Cairo University
Faculty of Economics and Political Science
Department of Statistics

On Handling Missing Values in Multivariate Statistical Process Control Via Control Charts

By
Doaa Faik Madbuly

Supervised by
Prof. Nadia Makary
Prof. Sameer Sharawy
Prof. Mahmoud AlSaid
Department of Statistics
Faculty of the Economics and Political Science
Cairo University

A Thesis Submitted to the Faculty of the Economics
and Political Science (Department of Statistics)
In Partial Fulfillment of the Requirements
For the Degree of Master of Science
2009

Abstract

Control charts are graphical tools widely used to monitor manufacturing process to quickly detect any change in a process that may result in a change in a product quality. The most well known and widely used measure of the performance of the control chart is the average run length. The presence of missing values in a data set used in building the control chart technique is a serious problem that may face the investigator when applying a control charts to practical situations. Many procedures have been developed to handle missing values, but which one is the most suitable for the control charts technique? This study compares the different methods of handling missing values in case of quality control charts using simulation. The comparison based on the criteria of the average run length. Beside that we study the effect of parameter estimation on the performance of the MEWMA chart as a side point.

Key Words

Multivariate control chart, missing values, average run length and Monte Carlo simulation.

Supervised by

Prof. Nadia Makary

Prof. Sameer Sharawey

Prof. Mahmoud Alsaid

Department of Statistics

Faculty of the Economics and Political Science

Cairo University

Name: Doaa Faik Madbuly.
Nationality: Egyptian.
Date and place of birth: 6/11/1982, Cairo.
Degree: Master.
Specialization: Statistics.
Supervisor: Prof. Nadia Makary
Prof. Sameer Sharawey
Prof. Mahmoud Alsaid.

Department of Statistics
Faculty of the Economics and Political Science
Cairo University

Title of the thesis: On Handling Missing Values in Multivariate Statistical Process Control via Control Charts.

Summary

This study is concerned with the handling of missing values in multivariate statistical process control via control charts. We discuss the complete case, mean substitution, regression, stochastic regression and the expectation maximization algorithm methods for handling missing values. We apply those method to a phase I data set assumed to be in-control. Estimates of mean and variance from the treated data set are used to build the Multivariate Exponentially Weighted Moving Average (MEWMA) control chart. Based on a Monte Carlo simulation study, the performance of each of the five methods is investigated in term of its ability to obtain good IC and OOC ARL. We consider three sample sizes, five levels of the percentage of missing values and three types of variable numbers.

For the in-control case: the stochastic regression method has the best overall performance among all the other methods.

For the out-of-control case: the regression method has the best overall performance among all the other methods.

The study consists of four chapters. In chapter one, first, we define the control chart procedure, spots the light on the average run length as a tool used to measure the performance of the control charts or to compare competing control charts. phase I and phase II analysis of the control charts and the goals of each phase are discussed in details.

A description of the Hotelling's T^2 chart and the MEWMA chart, the two multivariate control charts that are used in this thesis is given. Second, we introduce the problem of having missing values in the quality control data sets. We also review the literature on the comparative studies of handling missing values methods in the other statistical multivariate analysis procedures like the time series and discriminant analysis. Chapter (2) is devoted to the problem of missing values in a phase I data set. Some examples of the patterns of missing values as well as the mechanisms that lead to their presence are also given. The procedures of handing missing values suitable for multivariate data analysis are presented. Performance comparisons and the effect of parameter estimation on the performance of the MEWMA chart are explained in chapter (3). The future recommended works are presented in chapter (4). The appendix contains the SAS codes used for simulation.

Acknowledgment

I am indebted to prof. Mahmoud Alsaid for the considerable help, encouragement, patience and valuable advice he has given me during the research. I would like also to express my gratitude to Prof. Nadia Makary and Prof. Sameer Sharwey.

Special thanks go to my husband and colleague Mustafa Kamal AbdelAziz for his support.

Of course, I am grateful to my mother and my daughters Ganah and Areeg for their patience love and praying at all times.

I owe my deepest gratitude to my father, without his support and help this work would never have come into existence. This thesis is dedicated to his soul.

Doaa Faik.

Contents

List of Figures

List of Tables

List of Abbreviations

- SPC: the Statistical process control.
- UCL: the upper control limit of a control chart.
- LCL: the lower control limit of a control chart.
- IC: an In- control.
- OOC: an Out-Of-control.
- ARL: the average run length.
- EWMA: the Exponentially Weighted Moving Average control chart.
- MEWMA: the Multivariate Exponentially Weighted Moving Average control chart.
- MCAR: the missing completely at random mechanism of missing values.
- MAR: the missing at random mechanism of missing values.
- NMAR: the missing not at random mechanism of missing values.
- CC: the complete case method.
- AC: the available case method.
- MI: the multiple imputations.
- MS: the mean substitution method.
- RG: the regression method.
- SRG: the stochastic regression method.
- ML: the maximum likelihood.
- EM: the expectation maximizations.

