UNIVERSITI PUTRA MALAYSIA

THE PERFORMANCE OF BOOTSTRAP CONFIDENCE INTERVALS OF Cpk INDEX BASED ON MM·ESTIMATOR

HANISSAH BT MOHAMAD @ SULAIMAN

FSAS 2002 2

THE PERFORMANCE OF BOOTSTRAP CONFIDENCE INTERVALS OF Cpk INDEX BASED ON MM-ESTIMATOR

By

HANISSAH BT MOHAMAD @ SULAIMAN

Project Submitted in Partial Fulfillment of the Requirement for the
Degree of Master of Science (Applied Statistics) in the
Faculty of Science and Environmental Studies
Universiti Putra Malaysia

May 2002

UPM

CERTIFICATION OF SUPERVISOR

This project paper titled

THE PERFORMANCE OF BOOTSTRAP CONFIDENCE INTERVALS OF Cpk INDEX BASED ON MM-ESTIMATOR

was submitted by

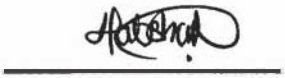
HANISSAH BT MOHAMAD @ SULAIMAN
(GS07805)

In partial fulfillment of the requirements for

Master of Science (Applied Statistics)

This report was accepted and examined

Certified by:



(DR. HABSHAH BT MIDI)
Project Supervisor
Date: 21 May 2002

DEPARTMENT OF MATHEMATICS
FACULTY OF SCIENCE AND ENVIROMENTAL STUDIES
UNIVERSITY PUTRA MALAYSIA
SERDANG, SELANGOR

THE PERFORMANCE OF BOOTSTRAP CONFIDENCE INTERVALS OF Cpk INDEX BASED ON MM-ESTIMATOR

ABSTRACT

A C_{pk} index is used to measure whether a production process is capable of producing items that satisfy a customer requirements(i.e specification limits). The c index is based on the sample mean, x and sample standard deviation, s which are known to be very sensitive to the presence of outliers. As an alternative, we may turn to the robust location and scale estimate based on a robust MM estimates which are less affected by outliers.

A major step toward the correct understanding and interpretation of C_{pk} index is by constructing it's confidence interval. The construction of such intervals assume that the measurement process having a normal distribution. However, many process are not normal and have a fat-tailed distribution which are prone to produce outliers. An alternative approach is to use bootstrap method such as the Percentile (P) and Bias-Corrected and Acceleration (Bca) for calculating approximates confidence intervals of C_{pk} index. It is computer intensive based method that can be utilized without relying any assumption on the underlying distribution. The results of the studies reveal that the Bca method seems to perform better than the Percentile method for both normal and skewed process.

The performance of the C_{pk}-MM estimates were investigated for further by comparing the bootstrap confidence interval for C_{pk} index MM estimates and the well-known classical C_{pk} estimates. Based on simulation studies, show that the MM estimates produced more reliable confidence interval compared to the classical C_{pk} estimates.

ABSTRAK

C_{pk} indeks digunakan untuk mengukur samada proses pengeluaran sesuatu barangan berupaya memenuhi kehendak pelanggan (had spesifikasi). Pengiraan indeks C_{pk} adalah berasaskan min sampel, x dan sisihan piawai sampel, s yang mana ukuran ini sangat sensitif terhadap kewujudan titik terpencil. Sebagai pilihan alternatif, lokasi teguh dan anggaran skala berasaskan kepada anggaran MM dipilih kerana ia kurang sensitif kepada titik terpencil.

Langkah utama untuk mendapatkan kefahaman dan pentafsiran yang betul bagi indeks C_{pk} ialah dengan membina selang keyakinan baginya. Pembinaan selang ini berdasarkan anggapan pengukuran proses mempunyai taburan normal. Bagaimanapun banyak proses adalah tidak normal dan mempunyai taburan yang berhujung tebal yang menghasilkan titik terpencil. Pendekatan alternatif ialah dengan menggunakan kaedah bootstrap seperti Persentil (P) dan Bca untuk mengira selang keyakinan bagi indeks C_{pk}. Kaedah ini berasaskan penggunaan komputer secara intensif dan boleh digunakan tanpa bergantung kepada sebarang taburan sesuatu proses. Keputusan kajian menunjukkan bahawa kaedah Bca mempunyai prestasi yang lebih baik daripada kaedah Persentil untuk kedua-dua proses normal dan pencong.

