Bootstrap Confidence Interval for Robust C_{pk} Index

[1]Habshah Midi & [2]Foo Lee Kien

[1]Laboratory of Applied and Computational Statistics
Institute for Mathematical Research
Universiti Putra Malaysia

[2]Multimedia University, Malaysia

[1]habshahmidi@hotmail.com

Abstract

A C_{pk} index is the most widely used measure of process capability. A better understanding and interpretation of C_{pk} is by constructing its confidence interval. The construction of such intervals, are normally based on the assumption that the measurements process can be treated as samples from a normal distribution. Nevertheless, many processes are not normal and have a heavy tail distribution, which probably due to the presence of outliers in the data. An alternative approach is to use a bootstrap confidence interval estimate of C_{pk} index which is based on robust method. The advantage of this estimate is that its usage does not depend on any distribution. A numerical example is presented to evaluate the performance of the bootstrap confidence interval of the robust C_{pk} index. The result indicates that the bootstrap confidence interval estimate based on the robust C_{pk} is better than the classical C_{pk} index.

Introduction

The fundamental tasks of process capability indices are to determine whether a manufacturing production process is capable of producing items within specification limit (see Grant & Leavenworth, 1988; Koronacki, 1993; Montgomery, 1997; Thompson, 1993). Process capability indices are used widely throughout industries, to give a relatively quick indication of process capability in a format that is easy to compute and understand. The most widely used process capability indices are the C_{pk} index. The C_{pk} index is developed to indicate how process conforms to two sided specification limits and it is used to measure the actual capability of a process.

Suppose, the lower (LSL) and the upper specification limits (USL) has been set by the manufacturer, then the C_{pk} is defined as

$$C_{pk}=\min\left[\frac{USL-\mu}{3\sigma}\;,\;\frac{\mu-LSL}{3\sigma}\right] \tag{1}$$

Usually, the value of the process mean, μ and process standard deviation, σ, is unknown and must be estimated using the sample mean, $\bar{x}$ and the sample standard deviation, s from a stable process. Then the estimated C_{pk} index are written as

$$\hat{C}_{pk}=\min\left[\frac{USL-\bar{x}}{3S}\;,\;\frac{\bar{x}-LSL}{3S}\right] \tag{2}$$

The drawback of using $\bar{x}$ and s as an estimator is that it is very sensitive to outliers. Indeed, one single outlier can have an arbitrarily large effect on the estimate. As an alternative, a robust location and scale estimates which are less affected by outliers are proposed to replace $\bar{x}$ and s in (2).

Once the C_{pk} index is estimated, the production engineer will compare this value with the recommended minimum value and will consider the process to be capable if the estimated value is larger than or equal to the minimum value. A manufacturing engineer recommended a value of C_{pk} equals to 1.33 is the minimum value that should be observed for the acceptable process capability. It is very important to point out that such practice ignores the fact that C_{pk} index is a random variable because it is calculated from different samples of the same size taken independently from the same stable distribution. Unfortunately, convenient mathematical formulas for estimating sampling variability of C_{pk} have not been developed for the widely differing distributions that are encountered in industrial applications. Fortunately, Chou, Owen and Borego (1990) have provided correct understanding and interpretation of the C_{pk} index by constructing its confidence limits estimates. Nevertheless, the calculation of such intervals has been based on the characteristics that the process having a normal distribution. In practice, many processes are not normal, perhaps due to the presence of outliers in the data set. Inferences based on assumed normal process can be very misleading especially when the underlying distribution is not normal. In order to obviate from this problem, a confidence interval estimation technique for C_{pk} which is free from any normality assumptions of the process is utilized. Such technique is called a bootstrapping method which was first introduced by Efron (1979).

In this paper we want to investigate the performance of the bootstrap confidence intervals of the C_{pk} index based on M-estimator in terms of their coverage probability. We proposed to replace the value of $\bar{x}$ and s in (2) with more robust estimates such as M estimator as proposed by Huber (1973). A 'good' confidence interval is one which possesses a reasonably accurate coverage probability and 'good' equitailness. Equitailness means that a confidence interval for θ of level (1-2 α) is such that the percentage for θ to lie outside the interval is divided equally between the lower and upper limits of the intervals.

