollingdale, R.J., trans., 1996).

But maybe Nietzsche was wrong. That may be a random error, too...

3.2 “Classical versus Robust” Estimation of Scale

Rookie remembered the fact that sample mean was the best estimator for the population mean under normality assumption. Then, he wondered if it was the case for the population variance. It was a good starting point for future wonders...

As stated before, $W = \frac{(n-1)s^2}{\sigma^2}$ follows a Chi-Square distribution with degrees of freedom $n - 1$. Mean of W is $n - 1$ and its variance is $2(n - 1)$. (Walck, 2007) Then, if $\hat{\theta} = s^2$ is used to estimate the parameter $\theta = \sigma^2$ we have:

$$E(W) = E\left(\frac{(n-1)s^2}{\sigma^2}\right) = n - 1$$

$$E(s^2) = \frac{\sigma^2(n-1)}{(n-1)} = \sigma^2$$

$$E(\hat{\theta}) = \theta \tag{3.14}$$

$$Var(W) = Var\left(\frac{(n-1)s^2}{\sigma^2}\right) = 2(n-1)$$

$$Var(s^2) = \frac{2\sigma^4}{n-1}$$

$$Var(\hat{\theta}) = \frac{2\theta^2}{n-1} \tag{3.15}$$

Fisher information of Gaussian distribution's variance $\theta = \sigma^2$ is:

$$f(x; \theta) = \frac{1}{\sqrt{2\pi\theta}} e^{-\frac{(x-\mu)^2}{2\theta}}, \quad -\infty < x < \infty$$

$$\ln f(x; \theta) = -\frac{1}{2} \ln(2\pi\theta) - \frac{(x-\mu)^2}{2\theta}$$

$$\begin{aligned}\frac{\partial \ln f(x; \theta)}{\partial \theta} &= \frac{(x - \mu)^2}{2\theta^2} - \frac{1}{2\theta} \\ \frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} &= -\frac{(x - \mu)^2}{\theta^3} + \frac{1}{2\theta^2} \\ I(\theta) &= -E \left[\frac{\partial^2 \ln f(x; \theta)}{\partial \theta^2} \right] = \frac{\theta}{\theta^3} - \frac{1}{2\theta^2} = \frac{1}{2\theta^2} \\ I_n(\theta) &= nI(\theta) = \frac{n}{2\theta^2}\end{aligned}\tag{3.16}$$

By Cramer Rao Inequality, we have:

$$\text{Var}(\hat{\theta}) \geq \frac{1}{I_n(\theta)} = \frac{2\theta^2}{n}\tag{3.17}$$

Therefore, $\text{Var}(\hat{\theta})$ is only a little greater than Cramer Rao lower bound and for large samples, this difference disappears. $\hat{\theta} = s^2$ is unbiased and asymptotically efficient estimator for the parameter $\theta = \sigma^2$ of Gaussian distribution.

As for the case in estimation of location, we will use a second estimator for the Gaussian standard deviation, which is $\hat{\theta}_2 = MAD$. Thus, its square is an estimator of variance, but we will define $\hat{\theta}_1 = s$ and compare the estimators for $\theta = \sigma$. MAD is defined as:

$$MAD_n = b * \text{med}_i |x_i - \text{med}_j x_j|\tag{3.18}$$

Calculation of MAD is a two-step procedure. At first, the absolute value of the differences from median of the data is found. Second, the median of these numbers is calculated and multiplied with the constant b to make the estimator consistent. Like that of median, MAD has the best possible breakdown point 50% and its influence function is bounded with the sharpest possible bound among all scale estimators. Moreover, its gross error sensitivity is 1.167 for Gaussian distribution, which is the

minimum a scale estimator can have. These excellent properties make MAD a very robust estimator of population standard deviation (Rousseeuw and Croux, 1993).

Although being very robust, “Median Absolute Deviation” has an efficiency of only 37%. Due to this low efficiency, young enthusiasts of Glorious Statistics thought that they may need some other estimators to study for their control purpose. They came up with two new estimators S_n and Q_n , and made a plan to study them. Pupil was going to study S_n and Rookie the other. Then they would compare the properties these estimators together.

The third estimator $\hat{\theta}_3 = S_n$ is defined as:

$$S_n = c * med_i\{med_j|x_i - x_j|\} \quad (3.19)$$

The two-step procedure of calculating S_n is as follows. For each i , the median of $\{|x_i - x_j|; j = 1, 2 \dots n\}$ is calculated, which yields n numbers. Median of these n numbers, multiplied with c for consistency, is the final estimate S_n . Like that of MAD, S_n also has 50% breakdown point and bounded influence function. Luckily, it is more efficient than MAD with a value of 58.23% under Gaussian distribution, but unfortunately, its gross error sensitivity is 1.625, which is larger than that of MAD (Rousseeuw and Croux, 1993).

Finally, the fourth and the last estimator that will be studied is $\hat{\theta}_4 = Q_n$:

$$Q_n = d * \{|x_i - x_j|; i < j\}_{(k)} \quad (3.20)$$

The estimator Q_n resembles S_n , but median is replaced by another order statistics k . Here, d is a constant factor again and $k = \binom{h}{2}$, where $h = \lfloor \frac{n}{2} \rfloor + 1$ is almost half of the observations. Q_n 's breakdown point is again 50% and its influence function is

again bounded. These two properties are the same for the three estimators suggested as alternatives to sample standard deviation. However, their efficiency and gross error sensitivity for Gaussian distribution change. Q_n is the most efficient one among the three with nearly 82% efficiency value, and it has the largest (worst) gross error sensitivity value of 2.069 (Rousseeuw and Croux, 1993). Therefore, studying these three estimators will construct a very good set for the purpose of performance comparison.

The following figure is the empirical sensitivity curves for the scale estimators under study. Same data set is used with that of location estimators. It can be followed that standard deviation has an unbounded influence curve, and therefore is not robust. The other three robust estimators' curves are bounded.

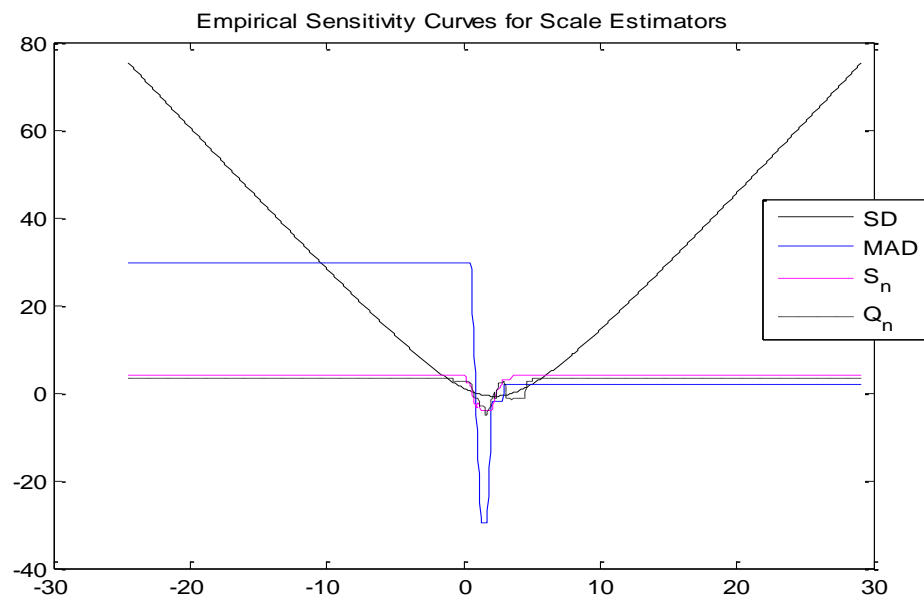


Figure 3.3 Empirical Sensitivity Curves for the scale estimators: standard deviation, median absolute deviation, S_n and Q_n using the simulated $n = 10$ sample data set.

Pupil was grateful to Rookie for teaching her new concepts. She got the theory but there was a question in her mind. How large is a large sample? Would the asymptotic results for unbiasedness and efficiency hold for small samples? They decided to conduct a simulation study to see practical counterpart of the theoretical results.

The following table shows simulated mean of the scale estimators for different sample sizes and their expected values. 10000 samples of each n are run using standard normal random data. Last row shows expected values and since all estimators are unbiased, their expected value is standard deviation of the generated data. Besides some negligible sampling errors, all samples have mean values that are close to 1.

Table 3.1 Simulated mean values of scale estimators MAD , S_n , Q_n and Standard Deviation for different sample sizes at each row, and their theoretical expected values are at the bottom row.

Average Estimated value of Scale Estimators for Gaussian data				
n	MAD_n	S_n	Q_n	SD
5	0.9957	1.0063	1.0064	0.9399
10	0.9945	0.8705	1.0090	0.9748
20	0.9992	0.9343	1.0008	0.9875
50	1.0024	0.9760	1.0010	0.9957
100	1.0016	0.9876	1.0005	0.9971
inf	1.0000	1.0000	1.0000	1.0000

It was stressed that sample standard deviation is the best scale estimator for Gaussian data, and therefore, it has minimum variance. The efficiencies for other estimators are the ratio of their variances to the variance of standard deviation. Last row of the following table gives the theoretical variances, and other rows' values are obtained by the standardized variance formula:

$$Var(\widehat{\theta}_j)_{std} = \frac{n * Var(\widehat{\theta}_j)}{(E(\widehat{\theta}_j))^2} \quad (3.21)$$

(Rousseeuw and Croux, 1993).

Table 3.2 Simulated standardized variance values of scale estimators MAD, S_n , Q_n , and Standard Deviation for different sample sizes at each row, and their theoretical variances are at the bottom row.

Standardized Variance of Scale Estimators for Gaussian Data				
n	MAD_n	S_n	Q_n	SD
5	1.6767	1.4729	1.3955	0.6550
10	1.3667	1.0020	0.8915	0.5687
20	1.3544	0.8867	0.7797	0.5325
50	1.3679	0.8578	0.6885	0.5193
100	1.3525	0.8533	0.6526	0.5157
inf	1.3610	0.8570	0.6080	0.5000

It can be deduced from the table that asymptotical efficiency values do not hold for small samples and especially for $n = 5$, almost all robust estimators are suffering from being totally inefficient. Hopefully, they converge to their theoretical values considerably fast and starting from $n = 20$, theoretical efficiencies can be stated as acceptable. This result is especially important for control purpose because sample size is one of the very important facts that determine estimation performance.

Pupil was satisfied since the studies they made supplied a necessary theoretical background to go further. Still, the background was not sufficient because she still had no idea about construction of the control chart limits using the robust scale estimators.

Since the time they met, Rookie possessed a kind of a teacher role by training her on estimation theory. But now, it was Pupil's turn. She was going to teach him the basics of Quality Control and they would be able to search robust scale control charts together.

3.3 A search for Robust Scale Control Charts

Before studying the theory, it would be a good idea to see the pattern of robust statistics on real data. Pupil showed Rookie the first leaf data she collected and the *Shewart S – Chart* she constructed. For the purpose of comparison, the same chart is given in the following figure but this time, its legend is not to the right of the chart but inside the chart. Rookie claimed that this would seem better. Pupil didn't think so, but she just smiled.

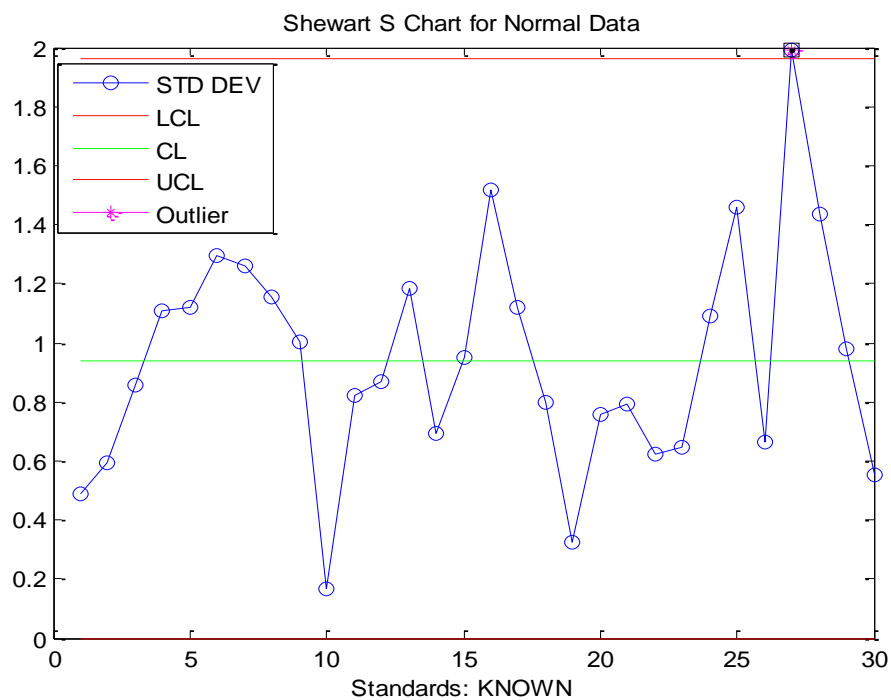


Figure 3.4 Shewart *S* Control Chart for leaves data given standard $\sigma = 2.5$

The following figures are the robust scale estimator charts using the same control limits with that of *Shewart S – Chart*. Their patterns are very similar because the estimators have similar characteristics. Like in the *Shewart S – Chart*, there is a single out of control value, which is the 27th observation, but this value is much higher than UCL at robust charts. Additionally, there are some points close to UCL, especially at MAD chart.

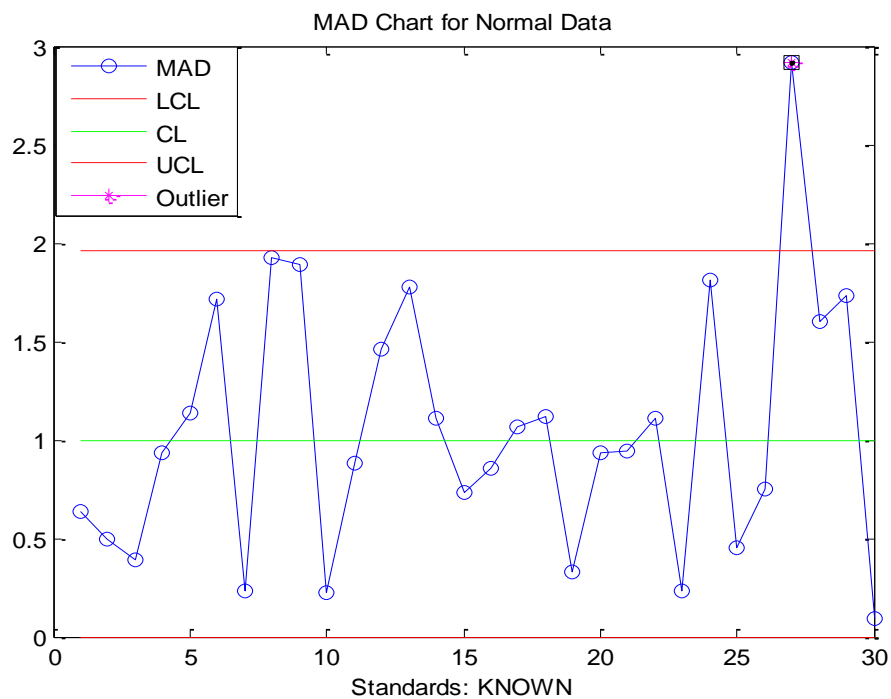


Figure 3.5 Median Absolute Deviation Control Chart for leaves data given standard $\sigma = 2.5$, using control limits of *Shewart S Chart*

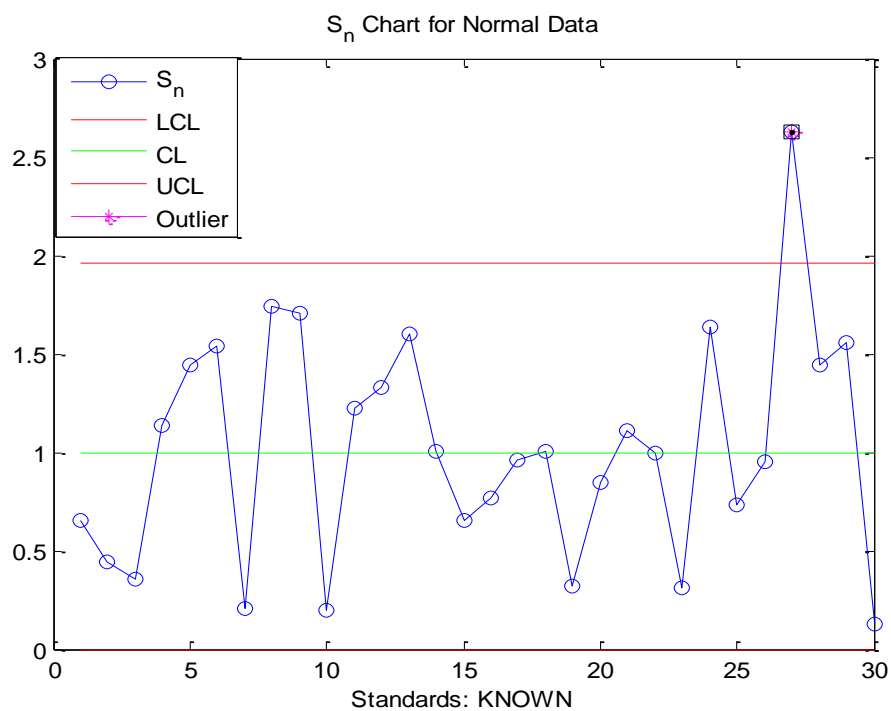


Figure 3.6 S_n Control Chart for leaves data given standard $\sigma = 2.5$, using control limits of *Shewart S Chart*

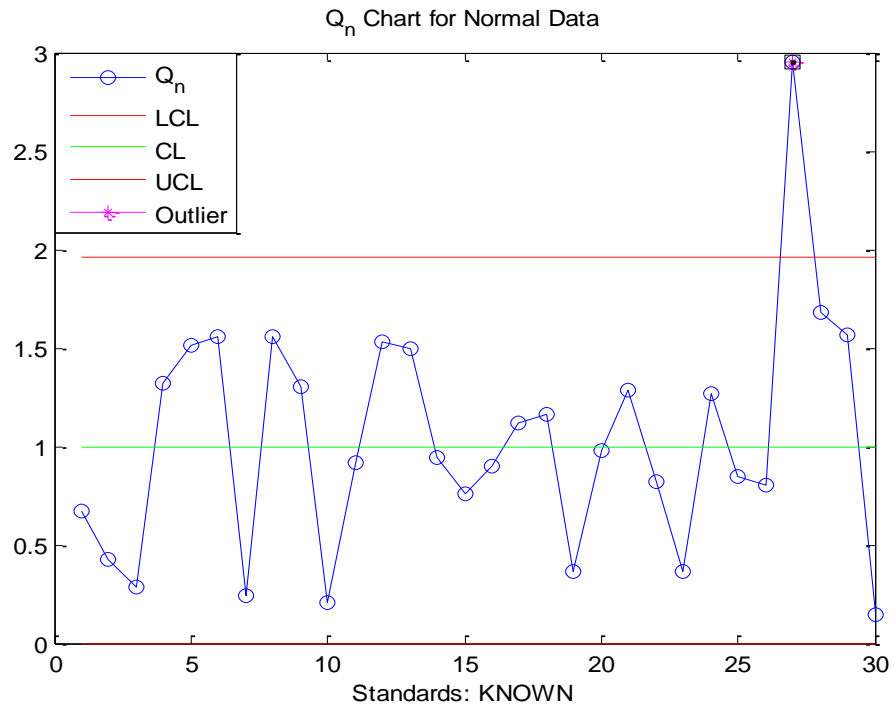


Figure 3.7 Q_n Control Chart for leaves data given standard $\sigma = 2.5$, using control limits of *Shewart S Chart*

At first glance, the robust charts may give such an opinion that they will not perform well. However, this may not be the case because this data is Gaussian and it is already expected that *Shewart S – Chart* is the best. Furthermore, we have already observed that sample size $n = 5$ is too small for our robust charts, and finally control limits are not updated yet.

To discover new control limits for each of the robust estimator control charts, a similar pattern following that of *Shewart S – chart* can be tried. Mean and standard error of s were given by the formulas:

$$E(s) = c_4\sigma = \sigma\sqrt{1 - c_4^2} \quad (3.22)$$

$$SE(s) = \sigma\sqrt{1 - c_4^2} \quad (3.23)$$

Like for the case of standard deviation, MAD also needs a constant for finite n values to be unbiased. We have $E(MAD) = \frac{\sigma}{b_n}$ but unfortunately, we do not have σ_{MAD} for finite n yet. Under Gaussian assumption, we may consider an efficiency constant f_n and define it as follows:

$$f_n = \frac{1}{RE_n(MAD)} \quad (3.23)$$

Values of f_n for changing n will be calculated from the simulation results given in Table 3.2. Since this coefficient will give the variance ratio of MAD to s for finite sample sizes, we may at least hope that standard error of MAD may be $\sigma_s = \sigma\sqrt{f_n(1 - c_4^2)}$. If this is the case, then 3σ control limits for MAD-Chart will be:

$$UCL = \frac{\sigma}{b_n} + 3\sigma\sqrt{f_n(1 - c_4^2)} \quad (3.24)$$

$$LCL = \frac{\sigma}{b_n} - 3\sigma\sqrt{f_n(1 - c_4^2)} \quad (3.25)$$

The following two constants are defined to reduce the formulas:

$$B_{61} = \frac{1}{b_n} + 3\sqrt{f_n(1 - c_4^2)} \text{ and } B_{51} = \frac{1}{b_n} - 3\sqrt{f_n(1 - c_4^2)} \quad (3.26)$$

Then, the control limits of *MAD chart* becomes:

$$UCL = B_{61}\sigma \quad (3.27)$$

$$LCL = B_{51}\sigma \quad (3.28)$$

To check the validity of these formulas, a simulation study on ARL_0 performances of *MAD Chart* will be made, as that of *Shewart S – Chart*. To enable comparison easier, Table 2.5 is shown again below, but with a different table number:

Table 3.3 Simulated run lengths with different sample sizes for standard deviation control chart of Normal data and the average run length, when the process is in control. Standard deviation of the mean run lengths and the control limits given standards are at the bottom part of the table.

ARL₀ for standard deviation given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
252.1050	258.2923	372.5630	361.4899	381.1980	365.4255
258.6720		364.5900		344.9180	
252.1900		370.1940		367.1300	
260.7010		345.9530		360.7810	
262.0200		382.1870		361.1380	
254.5820		352.7760		380.3800	
248.5270		348.2580		378.1370	
256.9360		360.7160		347.4870	
258.8570		340.1970		365.6360	
278.3330		377.4650		367.4500	
8.2211	=stdev	14.2905	=stdev	12.5757	=stdev
		$\sigma =$	1.0000		
UCL =	1.9636	UCL =	1.4703	UCL =	1.2972
LCL =	0.0000	LCL =	0.5036	LCL =	0.6926

The simulation results using the estimator MAD and new control limits are shown in the following table:

Table 3.4 Simulated run lengths with different sample sizes for MAD control chart of Normal data and the average run length, when the process is in control. Standard deviation of the mean run lengths and proposed control limits with f_n constants given standards are at the bottom part of the table.

ARL₀ for MAD with control limits using proposed f_n constants given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
60.3980	62.1869	181.7730	176.4228	237.1750	241.6036
65.6660		188.7180		247.7430	
60.6920		177.6150		243.7950	
63.7740		168.0890		231.1410	
60.7220		166.7830		236.7630	
60.8950		176.1770		239.3550	
64.0450		170.3120		246.6420	
62.9650		175.6420		249.4300	
62.8930		179.7530		245.6860	
59.8190		179.3660		238.3060	
1.9447	=stdev	6.6780	=stdev	5.9089	=stdev
		$\sigma =$	1.0000		
UCL =	2.4960	UCL =	1.7407	UCL =	1.4754
LCL =	0.0000	LCL =	0.1793	LCL =	0.4926

Unfortunately, poor ARL_0 performances are obtained. The idea seemed to be good but it wasn't, as we see. Anyway, here is the result: "Ideas are sometimes not as good as they seem." Here is another one: "Some bad ideas may create good ones later, so keep on trying."

