TOTAL QUALITY MANAGEMENT, VOL. 14, NO. 3, 2003, 341–354

Confidence intervals in repeatability and reproducibility using the Bootstrap method

FU-KWUN WANG[1] & ELDON Y. LI[2]

[1]*Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, 10626, Taiwan &* [2]*Orfalea College of Business, California Polytechnic State University, San Luis Obispo, USA*

ABSTRACT *The repeatability and reproducibility (R&R) study—also called a gauge capability study—has been employed as part of the statistical process control program in many organizations. The objective of the study is to determine whether a measurement procedure or instrument is adequate for monitoring the performance of a process. The classical control chart method can be easily performed and calculations can be done with a spreadsheet or statistical computer software. However, this approach only provides the point estimates on the variance components of the measurement error study. In many situations, confidence intervals are more useful than point estimates, because an interval estimate enables an engineer to see both how small and how large an effect may be. In this paper, the Bootstrap method is used for obtaining the confidence intervals of the gauge variability when the control chart method is use for finding the point estimates. One real-life example is used to show the application of this control chart with the Bootstrapping method and comparisons are made with three experimental design procedures in terms of point estimates and confidence intervals for repeatability, reproducibility and total gauge variability.*

Introduction

Total quality management (TQM) implementation is one of the most complex tasks that a company can attempt (Yusof & Aspinwall, 2000a). In addition, success in TQM implementation is still very difficult to come by for many organizations (Taylor, 1996). The main reason is that the implementation involves a dramatic change in corporate culture and affects all members in an organization (Kanji & Barker, 1990). Notwithstanding, many organizations are introducing some form of TQM into their operations (Martin, 1997). A recent survey of 402 US quality assurance professionals reveals that the existence of quality improvement measurement systems is among the top five important critical success factors (CSFs) for TQM implementation (Dayton, 2001). These top five CSFs are strategic quality management, corporate quality culture, people and customer management, quality improvement measurement systems, and customer satisfaction orientation (Black & Porter, 1996).

One of the major components of quality improvement measurement systems is the statistical process control (SPC) program. In today's manufacturing industry, many companies use SPC to ensure and improve quality (Yusof & Aspinwall, 2000b). The SPC program

Correspondence: F.K. Wang, Department of Industrial Engineering and Management, National Taipei University of Technology, Taipei, No. 1, Chung Hsiao East Rd., Sec.3, Taipei, 10626, Taiwan.

ISSN 1478-3363 print/ISSN 1478-3371 online/03/030341-14
DOI: 10.1080/1478336032000046643

relies extensively on measurement and test data as the primary inputs to the SPC system. The goal of testing and measuring is to provide results that are both accurate and precise. A variance or a standard deviation is commonly used to quantify these results. The repeatability and reproducibility (R&R) study, also called the gauge capability study, is usually employed to provide quantitative information about the performance of a measurement process. According to Taylor & Kuyatt (1994), a repeatability condition occurs when the standardized measuring procedure is carried out on identical test material under the same conditions (same operator, same apparatus, same laboratory, and short intervals of time). A reproducibility condition occurs when the standardized measuring procedure is carried out on identical test material but under different conditions such as operators, apparatus, laboratories and times. With standardized test methods, the standard deviations obtained under 'repeatability conditions' and 'reproducibility conditions' can be distinguished easily. We can define $\sigma^2_{\text{repeatability}}$ as the variability due to repeated measurements on a single part made by the same operator. This variability represents the basic inherent precision of the gauge itself. Similarly, we can define $\sigma^2_{\text{reproducibility}}$ as the variability due to different operators under the same gauge condition. This variability represents the product uniformity resulting from a process that is in the state of statistical control. Therefore, the total gauge variability is defined as the sum of the repeatability variability and the reproducibility variability, $\sigma^2_{\text{repeatability}} + \sigma^2_{\text{reproducibility}}$. The objective of the R&R study is to determine whether a measurement procedure or instrument is adequate for monitoring a process, which is relative to the precision of the gauge. If the measurement error is small relative to the total process variation, then the measurement procedure is deemed 'adequate'. Similarly, if the total process variation is small, it implies that a process has more stability.