Chapter (1)
Control Chart

1.1 Overview

If a product is to meet or exceed customer expectations, generally it should be produced by a process that is stable or repeatable. More precisely, the process must be capable of operating with little variability around the target or nominal dimensions of the product's quality characteristics. Statistical process control (SPC) is a powerful collection of problem solving tools useful in achieving stability and improving capability through the reduction of variability. SPC has seven major tools one of them is control chart [14].

All people agree that everything varies at least a little bit. How can we tell when a process is just experiencing normal variation versus when something out of the ordinary is occurring? Control Charts were designed to make that distinction. Control charts are graphical tools widely used to monitor manufacturing process to quickly detect any change in a process that may result in a change in a product quality.

Control charts are used in two phases of analysis, phases I and II. The main interest in phase I is to analyze a historical set of process data to understand the variation and to determine the stability of the process. Then, once samples associated with assignable causes are removed, one estimates the in-control values of the process parameters. On the other hand, the main concern in the analysis of phase II is to quickly detect shifts in the process from the in-control parameter values estimated in phase I. For phase I applications, one compares the competing control chart methods in terms of the probability of deciding whether or not the process is stable. This is the probability of obtaining at least one statistic outside the control limits when performing control charting using a set of historical observations. In phase II, however, one compares the competing methods in terms of the run length distribution, where the run length is defined as the number of samples taken until an out-of-control signal is given.

In some practical situations a problem arises when some observations in the historical data set are missing. Many methods are used to handle missing values, but which method gives the estimates that lead to the measure of performance of the control chart close to the one obtained when using estimates from a data without missing values?

The main concern of this study is to evaluate the effect of using the different methods of handling the missing values of a phase I data set on the in-control and out-of - control average run length of a control chart used in monitoring a phase II data set. In this study, the average run length is calculated using means of simulation. In other words, we want to study the effect of estimating the parameters from a phase I data set with missing values on the performance of the control chart in phase II.

Since the quality of many products is usually determined by more than one correlated variables, the multivariate control chart is more suitable than separate univariate ones. Therefore we assume that $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ are $1 \times k$ random vectors taken at regular time intervals, each representing the k variable quality characteristics to be monitored. We assume that $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}$ are independent identically distributed (i.i.d) multivariate normal random vectors with mean $\boldsymbol{\mu}$ and constant covariance matrix $\boldsymbol{\Sigma}$. Our main concern in the multivariate statistical process control is to detect changes in $\boldsymbol{\mu}$ from a target value $\boldsymbol{\mu_0}$. All the multivariate control charts considered in this study are directionally invariant. The performance of a directional invariant control chart can be determined solely by the non-centrality parameter D, where

$$D^2 = (\boldsymbol{\mu}-\boldsymbol{\mu_0})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}-\boldsymbol{\mu_0}). \tag{1.1}$$

1.2 Control chart basics

It is important to know when to leave things alone as they are and when to take action. Control charts tell us when we need to take action and when to leave the process alone.

Generally there are two types of statistical control charts, univariate control charts and multivariate control charts which are used for different scenarios. The univariate control charts apply to the processes which has only one process output variable or quality characteristic measured and tested. One of the disadvantages of this scheme is that for a single process, many variables may be monitored and even controlled. Multivariate statistical process control methods overcome this disadvantage by monitoring several variables simultaneously. This work is devoted to the multivariate control charts, but in this sub section we discuss the statistical basis of the control charts in general.

This sub section consists of four sub sections; the first sub section gives definition of the control chart. Sub Section 2 displays the measures of performance used to compare the competing control chats. Sub Section 3 discusses the phase I and phase II analyses. The last Sub section gives an introduction to the multivariate control charts, and then describes two of the most well known multivariate control charts for monitoring the mean vector of the process; the Hotelling's T^2 chart and the Multivariate Exponentially Weighted Moving Average (MEWMA) chart.

1.2.1 Definition of the control chart

The control chart is a graphical display of a quality characteristic that has been measured from a sample versus the sample number or time. Most control charts consist of three parts, which are, the center line that represents the average value of the quality characteristic corresponding to the in-control state, and two horizontal lines named the upper control limit (UCL) and the lower control limit (LCL). Figure 1.1 shows a control chart for monitoring the mean of a univariate variable.

Figure 1.1: A control chart for the mean of a univariate characteristic