Robust C_{pk} Index based on M-estimators

One reason for the need of robust statistical procedures is the existence of outliers in the data set. Outliers are data which are far from the bulk of the data. They can be due to genuinely long-tailed distribution or gross errors. Examples for gross errors may be copying or punching errors, wrong decimal point, wrong scales of measurements taken over a period of time, and others. This means that even when data is sampled from a normally distributed population, existence of outliers in sample data set may cause an incorrect capability index value.

In general, two parameters are involved in calculating the process capability index. The two parameters are μ and σ. μ is usually estimated by sample mean ($\bar{x}$) and σ is usually estimated by sample standard deviation(s). When the distribution of the sampled population deviates greatly from normality, such estimators may not have the desirable properties like unbiasedness and minimum variance. As an alternative, we may turn to robust estimator, for example, the M-estimator.

In calculating C_{pk} index, we want reliable estimators but would like to avoid using strong assumption about the behavior of the estimators. M-estimators can be applied to a location model. It compares favorably 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. Instead of looking at estimator $\hat{\theta} = T_n(x_1, x_2, ..., x_n)$ which minimize the objective function $\sum_{i=1}^{n}(x_i-\theta)$ for sample mean, a scale equivariant M estimates of location is obtain such that :

$$\hat{\mu} = \underset{\mu}{\arg\min} \sum \rho\left(\frac{x_i-\mu}{\hat{\sigma}}\right).$$

Since $\hat{\sigma}$ does not depend on μ, $\hat{\mu}$ is a solution of $\sum \psi\left(\frac{x_i-\mu}{\hat{\sigma}}\right)=0$ where ψ is the derivative of ρ with respect to μ.

In this paper, the Huber M-estimator will be used with c as the tuning constant ($c>0$). For Huber M-estimator (see Huber, 1981), the ρ , function is defined as

$$\rho(x;\theta)=\begin{cases} \dfrac{x^2}{2} & ; |x| \le c \\ c|x|-\dfrac{1}{2}c^2 & ; |x| > c \end{cases}$$

thus, the derivative of this ρ function is

$$\psi(x;\theta)=\begin{cases} x & ; \ |x| \le c \\ c\,\mathrm{sgn}(x) & ; \ |x| > c \end{cases}$$

This Huber ρ and ψ function will down-weights outliers.

Instead of using sample standard deviation s to estimate C_{pk} in (2), we proposed using the median absolute deviation (MAD) as suggested by Andrew *et al.* (1972). The median absolute deviation is a robust estimate of scale where

$$\mathrm{MAD} = \mathrm{median}\left|x_i - \mathrm{median}(x)\right|$$

To make the MAD comparable to the standard deviation, we usually consider the normalized MAD defined as MADN=MAD(x)/0.6745

Bootstrap Method

The bootstrap method was introduced by Efron in 1979. It can be utilized to calculate confidence interval without relying on assumption (such as normality) for the underlying population distribution. Bootstrap method is a computer intensive method that can replaced theoretical assumptions and analysis with considerable amount of computation. This means it can compute the estimated standard error ($\hat{\sigma}$), biases, confidence intervals, etc., in an unfamiliar way; purely by computational means, rather than through the use of mathematical formulas.

In this paper, bootstrap method will be adopted to estimate the confidence intervals for process capability index. In order to obtain the confidence interval for C_{pk} the standard errors of the estimates are required. In practice, the estimated standard error $\hat{\sigma}$, are usually employed to form approximate confidence intervals for C_{pk}. The usual (1-2α) 100% confidence interval for C_{pk} is, $\hat{C}_{pk} \pm \hat{\sigma}\, Z_{\alpha/2}$ where $Z_{\alpha/2}$ is the 100 (α/2) percentile point of a standard normal distribution. The validity of this interval depends on the assumption that $\hat{C}_{pk}$ is normally distributed. Otherwise, the approximate confidence interval will not be very accurate. Bootstrap confidence intervals do not rely upon the usual assumption of normality; it can give the standard error of the estimates automatically via a Monte Carlo algorithm. This algorithm makes use of re-sampling schemes where bootstrap samples are obtained. Bootstrap samples are repeated samples of the same size as the observed sample taken with replacement from the observed sample. Procedure to obtain the Bootstrap standard error is as follows:

1) Let a sample of size n is taken from a process that has distribution F,
   $x_1, x_2, \ldots, x_n \sim F$
2) A bootstrap sample is drawn from the above sample data set.
   $x_1^*, x_2^*, \ldots, x_n^* \sim F$
   Estimate the process performance index for this bootstrap sample
   $\hat{C}_{pk}^* = \hat{C}_{pk}(x_1^*, x_2^*, \ldots, x_n^*)$.
3) Independently repeat step 2 for B times, obtaining bootstrap replication for process capability index
   $\hat{C}_{pk}^*(1), \hat{C}_{pk}^*(2), \ldots, \hat{C}_{pk}^*(B)$

Hence, calculate

$$\hat{C}_{pk}^*(.) = \sum_{b=1}^{B} \frac{\hat{C}_{pk}^*(b)}{B}$$

Then, the estimated standard error

$$\hat{\sigma} = \left\{ \frac{1}{B-1} \Sigma \left[\hat{C}_{pk}^*(b) - \hat{C}_{pk}^*(.) \right]^2 \right\}^{\frac{1}{2}}$$

The number of bootstrap replications, B, depends on the application. For standard error estimates, Efron and Tibshirani (1993) suggested B to be between 25 and 200. However, Efron and Tibshirani (1993) pointed out that for confidence intervals construction, B should be 500 or 1000 in order to make the variability of the estimates acceptably low.

Franklin and Wasserman (1991) has employed the standard, percentile and bias corrected percentile confidence interval on the classical C_{pk} index. In this paper, we will employ the Bias Corrected and Accelerated confidence interval (also known as Bca) for the classical and the robust C_{pk} index. Efron and Tibshirani (1993) enumerated that the Bca method has two attractive properties, i.e. it is transformation respecting and second-order accurate.

A Simulation Study

In an effort to evaluate the performance of the bootstrap confidence intervals of the classical C_{pk} index and the bootstrap confidence intervals of the C_{pk} index based on M-estimator, a series of simulations were carried out, one on a normal process and another on a highly skewed process. The value for the upper specification limit and the lower specification limit are chosen to be 3 and -3, respectively. For the normal process, the data was generated from a standard normal distribution, i.e. $x \sim N(0,1)$. Following Kotz and

Johnson (1993), two non-normal processes were generated from a skewed distribution with finite lower boundary, namely χ^2 distribution with 4.5 degree of freedom and a heavy-tailed distribution, namely a student-t distribution with 8 degree of freedom. In each case, the distribution is standardized by suitably shifted and scaled the distribution to produce common mean (μ=0) and common standard deviation (σ=1). We also investigated the properties of the bootstrap confidence intervals by considering a data with 10% outliers. At each step, one 'good' observation from a $N(0,1)$ was deleted and replaced with a 'bad' datum point until 10% outliers are obtained. The outliers are generated from a normal distribution with μ=20 and σ=5. For each process, we considered two samples of size n=20 and n=40.

The performance of the bootstrap confidence interval is assessed by it's coverage probability and equitailness. The Bias Corrected and Accelerated (Bca) bootstrap confidence method is applied to the classical C_{pk} index and also the C_{pk} index based on M-estimator. In this paper, the value of the tuning constant used for the ρ function is equal to 1.45. 1000 bootstrap samples were drawn from a sample of size 20 and 40 and a bootstrap 95% confidence interval was constructed for the classical and the robust C_{pk} index. This procedure was conducted on normal process, skewed process and process which have 10% outliers. For each confidence interval of each type, it was then determined whether the confidence interval actually contained the true value C_{pk}. The simulation runs were replicated 500 times to obtain the coverage probability, i.e. the percentage of times the actual C_{pk} was contained in the intervals out of 500 could be calculated. The results of the simulation studies are summarized in Table 1-2.

It is interesting to note that for normal process, both type of confidence intervals have coverage which are reasonably closed to each other. The values are consistently near the expected value of 0.95 for both sample sizes of n=20 and 40. Nonetheless, the performance of the classical C_{pk} is remarkably deteriorating for skewed process, i.e. for chi-square and t distribution. Its coverage probability was lower than the expected value of 0.95 by 5 and 11 for t and χ^2 process, n=20, respectively. The performance of the C_{pk} index based on M-estimator seem to be less affected by the skewed process indicated by a slight decreased in the coverage probability.