For the purpose of statistical process control, confidence intervals must be constructed. Confidence intervals are based on the relative frequency property of probability. They can provide a clear and concise expression of the effect size whilst allowing for random variation (sampling variation). This gives us an idea about the relative uncertainty associated with the estimates and provides some basis for sample size decisions. A review of the literature reveals that there are many R&R studies proposing various ways of estimating variances and confidence intervals (Montgomery & Runger, 1993, Hoguet, 1994, Burdick & Larsen, 1997). However, none of them give these estimates for the gauge variability using a mixture of the control chart method and the Bootstrap method. This study is the first to fill the void. It uses one real-life example to show the application of the Control Chart With Bootstrapping (CCWB) method to constructing point estimates and confidence intervals for components of repeatability, reproducibility and total gauge variability. The results are compared with those from three experimental design procedures: the maximum likelihood, the restricted maximum likelihood, and the modified large sample method.

This article is organized as follows. First, several related previous studies are reviewed, followed by the measurement error study. Second, the control chart method and the Bootstrap method are discussed. The experimental design method is then reviewed and a real-life example is illustrated to compare the performance. Finally, conclusions and recommendations are drawn from the findings in this study.

Previous studies

In an R&R study, Hoguet (1994) discussed three different gauge methods: a quick-look procedure for a gross estimate of capability, the IBM Formal Gauge Repeatability and Reproducibility Study, and the Miles Short Method. The quick-look procedure can be performed quickly using two operators, both of whom measure a set of five parts once each time. It can quickly indicate whether a measurement is suitable or not. However, there is no suggested direction for improvement from this study if the capability consumes a significant percentage of the specification, or perhaps all of it. The Miles Short Method is particularly suited for the processing industry and can be used for comparing any two groups, such as two machines, two processes, two operators, etc. The IBM Formal Gauge Repeatability and Reproducibility Study is a control chart approach that can be easily performed and calculated with a spreadsheet or statistical software. However, this approach only provides point estimates on the variance components of a measurement error study.

There are many ways of constructing a confidence interval that shows the precision of the estimate under a particular process at a certain confidence level. Deutler (1991) described the Grubb-Type estimators for reproducibility variances. Montgomery & Runger (1993) treated the classical gauge R&R study as a designed experiment and discussed how to construct confidence intervals on the variance components. Burdick & Larsen (1997) discussed several methods based on the ANOVA approach for constructing confidence intervals on measures of variability in a classical R&R study. These methods included the modified large sample (MLS), the Satterthwaite's approach (SATT), the Automotive Industry Action Group (AIAG), and the restricted maximum likelihood (REML). The simulation results show that the MLS intervals can be applied for all the parameters. They maintain the stated confidence level and are comparable in length to alternative methods that maintain the stated confidence level. Borror *et al.* (1997) presented a comparison between the restricted maximum likelihood (REML) method and the modified large sample (MLS) method for constructing confidence intervals on gauge capability studies. They found these two methods give similar results when the number of operators is sufficiently large. Dolezal *et al.* (1998) presented a method for constructing confidence intervals on the measurement variability in a two-factor experiment in which one factor was fixed and the other factor was randomized. They found that the mixed model is a better representation of the R&R; failure to recognize this fact will result in confidence intervals that are much wider than necessary to obtain the desired level of confidence. Recently, Wang & Chen (1998) applied the Bootstrap method to obtaining the confidence intervals for multivariate capability. Their method appeared to be a useful procedure for constructing confidence intervals with high accuracy.