Pupil was sad for this result, but Rookie was cool and hopeful. He thought that extending the confidence limits might work for variance chart obtained using Chi-Square statistics. Pupil felt desperate and Rookie tried to change her feelings for a time. He told her: "It may not work. But how can we know, if we don't try?" Then he remembered the beautiful song of Mary-Mary, which is "Can't give up now." A part of its lyrics is given below.

*There will be mountains that I will have to climb
 And there will be battles that I will have to fight
 But victory or defeat, it's up to me to decide
 But how do I expect to win if I never try*

Listening to music made them happy and they performed their first dance. Then, they studied the extended confidence limits version of variance chart for *MAD chart*. The confidence limits for variance chart were:

$$UCL = \frac{\chi_{\alpha/2}^2}{n-1} \sigma^2 \quad ; \quad LCL = \frac{\chi_{1-\alpha/2}^2}{n-1} \sigma^2 \quad (3.29)$$

Variance chart simulation gave ARL_0 values that are around 370, its table will not be given here again. If UCL value is multiplied by a constant and LCL is divided to the same constant, the simulation results can be adjusted by choosing constants so that MAD chart ARL_0 simulation values are also close to 370. Then, these limits can be tried for data with some other distributions. Specifically, performances for Logistic, Laplace, and Cauchy distributions will be analyzed. Letting the extending constant be g_n , the newer control limits for *MAD Chart* is given as follows:

$$UCL = g_n \sqrt{\frac{\chi_{\alpha/2}^2}{n-1}} \sigma^2 \quad (3.30)$$

$$LCL = \frac{1}{g_n} \sqrt{\frac{\chi_{1-\alpha/2}^2}{n-1}} \sigma^2 \quad (3.31)$$

The values of g_n found via simulation for the sample sizes 5, 20 and 50 are equal to 4.11, 1.445 and 1.234, respectively. The following table contains the simulation results:

Table 3.5 Simulated run lengths with different sample sizes for MAD control chart of Normal Data and the average run length, when the process is in control. Standard deviation of the run lengths and proposed extended control limits with g_n constants given standards are at the bottom part of the table.

ARL₀ for MAD extended limits using proposed g_n constants given standards (Normal Distribution)					
n=5		n=20		n=50	
sims	ARL₀	sims	ARL₀	sims	ARL₀
396.9440	370.2951	365.0250	369.2875	377.3890	371.0283
370.8570		372.3360		363.1560	
360.5000		359.8560		372.4810	
358.1620		364.1760		372.0430	
368.6970		365.5690		365.5320	
368.3760		375.4910		361.5460	
374.6660		365.9700		396.5600	
379.7220		356.1640		373.0950	
354.7230		395.2920		344.2490	
370.3040		372.9960		384.2320	
12.0662	=stdev	10.8938	=stdev	14.0369	=stdev
$g_5 =$	4.1100	$g_{20} =$	1.4450	$g_{50} =$	1.2340
UCL =	8.6696	UCL =	2.1705	UCL =	1.6163
LCL =	0.0396	LCL =	0.3771	LCL =	0.5734

The purpose of this way of formulation was to see its performance for other distributions under study. For that reason, performance differences in changing sample sizes and their ARL₀ simulations will not be conducted any more. The following table demonstrates the simulated performances for Logistic data at specific shifts in population standard deviation. As asserted before, a shift is defined as the population standard deviation becoming $\lambda\sigma$.

Table 3.6 Simulated run lengths for g_n extended MAD control chart of Logistic data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the mean run lengths for different λ values are at the bottom row of the table.

ARL for MAD extended limits using proposed g_n constants with $n=50$ (Logistic Distribution)							
$\lambda = 1$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
83.1600	82.9906	232.9310	231.8269	9.4950	9.1201	1.4550	1.4566
85.4640		246.2950		9.3720		1.4650	
83.5510		234.5400		8.8930		1.4510	
81.8200		219.7800		8.6510		1.4440	
83.1120		233.5450		9.2480		1.4840	
82.0390		230.0410		9.2390		1.4400	
78.1160		223.6290		9.0350		1.4760	
87.5660		232.6620		8.8510		1.4330	
83.9570		241.0330		9.2850		1.4960	
81.1210		223.8130		9.1320		1.4220	
2.5359	=stdev	8.0695	=stdev	0.2610	=stdev	0.0235	=stdev

MAD is considerably increasing at the first shift at our formulation. Consequently, it didn't work here. We actually want to detect shifts and so want to have ARL a decreasing function of λ . It means that Rookie's idea for *MAD Chart* control limits has failed at Logistic distribution. May the chart have a chance for other distributions? We guess not, but it may still be worthy trying for our distributions under study. For sure, Glorious Statistics has created a trial with a distribution.

Table 3.7 Simulated run lengths for g_n extended MAD control chart of Laplace data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the mean run lengths for different λ values are at the bottom row of the table.

ARL for MAD extended limits using proposed g_n constants with $n=50$ (Laplace Distribution)							
$\lambda = 1$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
7.1390	6.8513	33.3950	32.4678	57.4720	54.6408	3.5910	3.6729
6.6860		31.9680		56.8790		3.7420	
6.6520		31.4200		53.8250		3.6960	
6.4270		32.4630		52.9610		3.6650	
6.9110		33.0550		53.2880		3.8140	
7.1130		33.3500		53.7220		3.5560	
6.6410		31.3630		55.2310		3.7760	
7.2850		32.7350		52.5520		3.4940	
6.8140		30.9170		56.1590		3.6820	
6.8450		34.0120		54.3190		3.7130	
0.2658	=stdev	1.0226	=stdev	1.7097	=stdev	0.1000	=stdev

The case is more terrible than Logistic case. ARL is not a decreasing function of λ again and it increases at 1.5σ shift also.

Table 3.8 Simulated run lengths for g_n extended MAD control chart of Cauchy data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the mean run lengths for different λ values are at the bottom row of the table.

ARL for MAD extended limits using proposed g_n constants with $n=50$ (Cauchy Distribution)							
$\lambda = 1$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
2.9920	2.9644	1.5190	1.5301	1.0890	1.0860	1.0080	1.0046
2.9220		1.5410		1.0750		1.0030	
2.9320		1.5310		1.0680		1.0000	
3.0560		1.5150		1.0850		1.0080	
2.8410		1.5150		1.0900		1.0070	
2.9390		1.4890		1.0660		1.0020	
2.9170		1.5470		1.0990		1.0060	
2.9820		1.5450		1.1010		1.0040	
3.0660		1.5950		1.1200		1.0040	
2.9970		1.5040		1.0670		1.0040	
0.0682	=stdev	0.0294	=stdev	0.0176	=stdev	0.0026	=stdev

There may be a slight improvement for Cauchy case since ARL_0 is higher than that of the sample variance and ARL is a decreasing function of λ , but the improvement is not satisfactory at all and so it isn't worthy...

CHAPTER FOUR

CONTROL CHARTS USING ROBUST SCALE ESTIMATORS

Autumn was approaching and the forest started to turn yellow. Rookie couldn't decide whether it was the forest having new colors or if it was the soul of variation that forces its existence to change. His current way of thinking was accustomed to supporting the former idea, but he couldn't resist against the provocation of the latter. What was the reason for not modeling the creation of variation as the main reason in itself that brings the life into existence? He couldn't give an answer and felt that the best decision was just waiting for an answer, doing nothing...

Autumn was coming back to recall that it has never gone. Pupil just couldn't understand her feelings that distinguish existence and realization. It may be that not realizing the autumn is just a permission for realizing other seasons. Moreover, Glorious Statistics might have created a stochastic process for realization of four seasons. "To illustrate" she thought, "let a fair pair of dice come up 2:1. This cannot mean that the fair pair possesses a 6:6." She was allowing herself to become a statistician...

Autumn was coming to realize the creation against the deterministic beautiful ideas swallowed in the summer. Its coming was just an offence for being forgetful in contrast to its enabling creativeness...

Rookie and Pupil were sharing their feelings and they were a little upset for not having started to collect data before autumn. They thought that their studies for understanding the basics went too long. While one of their morning walks, Pupil realized the coming of autumn and remembered the previous ones. She then silently dived into the moment they performed a lovely dance and strongly felt that it hadn't gone. That dance was still existing...

She didn't answer Rookie when he asked her the reason why she was smiling. Like Rookie, she waited for the time by doing nothing and unlike Rookie, she got the answer. Love was the answer, but what was the question?

Then, they turned back to their study. It would become a way soon to strengthen the unquestionable tie between...

4.1 Bootstrap Confidence Intervals

The brilliant statistician Brad Efron created an idea in 1989. His way of thinking was so simple but quite efficient. The idea was: "Why do not we treat the sample on hand as the whole population and take repeated samples from our sample with replacement?" This is brilliant because the only thing we practically have is the sample in hand.

In fact, the primary task of us as statisticians is to summarize a sample based study, and generalize the findings in order to make inferences for the whole population. In the early years of our Statistics education, we learned that there are populations –as if they really were– having a specified distribution and we take samples from them in order to understand where it resides (location) and how far its members can go away (dispersion) from the main base of resident (mean). Moreover, we are taught the following fable. If it was possible to draw all possible samples from the considered population and we would calculate a specific statistics for each, we obtain another population, whose distribution is the sampling distribution.

As dealings with Statistics passed on with years of my life, I was able to understand the real story. Truly, distribution is just a mathematical formula. It is a tool to fit our way of thinking and experiences. For example, if the experiences through a specific subject are considered as a population, a proverb about that subject can be thought as a summary statistics. However, all possible experiences about the subject do not really exist.

For some statistics such as mean and under specific distributions such as Gaussian, our current knowledge of mathematics lets us reach the mathematics of further inferences. But sometimes, it does not. Furthermore, much of us are not fans of boring and long theoretical searches. As one of our proverbs expresses, “A good example has twice the value of a good advice.”

Anyway, he introduced the bootstrap method. This method practically makes the thing we ideally or theoretically learn. Basically, we draw new random samples (of same size) with replacement from our original sample. Here, our sample in hand is simply replaced with the theoretical distribution, and the “bootstrap sample” is treated as a sample. By taking many bootstrap samples, we reach some inferential idea about the sampling distribution. Some theoretical support, notations, and formulas are as follows:

Remember the standard score formula:

$$Z = \frac{X - \mu}{\sigma} \quad (4.1)$$

And its sample mean version:

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \quad (4.2)$$

where the denominator is the standard error of the sample mean and can be expressed as $SE(\bar{X})$. Replacing μ with a general population parameter θ and sample mean with a general statistics $\hat{\theta}$, by Central Limit Theorem (CLT), we have a statistics whose limiting distribution is standard Gaussian:

$$Z = \frac{\hat{\theta} - \theta}{SE(\hat{\theta})} \quad (4.3)$$

Now, consider the random sample of the size n with data (X_1, X_2, \dots, X_n) . If this is the population, its parameter under consideration is now $\hat{\theta}$. We are going to draw bootstrap samples from the –population treated- sample and calculate the statistics $\hat{\theta}_B$. Theoretical studies show that for most of the commonly used statistics, Z also has a limiting Gaussian distribution with the following formula:

$$Z = \frac{\hat{\theta}_B - \hat{\theta}}{SE(\hat{\theta}_B)} \quad (4.4)$$

It is noticeable to mention that the power of the bootstrap does not stem from CLT. In fact, CLT is just an evidence for the use of bootstrap method for inferential purposes. As explained in the previous chapters, the main idea of the control chart is to construct a confidence interval for population parameter. Specifically, we search for a confidence interval for population standard deviation to be used as Upper and Lower control limits.

The bootstrap method allows us to estimate sampling distribution of a statistics. Then, why do not we try to construct confidence intervals for our robust statistics using bootstrap method? Here is the answer, why not? Here are some standard brands of confidence intervals constructed using bootstrap:

Consider a 90% confidence interval (L, U) of θ . We basically infer that, 90% of the time we obtain these two numbers, θ will be within them. Therefore, it makes sense to map L to the 5th percentile and U to the 95th percentile.

Suppose a random sample of size 100. It is customary (or maybe sufficient) to take n^2 bootstrap samples, hence our young enthusiasts of Statistics may settle 10000 bootstrap replications of the sample. For Pupil's considered parameter θ , they used the estimator $\hat{\theta}$ and calculated this statistics for each of the bootstrap sample yielding the tuple $(\theta_1^*, \theta_2^*, \dots, \theta_{10000}^*)$. When they make an ascending order of the realized statistics, the tuple will become:

$$(\theta_{(1)}^*, \theta_{(2)}^*, \dots, \theta_{(10000)}^*) \quad (4.5)$$

Then the 90% confidence interval for population parameter θ will be:

$$(\theta_{(500)}^*, \theta_{(9500)}^*) \quad (4.6)$$

This is called “*Bootstrap Percentile Method*” (Singh & Xie, 2010).

If she has some doubt about the symmetry of the sampling distribution, she might want to change the places of L and U in symmetry with $\hat{\theta}$. The new confidence by this consideration is:

$$(2\hat{\theta} - \theta_{(9500)}^*, 2\hat{\theta} - \theta_{(500)}^*) \quad (4.7)$$

This is called “*Centered Bootstrap Percentile Method*” (Singh & Xie, 2010).

The usual confidence interval idea based on CLT may also be applied provided that there is not a strong evidence for non-normality of the sampling distribution. It is simply the interval:

$$(\hat{\theta} - b_{0.95}SE(\hat{\theta}), \hat{\theta} - b_{0.05}SE(\hat{\theta})) \quad (4.8)$$

where $SE(\hat{\theta})$ is estimated from the bootstrap samples.

The b coefficients of this interval are explained as follows. Let $T_B = (\hat{\theta}_B - \hat{\theta}) / SE(\hat{\theta}_B)$ and consider the statement T lies within $[b_{0.05}, b_{0.95}]$. This is called “*Bootstrap-t Method*” (Singh & Xie, 2010).

4.2 Robust Control Charts

The bootstrapping method is applied to our four distributions under study, which are: Gaussian (Normal), Logistic, Laplace (Double Exponential) and Cauchy distributions. “Sample Variance” statistics is the control statistics to compare performances of the other alternative robust statistics, because it is the classical statistics used under the assumption of normality. The robust statistics, whose performances are analyzed via Average Run Length (ARL) values, are MAD, S_n , and Q_n .

The values for $\lambda = 1.0$ are ARL_0 values and ARL_1 values for $\lambda = 1.2, 1.5, 2.0$ are also given. The method used for confidence intervals is the “percentile” method. For comparison purposes, “centered percentile” method is also shown for Q_n statistics. Upper limit of the confidence interval is equated to UCL and that of lower is likewise to LCL. As we know, for 3 sigma confidence limits, type one error level is $\alpha = 0.0027$. However, this value of α has given too high values of ARL_0 for MAD and S_n , and too low ARL_0 values for Q_n . For practical purposes, the significance levels of the confidence intervals are 0.0075 for MAD and S_n . That value is 0.0010 for Q_n statistics. The idea here is like that of the second trial for MAD chart in Chapter Three.

As mentioned before, Run Length is a Geometric random variable. When the process is in control, R_0 has mean $\frac{1}{\alpha}$ and standard deviation is also approximately $\frac{1}{\alpha}$. Likewise, when the process is out of control, R_1 is a Geometric random variable with parameter $1 - \beta$. 10 values of \bar{R} statistics are calculated, each of which is mean of $r = 1000$ runs for corresponding R. Results are shown in the “sims” column of the tables, to observe the variation in \bar{R} . Their mean is used as a final estimator for ARL. Their standard deviation is also shown. Since LCL and UCL values are calculated using bootstrapping method, they are also random variables whose values change with respect to bootstrap samples taken. LCL and UCL are calculated using n^2 bootstrap samples, and this operation is performed 100 times each time to see the

variability in Control Limits. Their mean is used as final Control Limits. The standard deviations of 100 corresponding limits are also shown in the tables but 100 bootstrap confidence intervals are not shown.

Sample sizes used in simulations are $n = 50$ for Variance and MAD, and $n = 20$ for S_n and Q_n . The aim was to see the difference between small sample and large sample cases (practically less than and more than 30) and there seems no practical difference between these two sample sizes. It has been already shown that our robust statistics reaches their asymptotical efficiencies at sample size $n = 20$. The following subchapters are devoted to analyze the simulation statistics obtained by the four distributions under the study and to a search which aims to understand basic characteristics of the robust statistics used.

4.2.1 Gaussian Distribution

4.2.1.1 Sample Variance

The following table shows the ARL values of “Sample Variance.” Random samples of size 50 are taken from Standard normal distribution.

Table 4.1 Simulated run lengths for variance control chart of Normal Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the run lengths for different λ values are at the bottom row of the table.

ARL for Sample Variance with n=50 (Normal Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
388.8490	372.2155	5.8730	5.9524	1.1230	1.1241	1.0000	1.0002
352.5830		5.6870		1.1200		1.0010	
373.1640		5.9020		1.1380		1.0000	
364.4410		5.7400		1.1120		1.0000	
358.9310		6.1590		1.1350		1.0000	
392.8820		6.2800		1.1390		1.0010	
384.8150		5.7060		1.1270		1.0000	
352.8140		6.1540		1.0940		1.0000	
371.5570		6.0060		1.1260		1.0000	
382.1190		6.0170		1.1270		1.0000	
14.7600	=stdev	0.2062	=stdev	0.0134	=stdev	0.0004	=stdev

As might be expected, “Sample Variance” statistics performs very well for Normal distribution. Even for small shifts in the population parameter, shift can be quickly detected. It might be stressed that standard error of the ARL₀ statistics is relatively high.

To give an idea; under Normal Approximation to Geometric Distribution, (even if this approximation may not be well enough) 95% Confidence Interval for ARL₀ is almost (340 , 400). Namely, even if the process is under control, we may have a false alarm signal between 340 to 400 runs. Standard deviations for ARL₁ statistics are relatively small. On the average, a 1.2 shift in true standard deviation is expected to be detected in 6 runs and higher shifts are expected to be detected immediately.

4.2.1.2 Median Absolute Deviation

The following is a histogram of MAD for 2500 bootstrap samples constructed by a random sample of size 50 taken from Standard normal distribution.

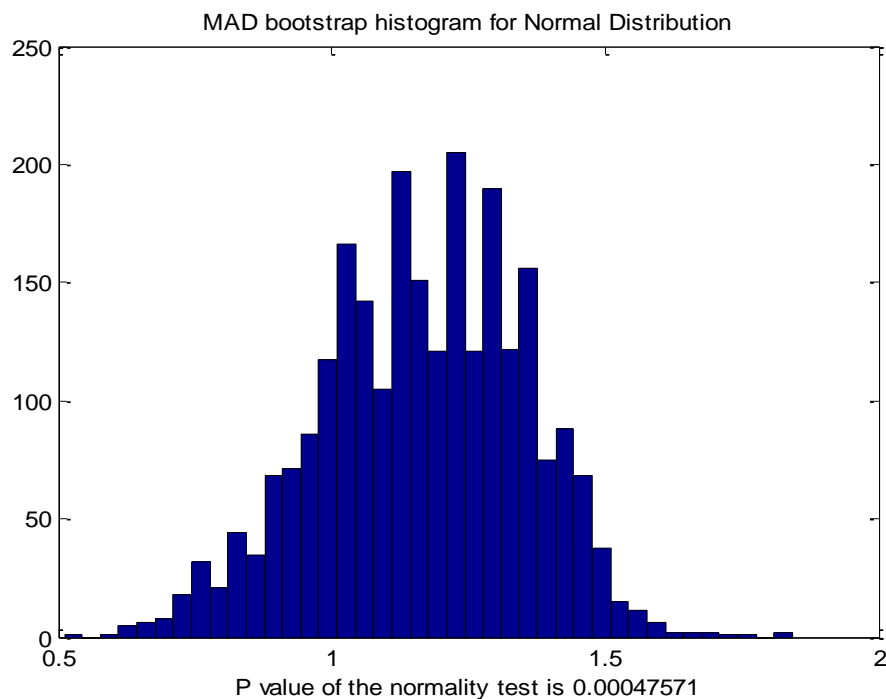


Figure 4.1 Histogram of Sampling distribution of MAD, based on bootstrap samples, when samples are taken from Gaussian distribution.

First of all, KS-test for normality has a p-value that can be rejected at any acceptable significance level. Therefore, sampling distribution of MAD statistics is not Normal. Then, confidence intervals based on Normal distribution will not be valid for MAD. The used confidence interval method which is based on bootstrapping is “Bootstrap Percentile Method.”

The relevant hypothesis testing is as follows:

H_0 : Sampling distribution fits Normal distribution

H_A : The distribution is not Normal

Test statistics is Kolmogorov Smirnov (KS) test

Reject H_0 if $p - value < \alpha$

$p - value = 0.00047$ (4.9)

Reject H_0 at any acceptable level of α . Sampling distribution of MAD is not Normal.

Moreover, the histogram does not like to seem to fit any known distribution, due to the peaks in the middle and dips between them. This fact makes hard to implement a theoretical study on MAD.

The following table shows the ARL values of “MAD.” Random samples of size 50 are taken from Standard normal distribution.

Table 4.2 Simulated run lengths for MAD control chart of Normal Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Median Absolute Deviation with n=50 (Normal Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
357.3500	355.3918	17.3260	17.4950	2.2470	2.2456	1.0750	1.0733
350.8000		17.6500		2.2800		1.0630	
338.6770		16.9940		2.2210		1.0760	
352.3980		17.6180		2.2140		1.0660	
374.1070		17.6210		2.2650		1.0740	
348.3770		17.4280		2.2550		1.0790	
365.9910		17.5660		2.2250		1.0680	
357.8550		16.9950		2.2040		1.0610	
360.2180		18.1840		2.3170		1.0800	
348.1450		17.5680		2.2280		1.0910	
10.0385	=stdev	0.3455	=stdev	0.0347	=stdev	0.0091	=stdev
		LCL	UCL				
	MEAN =	0.5386	1.4993				
	STDEV =	0.0878	0.1592				

Confidence interval significance level is 0.0075 for practical purposes, as mentioned before. This significance level achieves an ARL₀ value that is close to 370, which enables the MAD’s power comparable with that of “Sample Variance.”

The difference between estimated ARL_0 values (372.22 and 355.39) is not statistically significant.

Under Normal distribution, “Sample Variance” is more powerful than MAD, especially for detecting small shifts. ARL_1 of a 1.2 shift has an average value of 17 for MAD, which was only 6 for “Sample Variance.” Although having less power, MAD is not bad at all especially in detecting moderate or large shifts.

In fact, although detection of small shifts gives idea about the performances of the statistics under study, it is not a very important practical problem. In general, since Shewart control charts are not good enough to detect small shifts (because of their memoriless property), some other charts such as Cumulative Sum (CUSUM) chart are used simultaneously for this purpose (Montgomery, 2009). Additionally, a new procedure will be proposed in this chapter, which has an equal power to the sample variance chart.

4.2.1.3 S_n

The following is a histogram of S_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from Standard normal distribution. (This is for illustration purpose. Confidence intervals are performed using 400 bootstrap samples)

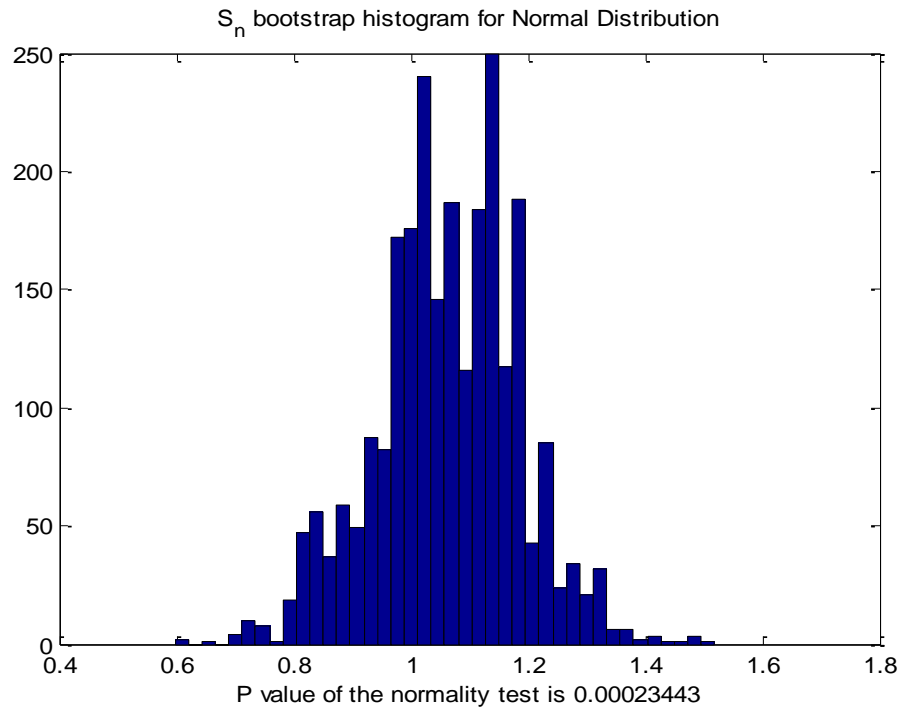


Figure 4.2 Histogram of Sampling distribution of S_n , based on bootstrap samples, when samples are taken from Gaussian distribution.