Prestasi bagi penganggar C_{pk} -MM seterusnya dikaji lagi dengan membandingkan selang keyakinan bootstrap bagi indeks C_{pk} anggaran MM dan anggaran klasik. Kajian simulasi menunjukkan bahawa anggaran MM menghasilkan selang keyakinan yang mempunyai kebolehpercayaan yang lebih tinggi berbanding dengan anggaran C_{pk} klasik.

ACKNOWLEDGEMENTS

First of all, I would like to thank my supervisor Dr Habshah Midi for her invaluable help, guidance and support towards the completion of this project. I deeply appreciate her patience and generosity her vast experience and knowledge.

I would like to thank all my other lecturers and friends for their kind assistance, motivation and contribution towards my studies and obtaining my master Degree.

Last but not least, I would like to thank my parents who were incredibly supportive and always give me a lots of encouragement towards my education.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Capability analysis is a set of statistical calculations performed on a set of data in order to determine the capability of the process. The capability of the process refers to the ability of the process to perform in comparison to its specification limits. A process is said to he 'capable' if it is producing approximately 100% within specification limits. Specification limits are set by customers, engineers or management. There are sometimes called requirements, goals, objectives or standards.

Normally, the set of calculations for capability analysis presented based on the assumption that the data is normally distributed. The shape given by the histogram will tell you if the data is normally distributed. Many, but not all dimensional values in manufacturing behave normally. Another assumption for capability analysis is that the process being studied is stable. In other word there are no special causes of variation present. A control chart of the process should be made to determine stability before capability analysis is performed.

Process capability analysis can be carried out by several methods such as histogram, probability plot and control chart or process capability indices.

In this project, we concern with the confidence intervals for a process performance index, C_{pk} which is one of process capability indices.

Objective of this study

1. Compare the performance of the classical C_{pk} index and robust C_{pk} index from normal process and normal process with 5% contamination.
2. Compare the performance of the Percentile and Bca bootstrap confidence intervals for C_{pk} based on classical estimators from normal process, skewed process and normal process with 5% contamination.
3. Compare the performance of the Percentile and Bca bootstrap confidence intervals for C_{pk} based on robust estimators from normal process, skewed process and normal process with 5% contamination.
4. Compare the performance of Bca confidence intervals based on classical and robust estimators from normal process, skewed process and normal process with 5% contamination.

Simulation data will be used for the above objectives.

Organization of the Project

This project comprises of five chapters

Chapter I, introduces to capability analysis and objectives of the study.

Chapter II, describes several process capability indices that are usually used in process capability analysis, formula for calculating these indices and discussion on their properties.

Chapter III, discusses the robust M, MM and Median Absolutes Deviation(MAD) estimators. In this chapter several bootstrap methods in constructing confidence intervals are also depicted.

A complete discussion on the simulation studies that has been carried out in order to investigate those mentioned objectives of this project are exhibited in Chapter IV.

Finally, the conclusion and recommendation for further study are documented in Chapter V.

CHAPTER II

PROCESS CAPABILITY INDICES

Introduction

Capability indices allow one to monitor and report the improvement of the process over time. Among the most widely used process capability indices are the process capability index (C_p) and process performance index (C_{pk}).

Process Capability Index

Process capability index, C_p is an index assumes that the process is properly centered and is in state of statistical control. It defined as the allowable spread (specification range) over the actual spread(natural variability). The allowable spread or specification range is the range between the upper specification limit (USL) and the lower specification limit (LSL). It will be compare with six- sigma (6σ) range of the process, which is the actual spread or the natural tolerance of the process.

$$C_p = \frac{USL - LSL}{6\sigma}$$

The actual spread is determine from the process data collected and is calculated by multiplying six times the standard deviation, s. The standard deviation quantifies a process's variability. Then the C_p can be estimate

$$\hat{C}_p = \frac{USL - LSL}{6s}$$

Standard deviation also can be estimate from R chart by $\hat{\sigma} \simeq \frac{\bar{R}}{d_2}$ when control charts are used in the capability study. Let $R_1, R_2 \ldots, R_m$ be the ranges of m samples that

has been used to construct the R control chart, then R is the average range, where $\bar{R}$ $\frac{R_1 + R_2 + ... + R_m}{m}$, or $\bar{R}$ is the center line from the sample range R control chart and d_2 is the mean of the relative range.