Measurement error study

An important assumption of many SPC implementations is the adequate capability of the gauge and the inspection system. In any process involving measurement of manufactured products, some of the observed variability will be due to variability in the product itself, and some will be due to measurement error or gauge variability. Expressed mathematically,

$$\sigma^2_{\text{total}} = \sigma^2_{\text{product}} + \sigma^2_{\text{measurement error}} \tag{1}$$

where σ^2_{total} is the total variance of the observed process, $\sigma^2_{\text{product}}$ is the component of variance due to the product, and $\sigma^2_{\text{measurement error}}$ is the component of variance due to measurement error. The measurement error and the product measurement are assumed to be independent of each other. Furthermore, we used the previous definitions of repeatability, reproducibility and total gauge variability. Thus, the variability $\sigma^2_{\text{measurement error}}$ is the sum of two variance components, say

$$\sigma^2_{\text{measurement error}} = \sigma^2_{\text{gauge}} = \sigma^2_{\text{repeatability}} + \sigma^2_{\text{reproducibilty}} \tag{2}$$

In order to provide useful information of gauge capability, we need to answer the following questions.

- How much variability exists among repeated measurements on a single sample made by one operator?
- How much variability is attributed to using different operators to measure the same sample?
- What is the ratio of the process variability to the measurement variability?

Usually, the control chart methods and the experimental design method can be used to separate these components of variance, as well as to give an assessment of R&R capability. Measurement error reduces the ability to sense (thus to control) the variation of the process and the reliability of product measurements and inspections. To have a good process control, the variance of the measurement components must be small compared with the process component. Therefore, finding the R&R of a gauge is one of the important steps in implementing SPC programs.

The control chart method

Variables control charts such as *X*-bar/*R* charts and *X*-bar/*S* charts are important and useful online statistical process control techniques. The *X*-bar/*R* charts or *X*-bar/*S* charts can be used to derive estimates of the variance components on an R&R study. When used to summarize data from repeated measurements of a set of product test samples (as shown below), the *X*-bar chart shows the discriminating power of the measurement instrument; that is, the ability of the gauge to distinguish between units of the product. The variability of the gauge reproducibility arises because of differences among the operators. Suppose that l operators are involved in a study with each operator conducting the h trials on n sample parts. Let x_{ijk} denote the measurement by operator i on part j at trial k. Then, $\bar{X}_i$ is obtained as the average of all trials on all sample parts by the operator i. That is,

$$\bar{X}_i = \frac{\sum_{j=1}^{n} \sum_{k=1}^{h} x_{ijk}}{nh}, \forall i = 1, 2, \ldots, l$$

If the $\bar{X}_i$ values differ, the reason will be operator bias, since all operators measure the same parts. Therefore, the estimated gauge reproducibility is defined as

$$\hat{\sigma}_{\text{reproducibility}} = \frac{\max(\bar{X}_1, \ldots, \bar{X}_l) - \min(\bar{X}_1, \ldots, \bar{X}_l)}{d_2} = \frac{X_{\text{diff}}}{d_2}$$

where values of d_2 denote various sizes of the operators. The R chart directly shows the magnitude of measurement error, or the gauge capability, often called gauge repeatability. The values of R represent the difference between measurements made on the same product using the same instrument. Therefore, the estimated gauge repeatability is defined as

$$\hat{\sigma}_{\text{repeatability}} = \frac{\sum_{i=1}^{l} \bar{R}_i / l}{d_2} = \frac{\bar{\bar{R}}}{d_2}$$

where $\bar{R}_i$ is calculated by summing the values of the individual subgroup ranges and dividing by the number of sample parts by the operator i; that is,

$$\bar{R}_i = \frac{\sum_{j=1}^{n} R_j}{n}$$

and values of d_2 denote various numbers of trials. Out-of-control points on the R chart would indicate that the operator was having difficulty using the instrument. Otherwise, if the R chart were in control, the operator should have no difficulty using the instrument. From the above discussion, the estimated total gauge variability can be obtained as in equation (2). However, the R chart method for estimating $\hat{\sigma}_{\text{repeatability}}$ loses statistical efficiency for moderate to large samples. If the sample size is moderately large, say $n > 10$, the S chart should generally be used instead of the R chart, where S is the sample standard deviation. Thus, the estimated gauge repeatability can be defined as

$$\hat{\sigma}_{\text{repeatability}} = \frac{\sum_{i=1}^{l} \bar{S}_i / l}{c_4} = \frac{\bar{\bar{S}}}{c_4}$$

where $\bar{S}_i$ is calculated by summing the values of the individual subgroup standard deviations and dividing by the number of sample parts from the operator i; that is,

$$\bar{S}_i = \sum_{j=1}^{n} S_j / n.$$

Generally speaking, X-bar/S charts are preferable to X-bar/R charts, when either the sample size n is moderately large or the subgroup size is variable. The detailed development of the control chart method can be found in Montgomery (2001).