The interpretation of the histogram of S_n is very similar to that of the MAD. Again, the sampling distribution is not Normal (p-value is 0.00023) and there are apparent dips in the middle.

The following table shows the ARL values of S_n . Random samples of size 20 are taken from Standard normal distribution.

Table 4.3 Simulated run lengths for S_n control chart of Normal Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for S_n with $n=20$ (Normal Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
278.3560	264.2448	16.9130	16.7496	2.7810	2.7793	1.2080	1.2123
267.9220		16.4990		2.7870		1.2010	
274.3580		16.8390		2.7440		1.2600	
260.7650		16.6800		2.8340		1.2110	
267.7780		16.9160		2.7980		1.1980	
262.0720		16.6090		2.8130		1.2090	
247.7410		16.9550		2.6870		1.2210	
264.2190		17.7870		2.7820		1.1960	
259.7380		16.3320		2.7990		1.2130	
259.4990		15.9660		2.7680		1.2060	
8.5599	=stdev	0.4780	=stdev	0.0406	=stdev	0.0183	=stdev
		LCL	UCL				
	MEAN =	0.3075	1.5015				
	STDEV =	0.1106	0.2855				

Confidence interval of S_n is relatively wider than that of MAD. The control limits' standard deviations are close to each other for both statistics. Therefore, it can be inferred that change in sample size from 50 to 20 does not practically effects the inference on performance measures.

Confidence interval significance level is again 0.0075, in order to enable comparison. This significance level achieves ARL₀ value of 264, which seems to be less than that of MAD. Moreover, ARL₀ statistics of S_n seems to be less variable than that of MAD. Surely, formal tests are required to make these inferences. The "Anderson-Darling test of Normality" supports the Normality of the both \bar{R}_0 values with corresponding p-values of 0.867 and 0.574. Namely we can safely assume that the values come from a Normal distribution, and therefore, F test and t test are valid. Corresponding tests are made as follows:

$$H_0: \text{VAR}(\overline{R}_0 \text{ of } S_n) = \text{VAR}(\overline{R}_0 \text{ of MAD})$$

$$H_A: \text{ARL}_0 \text{ of MAD has higher variance}$$

Test statistics is F statistics

Reject H_0 if $p\text{-value} < \alpha$

$$p\text{-value} = 0.3215 \quad (4.10)$$

Do not Reject H_0 at $\alpha=0.10$. Variance of ARL_0 statistics of S_n is not less than that of MAD.

$$H_0: E(\overline{R}_0 \text{ of } S_n) = E(\overline{R}_0 \text{ of MAD})$$

$$H_A: \text{ARL}_0 \text{ of MAD is higher}$$

Test statistics is t statistics

Reject H_0 if $p\text{-value} < \alpha$

$$p\text{-value} = 0.0000 \quad (4.11)$$

Reject H_0 at any acceptable type one error level. ARL_0 statistics of MAD has a higher mean than that of S_n .

The “two sample t test” strongly evident indicates that MAD outperforms S_n when there is no shift. Moreover, ARL_0 statistics of S_n is not significantly less variable than that of MAD. Performances of two statistics are quite similar in detecting shifted population standard deviation.

4.2.1.4 Q_n

The following is a histogram of Q_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from Standard normal distribution:

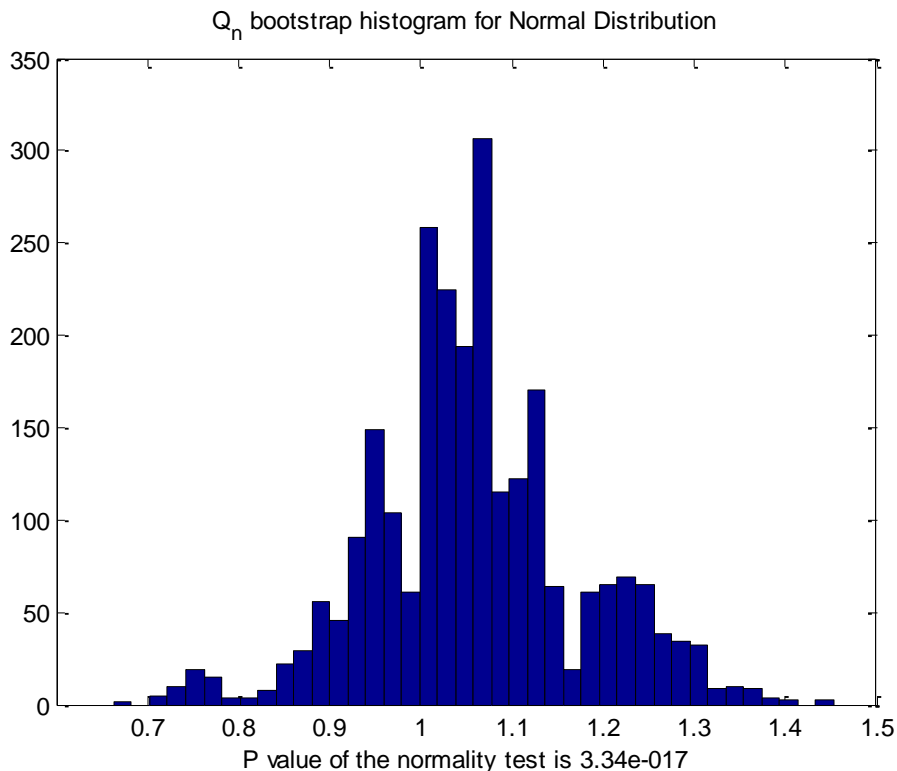


Figure 4.3 Histogram of Sampling distribution of Q_n , based on bootstrap samples, when samples are taken from Gaussian distribution

The sampling distribution of Q_n is again not Normal (p -value = 0.0000) but the histogram of Q_n exhibits different characteristics than the previous statistics. First of all, the histogram has much more dips and the dips are not only at the middle part. Secondly, the histogram is not bell shaped at all and except from the peaks, histogram has uniform characteristics from 0.7 to 0.9 and from 1.2 to 1.3. Finally, too many outliers are apparent at both sides, but especially at the upper side. The reason may be that Q_n has a higher GES compared to other estimators, MAD and S_n .

These characteristics -at least visually- support the bad performance of Q_n for “Normally distributed data’s quality control.” However, the histogram is surprisingly more skewed than the other two robust statistics. This fact may help to improve performance by using “Centered Bootstrap Percentile Method” for constructing the confidence intervals.

The following table shows ARL values of Q_n . Random samples of size 20 are taken from Standard normal distribution.

Table 4.4 Simulated run lengths for Q_n control chart of Normal Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Percentile" with $n=20$ (Normal Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
12.2000	12.1767	2.8130	2.8063	1.3140	1.3051	1.0430	1.0290
11.9590		2.7760		1.3010		1.0370	
13.0860		2.7980		1.3370		1.0260	
12.5000		2.9750		1.2980		1.0230	
12.1260		2.8390		1.2770		1.0300	
11.7120		2.7700		1.2930		1.0240	
11.8070		2.7880		1.3310		1.0330	
11.2280		2.8480		1.2850		1.0330	
12.4840		2.7740		1.3090		1.0240	
12.6650		2.6820		1.3060		1.0170	
0.5341	=stdev	0.0749	=stdev	0.0188	=stdev	0.0077	=stdev
		LCL	UCL				
	MEAN =	0.3008	1.5243				
	STDEV =	0.1166	0.2943				

Q_n has a very bad ARL₀ performance, and clearly cannot be used for control purposes. However, there may be an interesting idea here, based on the differences in performance measures.

Performance of Q_n is very good for ARL₁ statistics, even better than that of "Sample Variance." This means that, if Q_n and one of the other two robust statistics are used simultaneously, variability of the process can be screened in a perfect manner. Frequency of "out of control signals" around 12 for Q_n is not important because it naturally occurs and decision on "Process is in Statistical Control" can be based on S_n or MAD statistics, whichever is used in process. On the other hand, if Q_n

is out of control for 1st, 2nd or 3rd sample after the last signal, then a conclusion as “Process is out of Statistical Control” can be reached safely.

It is valuable to check the control limits using “Centered Bootstrap Percentile Method” for constructing the confidence intervals, since the bootstrap samples have exhibited a skewed pattern. This is basically for illustrative purposes. Method is also tried for other statistics, but the results were worse since their bootstrap samples are more symmetric than that of Q_n . The following table shows the ARL values of Q_n .

Table 4.5 Simulated run lengths for Q_n control chart of Normal Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on centered bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Centered Percentile" with $n=20$ (Normal Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
42.6250	43.1550	5.4380	5.3979	1.6790	1.6617	1.0870	1.0708
44.3250		5.4650		1.6400		1.0760	
44.5590		5.3050		1.7500		1.0740	
42.3750		5.4550		1.6420		1.0670	
42.9020		5.3490		1.6810		1.0710	
41.0730		5.3690		1.6480		1.0520	
44.0310		5.3480		1.6410		1.0760	
43.3470		5.6020		1.6190		1.0670	
43.0300		5.3260		1.6730		1.0780	
43.2830		5.3220		1.6440		1.0600	
1.0255 =stdev		0.0921 =stdev		0.0368 =stdev		0.0099 =stdev	
		LCL	UCL				
MEAN =		0.4588	1.6823				
STDEV =		0.4018	0.4842				

The results are better than that of the Percentile Method's. There is an improvement in the performance measure of ARL₀, but the performance is not better than either that of MAD's or S_n 's. For that reason, this improvement makes no help.

4.2.2 Logistic Distribution

4.2.2.1 Sample Variance

The following table shows the ARL values of “Sample Variance.” Random samples of size 50 are taken from Logistic distribution centered at zero (mean, and also median is zero) with variance 1.

Table 4.6 Simulated run lengths for variance control chart of Logistic Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the run lengths for different λ values are at the bottom row of the table.

ARL for Sample Variance with n=50 (Logistic Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
62.7940	61.4027	4.9420	4.8613	1.2120	1.2196	1.0010	1.0008
59.7870		4.7470		1.2230		1.0000	
59.6100		4.8270		1.2030		1.0010	
61.2240		4.7930		1.2170		1.0010	
65.0210		4.9860		1.2140		1.0000	
64.6460		4.8420		1.2240		1.0010	
60.7360		4.7100		1.2210		1.0030	
63.5340		4.7770		1.2420		1.0010	
55.7440		5.1160		1.2170		1.0000	
60.9310		4.8730		1.2230		1.0000	
2.7615	=stdev	0.1231	=stdev	0.0101	=stdev	0.0009	=stdev

Performance of “Sample Variance” for Logistic distribution is much worse than its performance for Normal Distribution. This shows that “Sample Variance” is not robust with respect to change in distribution. Its ARL₀ value dropped dramatically from 372 to 61. To make inference about variations of two distributions’ ARL₀ statistics, we need to compare the “Coefficient of Variation” values because their means differ too much. For Normal case, $Cov = \frac{s}{\bar{x}} = \frac{14.8}{372.2} = 4.0\%$ and for Logistic case, $Cov = \frac{s}{\bar{x}} = \frac{2.76}{61.4} = 4.5\%$. Then, the variability in ARL₀ for these two distributions is close to each other.

4.2.2.2 Median Absolute Deviation

The following is a histogram of MAD for 2500 bootstrap samples constructed by a random sample of size 50 taken from standard Logistic distribution. Unlike the Normal case, histogram for bootstrap samples of MAD has no apparent dips. The sampling distribution seems to be more symmetric and more bell-shaped in tails. However, the sampling distribution is not Normal again, since p-value for Normality test is 0.0001.

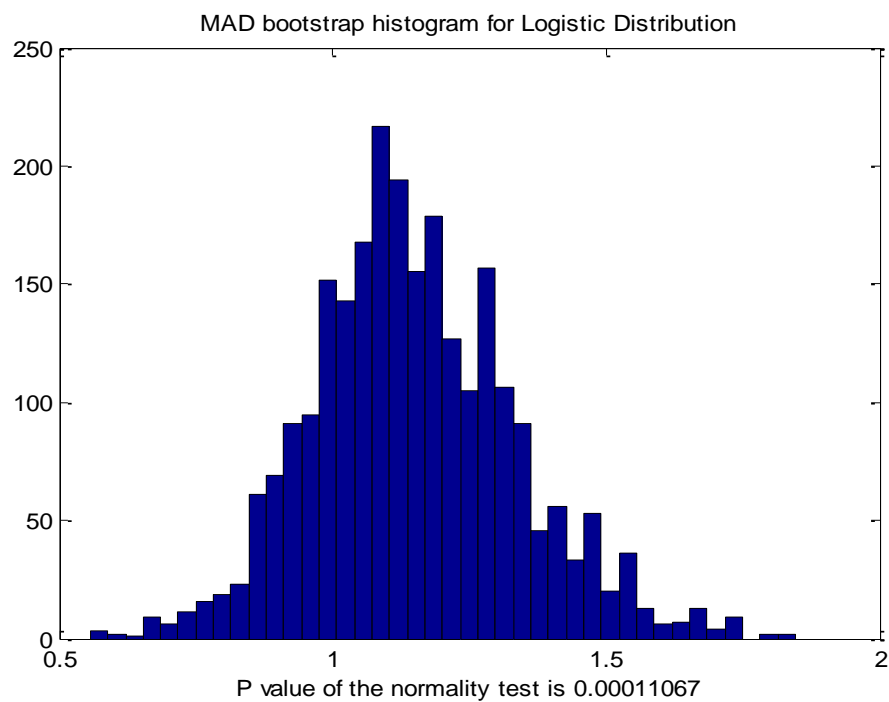


Figure 4.4 Histogram of Sampling distribution of MAD, based on bootstrap samples, when samples are taken from Logistic Distribution

The following table shows the ARL values of “MAD.” Random samples of size 50 are taken from standard Logistic distribution.

Table 4.7 Simulated run lengths for MAD control chart of Logistic Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Median Absolute Deviation with n=50 (Logistic Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
301.9620	294.5157	25.9690	25.2847	2.7980	2.8065	1.1240	1.1269
292.7750		24.0920		2.7290		1.1290	
294.1690		24.0510		2.9210		1.1250	
287.6920		25.6990		2.7710		1.1280	
288.9780		24.2520		2.8820		1.1320	
294.7600		25.7720		2.8130		1.1290	
295.7280		26.2250		2.7610		1.1310	
305.3920		25.3580		2.7480		1.1140	
289.4260		25.6920		2.7680		1.1390	
294.2750		25.7370		2.8740		1.1180	
5.5938	=stdev	0.8265	=stdev	0.0647	=stdev	0.0071	=stdev
		LCL	UCL				
	MEAN =	0.5041	1.4055				
	STDEV =	0.1054	0.1848				

Compared to the Normal case, the decrease in ARL_0 value of MAD from 355 to 294 is statistically significant, but it is not as dramatic as that of “Sample Variance.” Still, ARL_0 value of 294 can be interpreted as “practically good.” Variability of ARL_0 statistics is reduced since Coefficient of Variation decreases here to 1.9%. Similar to the Normal case, MAD is efficient in detecting moderate or high shifts but not efficient enough in detecting small shifts.

4.2.2.3 S_n

The following is a histogram of S_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Logistic distribution.

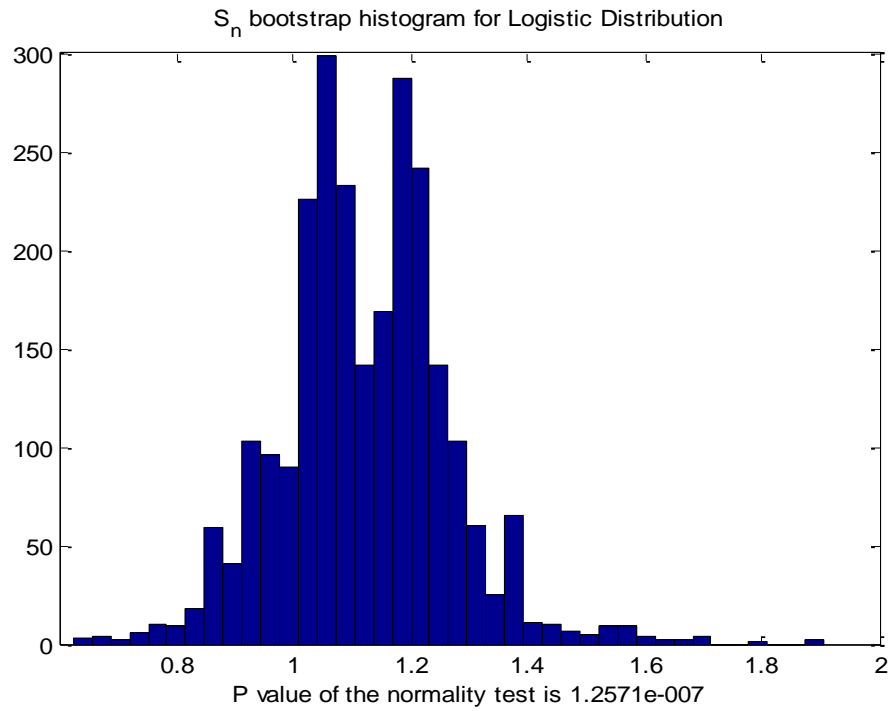


Figure 4.5 Histogram of Sampling distribution of S_n , based on bootstrap samples, when samples are taken from Logistic Distribution

The bootstrap samples histogram of S_n for Logistic case is similar to the Normal case, except that there seems to be more outliers. The p-values of Normality test is again close to zero and sampling distribution is not normal again.

The following table shows the ARL values of S_n Random samples of size 20 are taken from standard Logistic distribution.

Table 4.8 Simulated run lengths for S_n control chart of Logistic Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for S_n with $n=20$ (Logistic Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
322.4760	324.4974	26.5070	26.4653	4.0670	4.0856	1.3440	1.4076
330.8230		26.8720		4.3120		1.4560	
328.3160		27.2330		3.9740		1.3900	
322.3340		25.3840		3.9040		1.4110	
327.2260		25.4780		4.1340		1.3790	
316.9670		26.9480		4.1790		1.4730	
315.6380		27.1880		4.0450		1.4180	
329.1860		24.9030		4.1760		1.3920	
332.4770		26.0280		4.0430		1.3980	
319.5310		28.1120		4.0220		1.4150	
5.9285 =stdev		1.0000 =stdev		0.1173 =stdev		0.0370 =stdev	
		LCL	UCL				
MEAN =		0.2749	1.4964				
STDEV =		0.1093	0.3164				

Confidence interval of S_n is slightly wider than that of MAD. ARL₀ performance of S_n is interesting. There is a significant increase in ARL₀ from 264 and 324, which is contrary to our expectations. Can S_n be more robust than MAD? It is an early inference but it can be. We need to see the performances for other distributions to infer this. The variation in the \bar{R}_0 statistics is slightly different since CoV is 3.2% for Normal and 1.8% for Logistic distributions.

4.2.2.4 Q_n

The following is a histogram of Q_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Logistic distribution. Its shape is very similar to the Normal case and the sampling distribution is not Normal again (p-value = 0.0000).

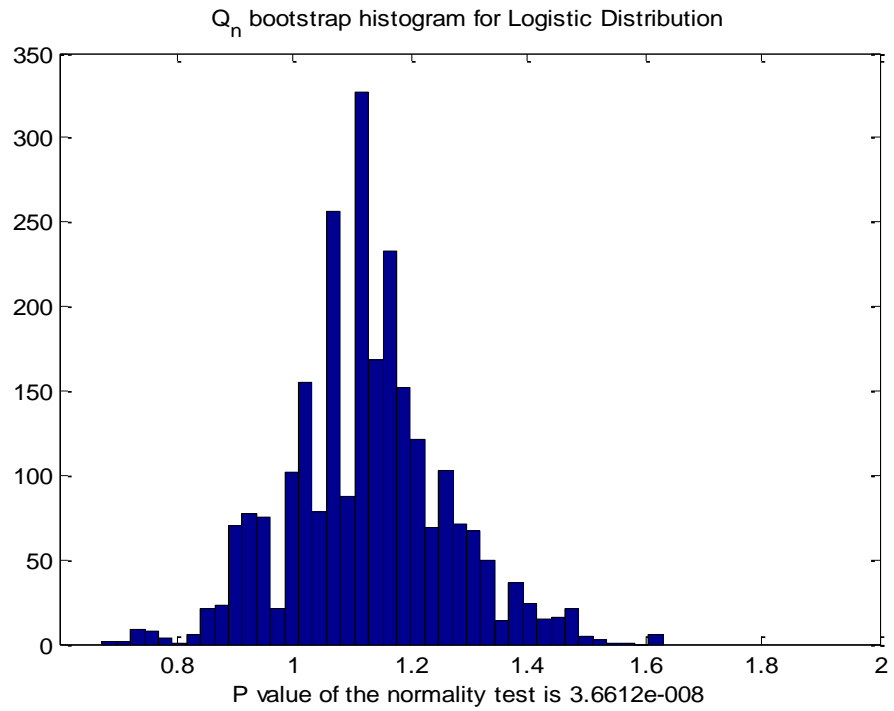


Figure 4.6 Histogram of Sampling distribution of Q_n , based on bootstrap samples, when samples are taken from Logistic Distribution

The following table shows the ARL values of Q_n . Random samples of size 20 are taken from standard Logistic distribution.

Table 4.9 Simulated run lengths for Q_n control chart of Logistic Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Percentile" with $n=20$ (Logistic Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
16.1400	15.5633	3.7290	3.7112	1.4590	1.5296	1.0650	1.0608
15.1300		3.7730		1.6000		1.0680	
15.5160		3.6940		1.4930		1.0590	
15.7890		3.5870		1.5150		1.0570	
16.2740		3.6120		1.4930		1.0590	
15.4000		3.8010		1.6000		1.0600	
15.5660		3.7540		1.5140		1.0590	
15.3840		3.7300		1.5280		1.0610	
14.9780		3.6890		1.5450		1.0670	
15.4560		3.7430		1.5490		1.0530	
0.4071	=stdev	0.0679	=stdev	0.0455	=stdev	0.0046	=stdev
		LCL	UCL				
	MEAN =	0.2630	1.5129				
	STDEV =	0.1150	0.3120				

The following table shows the ARL values of Q_n using "Centered Bootstrap Percentile Method" for constructing the confidence intervals.

Table 4.10 Simulated run lengths for Q_n control chart of Logistic Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on centered bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Centered Percentile" with $n=20$ (Logistic Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
32.8450	33.7421	5.7120	5.8122	1.7710	1.8591	1.0950	1.0978
32.5250		5.8330		1.9300		1.0980	
33.9240		5.7130		1.8080		1.1080	
34.1880		5.8070		1.8430		1.0960	
34.8690		5.6650		1.7780		1.1050	
32.6480		5.8640		1.9330		1.0960	
34.4450		6.0310		1.9000		1.0780	
33.7290		6.0420		1.8540		1.1110	
35.2000		5.5750		1.8760		1.0930	
33.0480		5.8800		1.8980		1.0980	
0.9479	=stdev	0.1514	=stdev	0.0589	=stdev	0.0092	=stdev
		LCL	UCL				
	MEAN =	0.3700	1.6199				
	STDEV =	0.4108	0.5028				

Practically, there is almost no additional comment for the performance of Q_n . Small shift average detection has increased from 3.7 to 5.8 for the two different confidence interval methods, and the difference is significant. It may be interesting to note that performance of Q_n is increased again by using "Centered Percentile Method" but the increase was higher for the Normal case.