The addition of 10% outliers to the 'clean' process prominently reduced the coverage of the probability by almost 100%. It gives erroneous results not only from the point of view of equitailness but also from the point of view of coverage probability. On the other hand, the C_{pk} based on M-estimator has slightly effected by the outliers.

Table 1: Coverage Probabilities for the bootstrap 95% Confidence Intervals, n=20

Process Distribution	Estimate C_{pk} using:	Coverage (%)	Lower Coverage (%)	Upper Coverage (%)
Normal	$\bar{x}, s$	93	3	5
(0,1)	robust location and scale	95	1	4
Student-t	$\bar{x}, s$	90	5	5
(ν=8)	robust location and scale	96	2	2
χ^2	$\bar{x}, s$	84	12	4
(ν=4.5)	robust location and scale	94	6	0
10%	$\bar{x}, s$	35	0	65
outliers	robust location and scale	94	0	6

Table 2: Coverage Probabilities For the bootstrap 95% Confidence Intervals, n=40

Population Distribution	Estimate C_{pk} using:	Coverage (%)	Lower Coverage (%)	Upper Coverage (%)
Normal	$\bar{x}, s$	93	2	5
(0,1)	robust location and scale	93	2	5
Student-t	$\bar{x}, s$	91	4	5
(ν=8)	robust location and scale	94	4	2
χ^2	$\bar{x}, s$	88	8	4
(ν=4.5)	robust location and scale	93	7	0
10%	$\bar{x}, s$	0	0	100
outliers	robust location and scale	89	0	11

Conclusion

The advantage of using bootstrap confidence interval for constructing the C_{pk} index is that its usage does not require the usual assumption of normality. We do not have to worry whether the process is normal or not because it can be calculated from any underlying process distribution. The empirical studies suggest that for a normal process, the classical C_{pk} and the robust C_{pk} have coverage probabilities which are reasonably closed to the nominal one. It is quite sufficient for the stated "95% confidence interval" to perform as a 95% confidence interval. However, the performance of the classical C_{pk} in a skewed process and 10% outliers cannot be trusted because its coverage decreased remarkably low and has a very poor equitailness. The performance of the C_{pk} index based on the M estimator seems to be better than the classical C_{pk} since its coverage probability appears to be closer to the nominal values in a skewed process than the classical coverage probability.

References

[1] Andrews, D. F., Bickel, P.J., Hampel, F.R., Huber, P.J., Rogers, W.H. and Turkey, J.W. 1972. *Robust Estimates of Location: Survey and Advances*. New Jersey: Princeton University Press.

[2] Chou, Y., Owen, D.B. and Borrego, S.A. 1990. Lower Confidence Limits on Process Capability Indices. 22 *Journal of Quality Technology*, **22**(3): 223-229.

[4] Efron, B. 1979. Bootstrap methods: another look at the jackknife. *Ann. Statist*, **7**: 1-26

[5] Efron, B. and Tibshirani, R.J. 1993. *An Introduction to the Bootstrap*. New York: Chapman and Hall.

[6] Franklin and Wassermann. 1991. Bootstrap Confidence Interval Estimation of C_{pk}: An Introduction. *Commun. Statist-Simula.* **20**(1). 231-242.

[7] Grant, E.L. and Leavenworth, R.S. 1988. *Statistical Quality Control*. New York. McGraw-Hill Book.

[8] Huber, P.J. 1981. *Robust Statistics*. New York: John Wiley.

[9] Koronacki, J. 1993. *Statistical Process Control for Quality Improvement*, New York.

[10] Grant, E.L. and Leavenworth, R.S. 1980. *Statistical Quality Control*, New York: McGraw- Hill.

[11] Kotz, S. and Johnson, N.L. 1993. *Process Capability Indices*. London: Chapman and Hall.

[12] Montgomery, D.C. 1997. *Introduction to Statistical Quality Control*. New York: John Wiley.

[13] Thompson, J.R. 1993. *Statistical Process Control for Quality Improvement*. New York: Chapman and Hall.