For demonstration purpose, let us consider the R&R study on the measurement of a shaft diameter by Hoguet (1994). The study involved ten shafts, three operators, and three replicate measurements; the data for this example are given in Table 1.

Table 1. *Engine shaft measurement data*

Part Name: Shaft	Gauge Name: Micrometer	Part No.: 1123456
Characteristic: Diameter	Gauge No.: Model #599	Measurement: 0.0001
Specification: 0.375 ± 0.002	Gauge Type: 1 inch	Zero Equals: 0.37

Operator	A				B				C			
Sample No.	1st	2nd	3rd	Range	1st	2nd	3rd	Range	1st	2nd	3rd	Range
1	56	55	57	2	57	58	56	2	56	57	56	1
2	63	62	62	1	64	64	64	0	62	64	64	2
3	56	54	55	2	57	55	56	2	55	55	55	0
4	57	55	56	2	56	57	55	2	56	57	55	2
5	58	58	57	1	59	60	60	1	57	60	60	3
6	56	55	54	2	60	59	57	3	55	57	56	2
7	56	55	56	1	58	56	56	2	55	55	57	2
8	57	56	56	1	57	58	57	1	57	58	57	1
9	65	64	64	1	64	64	65	1	65	64	65	1
10	58	57	57	1	61	60	60	1	58	59	60	2
Total	582	573	574	14	593	591	586	15	576	586	585	16
				$\bar{R}$A = 1.4				$\bar{R}$B = 1.5				$\bar{R}$C = 1.6
	Sum = 1729				Sum = 1770				Sum = 1747			
	$\bar{X}$A = 57.63				$\bar{X}$B = 59.00				$\bar{X}$C = 58.23			

Source: Adapted from Hoguet (1994).

Using the X-bar and R charts, the estimates of the repeatability, reproducibility and total gauge variability are given as follows:

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{\bar{R}}}{d_2} = \frac{(1.4 + 1.5 + 1.6)/3}{1.693} = 0.89$$

$$\hat{\sigma}_{\text{reproducibility}} = \frac{X_{\text{diff}}}{d_2} = \frac{59 - 57.63}{1.693} = 0.81$$

$$\hat{\sigma}_{\text{gauge}} = \sqrt{\hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}^2_{\text{reproducibility}}} = \sqrt{0.89^2 + 0.81^2} = 1.204$$

The Bootstrap method

Efron introduced the Bootstrap in 1979 as a computer-based method for estimating the standard error of a summary statistic (Efron & Tibshirani, 1993). The Bootstrap estimate of standard error requires no theoretical calculations, and is available no matter how mathematically complicated the summary statistic may be. Several types of Bootstrap method for constructing confidence intervals have been developed: the standard method, the percentile method, the bias-corrected percentile method, and the accelerated bias-corrected percentile method.

The standard method

Let $\mu_{Y(i)}$ be the mean of the $Y(i)$, $i = 1, 2 \ldots, B$, where $Y(i)$ is the ith resample data and B is the number of resampling data. Then, we have

$$\bar{\mu}_{\text{bootstrap}} = \frac{\sum_{i=1}^{B} \bar{\mu}_{Y(i)}}{B}$$

and

$$S_{\text{bootstrap}} = \sqrt{\frac{\sum_{i=1}^{B} (\bar{Y}(i) - \bar{\mu}_{\text{bootstrap}})^2}{B - 1}}$$

Therefore, a $100(1-\alpha)\%$ confidence interval for μ_Y is $\bar{\mu}_{\text{bootstrap}} \pm Z_{\alpha/2} S_{\text{bootstrap}}$ where $Z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

The percentile method

From the increasingly ordered $\bar{\mu}_Y[i]$, $i = 1, 2, \ldots, B$, a $100(1-\alpha)\%$ confidence interval for μ_Y is $\bar{\mu}_Y[(\alpha/2) \times B] \leqslant \mu_Y \leqslant \bar{\mu}_Y[(1 - \alpha/2) \times B]$ where $\bar{\mu}_Y[r]$ is the rth ordered element of the $\bar{\mu}_Y[i]$.