The idea for simultaneous use of S_n and Q_n is still valid here looking at the performance measures. Moreover, for the Normal case, it was an alternative idea but for Logistic case, the new idea's overall detection performance is expected to be much better.

4.2.3 Laplace Distribution

4.2.3.1 Sample Variance

The following table shows the ARL values of “Sample Variance.” Random samples of size 50 are taken from Laplace (Double Exponential) distribution centered at zero with variance 1.

Table 4.11 Simulated run lengths for variance control chart of Laplace Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the run lengths for different λ values are at the bottom row of the table.

ARL for Sample Variance with n=50 (Laplace Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
20,6820	20,4096	4,3300	4,2228	1,3120	1,3112	1,0040	1,0064
20,1790		4,3080		1,2880		1,0040	
21,1100		4,2760		1,3000		1,0120	
20,0790		4,1230		1,3050		1,0060	
20,0550		4,0930		1,3000		1,0050	
21,0220		4,3820		1,3610		1,0060	
19,6310		4,1940		1,3570		1,0070	
20,4900		4,0770		1,2830		1,0050	
20,1620		4,2090		1,2780		1,0050	
20,6860		4,2360		1,3280		1,0100	
0,4689	=stdev	0,1034	=stdev	0,0290	=stdev	0,0026	=stdev

Things are getting worse for “Sample Variance” compared to the previous distributions’ performances. It has no practical use for Quality Control purposes since 20 is a very low value for ARL₀ statistics.

4.2.3.2 Median Absolute Deviation

The following is a histogram of MAD for 2500 bootstrap samples constructed by a random sample of size 50 taken from standard Laplace distribution.

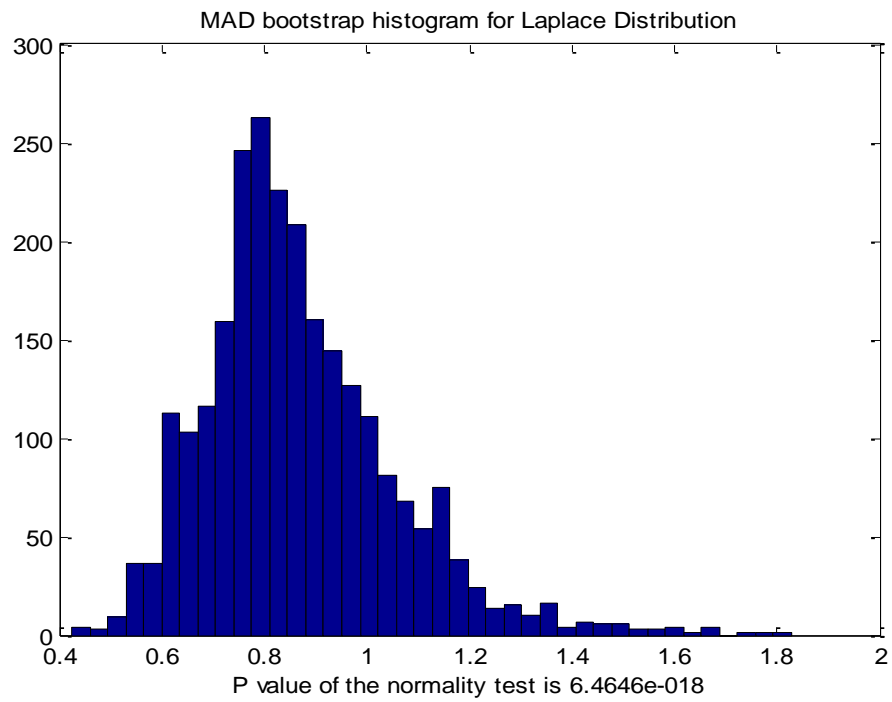


Figure 4.7 Histogram of Sampling distribution of MAD, based on bootstrap samples, when samples are taken from Laplace Distribution.

The sampling distribution of the Bootstrap samples is clearly right skewed, unlike that of the Normal case. Due to the right tail of the histogram, outliers from the upper side can be expected more.

The following table shows the ARL values of MAD. Random samples of size 50 are taken from standard Laplace distribution.

Table 4.12 Simulated run lengths for MAD control chart of Laplace Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Median Absolute Deviation with n=50 (Laplace Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
213.1170	214.7750	33.3150	32.3486	3.9530	4.0626	1.3040	1.3124
216.6350		33.2420		4.1200		1.3490	
220.4300		31.7080		4.0860		1.3260	
206.5180		31.6020		3.9870		1.3270	
217.8650		32.6420		4.1890		1.2900	
212.1960		33.0540		4.0060		1.3380	
220.7480		32.7250		4.1060		1.3090	
214.5800		30.5410		4.0030		1.2880	
211.6170		32.1690		4.0900		1.3040	
214.0440		32.4880		4.0860		1.2890	
4.3306	=stdev	0.8657	=stdev	0.0726	=stdev	0.0216	=stdev
		LCL	UCL				
	MEAN =	0.3877	1.2305				
	STDEV =	0.0933	0.1892				

In fact, the confidence in ARL_0 performance is considerably lost here but it is still much better than “Sample Variance” statistics’ ARL_0 values, and estimated ARL_0 215 is practically not too bad. CoV has decreased to 2.0%. Standard deviations for LCL and UCL are very close to those for Gaussian and Logistic cases.

4.2.3.3 S_n

The following is a histogram of S_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Laplace distribution. The histogram for S_n is alike with that of MAD for Laplace case. The sampling distribution is not normal since p-value for Normality test is 0.0000.

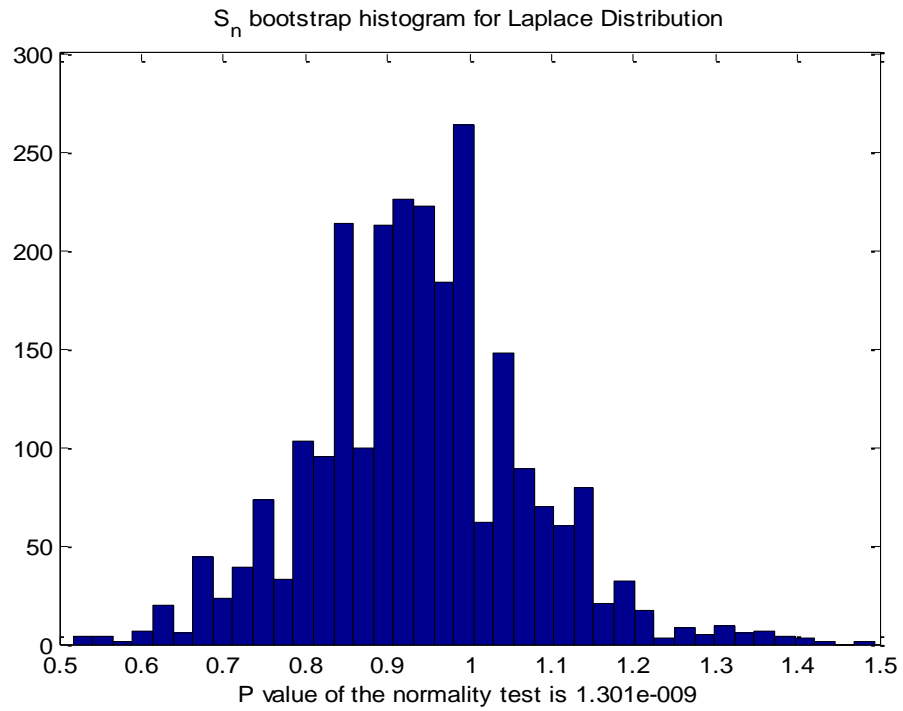


Figure 4.8 Histogram of Sampling distribution of S_n , based on bootstrap samples, when samples are taken from Laplace Distribution.

The following table shows the ARL values of S_n . Random samples of size 20 are taken from standard Laplace distribution.

Table 4.13 Simulated run lengths for S_n control chart of Laplace Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for S_n with $n=20$ (Laplace Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
313.3800	327.2173	40.7300	40.7901	6.8500	6.7097	1.9180	1.9237
328.2700		39.8630		6.8490		1.9110	
337.3530		40.2180		6.5270		1.9130	
331.5720		42.6660		7.0290		1.9500	
328.6380		41.4240		6.6690		1.8860	
308.9050		40.8240		6.5480		1.9180	
323.5030		39.4370		6.5160		1.8700	
322.8920		40.9600		6.5140		1.9040	
342.8500		39.3370		6.6540		1.9560	
334.8100		42.4420		6.9410		2.0110	
10.4678	=stdev	1.1456	=stdev	0.1928	=stdev	0.0400	=stdev
		LCL	UCL				
	MEAN =	0.2321	1.4443				
	STDEV =	0.0945	0.3393				

Simulation results so far show that S_n is a very robust statistics since unlike the other statistics, its ARL_0 performance do not change too much. The difference between 327 and 324 is not significant at all. Difference between 327 and 264 is significant but this change was much more for MAD. Cov is 3.2%; slightly more disperse \bar{R}_0 values compared to Logistic case, and a similar variability characteristic to Normal Case.

4.2.3.4 Q_n

The following is a histogram of Q_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Laplace distribution.

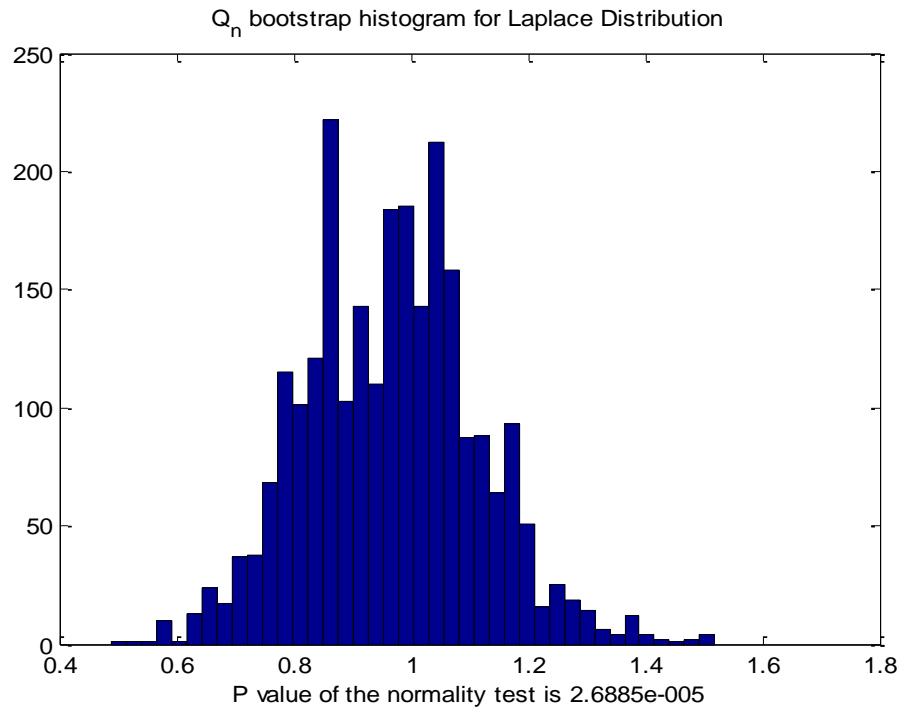


Figure 4.9 Histogram of Sampling distribution of Q_n , based on bootstrap samples, when samples are taken from Laplace Distribution.

The following table shows the ARL values of Q_n . Random samples of size 20 are taken from standard Laplace distribution.

Table 4.14 Simulated run lengths for Q_n control chart of Laplace Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Percentile" with $n=20$ (Laplace Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
26.8260	26.7597	6.3430	6.2592	2.1560	2.1775	1.2020	1.2120
25.4610		6.3880		2.1830		1.2100	
28.1450		6.1160		2.1700		1.2080	
26.8650		6.3030		2.2020		1.2260	
25.6230		6.2610		2.1650		1.2260	
26.6980		6.2750		2.1450		1.2080	
27.0310		6.2150		2.1570		1.1940	
25.8940		6.2240		2.1370		1.2200	
27.0790		6.0400		2.1790		1.2130	
27.9750		6.4270		2.2810		1.2130	
0.9018 =stdev		0.1182 =stdev		0.0410 =stdev		0.0101 =stdev	
		LCL	UCL				
MEAN =		0.2319	1.4993				
STDEV =		0.0947	0.3798				

The following table shows the ARL values of Q_n using "Centered Bootstrap Percentile Method" for constructing the confidence intervals.

Table 4.15 Simulated run lengths for Q_n control chart of Laplace Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on centered bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Centered Percentile" with $n=20$ (Laplace Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
17.9070	18.2837	4.9560	4.9164	1.9040	1.9288	1.1580	1.1615
16.9990		4.9450		1.9090		1.1600	
19.1460		4.9530		1.9200		1.1500	
18.9000		4.9690		1.9260		1.1730	
17.7890		4.9500		1.8960		1.1750	
18.2440		4.8570		1.8860		1.1530	
18.1160		4.8560		1.9360		1.1480	
18.1800		4.7350		1.9280		1.1620	
18.4860		4.9040		1.9550		1.1660	
19.0700		5.0390		2.0280		1.1700	
0.6549 =stdev		0.0836 =stdev		0.0402 =stdev		0.0095 =stdev	
		LCL	UCL				
MEAN =		0.1735	1.4409				
STDEV =		0.4217	0.5596				

The Q_n statistics has very similar performance to the previous distributions. Unlike the previous ones, ARL_0 value is less for the "Centered Percentile Method" than the "Percentile Method" but both are insufficient for practical use. Here, simultaneous use of S_n and Q_n -“Centered Percentile” may be a better idea than simultaneous use of S_n and Q_n -“Percentile” but since we do not know the distribution in practice, we can pass over this slight change and keep the idea, which is the simultaneous use of S_n and Q_n -“Percentile.”

4.2.4 Cauchy Distribution

Cauchy Distribution is a very special distribution with its interesting properties. If we let Z_1 and Z_2 be standard normal random variables, $C = \frac{Z_1}{Z_2}$ is a standard Cauchy random variable. Since Z_1 has mean 0, standard Cauchy random variable is said to be

“Centered at zero” (Actually, the median is zero). However, it has no mean (and so, no variance) because as the denominator term values get close to zero C tends to go plus or minus infinity. In fact, by a shift for Cauchy random variable, it is not meant a shift in standard deviation units, but a shift in the variable itself. Namely, the shift is not $\lambda * \sigma$ but $\lambda * C$

4.2.4.1 Sample Variance

The following table shows the ARL values of “Sample Variance.” Random samples of size 50 are taken from standard Cauchy distribution.

Table 4.16 Simulated run lengths for variance control chart of Cauchy Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Standard deviation of the run lengths for different λ values are at the bottom row of the table.

ARL for Sample Variance with n=50 (Cauchy Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
1.0000	1.0001	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0010		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
1.0000		1.0000		1.0000		1.0000	
0.0000	=stdev	0.0000	=stdev	0.0000	=stdev	0.0000	=stdev

The “Sample Variance” chart does not work for Cauchy Distribution because its control limits are designed as relatively small constants based on Normal distribution. However, Cauchy random variable will have a few very large values in its sample which will result in very high “Sample Variance” values. As the simulation table shows, all the values are out of control limits and all the ARL values are equal to 1.

4.2.4.2 Median Absolute Deviation

The following is a histogram of MAD for 2500 bootstrap samples constructed by a random sample of size 50 taken from standard Cauchy distribution.

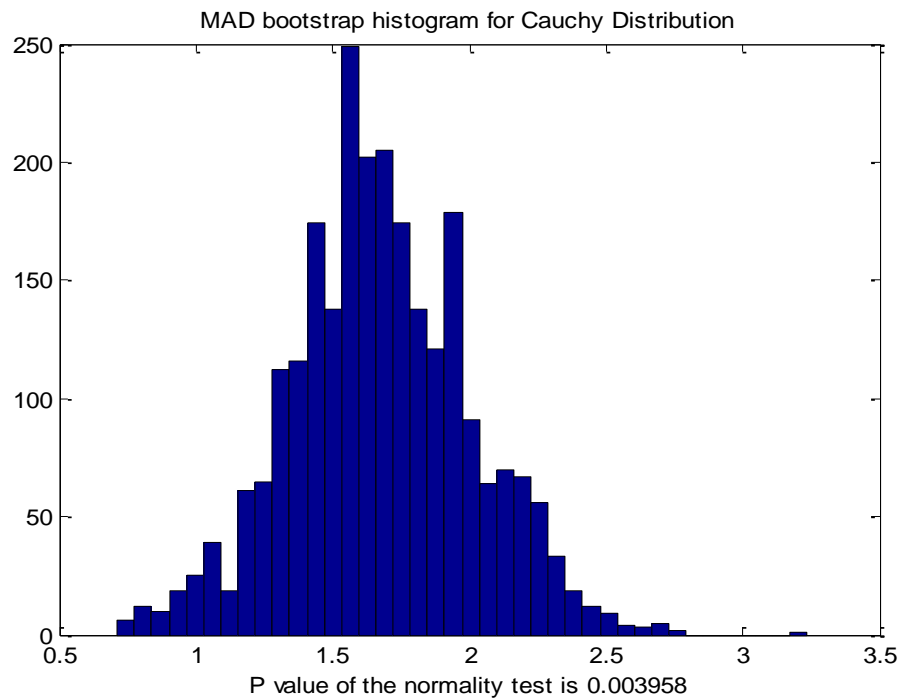


Figure 4.10 Histogram of Sampling distribution of MAD, based on bootstrap samples, when samples are taken from Cauchy Distribution.

The bootstrap samples histogram seems to be almost symmetric and bell shaped but its kurtosis is much higher than a usual normal distribution. Its right tail seems to be little long, allowing outliers there. Sampling distribution is not normal with p-value = 0.0040.

The following table shows the ARL values of MAD. Random samples of size 50 are taken from standard Cauchy distribution.

Table 4.17 Simulated run lengths for MAD control chart of Cauchy Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Median Absolute Deviation with n=50 (Cauchy Distribution)							
Lambda = 1.0		Lambda = 1.2		Lambda = 1.5		Lambda = 2.0	
sims	ARL_1	sims	ARL_1	sims	ARL_1	sims	ARL_1
233.2240	232.7874	57.1200	60.8988	7.9670	8.0270	1.8040	1.7895
243.2990		60.7920		7.8260		1.7670	
229.6540		61.9300		7.4420		1.7750	
235.3870		61.9790		8.0610		1.7810	
218.6470		59.0230		7.8680		1.7660	
237.4540		61.7520		8.4090		1.7410	
233.9540		56.7520		7.9520		1.7990	
232.6520		64.9990		8.1510		1.8030	
228.7120		60.5060		8.2010		1.8820	
234.8910		64.1350		8.3930		1.7770	
6.4239 =stdev		2.6957 =stdev		0.2875 =stdev		0.0378 =stdev	
		LCL UCL					
MEAN =		0.7786 2.8496					
STDEV =		0.1543 0.6053					

The ARL_0 value of 233 is between Logistic distribution and Laplace distribution, and can be accepted as a good performance. CoV of ARL_0 is 2.5%. However, small shift ARL_1 value of 61 is extremely poor in detecting small shifts. This value was around 20 for other distributions. Large shift ARL_1 values are also higher than the previous ones, but they are not bad at all and may be considered as practically acceptable.

Since MAD is a robust statistics with 50% breakdown point, it is not affected by a few very large values whereas “Sample Variance” does. The confidence interval is only a little wider than the other distributions studied. However, standard deviations of UCL and LCL are almost two to three times of the previous ones.

4.2.4.3 S_n

The following is a histogram of S_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Cauchy distribution. Histogram of S_n looks similar to MAD in shape.

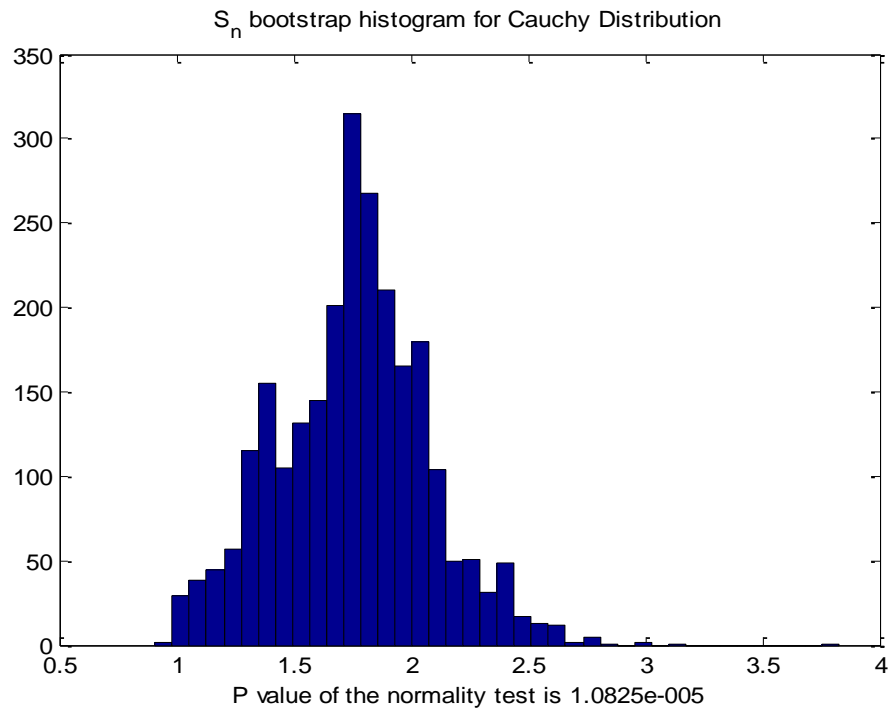


Figure 4.11 Histogram of Sampling distribution of S_n , based on bootstrap samples, when samples are taken from Cauchy Distribution.

The following table shows the ARL values of S_n . Random samples of size 20 are taken from standard Cauchy distribution.

Table 4.18 Simulated run lengths for S_n control chart of Cauchy Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for S_n with $n=20$ (Cauchy Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
821.7320	798.4701	237.7200	239.3869	45.9710	49.0875	9.1160	9.1280
812.1120		245.4310		49.7350		8.8850	
767.3120		246.1570		48.9500		8.9740	
759.1600		237.1630		48.9590		9.0920	
791.8170		231.8050		50.1740		9.1130	
802.4860		228.3210		48.2340		9.3410	
814.6000		240.4530		48.6760		9.0930	
793.7060		247.2680		50.7110		9.1420	
818.9840		237.8280		49.6750		9.4970	
802.7920		241.7230		49.7900		9.0270	
21.1505	=stdev	6.1569	=stdev	1.3216	=stdev	0.1756	=stdev
		LCL	UCL				
	MEAN =	0.4666	4.8729				
	STDEV =	0.2009	2.8635				

ARL values of S_n are very interesting for Cauchy distribution in that they are quite different from the previous distributions, and MAD of Cauchy case. Although CoV for ARL₀ 2.7% is not too large, standard deviation of UCL is more than half of mean UCL.

The overall performance can be considered as bad, but from the opposite point of view. That is, they are unacceptably large and give no idea about a possible shift in a short run. To observe such a high shift as 1.5σ , one should make an additional 49 observations on the average and its cost may be too high in practice.

4.2.4.4 Q_n

The following is a histogram of Q_n for 2500 bootstrap samples constructed by a random sample of size 20 taken from standard Cauchy distribution. The histogram is similar to that of S_n except that histogram is considerably right skewed.