As the standard deviation increases in a process, the C_p decreases in value. standard deviation decreases, the C_p increases in value.

By convention when a process has a C_p value less than 1.0, it is considered potentially incapable of meeting specification requirements. Conversely when a process C_p is greater than or equal to 1.0, the process has the potential of being capable. True value of C_p= 1.0 means that the specification range (USL-LSL) equal to the natural tolerance (six sigma) of the process.

product will fail to meet the specification for a process produce normally distributed output and is centered at the midpoint of the specification limits, since

$C_P = 1.0$ $\quad USL-LSL = 6\sigma$

$LSL - \mu = 3\sigma$ $\qquad USL - \mu = 3\sigma$

$$P(LSL<x<USL) = P\left(\frac{LSL-\mu}{\sigma} < Z < \frac{USL-\mu}{\sigma}\right)$$

$$= P\left(\frac{-3\sigma}{\sigma} < Z < \frac{3\sigma}{\sigma}\right)$$

$$= P(-3 < Z < 3)$$

$$= 0.9973$$

Probability of product that fail to meet the specification limits

$= P(X<LSL) + P(X>USL)$

$= 1 - P(LSL<X<USL)$

$= 1 - 0.9973$

$= 0.0027$

$= 0.27\%$

If the underlying distribution for the process output is not normal, or it is not center at the midpoint of the specification limits, then the percentage of product failing to meet specifications can be much longer than 0.27%.

One of a drawback of the C_p index is that it really evaluates only process spread and ignores the process average. If the process is not centered at the middle of the specifications, the C_p index may be misleading.

Ideally the C_p should be as high as possible. The higher the C_p, the lower the variability with respect to the specification limit. However a high C_p value does not guarantee a production process falls within specification limits because the C_p value doesn't imply that the actual spread coincided with the allowable spread (i.e the specification limits).

Process Performance Index, C_{pk}

The process capability index or C_{pk} is an appropriate measure a process ability to create product within specification limits. The C_{pk} tells how well a process can meet specification limit while accounting for the location of the average or center. C_{pk}

represents the difference between the actual process average and the closest specification limit over the standard deviation times three.

A C_{pk} index is defined as follows :

$$C_{pk} = \text{mimimum}\frac{[USL - \mu, LSL - \mu]}{3\sigma}$$

A C_{pk} of 1 indicates that the system is producing at least 99.73% within the specification limits. If the process is centered on its target value, the values for C_p and C_{pk} will be equal. When the C_{pk} is less than one, the process is referred to as incapable.

When the C_{pk} is greater than or equal to one, the process is considered capable of producing items within specification limits. The C_{pk} is inversely proportional to the standard deviation, or variability, of a process. The higher the C_{pk}, the narrower the process distribution as compared with the specification limits, and the more uniform the product. As the standard deviation increase, the C_{pk} index decreases. At the same time, the potential to create product outside the specification limits increases.

CHAPTER III

ROBUST ESTIMATOR AND BOOTSTRAP METHOD

Introduction

An important assumption underlying process capability analysis and process capability indices is that their usual interpretation is based on a normal distribution of process output, if not normal then the estimate of process capability index is unlikely to be correct. One approach to dealing with this situation is to transform the data so that the new transformed data has a normal distribution appearance. Other approach may be to extend the definition of the standard capability indices to the non-normal case. In this project, we proposed replacing the classical estimator in calculating the capability index with the robust estimator..

The data set which consist unusual observations are called outliers. They can be due to genuinely long-tailed distribution or gross errors. An outlier will influence any estimator such as the location estimator, specifically the mean. These explain why we need a robust statistical procedures to overcome this problem. The existing of outliers in sample data set may cause an incorrect capability index value even when data is sampled from a normally distribution population.

Robust estimator

To calculate the process capability index, there are two estimators involve μ and σ. μ is estimated by sample mean($\bar{x}$) and σ by sample standard deviation(s). $\bar{x}$ and s are point estimators that based on assumption where the sampled population is

approximately normal distributed. For the distribution which is not normal or the sample have outliers, such estimators may not have the desirable properties like unbiasedness and minimum variance. So, we need a location estimator which is insensitive to the outliers. These resistant estimators are called robust estimator. The robust estimators required two properties which are resistance(Breakdown-point by Hampel (1971)) and robustness of efficiency. These estimators can perforns well for a very range of probability distribution. There are many robust estimators that can be used to estimate μ and σ such as MAD(Median Absolute Deviation)(Andrew at al ,1972), M-estimator(Huber, 1981) and Mm-estimator(Yohai, 1987). In this project, we consider two robust estimators namely, MM and MAD estimators which can replace $\bar{x}$ and s in traditional estimators and compare the performance of these two estimators.