The bias-corrected percentile method

The Bootstrap distribution may be a biased distribution. Therefore, Efron suggested a method to correct for this bias problem. First, determine i such that the estimate of μ_Y from the original sample data is between $\bar{\mu}_Y[i]$ and $\bar{\mu}_Y[i+1]$. The cumulative probability that the Bootstrap estimates of μ_Y is less than (or equal to) the original index is $P_0 = i/B$. Then, compute the inverse of the cumulative distribution function of a standard normal distribution based upon P_0 as $Z_{P_0} = \Phi^{-1}(P_0)$. The bias-corrected percentile endpoints of the standard normal distribution are given by $P_L = \Phi(2Z_{P_0} + Z_{\alpha/2})$ and $P_U = \Phi(2Z_{P_0} + Z_{1-\alpha/2})$. Then, a $100(1-\alpha)\%$ confidence interval for μ_Y is $\bar{\mu}_Y[P_L B] \leqslant \mu_Y \leqslant \bar{\mu}_Y[P_U B]$.

The accelerated bias-corrected percentile method

Efron suggested this method to improve the bias-corrected percentile method. The first two steps are the same as the bias-corrected percentile method. In the third step, accelerated corrected percentile endpoints of the standard normal distribution are computed as

$$P_{AL} = \Phi\left(Z_{P_0} + \frac{Z_{P_0} + Z_{\alpha/2}}{1 - a(Z_{P_0} + Z_{\alpha/2})} \right)$$

and

$$P_{AU} = \Phi\left(Z_{P_0} + \frac{Z_{P_0} + Z_{1-\alpha/2}}{1 - a(Z_{P_0} + Z_{1-\alpha/2})} \right)$$

where

$$a = \frac{\sum_{i=1}^{n} (\bar{\mu}_{Y(\cdot)} - \dot{\mu}_{Y(i)})^3}{6\left\{ \sum_{i=1}^{n} (\bar{\mu}_{Y(\cdot)} - \mu_{Y(i)})^2 \right\}^{3/2}}$$

$\dot{\mu}_{Y(i)}$ is computed from the original sample with the ith point deleted, and

$$\bar{\mu}_{Y(\cdot)} = \sum_{i=1}^{n} \dot{\mu}_{Y(i)}/n$$

Then, a $100(1-\alpha)\%$ confidence interval for μ_Y is $\bar{\mu}_Y[P_{Al} B] \leqslant \mu_Y \leqslant \bar{\mu}_Y[P_{AU} B]$.

Efron & Tibshirani (1993) concluded that the accelerated bias-corrected percentile method has endpoints that transform correctly and have second-order accuracy. Its errors go to zero at rate $1/n$ in terms of the sample size n. Therefore, the accelerated bias-corrected percentile method is recommended for general use. In order to construct the confidence intervals for $\bar{\bar{R}}$ or $\bar{\bar{S}}$ and X_{diff}, a standard procedure must be taken. The steps are as follows.

Step 1. Determine the number of sample parts, operator and trials, and collect the measurement data.
Step 2. Using the accelerated bias-corrected percentile method, generate a Bootstrap resample ($B = 2000$) from the original ones.
Step 3. Compute the values of $\bar{\bar{R}}$ or $\bar{\bar{S}}$ and X_{diff}, respectively.
Step 4. Order the 2000 values of $\bar{\bar{R}}$ or $\bar{\bar{S}}$ and X_{diff}, respectively.
Step 5. Compute the confidence intervals for $\bar{\bar{R}}$ or $\bar{\bar{S}}$ and X_{diff} at the 95% confidence level by the accelerated bias-corrected percentile method.