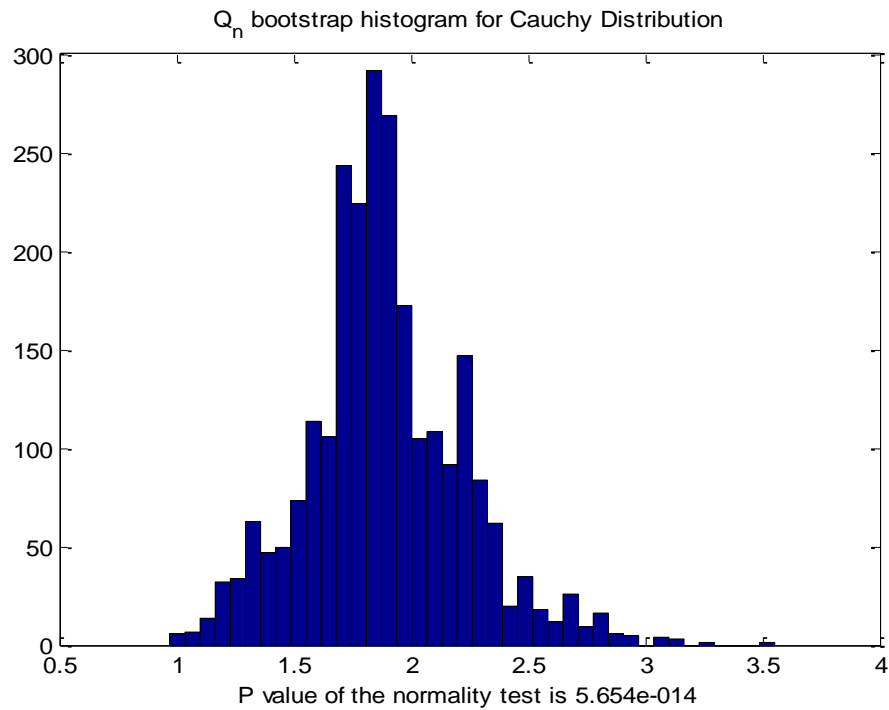


Figure 4.12 Histogram of Sampling distribution of Q_n , based on bootstrap samples, when samples are taken from Cauchy Distribution.

The following table shows the ARL values of Q_n . Random samples of size 20 are taken from standard Cauchy distribution.

Table 4.19 Simulated run lengths for Q_n control chart of Cauchy Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Percentile" with $n=20$ (Cauchy Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL ₀	sims	ARL ₁	sims	ARL ₁	sims	ARL ₁
191.6200	191.3837	50.7930	53.8334	13.8570	14.1320	3.9270	3.8517
192.2360		56.0900		13.8960		3.7170	
174.7040		53.3410		14.1040		3.8110	
181.4880		54.1980		14.6150		3.7400	
200.1790		56.3620		14.2040		3.8310	
202.6560		53.2080		14.1170		3.8530	
189.3920		54.0570		14.1010		3.8960	
183.4160		53.3890		14.1360		4.0980	
208.5250		53.8810		14.2930		3.8810	
189.6210		53.0150		13.9970		3.7630	
10.2509 =stdev		1.5782 =stdev		0.2150 =stdev		0.1104 =stdev	
		LCL UCL					
MEAN =		0.4719 5.5376					
STDEV =		0.2283 3.6103					

Q_n has serious problems to be used practically. Although 191 is an acceptable mean level for ARL_0 , ARL_1 values of 54, 14 and 4 for corresponding shifts of 1.2 1.5 and 2 are too high to be acceptable for detection. Moreover, the standard deviation of UCL is enormously large which means that sampling error is too high and performance measures are highly dependent on the bootstrap samples taken. This shows that, we cannot see the robust characteristics of Q_n for Cauchy distribution using "Percentile Method."

The following table shows the ARL values of Q_n using "Centered Bootstrap Percentile Method" for constructing the confidence intervals. Since the control limits are obtained from another random sample, their standard deviations are different from the previous Q_n table. (If same sample and same bootstrap samples were used, standard deviations of UCL and LCL would be exactly the same as those of "Percentile Method", or vice versa.)

Table 4.20 Simulated run lengths for Q_n control chart of Cauchy Data and the average run length, when the process is out of control with a $\lambda\sigma$ shift. Next, standard deviations of the run lengths for different λ values are given. The bottom part consists of the Control Limits based on centered bootstrap percentile confidence interval, and their standard deviation.

ARL for Q_n "Centered Percentile" with $n=20$ (Cauchy Distribution)							
$\lambda = 1.0$		$\lambda = 1.2$		$\lambda = 1.5$		$\lambda = 2.0$	
sims	ARL₀	sims	ARL₁	sims	ARL₁	sims	ARL₁
11.1280	11.4421	4.9370	4.9724	2.4280	2.3488	1.3340	1.3344
11.4340		4.9020		2.3900		1.3780	
11.3970		4.9610		2.3730		1.3120	
11.6010		4.9090		2.3420		1.3630	
11.2560		4.8800		2.3050		1.3710	
12.1720		5.0450		2.2970		1.3190	
11.2510		5.0890		2.2900		1.3040	
11.0540		5.2600		2.3370		1.2950	
11.3610		4.9070		2.3960		1.3480	
11.7670		4.8340		2.3300		1.3200	
0.3321	=stdev	0.1265	=stdev	0.0464	=stdev	0.0292	=stdev
		LCL	UCL				
	MEAN =	0.0000	3.5525				
	STDEV =	1.4590	3.8872				

Unlike the previous distributions, there is a dramatic decrease in ARL values for Q_n using the "Centered Percentile Method." This may be a result of the fact that the distribution of Q_n (based on the histogram) is right skewed.

Although the performance measures are bad and the standard deviations of the control limits are high, this is the only statistics of the kind that is proposed as a pair with "simultaneous use" for good detection performance. The ARL_1 values are quite good in detecting shifts, even if the shift is small.

Simultaneous use of MAD and Q_n "Centered" will yield good detection performance and low probability of type one error for control purposes. For the other distributions, the proposed charts were S_n and Q_n (per) respectively. It means that, if the distribution of the data is not known for Cauchy case, it will be very hard to design a good detector.

4.3 Proposed Control Designs

Rookie used to think that “Life is Random...”

He still thinks so, but Pupil has updated his way of thinking so as to infer “...but not that much!” ...

She offered him to consider the relationship between exponential distribution and Erlang distribution. A typical example to reveal this relationship is as follows:

Consider that a task’s finishing time T follows an exponential distribution.

We have,

$$T \sim \text{Exponential}(\lambda) \quad (4.12)$$

$$E(T) = \frac{1}{\lambda} \text{ and } \text{Var}(T) = \frac{1}{\lambda^2} \quad (4.13)$$

(Taylor & Karlin, 1998).

If the task can be divided into k equivalent exponential events, each of the events $T_i, i = 1, 2, \dots, k$ will follow an exponential distribution with rate parameter $k * \lambda$. Then,

$$T = T_1 + T_2 + \dots + T_k \quad (4.14)$$

The distribution of T becomes an Erlang distribution (Gamma distribution) with parameters k and $k * \lambda$ ($\alpha = k$ and $\beta = k * \lambda$).

$$T \sim E_k(\lambda) \quad (4.15)$$

Now, we have,

$$E(T) = \frac{1}{\lambda} \text{ and } Var(T) = \frac{1}{k\lambda^2} \quad (4.16)$$

(Taylor & Karlin, 1998).

The corollary of this fact is extremely interesting for me. Although mean finishing time does not change, if one can find a way to divide the task into parts without changing the lifetime distribution of each subtask, then this person can reduce the variability of the finishing time. Moreover, as the number of subtasks goes to infinity, the task itself becomes a deterministic one!

This philosophy makes me contemplate two crucial facts...

One of my conversations in the past comprises the former one. Once, I was talking to one of my legendary friends, and he is also one of the most creative scientists I've ever met. He told me that, if the life itself was random, then why did we need to make Statistics? My answer was somehow clear: "We need Statistics to cope with randomness!" Once upon a time, I was just a Rookie, with my strong feelings, but poor capability in explanations. To be able to give Exponential-Erlang relationship example took my ten years of deal...

The latter fact is that "the heaviness of tail" characteristic somehow resembles to the change of the number of subtasks: k . Letting k equal to 1 is similar to modeling a Cauchy distribution and higher values of k lets the distribution model be Laplace, Logistic and Gaussian correspondingly. The similarity here is not the decreasing variance, but the decreasing kurtosis.

Obviously, a true model for the population distribution is very important, but what if the population parameters are also random variables or the population itself is also changing?

Let's say, arbitrary 30% of the time, the data is generated by Laplace, 50% by Logistic and the remainder by Gaussian distributions. Even if this information is on hand, controlling the population parameters steadily is very difficult in classical sense. That's why we need distribution free, namely robust estimators.

The simulations performed in the previous subchapter revealed that alternative robust charts to the usual variance chart perform much better for the non-normal distributions, and the performance is also close for the normal case. Following table shows "summary statistics" of all the works done, which is based on the result of this research. Gray shaded rows show the best performing charts for each distribution.

Table 4.21 Summary of performance measures for the four distributions and the five charts used.

Performances Summary							
Distribution	Statistics	$\lambda = 1.0$	$\lambda = 1.2$	$\lambda = 1.5$	$\lambda = 2.0$	SE(R_0)	Cov of ARL ₀
Gaussian	Variance	372.22	5.95	1.12	1.00	14.76	3.97%
	MAD	355.39	17.50	2.25	1.07	10.04	2.82%
	S _n	264.24	16.75	2.78	1.21	8.56	3.24%
	Q _n -Per	12.18	2.81	1.31	1.03	0.53	4.39%
	Q _n -Cent	43.16	5.40	1.66	1.07	1.03	2.38%
Logistic	Variance	61.40	4.86	1.22	1.00	2.76	4.50%
	MAD	294.52	25.28	2.81	1.13	5.59	1.90%
	S _n	324.50	26.47	4.09	1.41	5.93	1.83%
	Q _n -Per	15.56	3.71	1.53	1.06	0.41	2.62%
	Q _n -Cent	33.74	5.81	1.86	1.10	0.95	2.81%
Laplace	Variance	20.23	4.20	1.30	1.01	0.53	2.62%
	MAD	214.78	32.35	4.06	1.31	4.33	2.02%
	S _n	327.22	40.79	6.71	1.92	10.47	3.20%
	Q _n -Per	26.76	6.26	2.18	1.21	0.90	3.37%
	Q _n -Cent	18.28	4.92	1.93	1.16	0.65	3.58%
Cauchy	Variance	1.00	1.00	1.00	1.00	0.00	0.00%
	MAD	232.79	60.90	8.03	1.79	6.42	2.76%
	S _n	798.47	239.39	49.09	9.13	21.15	2.65%
	Q _n -Per	191.38	53.83	14.13	3.85	10.25	5.36%
	Q _n -Cent	11.44	4.97	2.35	1.33	0.33	2.90%

4.3.1 Proposed Design for Finite Moment Symmetric Distributions

Under Gaussian distribution, sample variance chart performs the best, as might be expected. MAD's performance is close to sample variance in ARL₀, but is poor in detecting especially small shifts. MAD clearly outperforms S_n, but S_n is still satisfactory. In terms of ARL₀ performance, Q_n is not practically applicable for both methods, but interestingly, its detection performance is very good, especially for the "percentile method" case.

This fact gives the idea of simultaneous use of Q_n-per and one of the other two robust charts. In this design, intuitively, Q_n-per's response and the other chart's

response will be considered together to infer that “process is out of statistical control.”

For quality purposes, GES seems to be a more important characteristic than relative efficiency. Although Q_n is the most efficient robust statistics of the three, its poor GES makes it underperform compared to the others in terms of ARL_0 . On the other hand, S_n is the best neither in efficiency nor in GES, but its performance is the best for ARL_0 at heavy tailed distributions.

Looking at the results for Logistic and Laplace distributions, it is inferred that as heaviness of tail increases, MAD's and Q_n -cent's ARL_0 performances are moderately decreasing and both S_n 's and Q_n -per's are moderately increasing. Since S_n 's performance is also satisfactory under Gaussian data, the proposed design appears to be simultaneous use of S_n and Q_n -per. There is a considerable performance loss of sample variance chart, which clearly supports the non-robustness of this statistics. Its use in standard deviation control will result in a considerable increase in production costs in practice.

Before interpreting the results of Cauchy case, a formal definition of the proposed design, followed by the false alarm and detection probabilities in comparison with the control chart: “sample variance chart,” will be given. Since Cauchy is an extreme case, its results will not be included in this design, and the reason will be explained later again, but in more detail. Moreover, a new design will be proposed for Cauchy model, provided that the model is known to the designer.

An excerpted part of the previous table is given below, which shows the simulated average run length values for the cases when the process is in control and for the case when there is a $\lambda = 1.2$ shift in the standard deviation of the process. These values are shown for Gaussian, Logistic and Laplace distributions. Since the ARL_1 values are very close to each other for these three distributions, their average is calculated and this value is inserted in the corresponding value of each distribution. The last two column stands for the parameter of the corresponding run length random variable,

where $p_0 = \frac{1}{ARL_0}$ and $p_1 = \frac{1}{ARL_1}$. Clearly, these parameters, respectively, stand for the false alarm and detection probabilities of the decision that “Process is out of statistical control” for a single sample.

Table 4.22 ARL_0 and ARL_1 at $\lambda = 1.2$ values Mean and parameters of corresponding Run Length random variables of the sample variance chart. Mean ARL_1 column is the mean of three distribution's ARL_1 values.

Distribution	ARL_0	Mean ARL_1	p_0	p_1
Gaussian	372.22	5.00	0.0027	0.1998
Logistic	61.40	5.00	0.0163	0.1998
Laplace	20.23	5.00	0.0494	0.1998

The usual control process for sample variance chart (or its standard deviation counterpart, *Shewart S Chart*) detects a shift when a single observation gives an out of control signal. Considering the process standard deviation as $\sigma_p = \lambda\sigma$, the hypothesis testing of the control process is as follows:

$H_0: \lambda = 1$ (The process standard deviation is in statistical control)

$H_A: \lambda > 1$ (There is a shift in process standard deviation)

Test statistics is $R_V(t)$, $t = 0, 1, 2, \dots$

Reject H_0 if $R_V(t + 1) = 1$ (4.17)

$R_V(t)$ is the run length of the process, which is defined as the following discrete time Markov chain:

$\{R_V(t), \quad t = 0, 1, 2, \dots\}$

$SS = \{1, 2, \dots\}$

$$R_V(0) = 1$$

$$R_V(t + 1) = \begin{cases} 1, & \text{if sample variance at period } t \text{ is out of control} \\ R_V(t) + 1, & \text{otherwise} \end{cases} \quad (4.18)$$

Then, $R_V(t)$ is the number of “in control signals” after the last “out of control signal” at the beginning of sample period t . A stochastic process is a Markov chain if it holds Markovian property. Markovian property states that, if the process’ value at time t is known, probabilities of the possible values for time $t+1$ can be calculated independent from the previous information about the process (Taylor & Karlin, 1998).

To calculate false alarm and detection probabilities, we need to define run length variables, and relate the Markov chain to these variables. Let, R_{V0} and R_{V1} be the run length random variables for the cases, which are “process is in statistical control” and “there is a 1.2 shift in process standard deviation,” respectively. These are geometric random variables and each of their parameters can be estimated from the simulation results. Therefore,

$$R_{V0} \sim \begin{cases} \textit{Geometric}(p_0 = 0.0027), & \text{if data is Gaussian} \\ \textit{Geometric}(p_0 = 0.0163), & \text{if data is Logistic} \\ \textit{Geometric}(p_0 = 0.0494), & \text{if data is Laplace} \end{cases}$$

$$R_{V1} \sim \textit{Geometric}(p_1 = 0.1998) \quad (4.19)$$

Now, detection (power) and false alarm probabilities are calculated as follows:

$$P_D = 1 - \beta = P\{\textit{Out of control decision}; \textit{Process is out of control}\}$$

$$= P\{R_V(t + 1) = 1\} = P(R_{V1} = 1) = p_1 \quad (4.20)$$

$$P_{FA} = \alpha = P\{\text{Out of control decision; Process is in control}\}$$

$$= P\{R_V(t + 1) = R_v(t) + 1\} = P(R_{V0} = 1) = p_0 \quad (4.21)$$

The proposed design aims to perform an equal-power test with the current one, but with less false alarm probabilities. Since S_n and Q_n -Per charts have relatively close mean ARL values for the three distributions, mean of their ARL values can be safely used to estimate corresponding run length variable's parameter. The following table shows the excerpted part of Table 4.21 for S_n and Q_n -Per charts, where the mean is taken among the three distributions: Gaussian, Logistic and Laplace, and the p values are reciprocals of the corresponding means:

Table 4.23 ARL_0 and ARL_1 at $\lambda = 1.2$ values of mean and parameters of corresponding Run Length random variables to be used for proposed design.

Statistics	Mean ARL_0	Mean ARL_1	p_0	p_1
S_n	305.3198	28.0017	0.0033	0.0357
Q_n -Per	18.1666	4.2589	0.0550	0.2348

The proposed design asserts the following: Use S_n and Q_n charts simultaneously and decide the fact that “process is out of statistical control” when each chart's run length stochastic process is less than their corresponding pre-determined values. In other words, instead of the memoryless decision that takes each single sample statistics into account, the cumulative information obtained by both charts' Markov chains will be used as test statistics. The following hypothesis testing of the process represents design in a formal manner:

$$H_0: \lambda = 1 \text{ (The process standard deviation is in statistical control)}$$

$$H_A: \lambda > 1 \text{ (There is a shift in process standard deviation)}$$

$$\text{Test statistics are } R_S(t), t = 0,1,2, \dots \text{ and } R_Q(t), t = 0,1,2, \dots$$

$$\text{Reject } H_0 \text{ if } R_Q(t + 1) \leq C_Q \text{ and } R_S(t + 1) \leq C_S \quad (4.22)$$

Like $R_V(t)$, $R_Q(t)$ and $R_S(t)$ are the number of “in control signals” after the last “out of control signal” for the corresponding charts at the beginning of sample period t . Initial values are arbitrary large values, which state that initially, the process is in control. Following discrete time Markov chains show the formal definitions:

$$\{R_Q(t), \quad t = 0, 1, 2, \dots\}$$

$$SS = \{1, 2, \dots\}$$

$$R_Q(0) = 50$$

$$R_Q(t + 1) = \begin{cases} 1, & \text{if } Q_n \text{ value at period } t \text{ is out of control} \\ R_Q(t) + 1, & \text{otherwise} \end{cases} \quad (4.23)$$

$$\{R_S(t), \quad t = 0, 1, 2, \dots\}$$

$$SS = \{1, 2, \dots\}$$

$$R_S(0) = 50$$

$$R_S(t + 1) = \begin{cases} 1, & \text{if } S_n \text{ value at period } t \text{ is out of control} \\ R_S(t) + 1, & \text{otherwise} \end{cases} \quad (4.24)$$

A similar pattern is followed to calculate false alarm and detection probabilities of this proposed design. Let, R_{Q0} , R_{Q1} , R_{S0} and R_{S1} are the Q_n and S_n run length random variables for the cases, where “process is in statistical control” and “there is a 1.2 shift in process standard deviation”, respectively. Therefore,

$$R_{Q0} \sim \text{Geometric}(p_0 = 0.0550) \quad (4.25)$$

$$R_{Q1} \sim \text{Geometric}(p_1 = 0.2348) \quad (4.26)$$

$$R_{S0} \sim \text{Geometric}(p_0 = 0.0033) \quad (4.27)$$

$$R_{S1} \sim \text{Geometric}(p_1 = 0.0357) \quad (4.28)$$

What is the performance of this design? Its false alarm and detection probabilities (P_{FA} and P_D) for changing critical decision points should be calculated for comparison purposes. $F(c)$ is the corresponding geometric cumulative distribution function:

$$\begin{aligned} P_D &= 1 - \beta = P\{\text{Out of control decision; Process is out of control}\} \\ &= P\{R_Q(t+1) \leq C_Q \text{ and } R_S(t+1) \leq C_S\} \\ &= P\{R_{Q1} \leq C_Q \text{ and } R_{S1} \leq C_S\} = F_{Q1}(C_Q) * F_{S1}(C_S) \end{aligned} \quad (4.29)$$

$$\begin{aligned} P_{FA} &= \alpha = P\{\text{Out of control decision; Process is in control}\} \\ &= P\{R_Q(t+1) \leq C_Q \text{ and } R_S(t+1) \leq C_S\} \\ &= P\{R_{Q0} \leq C_Q \text{ and } R_{S0} \leq C_S\} = F_{Q0}(C_Q) * F_{S0}(C_S) \end{aligned} \quad (4.30)$$

Following table shows the detection probabilities with changing critical decision values of S_n and Q_n . The yellow shaded region consists of the detection probabilities that achieve to exceed the detection probability of sample variance chart for the corresponding proposed design parameters. The light blue shaded intersection point stands for the design in which the exceedance is achieved at minimum. The dark blue shaded point has slightly better detection performance than the light blue point. These two points have equal false alarm probabilities as will be shown in the next table.

Table 4.24 Detection probabilities for the proposed control design for finite moment symmetric distributions

P_D	Critical Q_n				
Critical S_n	1	2	3	4	5
1	0.0084	0.0148	0.0197	0.0235	0.0263
2	0.0165	0.0291	0.0387	0.0461	0.0517
3	0.0243	0.0428	0.0570	0.0679	0.0762
4	0.0318	0.0561	0.0747	0.0890	0.0999
5	0.0390	0.0689	0.0918	0.1093	0.1226
6	0.0460	0.0812	0.1082	0.1288	0.1446
7	0.0528	0.0931	0.1240	0.1477	0.1658
8	0.0593	0.1046	0.1393	0.1659	0.1862
9	0.0655	0.1157	0.1541	0.1834	0.2059
10	0.0716	0.1264	0.1683	0.2003	0.2249
11	0.0774	0.1366	0.1820	0.2167	0.2432
12	0.0830	0.1466	0.1952	0.2324	0.2609
13	0.0885	0.1561	0.2079	0.2476	0.2779
14	0.0937	0.1654	0.2202	0.2622	0.2943
15	0.0987	0.1743	0.2321	0.2763	0.3101

The following table reflects the heart of this research, which shows the false alarm probabilities for corresponding design parameters, and which allows the comparison with that of the sample variance chart for changing distributions.

Table 4.25 False alarm probabilities for the proposed control design for finite moment symmetric distributions

P_{FA}	Critical Q_n				
Critical S_n	1	2	3	4	5
1	0.0002	0.0004	0.0005	0.0007	0.0008
2	0.0004	0.0007	0.0010	0.0013	0.0016
3	0.0005	0.0010	0.0015	0.0020	0.0024
4	0.0007	0.0014	0.0020	0.0026	0.0032
5	0.0009	0.0017	0.0025	0.0033	0.0040
6	0.0011	0.0021	0.0030	0.0040	0.0048
7	0.0012	0.0024	0.0035	0.0046	0.0056
8	0.0014	0.0028	0.0040	0.0052	0.0064
9	0.0016	0.0031	0.0045	0.0059	0.0072
10	0.0018	0.0035	0.0050	0.0065	0.0080
11	0.0020	0.0038	0.0055	0.0072	0.0087
12	0.0021	0.0041	0.0060	0.0078	0.0095
13	0.0023	0.0045	0.0065	0.0085	0.0103
14	0.0025	0.0048	0.0070	0.0091	0.0111
15	0.0026	0.0051	0.0075	0.0097	0.0118

For the chosen levels of $C_Q = 3$ and $C_S = 13$ for the decision criteria (4.22), design achieves a detection performance $P_D = 0.2079$, which was $P_D = 0.1998$ for sample variance chart. It means that, simultaneous use of S_n and Q_n charts has an equal power with this choice of decision parameters (in fact, the proposed design is slightly more powerful) .

However, the changing false alarm probabilities with respect to Gaussian, Logistic, and Laplace distributions of “sample variance chart” were 0.0027, 0.0163, and 0.0494 respectively. Proposed design’s false alarm probability is 0.0065, which is slightly higher than Gaussian of sample variance chart, but is clearly outperforming for Logistic and Laplace cases. Moreover, it does not need a prior estimate for the distribution model.

4.3.2 Proposed Design for Cauchy Model

Having completed the proposal of quality design for Gaussian, Laplace, and Logistic distributions, it is time to interpret the results of the research for Cauchy distribution and propose another design for Cauchy model. To begin with, recalling the performances summary table for the response of sample variance chart to Cauchy distribution will be useful.