M-estimator

M-estimator is the robust estimator of the population center location. It compares favourably with the sample mean when the sampled population is normally distributed, and is considerably better than the sample mean when the underlying distribution for sampled population is heavy -tailed. The M-estimator with monotone psi-function by Huber (1973) have a breakdown-point which is equal to 0.

For a given set of observations $y_1, y_2, \ldots\ldots, y_n$. The M-estimate θ for the function ρ and the sample estimate value $\hat{\theta}$ that minimize the objective function

$$\sum_{i=1}^{n} \rho(y_i, \hat{\theta})$$

(see David, Frederick and Tukey(1983))

We also can find the value of $\hat{\theta}$ that satisfies the following equation

$$\sum_{i=1}^{n} \psi(y_i, \hat{\theta}) = 0$$

where ψ is the first derivative of ρ with respect to θ.

Mm-estimator

Yohai(1987) introduced a new estimator called as MM-estimator having simultaneously high breakdown-point = 0.5 and high efficiency under normal errors Yohai's estimators are defined in three stages.

Stage 1:

Using Least Median of Square(LMS) or LTS method or other initial estimate see (Yohai (1987)) to estimate value θ_0 with high breakdown-point, possibly 0.5.

Least Median of Squares (LMS) in one dimension

In one dimensional-location, the LMS reduces to

$$\underset{\hat{\theta}}{Minimize} \ \underset{i}{med} \ (y_i - \hat{\theta})^2$$

From (Rousseeuw and Leroy)(1987), the Theorem makes it easy to compute LMS in the location case. The $\hat{\theta}$ is the center of the half samples of the shortest length when $\hat{\theta}$ satisfies

$$\underset{\hat{\theta}}{Minimize} \ \underset{i}{med} \ (y_i - \hat{\theta})^2$$

This is done by finding the smallest of the differences

$$y_{(h)} - y_{(1)}, y_{(h+1)} - y_{(2)}, \ldots\ldots, y_{(n)} - y_{(n-h+1)}$$

where h=1 + [n/2] and $y_{(1)} \leq y_{(2)} \leq \ldots\ldots y_{(n)}$ are the ordered observations.

For example, if the $y_{(i)}$ and $y_{(j)}$ be the shortest length where $j = i + [n/2]$

Then the location estimator

$$\hat{\theta} = \frac{y_{(i)} + y_{(j)}}{2}$$

Stage 2:

Compute the residuals

$$r_i = y_i - \hat{\theta}_0 \qquad 1 \leq i \leq n$$

and compute the M-scale s(u). The estimates s(u) is defined as the value of s which is the solution of

$$\frac{1}{n}\sum_{i=1}^{n} \rho(r_i / c_0 s) = b \quad \text{----------(1)}$$

c_0 = *tuning constant*

where b may be defined by $E_{\Phi}(\rho(u)) = b$ and Φ stands for the standard normal distribution

Huber(1981) proved that the use of a function ρ_0 and a constant b such that

b/a = 0.5

where a = max ρ_0 (u) will implies that this scale estimate has the breakdown-point equal to 0.5.

Stage 3:

Let be another function such that

$$\rho_1(u) \leq \rho_0(u)$$

$$\sup \rho_1(u) = \sup \rho_0(u) = a$$

then the MM-estimate T_1 is defined as any solution of

$$\sum_{i=1}^{n} \psi_1(r_i / cs) x_i = 0$$

which verifies

$$S(T_1) \leq S(T_0)$$

where

$$S(\theta) = \sum_{i=1}^{n} \rho_1(r_i / cs)$$

and where $\rho_1(0/0)$ is defined as 0.

For the project, the calculation part of three stages of MM-estimator are shown below:

Stage 1:

Using Least Median of Square(LMS) to get initial location estimate and the initial scale estimate.

Stage 2

The solving of equation (1) using the root finding method.