Let us consider the data in Table 1. Using 2000 sampling replications, the Bootstrap intervals for $\bar{\bar{R}}$ and X_{diff} at the 95% confidence level are (0.70, 1.87) and (1.27, 1.50), respectively. Then, the approximate 95% confidence intervals for $\sigma_{\text{repeatability}}$, $\sigma_{\text{reproducibility}}$ and σ_{gauge} are given as follows:

$$0.41 \leqslant \sigma_{\text{repeatability}} \leqslant 1.06$$
$$0.75 \leqslant \sigma_{\text{reproducibility}} \leqslant 0.89$$
$$0.87 \leqslant \sigma_{\text{gauge}} \leqslant 1.41$$

The programs for computing the confidence intervals were implemented in FORTRAN 77 (available upon request from the first author). Note that the control chart method can simply find point estimates for the repeatability and reproducibility components; however, it does not lend itself to finding confidence intervals. Therefore, it is necessary for us to use the Bootstrap method to find the confidence intervals when the control chart method is adopted.

The experimental design method

The experimental design method is also known as the analysis of variance (ANOVA) method. A factorial experimental design can be used to investigate many factors simultaneously in an experiment and will allow us to estimate the contribution of each factor to measurement variability. For example, we can assign two operators and two different gauges as two levels for each of the four combinations of operators and gauges. After the important factors have been identified, further experiments can use more levels of the factors (if necessary) to quantify their effect on measurement variability. Fractional factorial designs could also be used in the initial phase of a gauge capability study to screen out the important sources of variability. Other experimental designs, such as nested designs, can be used to estimate the variance components of the gauge capability. The confidence intervals on the variability components of interest, $\sigma_{\text{repeatability}}$, $\sigma_{\text{reproducibility}}$ and σ_{gauge}, will give an idea about the relative uncertainty associated with the estimates and provide some basis for sample size decisions. The methods for constructing confidence intervals of the variance components can be found in Burdick & Graybill (1992). Burdick & Larsen (1997) provided the recommendations for selecting an appropriate method in R&R and for selecting the number of operators and samples. As a general recommendation, the modified large sample (MLS) method for confidence intervals can be applied for all of the parameters considered in their paper. Dolezal *et al.* (1998) discussed the construction of confidence intervals on measures of variability for a mixed model in which one factor is fixed and the other is random. Furthermore, the experimental design methods include normal probability plotting of residuals and plots of residuals versus the predicted responses.

Let us considered the data in Table 1. The measurement by operator j on part i at replication k can be denoted as Y_{ijk}, and the two-factor random effect model is defined as:

$$Y_{ijk} = \mu + P_i + O_j + (PO)_{ij} + e_{ijk} \quad i = 1, \ldots, 10; j = 1, \ldots, 3; k = 1, \ldots, 3 \tag{3}$$

where μ represents an overall mean and P_i, O_j, $(PO)_{ij}$, e_{ijk} are jointly independent normal random variables with means of zero and variances σ_p^2, σ_O^2, σ_{PO}^2 and σ_R^2, respectively. SAS PROC VARCOMP and MIXED provide excellent ways to fit the variance component model. These procedures allow the user to define an effect as a fixed or a random factor and the maximum likelihood estimation or the restricted maximum likelihood estimation can be selected. Note that the estimates for all variance components by the restricted maximum likelihood estimation method must be non-negative. The Part by Operator interaction resulted in an estimate of $\hat{\sigma}_{PO}^2 \cong 0$, indicating that the interaction is statistically insignificant, while the main effects of Parts and Operators are statistically significant. Therefore, the two-factor factorial model can be modified to a reduced model; that is, the interaction term can

Table 2. *Analysis of variance for the gauge capability studies using Table 1 data*

Source of variability	Degree of freedom	Sum of squares	Mean square	F_0	P-value
Parts	9	820.93	91.21	99.65	0.0001
Operator	2	28.16	14.08	15.38	0.0001
Repeatability	78	71.40	0.92		
Total	89	920.49			
Expected mean squares:			Estimated variance components:		
$E(MS_P) = \sigma_R^2 + 9\sigma_O^2$			$\hat{\sigma}^2_{\text{repeatability}} = 0.92$		
$E(MS_O) = \sigma_R^2 + 30\sigma_O^2$			$\hat{\sigma}^2_{\text{reproducibility}} = \sigma_O^2 = \dfrac{14.08 - 0.92}{30} = 0.44$		
$E(MS_R) = \sigma_R^2$			$\hat{\sigma}^2_{\text{gauge}} = \hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}^2_{\text{reproducibility}} = 1.36$		

be eliminated and the reduced model fit to the data. Table 2 demonstrates the analysis of variance and the estimation of the variance components of the data in Table 1. Normal probability plots of residuals, and plots of residuals versus the predicted responses do not indicate any violation of the assumptions. Thus, this model is adequate for summarizing the measurement data.