Table 4.26 Sample Variance Chart's response to Cauchy distribution

Distribution	Statistics	$\lambda = 1.0$	$\lambda = 1.2$	$\lambda = 1.5$	$\lambda = 2.0$	SE(R_0)	Cov of ARL_0
Cauchy	Variance	1.0001	1	1	1	0	0

Clearly, sample variance chart does not work for Cauchy case at all. The reason is that Cauchy distribution has no standard deviation. Standard Cauchy random variable is the ratio of two independent standard Gaussian random variables. For each of the other three distributions, we used the term “standard” implying that its mean is zero and standard deviation is one, but this is not the case for Cauchy distribution.

On the other hand, our proposed “ S_n and Q_n -per design” also does not work here. Moreover, their responses to Cauchy model are unacceptably high and adding these values to the mean for a more general design will be misleading. This can be explained as follows: The mean value obtained using the responses to the other three distributions, which are close to each other, is used as an estimator for geometric distribution's parameter. Since responses to the Cauchy model are extremely high, the addition of corresponding value to the mean will result in an inconsistent estimator.

Fortunately, we still have something to do for Cauchy model, using the well performed robust charts in this case. The same idea of previous design will be used in order to design a detector to cope with a Cauchy shift. Recall that the shift is not in terms of standard deviation units: $\lambda\sigma$ for a Cauchy random variable (C), but that of the variable itself: λC .

Due to the use of another design for Cauchy model, it is required that the model is known to the designer. In other words, the inference of the previous design is that it is robust for the symmetric distributions which have finite moments. However, the design here is unique to the Cauchy model. For that reason, some applications of the Cauchy distribution will be introduced before the interpretation of performance measures.

In science and engineering (especially the electrical engineering), Signal-to-Noise Ratio (SNR) is defined as the ratio of “the variance of the desired signal” to “the variance of the level of background noise.” Square root of this ratio is called the voltage SNR (Childers, 1997).

If $S(t)$ and $N(t)$ are independent discrete time Markov chains at time $t = 1, 2, \dots$, then $SNR(t)$ will also be a discrete time Markov chain. Moreover, if $S(t)$ and $N(t)$ are modeled as standard Gaussian distributions, then (voltage) $SNR(t)$ will be a Cauchy model. Here, if a ratio is higher than one, then we indicate the case that the signal is more than the noise. Thus, if one may want to control the process $1/SNR(t)$ (which also is a Cauchy model), then an out of control case can be defined as “the noise is shadowing the signal.”

An application in Physics can be given as another example for Cauchy model. Before giving this example, we need to explain Brownian motion: $B(t)$, which is a continuous-time, continuous-space stochastic process, and which models the position of a particle at time t . Einstein showed that the solution of the particle’s diffusion equation is a $B(t)$ process, which is a Gaussian random variable (Taylor & Karlin, 1998).

Belghin L., Sakhno L. and Orsingher E. (2010) have shown that many practical differential equations in physics, like wave equation, equation of vibration of rods and higher order heat equation are specific kinds of Brownian motions. They have also described the motion in more specific examples involving the composition of two independent Brownian motions $B_1(t)$ and $B_2(t)$, whose examples are diffusions

in cracks or the flow of a gas in a fracture. Finally, they reach a Cauchy model, which is the solution of the “space-fractional equation.”

The simulations of our research show that, under Cauchy distribution, the best performers of ARL_0 and ARL_1 are MAD Chart and Q_n -Cent chart, respectively. Hence, the proposed design for Cauchy model should involve the simultaneous use of MAD and Q_n -Cent. Since the formulation of the design is the same as the previous ones, corresponding stochastic processes will not be defined again. The following table gives the required results for these two charts under Cauchy distribution.

Table 4.27 ARL_0 and ARL_1 at $\lambda = 1.2$ values of mean and parameters of corresponding Run Length random variables to be used for Cauchy model design.

Statistics	ARL_0	ARL_1	p_0	p_1
MAD	232.7874	60.8988	0.0043	0.0164
Q_n-Cent	11.4421	4.9724	0.0874	0.2011

Considering “the in control process random variable” C , process random variable is $C_P = \lambda C$, and the hypothesis testing of the control process is as follows:

H_0 : $\lambda = 1$ (The process random variable is in statistical control)

H_A : $\lambda > 1$ (There is a shift in process random variable)

Test statistics are $R_{MAD}(t)$, $t = 0,1,2, \dots$ and $R_Q(t)$, $t = 0,1,2, \dots$

Reject H_0 if $R_Q(t + 1) \leq C_Q$ and $R_{MAD}(t + 1) \leq C_{MAD}$ (4.31)

The following table shows the detection performances for changing values of the control design parameters C_Q and C_{MAD} . Since sample variance chart does not work here, there is not such a comparable detector as in the previous case. Then, to illustrate the use of the Cauchy design, let the designer want to achieve a minimum detection probability of 10%, whose parameter space’s feasible region is shaded with yellow in the table and blue shaded point is the minimum exceedance point.

Table 4.28 Detection probabilities for the Cauchy model design

P_D	Critical Q_n				
Critical MAD	1	2	3	4	5
1	0.0033	0.0059	0.0080	0.0097	0.0111
2	0.0066	0.0118	0.0160	0.0193	0.0220
3	0.0097	0.0175	0.0238	0.0287	0.0327
4	0.0129	0.0232	0.0314	0.0380	0.0432
5	0.0160	0.0287	0.0389	0.0471	0.0536
6	0.0190	0.0342	0.0463	0.0560	0.0638
7	0.0220	0.0396	0.0536	0.0649	0.0738
8	0.0249	0.0449	0.0608	0.0735	0.0837
9	0.0278	0.0501	0.0679	0.0821	0.0934
10	0.0307	0.0552	0.0748	0.0904	0.1029
11	0.0335	0.0602	0.0816	0.0987	0.1123
12	0.0362	0.0652	0.0883	0.1068	0.1216
13	0.0389	0.0701	0.0949	0.1148	0.1306
14	0.0416	0.0748	0.1014	0.1226	0.1396
15	0.0442	0.0796	0.1078	0.1303	0.1484

The design parameters $C_Q = 3$ and $C_{MAD} = 14$ correspond to the solution point of the following optimization problem. Letting the false alarm and detection probability functions of design parameters be assigned as $P_{FA} = \psi_{FA}(C_Q, C_{MAD})$ and $P_D = \psi_D(C_Q, C_{MAD})$, we have:

$$\text{Minimize } z = \psi_{FA}(C_Q, C_{MAD})$$

Subject to:

$$\psi_D(C_Q, C_{MAD}) \geq 10\%$$

$$C_Q \in \mathbb{Z}^+ \text{ and } C_{MAD} \in \mathbb{Z}^+ \quad (4.32)$$

We have the solution $z = 0.0140$, which is shown as the blue shaded intersection point in the following table of false alarm probabilities (objective function) table:

Table 4.29 False alarm probabilities for the Cauchy model design

P_{FA}	Critical Q_n				
Critical MAD	1	2	3	4	5
1	0.0004	0.0007	0.0010	0.0013	0.0016
2	0.0007	0.0014	0.0021	0.0026	0.0031
3	0.0011	0.0021	0.0031	0.0039	0.0047
4	0.0015	0.0029	0.0041	0.0052	0.0063
5	0.0019	0.0036	0.0051	0.0065	0.0078
6	0.0022	0.0043	0.0061	0.0078	0.0094
7	0.0026	0.0050	0.0071	0.0091	0.0109
8	0.0030	0.0057	0.0081	0.0104	0.0124
9	0.0033	0.0064	0.0091	0.0116	0.0139
10	0.0037	0.0070	0.0101	0.0129	0.0155
11	0.0040	0.0077	0.0111	0.0142	0.0170
12	0.0044	0.0084	0.0121	0.0154	0.0185
13	0.0048	0.0091	0.0131	0.0167	0.0200
14	0.0051	0.0098	0.0140	0.0179	0.0215
15	0.0055	0.0105	0.0150	0.0192	0.0229

The corresponding probabilities of the former design were $P_{FA} = 0.065$ and $P_D = 0.2079$. Although the proposed design for Cauchy model does not perform as well as the previously proposed design for finite moment symmetric distributions, it is the best at hand for a Cauchy model, at least for now.

CHAPTER FIVE

CONCLUSION

In a production process of industry, product's quality is aimed to be optimized continuously. Such an optimization entails maximum available quality at minimum cost, which is achieved by standardizing the production level. In order to standardize the production level, say volume of the cola produced, controlling the standard deviation of the process cannot be underestimated.

Usual Shewart S control chart uses each of the periodically taken sample's standard deviation as an estimator for the process standard deviation. Equivalently, one can control the variance of the production process using sample variance chart.

Run Length (R) counts the number of periods between the last out of control signal and that of the recent one. In order to evaluate the performance of a chart, R is an important random variable, which enables us to make comparisons and to achieve relevant inferences. That is, a large value of R is desired to realize "when the process is in control," and that of a small one is required, "when the process is out of control." The reason is that, an "out of control signal" is a false alarm during an "in control case," and is a "detection of a shift," otherwise.

Statistical theory exhibits that sample variance is the best estimator for population variance under Gaussian distribution, but as is stated and supported with the simulation results, sample variance chart's performance is highly dependent on the assumption that the relevant data follows a Gaussian distribution. Its performance becomes too poor for heavy tailed symmetric distributions.

Some robust estimators of population standard deviation in the literature are: "Median Absolute Deviation" (MAD), S_n , and Q_n . These estimators are very robust since they have the maximum available breakdown point, 50%. However, their robustness characteristics under Gaussian distribution change with respect to

“relative efficiency” and “gross error sensitivity” (GES), Q_n being the most efficient, and MAD having the minimum available GES among these three.

For the construction of robust control charts using these robust scale estimators, their standard errors are required. Yet, there is a contradictory problem here, because standard error is a distribution-dependent-parameter.

At first, I tried two formulations using the ideas of *Shewart S chart*, and sample variance chart. Unfortunately, these trials resulted in poor performing charts.

Then, I used the help of another useful statistical method, which is the bootstrap method. Bootstrap method is used to estimate the sampling distribution by taking repeatedly samples from the sample in hand, with replacement. The proposed robust charts with their limits constructed by the bootstrap confidence intervals, perform really well for non-normal symmetric distributions. Moreover, since the robust estimators used have different characteristics in terms of relative efficiency and GES, their control charts' responses to the heaviness of a distribution's tail are different.

Although S_n chart's both ARL_0 and ARL_1 values significantly underperform those of the sample variance chart under Gaussian distribution, they still seem to be practically satisfactory. Interestingly, as the kurtosis of the distribution becomes higher, there is a significant loss in performance of the sample variance chart. However, S_n chart's ARL_0 increases with the increasing kurtosis of the distribution. A pitfall of S_n Chart is its poor performance in detecting shifts. Having seen the good detection performance of Q_n with low ARL_1 values for the finite moment symmetric distributions under study, I decided to consider the idea that S_n and Q_n charts can be designed to perform together.

The first proposed design of the research is the simultaneous use of S_n and Q_n charts, which infers an out of control decision when both charts' run length variables are less than their corresponding predetermined constants. Monte Carlo simulation

study is performed to solve the optimization problem, whose objective is to minimize the false alarm probability of the process, and whose constraint is that the design matches the power of the variance chart. Performance of this optimization scheme is slightly worse than the sample variance chart's false alarm probability under Gaussian case, but significantly outperforms that of the sample variance under the other two distributions: Logistic and Laplace.

Cauchy distribution is an extremely heavy tailed one and its moments are divergent. Since it has no variance, sample variance chart does not even work under Cauchy distribution at all. Cauchy model has some applications in the specific fields of Electrical Engineering and Physics. Provided that the model is known to the designer, simulation results of the research also reveals a control design for Cauchy model. This second proposed design makes simultaneous use of MAD and Q_n charts, whose logic is exactly the same as the previous one. Despite not performing as well as the former one, it is the best at hand for a Cauchy model, at least for now.

All these findings express me the following fact: Having sources limited, there is a natural tradeoff between "the false alarm based acts," and "missed events." To improve the probabilities of both without additional sources may be accomplished by controlling these via two different detectors. I guess the idea, suggesting the simultaneous use of two charts, is a good one...

For the prospective studies that may be ensuant responses to this research, the following suggestions are given:

First of all, MAD assumes a prior estimate for location, and therefore MAD is expected to be good at symmetric distributions. On the other hand, it might be expected that S_n and Q_n will not suffer from asymmetry. Studying the case on asymmetric distributions can be entertaining.

Secondly, any other creative ideas may help standard errors of the robust estimators used in this research to be further developed. Good theoreticians can perform theoretical studies to find such creative ideas. Moreover, real data applications of this research's findings might be helpful.

Finally, there are many fields of Statistics that robust estimation of scale hasn't been tried yet. One can try the estimators MAD, S_n , and Q_n in other applied fields. Honestly, "Statistical Quality Control" was just one of them.

Before finishing it up, I want to complete my story.

"And Pupil said "I love you"..."

Rookie thought, "I wish I could know why"...

But he got the case; his thoughts were dependent on the process itself, not on that of the existence. Also, she got the question, but she decided not to tell ever...

They shared the remaining of their lives through the wings of Glorious Statistics...

What a wonderful life!"

REFERENCES

- Aphorisms by Hippocrates*, (n.d.). Retrieved July 24, 2011, from <http://classics.mit.edu/Hippocrates/aphorisms.1.i.html>.
- Banks, J., Carson II, J. S., Nelson, B. L., & Nicol, D. M. (2005). *Discrete-event System Simulation* (3rd ed.). New Jersey: Prentice-Hall.
- Beghin, L., Orsingher, E. Sakhno, L., (5 August 2010). *Equations of Mathematical Physics and Compositions of Brownian and Cauchy processes*. Retrieved January 31, 2012, from <http://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf>
- Childers, D. G. (1997). *Probability and Random Processes*. USA: Irwin.
- Daszykowski, M., Serneels, S., Kaczmarek, K., Espen, P. V., Croux, C., & Walczak, B., (2007). TOMCAT: A MATLAB Toolbox for Multivariate Calibration Techniques. *Chemometrics and Intelligent Laboratory Systems*, 85 (2), 269-277.
- DeCarlo, L. T. (1997). On the Meaning and Use of Kurtosis. *Psychological Methods*, 2 (3), 292-307.
- Hogg, R. V. & Craig, A. T. (1995) *Introduction to Mathematical Statistics* (5th ed.). New Jersey: Prentice-Hall.
- Klawonn, F. (30 July 2009). *Robust Statistics*. Retrieved August 25, 2011, from http://www.cost-ic0702.org/summercourse/files/robust_statistics.pdf
- Martin, R. D. & Zamar, R. H. (1991). Efficiency Constrained Bias Robust Estimation of Location. *Annals of Statistics*, 21 (1), 338-354.
- Montgomery, D. C. (2009). *Statistical Quality Control* (6th ed.). Singapore: Wiley.

- Nietzsche, F. (1996). *Human, all too Human* (2nd ed.) (aphorism 149). (R. J. Hollingdale, Trans.). New York: Cambridge University Press. (Original work published 1878)
- Rousseeuw, P. J. & Croux. C. (1993). Alternatives to the Median Absolute Deviation. *Journal of the American Statistical Association*, 88 (424), 1273-1283.
- Singh, K. & Xie, M. (n.d.). Bootstrap: A statistical method. Retrieved September 25, 2011, from <http://www.stat.rutgers.edu/home/mxie/RCPapers/bootstrap.pdf>
- Taylor, H. M. & Karlin S. (1998). *An introduction to stochastic modeling* (3rd ed.) London: Academic Press.
- Wilcox, R. R. (2005). *Introduction to Robust Estimation and Hypothesis Testing*. Burlington: Elsevier Academic Press
- Walck, C. (30 September 2007). *Hand-book on Statistical Distributions for Experimentalists*. Retrieved August 21, 2011, from <http://www.fysik.su.se/~walck/suf9601.pdf>
- Quassia amara*, (n.d). Retrieved October 01, 2011, from http://en.wikipedia.org/wiki/Quassia_amara

APPENDIX-1**(Information for Quassia Amara tree)**

Quassia amara is a species in the genus *Quassia*, with some botanists treating it as the sole species in the genus. It is a shrub or rarely a small tree, growing to 3 m tall (rarely 8 m), native to Brazil. The leaves are compound and alternate, 15-25 cm long, and pinnate with 3-5 leaflets, the leaf rachis being winged. The flowers are produced in a panicle 15-25 cm long, each flower 2.5-3.5 cm long, bright red on the outside and white inside. The following figure is an example picture. (Wikipedia, n.d.)



Figure A1.1 *Quassia Amara* tree

APPENDIX – 2

(Factors for Constructing Variables Control Charts)

Observations in Sample, <i>n</i>	Factors for Constructing Variables Control Charts															
	Chart for Standard Deviations						Chart for Ranges									
	Chart for Averages			Chart for Center Line			Chart for Center Line			Chart for Ranges						
	A	A ₂	A ₃	A ₄	A ₅	A ₆	B ₃	B ₄	B ₅	B ₆	d ₂	d ₃	D ₁	D ₂	D ₃	D ₄
2	2.121	1.880	2.659	0.7979	1.2533	0	3.267	0	2.606	1.128	0.8865	0.853	0	3.686	0	3.267
3	1.732	1.023	1.954	0.8862	1.1284	0	2.568	0	2.276	1.693	0.5907	0.888	0	4.358	0	2.575
4	1.500	0.729	1.628	0.9213	1.0854	0	2.266	0	2.088	2.059	0.4857	0.880	0	4.698	0	2.282
5	1.342	0.577	1.427	0.9400	1.0638	0	2.089	0	1.964	2.326	0.4299	0.864	0	4.918	0	2.115
6	1.225	0.483	1.287	0.9515	1.0510	0.030	1.970	0.029	1.874	2.534	0.3946	0.848	0	5.078	0	2.004
7	1.134	0.419	1.182	0.9594	1.0423	0.118	1.882	0.113	1.806	2.704	0.3698	0.833	0.204	5.204	0.076	1.924
8	1.061	0.373	1.099	0.9650	1.0363	0.185	1.815	0.179	1.751	2.847	0.3512	0.820	0.388	5.306	0.136	1.864
9	1.000	0.337	1.032	0.9693	1.0317	0.239	1.761	0.232	1.707	2.970	0.3367	0.808	0.547	5.393	0.184	1.816
10	0.949	0.308	0.975	0.9727	1.0281	0.284	1.716	0.276	1.669	3.078	0.3249	0.797	0.687	5.469	0.223	1.777
11	0.905	0.285	0.927	0.9754	1.0252	0.321	1.679	0.313	1.637	3.173	0.3152	0.787	0.811	5.535	0.256	1.744
12	0.866	0.266	0.886	0.9776	1.0229	0.354	1.646	0.346	1.610	3.258	0.3069	0.778	0.922	5.594	0.283	1.717
13	0.832	0.249	0.850	0.9794	1.0210	0.382	1.618	0.374	1.585	3.336	0.2998	0.770	1.025	5.647	0.307	1.693
14	0.802	0.235	0.817	0.9810	1.0194	0.406	1.594	0.399	1.563	3.407	0.2935	0.763	1.118	5.696	0.328	1.672
15	0.775	0.223	0.789	0.9823	1.0180	0.428	1.572	0.421	1.544	3.472	0.2880	0.756	1.203	5.741	0.347	1.653
16	0.750	0.212	0.763	0.9835	1.0168	0.448	1.552	0.440	1.526	3.532	0.2831	0.750	1.282	5.782	0.363	1.637
17	0.728	0.203	0.739	0.9845	1.0157	0.466	1.534	0.458	1.511	3.588	0.2787	0.744	1.356	5.820	0.378	1.622
18	0.707	0.194	0.718	0.9854	1.0148	0.482	1.518	0.475	1.496	3.640	0.2747	0.739	1.424	5.856	0.391	1.608
19	0.688	0.187	0.698	0.9862	1.0140	0.497	1.503	0.490	1.483	3.689	0.2711	0.734	1.487	5.891	0.403	1.597
20	0.671	0.180	0.680	0.9869	1.0133	0.510	1.490	0.504	1.470	3.735	0.2677	0.729	1.549	5.921	0.415	1.585
21	0.655	0.173	0.663	0.9876	1.0126	0.523	1.477	0.516	1.459	3.778	0.2647	0.724	1.605	5.951	0.425	1.575
22	0.640	0.167	0.647	0.9882	1.0119	0.534	1.466	0.528	1.448	3.819	0.2618	0.720	1.659	5.979	0.434	1.566
23	0.626	0.162	0.633	0.9887	1.0114	0.545	1.455	0.539	1.438	3.858	0.2592	0.716	1.710	6.006	0.443	1.557
24	0.612	0.157	0.619	0.9892	1.0109	0.555	1.445	0.549	1.429	3.895	0.2567	0.712	1.759	6.031	0.451	1.548
25	0.600	0.153	0.606	0.9896	1.0105	0.565	1.435	0.559	1.420	3.931	0.2544	0.708	1.806	6.056	0.459	1.541

For *n* > 25.

$$A = \frac{3}{\sqrt{n}} \quad A_3 = \frac{3}{c_4\sqrt{n}} \quad c_4 = \frac{4(n-1)}{4n-3}$$

$$B_3 = 1 - \frac{3}{c_4\sqrt{2(n-1)}} \quad B_4 = 1 + \frac{3}{c_4\sqrt{2(n-1)}}$$

$$B_5 = c_4 - \frac{3}{\sqrt{2(n-1)}} \quad B_6 = c_4 + \frac{3}{\sqrt{2(n-1)}}$$

APPENDIX – 3

(MATLAB functions for Robust Scale Estimators: MAD, S_n , and Q_n)

The following codes are taken from MATLAB toolbox TOMCAT (Daszykowski, M., Serneels, S., Kaczmarek, K., Espen, P. V., Croux, C., & Walczak, B., 2007). These m-files calculate the robust scale estimators MAD, S_n and Q_n correspondingly for the given data.

MAD:

```
function m=madn(X)

% MAD computes the median absolute deviation of X. If X is
% a matrix, MAD is a row vector containing the MAD's of the
% columns of X.
%
% ! Includes correction for consistency !
%
% Written by S. Serneels, 17.12.2003

[n,p]=size(X);
Xmc=X-repmat(median(X),n,1);
m=1.4826*median(abs(Xmc));

bn=0;
switch n
    case 2
        bn=1.196;
    case 3
        bn=1.495;
    case 4
        bn=1.363;
    case 5
        bn=1.206;
    case 6
        bn=1.200;
    case 7
        bn=1.140;
    case 8
        bn=1.129;
    case 9
        bn=1.107;
    otherwise
        bn = n/(n-0.8);
end

m=bn*m;
```

S_n:

```

function s=sn(y)

% Sn scale estimator
% -----
% Input:  y, matrix of size (n,p)
% -----
% Output: s, vector of size (1,p) containing the Sn scale estimates
of the
% columns of y
% -----
% The Sn estimator is proposed in P.J. Rousseeuw, C. Croux,
Alternatives to
% the median absolute deviation, J. Am. Statist. Assoc., 88 (1993),
% 1273-1283

% Written by Sven Serneels, University of Antwerp

if size(y,2)>1
    if size(y,1)>1
        for i=1:size(y,2)
            s(:,i)=snsven(y(:,i));
        end
    else
        y=y';
        s=snsven(y);
    end
else
    s=snsven(y);
end;

% -----
function s=snsven(y)

n=length(y);
if n>1000
    sy=sort(y);
    nbins=floor(n/10);
    mys=zeros(nbins,1);
    ninbins=floor(n/nbins);
    for i=1:nbins
        if (mod(n,nbins)~=0 && i==nbins)
            mys(i)=median(sy((i-1)*ninbins+1:n));
        else
            mys(i)=median(sy((i-1)*ninbins+1:i*ninbins));
        end
    end
    y=mys;
    n=nbins;
end
pairwisediff=sort(abs repmat(y',n,1)-repmat(y,1,n)));
pairwisediff=pairwisediff(floor((n+1)/2),:);
pairwisediff=sort(pairwisediff);
s=1.1926*(pairwisediff(floor(n/2)+1));

```

```

cn=1;
switch n
    case 2
        cn=0.743;
    case 3
        cn=1.851;
    case 4
        cn=0.954;
    case 5
        cn=1.351;
    case 6
        cn=0.993;
    case 7
        cn=1.198;
    case 8
        cn=1.005;
    case 9
        cn=1.131;
    otherwise
        if (mod(n,2)==1)
            cn=n/(n-0.9);
        end
end
s=cn*s;

```

Q_n:

```

function s=qn(y)

% Qn scale estimator
% -----
% Input:  y, matrix of size (n,p)
% -----
% Output: s, vector of size (1,p) containing the Qn scale estimates
of the
% columns of y
% -----
% The Qn estimator is proposed in P.J. Rousseeuw, C. Croux,
Alternatives to
% the median absolute deviation, J. Am. Statist. Assoc., 88 (1993),
% 1273-1283

% Written by Sven Serneels, University of Antwerp

if size(y,2)>1
    if size(y,1)>1
        for i=1:size(y,2)
            s(:,i)=qnsven(y(:,i));
        end
    else
        y=y';
        s=qnsven(y);
    end
else
    s=qnsven(y);
end

```

```

end;

function s=qnsven(y)

n=length(y);
% Do binning for big n
if n>1000
    sy=sort(y);
    nbins=floor(n/10);
    mys=zeros(nbins,1);
    ninbins=floor(n/nbins);
    for i=1:nbins
        if (mod(n,nbins)~=0 && i==nbins)
            mys(i)=median(sy((i-1)*ninbins+1:n));
        else
            mys(i)=median(sy((i-1)*ninbins+1:i*ninbins));
        end
    end
    y=mys;
    n=nbins;
end
h=floor(n/2)+1;
k=0.5*h*(h-1);
pairwisediff= repmat(y,1,n)-repmat(y',n,1);
pairwisediff=sort(abs(pairwisediff(find(tril(ones(n,n),-1)))));
s=2.2219*(pairwisediff(k));

switch n
    case 1
        error('Sample size too small');
    case 2
        dn=0.399;
    case 3
        dn=0.994;
    case 4
        dn=0.512;
    case 5
        dn=0.844;
    case 6
        dn=0.611;
    case 7
        dn=0.857;
    case 8
        dn=0.669;
    case 9
        dn=0.872;
    otherwise
        if (mod(n,2)==1)
            dn=n/(n+1.4);
        elseif (mod(n,2)==0)
            dn=n/(n+3.8);
        end
end
s=dn*s;

```


APPENDIX – 4

(MATLAB functions used to create the figures and tables in this thesis)

At this final part, I want to give the MATLAB functions that I used in my study with a detailed explanation. I hope that this part will be useful for scientists who want to make similar studies in order to go further...