Let

$$f(u) = \frac{1}{n} \sum_{i=1}^{n} \rho(u) - b$$

where

$$u_i = \frac{y_i - \hat{\vartheta}}{c_0 s} \qquad c_0 = 1.56$$

The method of finding c_0 is taken from Yohai(1987), then find the root of the equation using Regula-Falsi Method as suggested by Asaithambi(1995). The value of a = 1/6 when using Tukey's bisquare function, so the constant b = 1/6 * 0.5

In this project, (see Yohai, 1987) we choose to use bisquare function because an outlier does not effect $\psi(u) = 0$ if $|u|$ is sufficiently large. Morever, the estimator is not as sensitive to small changes in the data values as is the median. Since $\psi(u) \approx u$ for small u, near the center of the sample the bisquare behaves like the mean.

The function tukey's bisquare is give by

$$\rho(u) = \begin{cases} u^2/2 - u^4/2 + u^6/6 & \text{if } |u| \le 1 \\ 1/6 & \text{if } |u| > 1 \end{cases}$$

and the corresponding bisquare psi-function

$$\psi(u) = \begin{cases} u(1-u^2)^2 & \text{if } |u| \le 1 \\ 0 & \text{if } |u| > 1 \end{cases}$$

Stage 3

Computation algorithm

The computational algorithm for MM-estimate is a modified version of the iterated weighted least-squares (IWLS) algorithm used for computing M-estimates.

Let

$$y_i = \theta + e_i \qquad 1 \le i \le n$$

where θ is an unknown one-dimensional parameter and is normal distributed with mean zero and standard deviation σ . In this situation , an estimator T_0 of θ is called a univariate location estimator and s of σ is called s scale estimator.

Suppose we have already computed the initial estimate T_0 (using the LMS method) and scale estimate s in stage 2. The weight function is define as

$$w_i(u_i) = \frac{\psi(u_i)}{u_i}$$

and in matrix [n x 1]

where the $w_i(u_i)$ for biweight or bisquare

$$w_i(u_i) = \begin{cases} (1 - u_i^2)^2 & \text{if } |u| \le 1 \\ 0 & \text{if } |u| \ge 1 \end{cases}$$

also define

$$g(u_i) = X_i^T W_i u_i$$

(dimension for $g(u_i)$ = [p x 1])

where X_i = matrix dimension [nxp]

X_i^T = matrix transposes dimension [pxn]

(p = number of parameters in the model. Since we want the location estimate, so p = 1)

$$W_i = [\, w_i(u_i)][w_i(u_i)]^T$$

(W_i is a square matrix with dimension = [nxn] and the element not diagonal is 0)

$$u_i = \frac{y_i - \hat{\theta}}{c_1 s} \qquad , c_1 = 4.68 \text{ see Yohai (1987)}$$

(dimension for u_i = [nx1])

and

$$M(u_i) = \frac{X_i^T W_i X_i}{c_1 s^2}$$

(dimension for M(u_i) = [pxp])

The recursion step of the ILWS is defined as follows. If $t^{(j)}$ is the value of the estimate in the j^{th} step, then $t^{(j+i)}$ is defined by

$$t^{(j+i)} = t^{(j)} + \Delta (t^{(j)})$$

where

$$\Delta (t) = M^{-1}(t)\, g(t)$$

MAD(Median Absolute Deviation) is a robust estimate of scale proposed by Andrew et al (1972), where

$$MAD = median \left| x_i - \hat{\theta} \right|$$

where $\hat{\theta}$ is MM estimator.

Bootstrap

In statistical analysis, the researcher is usually interested in obtaining not only a point estimate of a statistic but also an estimate of the variation in this point estimate, and a confidence interval for true value of the parameter. Traditionally , researcher have relied on the central limit theorem and normal approximation to obtain standard errors and confidence intervals. Now , with the availability of modem computing power, they may use resampling methods such as the bootstrap and jackknife which provide estimates of the standard error, confidence intervals and distributions for any statistic.

Bootstrap Method

The method of bootstrap was introduced by Efron (1979). The statistical method of Bootstrapping would be helpful to calculate confidence interval estimation technique for C_{pk} that is nonparametric or free from assumption of the distribution. Bootstrap method is a computer intensive based method that can replace the theoretical assumptions and analysis with considerable amount of computation. This means it can computes the estimated standard error ($\hat{\sigma}$), biases, confidence interval in an unfamiliar way; purely by computational means ,rather than through the use of mathematical formulas.

Suppose we have a random sample drawn independently from one member of a parametric family of distribution. The bootstrap procedure is to obtain repeated samples (each with the same size as observed data) with replacement (new data sets) from the observed data.