The large sample method is often used in constructing the confidence intervals. Specifically, if σ^2 is the variance component of interest estimated by $\hat{\sigma}^2$ and if $\hat{V}(\hat{\sigma}^2)$ is the variance of $\hat{\sigma}^2$, which is estimated by the corresponding diagonal element of the asymptotic covariance matrix from the maximum likelihood procedure, then the approximate $100(1-\alpha)\%$ confidence interval for σ is

$$\sqrt{\hat{\sigma}^2 - Z_{\alpha/2}\sqrt{\hat{V}(\hat{\sigma}^2)}} \leqslant \alpha \leqslant \sqrt{\hat{\sigma}^2 + Z_{\alpha/2}\sqrt{\hat{V}(\hat{\sigma}^2)}} \tag{4}$$

The approximated 95% confidence intervals for $\sigma_{\text{repeatability}}$, $\sigma_{\text{reproducibility}}$ and σ_{gauge} can be determined by the above equation.

Unfortunately, the large sample method does not perform well for small sample sizes. Burdick & Larson (1997) reported that the modified large sample (MLS) method tends to result in exact confidence intervals in some instances and is very close to the true intervals in most cases. This method is based on the relationship between the expected mean squares and their corresponding variance components. The resulting confidence intervals are functions of the expected mean squares, e.g. a linear combination. Considering the data in Table 1. The approximate 95% confidence intervals for $\sigma^2_{\text{repeatability}}$, $\sigma^2_{\text{reproducibility}}$ and σ^2_{gauge} are given as follows (see details in Burdick & Larsen, 1997):

$$\begin{gathered}(1 - G_5)\, S_E^2 \leqslant \sigma^2_{\text{repeatability}} \leqslant (1 + H_5)\, S_E^2 \\ \frac{(S_O^2 - S_E^2 - \sqrt{V_{RL}})}{30} \leqslant \sigma^2_{\text{reproducibility}} \leqslant \frac{(S_O^2 - S_E^2\sqrt{V_{RU}})}{30} \\ \frac{(S_O^2 + 29 \times S_E^2 - \sqrt{V_{GL}})}{30} \leqslant \sigma^2_{\text{gauge}} \leqslant \frac{(S_O^2 + 29 \times S_E^2 + \sqrt{V_{GU}})}{30}\end{gathered} \tag{5}$$

where

$$G_5 = 1 - \frac{78}{\chi^2_{0.0025,78}},\ H_5 = \frac{78}{\chi^2_{0.975,78}} - 1$$

$$G_2 = 1 - \frac{2}{\chi^2_{0.025,2}}, H_2 = \frac{2}{\chi^2_{0.975,2}} - 1$$

$$G_{25} = \frac{(F_{0.025,2,78} - 1)^2 - G_2^2 F^2_{0.025,2,78} - H_5^2}{F_{0.025,2,78}}, \; H_{25} = \frac{(1 - F_{0.975,2,78})^2 - H_2^2 F^2_{0.975,2,78} - G_5^2}{F_{0.975,2,78}}$$

$$V_{RL} = G_2^2 S_O^4 + H_5^2 S_E^4 + G_{25} S_O^2 S_E^2, \; V_{RU} = H_2^2 S_O^4 + G_5^2 S_E^4 + H_{25} S_O^2 S_E^2$$

$$V_{GL} = G_2^2 S_O^4 + G_5^2 \times 29^2 \times S_E^4, \; V_{GU} = H_2^2 S_O^4 + H_5^2 \times 29^2 \times S_E^2$$

Although these computations may appear complicated, the values can be obtained using Microsoft Excel or some statistical software packages. For the engine shaft measurement data, the approximate 95% confidence intervals for repeatability, reproducibility and total gauge variability by different methods are listed in Table 3.