Basically, two facts change in the tables and figures created. First one is the distribution, and the second one is the scale estimator used. Creating functions for these two purposes will make the works more efficient because otherwise, each table or figure would require another m-file. (This was what I'd done at the beginning). It will be a good starting point to give these two functions. The following m-files generator.m and estimator.m give the random number generator function and estimator used, respectively:

generator:

```
% [datam] = generator(distribution,mu,sigma,n,m)
%
% function generator generates random stream for the
% input distribution.
%
% INPUTS:
% distribution: Distribution Type is the input
% string of the function
% Input
%      'nor' for Normal (Gaussian)
%      'log' for Logistic
%      'de' for Double Exponential (Laplace)
%      'cau' for Cauchy
%
%                                     distributions
% distribution parameters:
% mean: mu
% standard deviation: sigma
%
% matrix size nxm
%
% OUTPUT:
% datam: Random data with given distribution
%
% Written by Alp Giray Özen, 2011

function [datam] = generator(distribution,mu,sigma,n,m)
```

```

switch lower(distribution)
    case('nor')
        datam = mu+sigma*randn(n,m);

    case('log')
        Xuni = rand(n,m);
        datam =mu+sigma*(-sqrt(3)/pi)*log(Xuni./(1-Xuni));

    case('de')
        Xuni = rand(n,m,2);
        datam = mu+sigma*(1/sqrt(2))*log(Xuni(:,:,1)./Xuni(:,:,2));

    case('cau')
        Xtemp = randn(n,m,2);
        datam = Xtemp(:,:,1)./Xtemp(:,:,2);

    otherwise
        disp('No match for this distribution type')
        return
end

```

estimator:

```

% stat = estimator(distribution,X)
%
% function estimator returns the input estimate of the input
% matrix X
%
% INPUTS:
% scale: Estimate of the vector to be returned
% Input
%     'mad' for Median Absolute Deviation
%     'sn' for Sn
%     'qn' for Qn
%     'sd' for Standard Deviation
%     'range' for Range
%     'var' for Variance
%     'mean' for Mean
%
%                                     estimators
%
% X: input matrix whose scale estimator will be calculated
% column wise
%
% OUTPUT:
% stat: a row vector containing the scale estimators of columns X
% if X is a vector (either row or column) stat will be a scalar.
%
% Written by Alp Giray Özen, 2011

function stat = estimator(scale,vector)

switch lower(scale)

```

```

case('mad')
    stat = madn(vector);
case('sn')
    stat = sn(vector);

case('qn')
    stat = qn(vector);

case('sd')
    stat = std(vector);

case('range')
    stat = range(vector);

case('var')
    stat = var(vector);

case('mean')
    stat = mean(vector);

otherwise
    disp('No match for this estimation type')
    return
end

```

Now, it is easy to generate such a table as leaves data of table (2.1), and to calculate the relevant statistics. For example, creation of a table and calculation of the relevant row statistics is as follows:

```

leaves = generator('nor',20,2.5,30,5);
stats = [estimator('mean',leaves); estimator('sd',leaves);...
        estimator('var',leaves); estimator('range',leaves)]];

```

The control charts of the thesis are obtained by the function `intro_charts.m`, which is shown as follows. The inputs and outputs of the function are defined before writing the actual code. Between the explanation and the actual code, there exists my name and my creation year of the function.

intro_charts:

```

% [datam] = intro_charts(distribution,standard,avg,sigma)
%
% function intro_charts draws the Control charts that I used
% in my thesis based on n=5 observations and m=30 subgroups,
% Shewart R Chart, Shewart S Chart, Variance Chart, Shewart

```

```

% X-bar Chart, MAD Chart, S_n Chart and Q_n chart will be plotted
%
%
%   INPUTS:
%   distribution: Distribution Type is the input string
% of the function
% Input
%       'nor' for Normal
%       'log' for Logistic
%       'de' for Double Exponential
%       'cau' for Cauchy
%
%                               distributions
%
%   standard: Standards Known/Unknown is the input string
% of the function
%
% Input
%       'yes' for Quality Standards Known case
%       'no' for Quality Standards estimated from the data case
%
%   avg: Mean and sigma: Standard Deviation of the data
% to be generated
%
%   OUTPUT:
%   data: Random data with given distribution,
% mean and standard deviation
%
%   Written by Alp Giray Özen, 2011

function [datam] = intro_charts(distribution,standard,avg,sigma)

n = 5; m = 30; %define n, m
seed = 1978; %state seed

% Enter constants for Shewart R Chart for n<26

if n<26
    d_2 = [0 1.128 1.693 2.059 2.326 2.534 2.704 2.847 2.970...
           3.078 3.173 3.258 3.336 3.407 3.472 3.532 3.588...
           3.640 3.689 3.735 3.778 3.819 3.858 3.895 3.931];
    d_3 = [0 0.853 0.888 0.880 0.864 0.848 0.833 0.820 0.808...
           0.797 0.787 0.778 0.770 0.763 0.756 0.750 0.744...
           0.739 0.734 0.729 0.724 0.720 0.716 0.712 0.708];
    D_3p = 1-3.*d_3./d_2;
    D_3 = max([D_3p ; zeros(1,25)]);%LCL coefficient, standard=='NO'
    D_4 = 1+3.*d_3./d_2; %UCL coefficient, standard=='NO'
    D_1p = d_2-3*d_3;
    %LCL coefficient, standard=='YES':
    D_1 = max([D_1p ; zeros(1,25)]);
    %UCL coefficient, standard=='YES':
    D_2 = d_2+3*d_3;
    clear D_3p;
end

% Enter constants for Shewart S Chart

c_4 = sqrt(2/(n-1))*gamma(n/2)/gamma((n-1)/2);
B_3p = 1-3*sqrt(1-c_4^2)/c_4;

```

```

B_3 = max([B_3p ; 0]);           %LCL coefficient
B_4 = 1+3*sqrt(1-c_4^2)/c_4;    %UCL coefficient
clear B_3p;

% Enter constants for Variance Chart

U_var = chi2inv(0.99865,n-1)/(n-1);
L_var = chi2inv(0.00135,n-1)/(n-1);

% Enter constants for X-bar Chart
A = 3/sqrt(n);
A_3 = 3/(c_4*sqrt(n));

% Enter constants for MAD Chart, S_n Chart and Q_n Chart

B_5p = (c_4-3*sqrt(1-c_4^2));
B_5 = max([B_5p ; 0]);         %LCL coefficient
B_6 = (c_4+3*sqrt(1-c_4^2));   %UCL coefficient
clear B_5p;

% Create random data for given distribution and return datam

randn('state',seed)
rand('twister',seed)

X = generator(distribution,avg,sigma,n,m);
datam = X';

% Calculation of chart statistics

obs = 1:m;
X_R = estimator('range',X);
X_std = estimator('sd',X);
X_var = estimator('var',X);
X_mean = estimator('mean',X);
X_mad = estimator('mad',X);
X_sn = estimator('sn',X);
X_qn = estimator('qn',X);

% Calculation of Control Chart Limits

switch lower(standard)

    case('no')

        R_bar = mean(X_R);
        std_bar = mean(X_std);
        var_bar = mean(X_var);
        mean_bar = mean(X_mean);
        mad_bar = mean(X_mad);
        sn_bar = mean(X_sn);
        qn_bar = mean(X_qn);

        if n<26
            LCL_R = D_3(1,n)*R_bar*ones(1,m);

```

```

        CL_R = R_bar*ones(1,m);
        UCL_R = D_4(1,n)*R_bar*ones(1,m);
    end

    LCL_std = B_3*std_bar*ones(1,m);
    CL_std = std_bar*ones(1,m);
    UCL_std = B_4*std_bar*ones(1,m);

    LCL_var = L_var*var_bar*ones(1,m);
    CL_var = var_bar*ones(1,m);
    UCL_var = U_var*var_bar*ones(1,m);

    LCL_mean = mean_bar-A_3*std_bar*ones(1,m);
    CL_mean = mean_bar*ones(1,m);
    UCL_mean = mean_bar+A_3*std_bar*ones(1,m);

    LCL_mad = B_5*mad_bar*ones(1,m);
    CL_mad = c_4*mad_bar*ones(1,m);
    UCL_mad = B_6*mad_bar*ones(1,m);

    LCL_sn = B_5*sn_bar*ones(1,m);
    CL_sn = c_4*sn_bar*ones(1,m);
    UCL_sn = B_6*sn_bar*ones(1,m);

    LCL_qn = B_5*qn_bar*ones(1,m);
    CL_qn = c_4*qn_bar*ones(1,m);
    UCL_qn = B_6*qn_bar*ones(1,m);

case('yes')

    if n<26
        LCL_R = D_1(1,n)*sigma*ones(1,m);
        CL_R = d_2(1,n)*sigma*ones(1,m);
        UCL_R = D_2(1,n)*sigma*ones(1,m);
    end

    LCL_std = B_5*sigma*ones(1,m);
    CL_std = c_4*sigma*ones(1,m);
    UCL_std = B_6*sigma*ones(1,m);

    LCL_var = L_var*(sigma^2)*ones(1,m);
    CL_var = (sigma^2)*ones(1,m);
    UCL_var = U_var*(sigma^2)*ones(1,m);

    LCL_mean = avg-A*sigma*ones(1,m);
    CL_mean = avg*ones(1,m);
    UCL_mean = avg+A*sigma*ones(1,m);

    LCL_mad = B_5*sigma*ones(1,m);
    CL_mad = sigma*ones(1,m);
    UCL_mad = B_6*sigma*ones(1,m);

    LCL_sn = B_5*sigma*ones(1,m);
    CL_sn = sigma*ones(1,m);
    UCL_sn = B_6*sigma*ones(1,m);

```

```

        LCL_qn = B_5*sigma*ones(1,m);
        CL_qn = sigma*ones(1,m);
        UCL_qn = B_6*sigma*ones(1,m);
    end

    % Determination of out of control values

    if n<26
        out_R = (double(X_R>UCL_R) +double(X_R<LCL_R)).*X_R;
        for i=1:30
            if out_R(1,i)==0
                out_R(1,i)=NaN;
            end
        end
    end

    out_std = (double(X_std>UCL_std) +double(X_std<LCL_std)).*X_std;
    for i=1:30
        if out_std(1,i)==0
            out_std(1,i)=NaN;
        end
    end

    out_var = (double(X_var>UCL_var) +double(X_var<LCL_var)).*X_var;
    for i=1:30
        if out_var(1,i)==0
            out_var(1,i)=NaN;
        end
    end

    out_mean = (double(X_mean>UCL_mean) +...
        double(X_mean<LCL_mean)).*X_mean;
    for i=1:30
        if out_mean(1,i)==0
            out_mean(1,i)=NaN;
        end
    end

    out_mad = (double(X_mad>UCL_mad) +double(X_mad<LCL_mad)).*X_mad;
    for i=1:30
        if out_mad(1,i)==0
            out_mad(1,i)=NaN;
        end
    end

    out_sn = (double(X_sn>UCL_sn) +double(X_sn<LCL_sn)).*X_sn;
    for i=1:30
        if out_sn(1,i)==0
            out_sn(1,i)=NaN;
        end
    end

    out_qn = (double(X_qn>UCL_qn) +double(X_qn<LCL_qn)).*X_qn;
    for i=1:30
        if out_qn(1,i)==0
            out_qn(1,i)=NaN;
        end
    end

```

```

end

clear i;

%Control Charts

if n<26
    plot(obs,X_R,'bo-',obs,LCL_R,'r',obs,CL_R,'g',obs,UCL_R,'r'...
        ,obs,out_R,'m*-',obs,out_R,'ks-');
    switch lower(distribution)

        case('nor')
            title('Shewart R Chart for Normal Data');

        case('log')
            title('Shewart R Chart for Logistic Data');

        case('de')
            title('Shewart R Chart for Double Exponential Data');

        case('cau')
            title('Shewart R Chart for Cauchy Data');
    end

    switch lower(standard)
        case('no')
            xlabel('Standards: UNKNOWN');
        case('yes')
            xlabel('Standards: KNOWN');
    end

    legend('RANGE','LCL','CL','UCL','Outlier',...
        'Location','NorthEastOutside')

else sprintf('For n>25, Range is an inefficient estimator for
sigma')
end

figure;

plot(obs,X_std,'bo-
',obs,LCL_std,'r',obs,CL_std,'g',obs,UCL_std,'r'...
,obs,out_std,'m*-',obs,out_std,'ks-');

switch lower(distribution)

    case('nor')
        title('Shewart S Chart for Normal Data');

    case('log')
        title('Shewart S Chart for Logistic Data');

    case('de')
        title('Shewart S Chart for Double Exponential Data');

```



```

        case('cau')
            title('Shewart S Chart for Cauchy Data');
    end

    switch lower(standard)
        case('no')
            xlabel('Standards: UNKNOWN');
        case('yes')
            xlabel('Standards: KNOWN');
    end

    legend('STD DEV', 'LCL', 'CL', 'UCL', 'Outlier', ...
           'Location', 'NorthEastOutside')

    figure;

    plot(obs, X_mean, 'bo-
', obs, LCL_mean, 'r', obs, CL_mean, 'g', obs, UCL_mean, 'r'...
, obs, out_mean, 'm*-', obs, out_mean, 'ks-');

    switch lower(distribution)

        case('nor')
            title('MEAN Chart for Normal Data');

        case('log')
            title('MEAN Chart for Logistic Data');

        case('de')
            title('MEAN Chart for Double Exponential Data');

        case('cau')
            title('MEAN Chart for Cauchy Data');
    end

    switch lower(standard)
        case('no')
            xlabel('Standards: UNKNOWN');
        case('yes')
            xlabel('Standards: KNOWN');
    end

    legend('MEAN', 'LCL', 'CL', 'UCL', 'Outlier'...
           'Location', 'NorthEastOutside')

    figure;

    plot(obs, X_var, 'bo-
', obs, LCL_var, 'r', obs, CL_var, 'g', obs, UCL_var, 'r'...
, obs, out_var, 'm*-', obs, out_var, 'ks-');

    switch lower(distribution)

        case('nor')
            title('Variance Chart for Normal Data');

```

```

        case('log')
            title('Variance Chart for Logistic Data');

        case('de')
            title('Variance Chart for Double Exponential Data');

        case('cau')
            title('Variance Chart for Cauchy Data');
    end

    switch lower(standard)
        case('no')
            xlabel('Standards: UNKNOWN');
        case('yes')
            xlabel('Standards: KNOWN');
    end

    legend('VARIANCE', 'LCL', 'CL', 'UCL', 'Outlier'...
        , 'Location', 'NorthEastOutside')

figure;

plot(obs, X_mad, 'bo-
', obs, LCL_mad, 'r', obs, CL_mad, 'g', obs, UCL_mad, 'r'...
, obs, out_mad, 'm*-', obs, out_mad, 'ks-');

    switch lower(distribution)

        case('nor')
            title('MAD Chart for Normal Data');

        case('log')
            title('MAD Chart for Logistic Data');

        case('de')
            title('MAD Chart for Double Exponential Data');

        case('cau')
            title('MAD Chart for Cauchy Data');
    end

    switch lower(standard)
        case('no')
            xlabel('Standards: UNKNOWN');
        case('yes')
            xlabel('Standards: KNOWN');
    end

    legend('MAD', 'LCL', 'CL', 'UCL', 'Outlier'...
        , 'Location', 'NorthEastOutside')

figure;

plot(obs, X_sn, 'bo-', obs, LCL_sn, 'r', obs, CL_sn, 'g', obs, UCL_sn, 'r'...
, obs, out_sn, 'm*-', obs, out_sn, 'ks-');

```

```

switch lower(distribution)

    case('nor')
        title('S_n Chart for Normal Data');

    case('log')
        title('S_n Chart for Logistic Data');
    case('de')
        title('S_n Chart for Double Exponential Data');

    case('cau')
        title('S_n Chart for Cauchy Data');
end

switch lower(standard)
    case('no')
        xlabel('Standards: UNKNOWN');
    case('yes')
        xlabel('Standards: KNOWN');
end

legend('S_n', 'LCL', 'CL', 'UCL', 'Outlier'...
, 'Location', 'NorthEastOutside')

figure;

plot(obs, X_qn, 'bo-', obs, LCL_qn, 'r', obs, CL_qn, 'g', obs, UCL_qn, 'r'...
, obs, out_qn, 'm*-', obs, out_qn, 'ks-');

switch lower(distribution)

    case('nor')
        title('Q_n Chart for Normal Data');

    case('log')
        title('Q_n Chart for Logistic Data');

    case('de')
        title('Q_n Chart for Double Exponential Data');

    case('cau')
        title('Q_n Chart for Cauchy Data');
end

switch lower(standard)
    case('no')
        xlabel('Standards: UNKNOWN');
    case('yes')
        xlabel('Standards: KNOWN');
end

legend('Q_n', 'LCL', 'CL', 'UCL', 'Outlier'...
, 'Location', 'NorthEastOutside')

```

Here, some notes may be considerable:

First of all, it is also possible to design `intro_charts.m` in the way that the estimator is also taken as an input.

The next point is that “seed” is stated in the function as a constant to obtain the same result when function is run again. If seed was deleted, each run would generate different streams. In fact, leaves data of table (2.1) is the output vector: “datam.” Its seed is my birth year, 1978.

Finally, if standards are set as “NO,” function does not take the avg and sigma values into account and estimates mean and standard deviation from the sample data. Moreover, it is possible to set another m and n values for a different dimension matrix.

The Logistic, Laplace, and Cauchy pdf graphs are drawn by the m-file: `plotpdf.m`. The m-file is as follows:

plotpdf:

```
% plotpdf.m draws the pdf figures for Logistic, Laplace
% and Cauchy distributions, with changing parameters.
%
% Written by Alp Giray Özen, 2011

x = -15:0.1:15;

f1 = zeros(1,size(x,2));
f2 = zeros(1,size(x,2));
f3 = zeros(1,size(x,2));
f4 = zeros(1,size(x,2));

% Logistic pdf

mu=0; s=1;

for i=1:size(x,2)
    f1(i)=exp(-(x(i)-mu)/s)/(s*((1+exp(-(x(i)-mu)/s)).^2));
end

mu=0; s=3;

for i=1:size(x,2)
```

```

        f2(i)=exp(-(x(i)-mu)/s)/(s*((1+exp(-(x(i)-mu)/s)).^2));
    end

mu=0; s=4;

for i=1:size(x,2)
    f3(i)=exp(-(x(i)-mu)/s)/(s*((1+exp(-(x(i)-mu)/s)).^2));
end

mu=5; s=1;

for i=1:size(x,2)
    f4(i)=exp(-(x(i)-mu)/s)/(s*((1+exp(-(x(i)-mu)/s)).^2));
end

plot(x,f1,'b',x,f2,'r',x,f3,'k',x,f4,'g');

title('Logistic pdf');
xlabel('x');
ylabel('p(x)');
legend('mu=0 s=1','mu=0 s=3',...
        'mu=0 s=4','mu=5 s=1','Location','NorthEast')

figure

% Laplace pdf

x = -15:0.1:15;

f1 = zeros(1,size(x,2));
f2 = zeros(1,size(x,2));
f3 = zeros(1,size(x,2));
f4 = zeros(1,size(x,2));

mu=0; s=1;

for i=1:size(x,2)
    f1(i)=(1/(2*s))*exp(-(abs(x(i)-mu))/s);
end

mu=0; s=2;

for i=1:size(x,2)
    f2(i)=(1/(2*s))*exp((-abs(x(i)-mu))/s);
end

mu=0; s=4;

for i=1:size(x,2)
    f3(i)=(1/(2*s))*exp((-abs(x(i)-mu))/s);
end

mu=5; s=4;

```

```

for i=1:size(x,2)
    f4(i)=(1/(2*s))*exp((-abs(x(i)-mu))/s);
end

plot(x,f1,'b',x,f2,'r',x,f3,'k',x,f4,'g');

title('Laplace pdf');
xlabel('x');
ylabel('p(x)');
legend('mu=0 b=1','mu=0 b=2',...
       'mu=0 b=4','mu=5 b=4','Location','NorthEast')

figure

% Cauchy pdf

x = -5:0.1:5;

f1 = zeros(1,size(x,2));
f2 = zeros(1,size(x,2));
f3 = zeros(1,size(x,2));
f4 = zeros(1,size(x,2));
f5 = zeros(1,size(x,2));

mu=0; s=1;

%f=(1/pi)*(s/(((x(i)-mu))^2+s^2));

for i=1:size(x,2)
    f1(i)=(1/pi)*(s/(((x(i)-mu))^2+s^2));
end

mu=0; s=0.5;

for i=1:size(x,2)
    f2(i)=(1/pi)*(s/(((x(i)-mu))^2+s^2));
end

mu=0; s=2;

for i=1:size(x,2)
    f3(i)=(1/pi)*(s/(((x(i)-mu))^2+s^2));
end

mu=2; s=1;

for i=1:size(x,2)
    f4(i)=(1/pi)*(s/(((x(i)-mu))^2+s^2));
end

plot(x,f1,'b',x,f2,'r',x,f3,'k',x,f4,'g');

```

```

title('Cauchy pdf');
xlabel('x');
ylabel('p(x)');
legend('mu=0 gamma=1','mu=0 gamma=0.5',...
      'mu=0 gamma=2','mu=2 gamma=1','Location','NorthEast')

clear

```

To obtain the ARL simulations of chapter two, the function `runlength_intro` has been written. The generator function is used to generate a random sample for given distribution and estimator function is used to perform the mean, the standard deviation, and the variance chart ARL simulations. Location parameter is set to 0 and scale parameter is set to 1, since they have no effect on simulation results. `new_seed` value 1405 is my birth day and month. M-file of the function is given below:

runlength_intro:

```

% [ARL_sims ARL_final UCL LCL] =
%   runlength_intro(distribution,statistics,shift,standard,n)
%
% Function runlength_intro performs ARL simulations
% for the given distribution and statistics.
%
% INPUTS:
% distribution: Distribution Type is the input string
% of the function
% Input
%   'nor' for Normal (Gaussian)
%   'log' for Logistic
%   'de' for Double Exponential (Laplace)
%   'cau' for Cauchy
%
%                               distributions
% distribution location parameter is 0 and scale parameter is 1.
%
% statistics: Estimator used in the ARL simulation
% Input
%   'sd' for Standard Deviation
%   'var' for Variance
%   'mean' for Mean
%
%                               estimators
%
% shift: The shift occurred in the process. For mean shift, shift
% is defined in standard deviation units and for scale shift,
% scale parameter of the distribution is multiplied by the shift.
% Then, enter 0 for no shift case in the mean and 1 for
% no shift case in scale estimators.
%
% standard: Standards Known/Unknown is the input string
% of the function
% Input

```

```

%      'yes' for Quality Standards Known case
%      'no' for Quality Standards estimated from the data case
%
%      n: Sample size of each period in control process.
%
%      OUTPUTS:
%      ARL_sims: Average run length for each of 1000 simulation run
%
%      ARL_final: Mean of ARL_sims values. The purpose here is to
%      see the variability of simulated ARL values.
%
%      UCL and LCL: Upper and Lower Control Limits of the Chart.
%
%      Written by Alp Giray Özen, 2011

function [ARL_sims ARL_final UCL LCL] = ...
    runlength_intro(distribution,statistics,shift,standard,n)

maxm = 500000;      %define maxm: max number of random stream
r = 1000;           %define r: number of run lengths
format short g      %state format
mu = 0;             %Mean of the data to be generated
sigma = 1;          %Standard deviation of the data to be generated

% Enter constants for Shewart S Chart

c_4 = sqrt(2/(n-1))*gamma(n/2)/gamma((n-1)/2);
B_3p = 1-3*sqrt(1-c_4^2)/c_4;
B_3 = max([B_3p ; 0]);           %LCL coefficient
B_4 = 1+3*sqrt(1-c_4^2)/c_4;     %UCL coefficient
clear B_3p;

B_5p = (c_4-3*sqrt(1-c_4^2));
B_5 = max([B_5p ; 0]);           %LCL coefficient
B_6 = (c_4+3*sqrt(1-c_4^2));     %UCL coefficient
clear B_5p;

% Enter constants for Variance Chart

U_var = chi2inv(0.99865,n-1)/(n-1);
L_var = chi2inv(0.00135,n-1)/(n-1);

% Enter constants for X-bar Chart

A = 3/sqrt(n);
A_3 = 3/(c_4*sqrt(n));

% Calculation of Control Limits

switch lower(standard)

    case('no') % No standards given for mean and standard deviation

        % Estimate mean and standard deviation from m=100 data

```



```

m = 100;
seed = 1978;

randn('state',seed)
rand('twister',seed)

X_data = generator(distribution,mu,sigma,n,m);

barbar_X = mean(mean(X_data));
std_X = mean(std(X_data));
var_X = mean(var(X_data));

% Calculate Control Limits based on sample statistics

switch lower(statistics)

    case('mean')

        UCL = barbar_X+A_3*std_X;
        LCL = barbar_X-A_3*std_X;

    case('sd')

        UCL = B_4*std_X;
        LCL = B_3*std_X;

    case('var')

        LCL = L_var*(var_X);
        UCL = U_var*(var_X);

    otherwise
        disp('Please enter mean sd or var for statistics')
        return

end

case('yes')% mean and standard deviation are assumed to be known

% Calculate Control Limits based on sample statistics

switch lower(statistics)

    case('mean')

        LCL = mu-A*sigma;
        UCL = mu+A*sigma;

    case('sd')

        LCL = B_5*sigma;

```

```

        UCL = B_6*sigma;

        case('var')

            LCL = L_var*(sigma^2);
            UCL = U_var*(sigma^2);

            otherwise
                disp('Please enter mean sd or var for statistics')
                return

        end

    otherwise
        disp('Please enter no or yes for standard')
        return

end

% Create new random data to obtain run lengths
% Obtain index vector of out of control values

new_seed = 1405;
randn('state',new_seed)
rand('twister',new_seed)

% Perform ARL simulations of 1000 out of control runs, 10 times

ARL_sims = zeros(10,1);

for i=1:10

    if strcmp(statistics,'mean')
        X = shift*sigma+generator(distribution,mu,sigma,n,maxm);
    else
        X = shift*generator(distribution,mu,sigma,n,maxm);
    end

    X_stat = estimator(statistics,X);
    out_stat = logical(X_stat>UCL) + logical(X_stat<LCL);
    out_index = find(out_stat==1);

    ARL = out_index(r)/r;
    ARL_sims(i,1) = ARL;
    clear out* X*
end

% Calculate mean of the simulations

ARL_final = mean(ARL_sims);

```

This is the end of chapter two and chapter three begins with figures of empirical influence functions. Since the codes for location and scale cases are very similar to each other, only the scale part function, which create the influence curves for the estimators, and their corresponding sensitivity curves, will be given. The following function `eif_scale` is written for this purpose. A figure is not an output for Sir MATLAB, so the function has no output. The only input is the sample data. It is also possible to design the function in the way that it takes estimator as an input.

eif_scale:

```
% [] = eif_scale(data)
%
% function eif_scale draws the figures of
% empirical influence functions and empirical sensitivity curves
% of scale estimators MAD, Sn, Qn and S for the given input data.
%
% Written by Alp Giray Özen, 2011

function [] = eif_scale(data)

%data: X = [0.43 1.27 1.44 1.52 1.75 2.09 2.96 3.80 3.83 4.22];

[m n] = size(data);
data = reshape(data,1,m*n);

add = mean(data)-...
      20*std(data):range(data)/100:mean(data)+20*std(data);
sd = zeros(1,size(add,2));
md = zeros(1,size(add,2));
s = zeros(1,size(add,2));
q = zeros(1,size(add,2));

for i=1:size(add,2)
    sd(i) = std([data add(i)]);
    md(i) = madn([data add(i)]);
    s(i) = sn([data add(i)]);
    q(i) = qn([data add(i)]);
end

plot(add,sd,'k')
hold on
plot(add,md,'b')
hold on
plot(add,s,'m')
hold on
plot(add,q,'k--')

legend('SD','MAD','S_n','Q_n');
title('Empirical Influence Functions for Scale Estimators');
```

```

sd_esc = (size(data,2)+1)*(sd-std(data));
md_esc = (size(data,2)+1)*(md-madn(data));
s_esc = (size(data,2)+1)*(s-sn(data));
q_esc = (size(data,2)+1)*(q-qn(data));

figure

plot(add,sd_esc,'k')
hold on
plot(add,md_esc,'b')
hold on
plot(add,s_esc,'m')
hold on
plot(add,q_esc,'k--')

legend('SD','MAD','S_n','Q_n');

title('Empirical Sensitivity Curves for Scale Estimators');

```

Average value and standardized variance of scale estimators used in the study were given in tables (3.1) and (3.2). These two tables are created (except the last column, which consists of the limiting results) by the following m-file: robust.m:

robust:

```

% robust.m is simulation of table 1 and table 2
% in "Alternatives to MAD, (Rousseeuw and Croux, 1993)" whose
% results are at tables (3.1) and (3.2) in my thesis.
%
% Written by Alp Giray Özen, 2011.

seed = 1978;
n=[5 10 20 50 100];

table_mean = zeros(size(n,2),4);
table_var = zeros(size(n,2),4);

for i=1:size(n,2)
    X = generator('nor',0,1,n(i),10000);

    mad_X = madn(X);
    var_mad = var(mad_X);

    mean_mad = mean(mad_X);
    stan_varmad = n(i)*var_mad/((mean_mad)^2);

    table_mean(i,1) = mean_mad;
    table_var(i,1) = stan_varmad;
    sn_X = sn(X);

```

```

var_sn = var(sn_X);

mean_sn = mean(sn_X);
stan_varasn = n(i)*var_sn/((mean_sn)^2);

table_mean(i,2) = mean_sn;
table_var(i,2) = stan_varasn;

qn_X = qn(X);
var_qn = var(qn_X);

mean_qn = mean(qn_X);
stan_varqn = n(i)*var_qn/((mean_qn)^2);

table_mean(i,3) = mean_qn;
table_var(i,3) = stan_varqn;

sd_X = std(X);
var_sd = var(sd_X);

mean_sd = mean(sd_X);
stan_varasd = n(i)*var_sd/((mean_sd)^2);

table_mean(i,4) = mean_sd;
table_var(i,4) = stan_varasd;
end

```

The ARL simulations of chapter three are obtained by the function `runlength_mad`. The design of this function is very much like to `runlength_intro` except that the only estimator here is MAD, since these trials aim a starting point for robust estimator control limits. Calculation of f_n values is performed by the sub function `mc`, but calculation of g_n values is not shown, since they are already consistent with the corresponding simulation results. Change the value of `maxm` from 500000 to 400000 if you suffer from “Out of memory error.”

runlength_mad:

```

% [ARL_sims ARL_final UCL LCL] =
%     runlength_mad(distribution,trial,shift,n)
%
% Function runlength_mad performs ARL simulations of
% MAD Control Chart trials for the given distribution.
%
% INPUTS:
% distribution: Distribution Type is the
% input string of the function
% Input

```

```

%      'nor' for Normal (Gaussian)
%      'log' for Logistic
%      'de' for Double Exponential (Laplace)
%      'cau' for Cauchy
%
%                                     distributions
%      distribution location parameter is 0 and scale parameter is 1.
%
%      trial: Estimator used in the ARL simulation
% Input
%      'fn' for the std_dev like trial version
%      'gn' for the variance like trial version
%
%      shift: The shift occurred in the process. Scale parameter
% of the distribution is multiplied by the shift.
% Then, enter 1 for no shift case.
%
%      n: Sample size of each period in control process.
%
%      OUTPUTS:
%      ARL_sims: Average run length for each of 1000 simulation run
%
%      ARL_final: Mean of ARL_sims values. The purpose here is to
% see the variability of simulated ARL values.
%
%      UCL and LCL: Upper and Lower Control Limits of the Chart.
%
%      Written by Alp Giray Özen, 2011

```

```

function [ARL_sims ARL_final UCL LCL] = ...
runlength_mad(distribution,trial,shift,n)

maxm = 500000;    %define maxm: max number of random stream
r = 1000;        %define r: number of run lengths
mu = 0;          %mean of the random stream to be generated
sigma = 1;       %std_dev of the random stream to be generated
format short g   %state format

% bn coefficients of MAD

switch n
case 2
    bn=1.196;
case 3
    bn=1.495;
case 4
    bn=1.363;
case 5
    bn=1.206;
case 6
    bn=1.200;
case 7
    bn=1.140;
case 8
    bn=1.129;
case 9
    bn=1.107;
otherwise

```

```

        bn = n/(n-0.8);
end

% ARL simulation of MAD, for fn and gn trials

switch(trial)

    case('fn')

        % Calculate control limits

        m = mc(n);

        c_4 = sqrt(2/(n-1))*gamma(n/2)/gamma((n-1)/2);
        B_61 = 1/bn+3*sqrt(m*(1-c_4^2));
        B_51 = max(1/bn-3*sqrt(m*(1-c_4^2)),0);

        UCL = B_61*sigma;
        LCL = B_51*sigma;

        % Create new data to obtain run lengths
        % Obtain index vector of out of control values

        new_seed = 1405;
        randn('state',new_seed)
        rand('twister',new_seed)

        ARL_sims = zeros(10,1);

        for i=1:10
            X = shift*generator(distribution,mu,sigma,n,maxm);
            X_mad = madn(X);
            out = logical(X_mad>UCL) + logical(X_mad<LCL);
            out_index = find(out==1);
            ARL = out_index(r)/r;
            ARL_sims(i,1) = ARL;
            clear out* X*
        end

    case('gn')

        % Calculate control limits

        constant_search = abs([n-5 n-20 n-50]);

        if min(constant_search)==constant_search(1)
            constant = 4.11;
        elseif min(constant_search)==constant_search(2)
            constant = 1.445;
        else constant = 1.234;
        end

        chi_coeff = [chi2inv(1-0.0013513,n-1) ...
                    chi2inv(0.0013513,n-1)]/(n-1);
        chi_u = chi_coeff(1);
        chi_l = chi_coeff(2);

```

```

B_61 = constant*sqrt(chi_u);
B_51 = (1/constant)*sqrt(chi_l);

UCL = B_61*sigma;
LCL = B_51*sigma;

% Create new data to obtain run lengths
% Obtain index vector of out of control values

new_seed = 1405;
randn('state',new_seed)
rand('twister',new_seed)

ARL_sims = zeros(10,1);

for i=1:10
    X = shift*generator(distribution,mu,sigma,n,maxm);
    X_mad = madn(X)/bn;
    out = logical(X_mad>UCL) + logical(X_mad<LCL);
    out_index = find(out==1);
    ARL = out_index(r)/r;
    ARL_sims(i,1) = ARL;
    clear out* X*
end

end

ARL_final = mean(ARL_sims);

function m = mc(n)

% Calculate standardized variance for MAD and standard deviation

seed = 1978; % Define seed of the random vector
randn('state',seed)
rand('twister',seed)
X = generator('nor',0,1,n,10000);
mad_X = madn(X);
var_mad = var(mad_X);
mean_mad = mean(mad_X);
stan_varmad = n*var_mad/((mean_mad)^2);

sd_X = std(X);
var_sd = var(sd_X);

mean_sd = mean(sd_X);
stan_varstd = n*var_sd/((mean_sd)^2);

% Calculate efficiency and constant m

eff = stan_varstd/stan_varmad;
m=1/eff;

```


This is the end of chapter three and the beginning of the fourth one. Every ending is a beginning of something other, isn't it? Anyway, the following function `bs_hist` draws the histogram of bootstrap samples for the input robust estimator. In other words, it is a graphical view for an estimate of sampling distribution.

bs_hist:

```
% [] = bs_hist(statistics, distribution, sample, bins)
%
% function bs_hist draws the histogram of bootstrap samples.
% Ideally, this is an estimation of sampling distribution of
% the statistics.
%
% INPUTS:
% Statistics: The robust estimator of scale
% Input
%   'mad' for Median Absolute Deviation
%   'sn' for Sn
%   'qn' for Qn
%
%                               estimators
%
% distribution: Distribution Type is the input
% string of the function
% Input
%   'nor' for Normal (Gaussian)
%   'log' for Logistic
%   'de' for Double Exponential (Laplace)
%   'cau' for Cauchy
%
%                               distributions
%
% sample: sample size of each bootstrap sample
%
% bins: Number of bins to be used in histogram
%
% There is no output since the figure does not require it.
%
% Additionally, a hypothesis testing is made:
%   Result is 1 if Ho is rejected and 0 otherwise.
%   p_value is the p-value of the test, which is shown
% as x-label of the histogram.
%
% The corresponding KS-test has a null hypothesis
%   Ho: Bootstrap distribution is Normal.
%
% For example, the command,
%   bs_hist('sn','cau',20,50)
% draws the bootstrap samples histogram (using 50 bins) of Sn
% that is obtained by a random sample of 20 taken from
```

```

% Cauchy Distribution (n^2=4000 bootstrap samples are used).
%
%   Written by Alp Giray Özen, 2011

function [] = bs_hist(statistics, distribution, sample, bins)

n = sample;
m = bins;

%   Create a random data with location parameter 0
%   and scale parameter 1

new_seed = 1405;
randn('state',new_seed)
rand('twister',new_seed)

X = generator(distribution,0,1,1,n);

%   Write the name of the distribution for title

switch lower(distribution)

    case('nor')
        D = 'Normal Distribution';

    case('log')
        D = 'Logistic Distribution';

    case('de')
        D = 'Laplace Distribution';

    case('cau')
        D = 'Cauchy Distribution';

    otherwise
        disp('No match for this distribution type')
        return
end

%   Collect statistics of the bootstrap samples,
%   draw histogram and write the name of the estimator for title

switch lower(statistics)

    case('mad')
        bootstat = bootstrp(n^2,@madn,X);
        hist(bootstat,m)
        S = 'MAD';

    case('sn')
        bootstat = bootstrp(n^2,@sn,X);
        hist(bootstat,m)
        S = 'S_n';

```

```

    case('qn')
        bootstat = bootstrp(n^2,@qn,X);
        hist(bootstat,m)
        S = 'Q_n';

    otherwise
        disp('No match for this statistics type')
        return
end

T = [S ' bootstrap histogram for ' D];
title(T);

% KS Normality test for sampling distribution of the statistics

z_bootstat = (bootstat-mean(bootstat))/std(bootstat);
[H,P] = kstest(z_bootstat);

clear H
p_value = num2str(P);
xlabel(['P value of the normality test is ' p_value]);

```

The one before the last function of the thesis is `runlength_bs`, which creates the bootstrap ARL simulation tables of chapter four. The first table of each distribution (sample variance ARL simulation) is constructed by the function “`runlength_intro`.” The aim is to simulate ARL using bootstrap confidence intervals. Its design is very similar to those of the other two chapters’ `runlength` functions, but I changed the algorithm for calculation of sims because the old algorithm was very slow for S_n and Q_n statistics. The difference in run time between two is extremely high. The former works in hours but the latter works in minutes.

Some additional cautions are required here. Reduce `maxm` to 300000 when you run `mad` for `log`, not to face with an “out of memory” error. Also increase `maxm` to 800000 when you run S_n for `cau` because the run length is close to 800 for this case.

runlength_bs:

```

% [ARL_sims ARL_final limits se_limits] =
% runlength_bs(distribution,statistics,center,shift,n)
%
% Function runlength_intro performs ARL simulations
% for the given distribution and statistics based on
% bootstrap confidence interval limits taken as control limits.

```

```

%
% INPUTS:
% distribution: Distribution Type is the input
% string of the function
% Input
%     'nor' for Normal (Gaussian)
%     'log' for Logistic
%     'de' for Double Exponential (Laplace)
%     'cau' for Cauchy
%
%                                     distributions
% distribution location parameter is 0 and scale parameter is 1.
%
% statistics: Estimator used in the ARL simulation
% Input
%     'mad' for Median Absolute Deviation
%     'sn' for Sn
%     'qn' for Qn
%
%                                     estimators
%
% center: States if "percentile method" or
% "centered percentile method" is used for confidence interval.
% Input
%     'no' for percentile method
%     'yes' for centered percentile method
%
% shift: The shift occurred in the process. Scale parameter
% of the distribution is multiplied by the shift.
% Then, enter 1 for no shift case.
%
%
% n: Sample size of each period in control process.
%
% OUTPUTS:
% ARL_sims: Average run length for each of 1000 simulation run
%
% ARL_final: Mean of ARL_sims values. The purpose here is to
% see the variability of simulated ARL values.
%
% limits: Upper and Lower Control Limits of the Chart.
%
% se_limits: Standard error of Control limits
%
% Written by Alp Giray Özen, 2011

function [ARL_sims ARL_final limits se_limits] = ...
    runlength_bs(distribution,statistics,center,shift,n)

maxm = 500000;    %define maxm: max number of random stream
r = 1000;        %define r: number of run lengths
format short g   %state format

% Calculate control limits

ci_per = zeros(2,25);
stat = zeros(25,1);
seed = 1978;
rand('twister',seed);

```

```

randn('state',seed);

switch lower(statistics)

    case ('mad')
        for i=1:100
            X_limit = generator(distribution,0,1,1,n);
            stat(i,1) = estimator(statistics,X_limit);
            ci = bootci(n^2,{@madn,X_limit},...
                'alpha',0.0075,'type','per');
            ci_per(:,i) = ci;
        end

    case ('sn')
        for i=1:100
            X_limit = generator(distribution,0,1,1,n);
            stat(i,1) = estimator(statistics,X_limit);
            ci = bootci(n^2,{@sn,X_limit},...
                'alpha',0.0075,'type','per');
            ci_per(:,i) = ci;
        end

    case ('qn')
        for i=1:100
            X_limit = generator(distribution,0,1,1,n);
            stat(i,1) = estimator(statistics,X_limit);
            ci = bootci(n^2,{@qn,X_limit},...
                'alpha',0.0010,'type','per');
            ci_per(:,i) = ci;
        end

end

ci_per = ci_per';

limits = mean(ci_per);
teta_head = mean(stat);
teta_head_var = var(stat)*ones(1,2);

se_limits = std(ci_per);

UCL = limits(2);
LCL = limits(1);

if strcmp(center,'yes')
    UCL = 2*teta_head - limits(1);
    LCL = max(0,2*teta_head - limits(2));
    limits = [LCL UCL];
    se_limits = sqrt(4*teta_head_var + var(ci_per));
end

% Create a random data of observations with a new seed
% Obtain index vector of out of control values

```

```

new_seed = 1405;
rand('twister',new_seed)
randn('state',new_seed);

ARL_sims = zeros(10,1);

for i=1:10

    X = shift*generator(distribution,0,1,n,maxm);

    j = 1;
    check = 0;

    while check<r

        X_stat = estimator(statistics,X(:,j))/coeff(statistics,n);
        out = logical(X_stat>UCL) + logical(X_stat<LCL);

        if out==1
            check = check+1;
        end

        j = j+1;

    end

    clear X X_stat out

    ARL = (j-1)/r;
    ARL_sims(i,1) = ARL;

end

ARL_final = mean(ARL_sims);

function cons = coeff(statistics,n)

switch lower(statistics)

    case('mad')
        cons = n/(n-0.8);

    case('sn')
        cons = 1;
        if (mod(n,2)==1)
            cons=n/(n-0.9);
        end

    case('qn')
        if (mod(n,2)==1)
            cons=n/(n+1.4);
        elseif (mod(n,2)==0)
            cons=n/(n+3.8);
        end
end

```

end

end

And finally, the last function of my thesis, which is “detect.m,” creates the false alarm and detection probabilities of “Chapter 4.3: Proposed Control Designs.” The only input is type, which takes the input ‘gau’ for the former proposed design: “simultaneous use of S_n and Q_n ” and the input ‘cau’ give the matrices for the proposed design for the Cauchy model. The outputs are D: detection probability matrix for changing decision variables of the design, and FA is the corresponding false alarm matrix.

detect:

```
% function [D FA] = detect(type)
%
% m-file detect.m gives the detection and false alarm
% probabilities
% for the new quality design proposed in the conclusion chapter.
%
% mean run lengths of the corresponding robust estimators
% for Cauchy model design are given in the following table.
%
% Statistics      Mean ARL0   Mean ARL1
% Qn-Cent         18,1666    4,2589
% MAD             305,3198   28,0017
%
% mean run lengths of the corresponding robust estimators
% for sn & qn-per design are given in the following table.
%
% Statistics      Mean ARL0   Mean ARL1
% Qn-Cent         11,4421    4,9724
% MAD             232,7874   60,8988
%
% INPUT:
% type: Type of the design whose probabilities will be calculated.
% Input:
%       'gau' for sn-qn design
%       'cau' for Cauchy model design
%
% Written by Alp Giray Özen, 2012

function [D FA] = detect(type)

switch lower(type)

case('gau')
    alpha = [1/18.1666;1/305.3198];
```

```

        beta = [1/4.2589;1/28.0017];

    case('cau')
        alpha = [1/11.4421;1/232.7874];
        beta = [1/4.9724;1/60.8988];

    otherwise
        disp('Invalid type entry')
        disp('Enter gau for sn & qn-per design ')
        disp('or cau for Cauchy model design')
        return
end

detection = zeros(6,16);
falarm = zeros(6,16);

for i=2:size(detection,1)
    detection(i,1)=i-1;
    falarm(i,1)=i-1;
end

for j=2:size(detection,2)
    detection(1,j)=j-1;
    falarm(1,j)=j-1;
end

for i=2:size(detection,1)
    for j=2:size(detection,2)
        detection(i,j)= geocdf(detection(i,1)-1,beta(1)) ...
            *geocdf(detection(1,j)-1,beta(2));
        falarm(i,j)= geocdf(falarm(i,1)-1,alpha(1)) ...
            *geocdf(falarm(1,j)-1,alpha(2));
    end
end

D = detection';
FA = falarm';

clear i j

```

The last command also means that it is over. That's all for now...

Goodbye everybody, I wish "Random Power" were with you...

I also wish each of your lives took its power from love, found its route in art, and formed its shape by scientific thought...