With respect to the point estimates, the estimates for repeatability of the control chart method and the ANOVA approach are similar, while the ANOVA reproducibility estimate is smaller than the control chart estimate. Since the largest component of variation is repeatability, it indicates that repeatability problems should be addressed first in this measurement system. Although these computations of the repeatability and reproducibility from the control chart method and the ANOVA method are not identical, the results are still comparable. The confidence intervals for the repeatability variance component from all methods are very similar. However, the confidence intervals for the reproducibility and total gauge variance components from the MLS method are much wider than those of the MLE, the RMLE and the CCWB methods. Burdick & Larsen (1997) adopted the MLS method in their gauge capability studies and determined that its actual level of confidence meets the nominal 95% level very consistently. The value of the mean squared operator (S_O^2) in equation (5) will affect the confidence intervals. That is, if the operators have differences, the confidence intervals on the variance components of reproducibility and total gauge variability from the MLS method will have a large width.

In addition, there is concern that the confidence intervals obtained by the Bootstrap method may not maintain the stated level of confidence. The actual level of confidence can

Table 3. *Comparison of repeatability, reproducibility and total gauge variability by different methods using Table 1 data*

	$\sigma_{\text{repeatability}}$		$\sigma_{\text{reproducibility}}$		σ_{gauge}	
Method	Point estimate	95% confidence interval	Point estimate	95% confidence interval	Point estimate	95% confidence interval
MLE	0.96	**(0.79, 1.10)**	0.66	(0, 1.10)	1.17	(0.68, 1.48)
RMLE	0.96	**(0.79, 1.10)**	0.66	(0, 1.17)	1.17	(0.62, 1.52)
MLS	0.96	**(0.83, 1.14)**	0.66	*(0.65, 4.30)*	1.17	*(0.97, 4.41)*
CCWB	0.89	*(0.41, 1.06)*	0.81	**(0.75, 0.89)**	1.20	**(0.87, 1.41)**

MLE = the maximum likelihood method.
RMLE = the restricted maximum likelihood method.
MLS = the modified large sample method.
CCWB = the control chart with Bootstrapping method.
Italic interval indicates the longest interval among the four methods.
Bold interval indicates the shortest interval among the four methods.

Table 4. *The measurement data of an automobile mini-motor*

Part Name: Automobile Mini-Motor			Part No. 19-358A			
Variable: Length			Unit: mm			
Operator	A		B		C	
Sample No.	1st	2nd	1st	2nd	1st	2nd
1	21	20	22	20	19	21
2	24	23	24	24	23	24
3	20	21	19	21	20	22
4	27	27	28	26	27	28
5	19	18	20	18	19	21
6	23	21	24	21	23	22
7	22	21	22	24	22	20
8	19	17	18	20	21	18
9	24	23	25	23	24	26
10	25	23	23	25	27	25
11	21	20	20	21	21	20
12	18	19	17	19	18	19
13	23	25	25	25	25	25
14	24	24	23	25	24	25
15	29	30	30	32	31	33
16	26	26	25	26	25	27
17	20	20	19	20	20	20
18	24	23	19	21	21	23
19	25	26	25	24	27	25
20	19	19	18	17	19	17
21	21	22	18	21	21	24
22	22	23	19	21	20	23
23	25	26	23	24	23	25
24	21	19	20	18	17	19
25	24	23	23	24	21	25

be much less than the stated level. However, for most practical applications, the actual level of confidence obtained from the Bootstrap method can be around 85% (Prada-Sanchez & Cotos-Yanez, 1997).

Example

This real-life example demonstrates a gauge capability study in the manufacturing process of an automobile mini-motor. One of the part's quality characteristics was length, which is used as the variable for demonstration. Twenty-five units of the product were obtained, and three process operators measured each unit of product twice. The data are shown in Table 4. By using the X-bar and S charts for these data, we found that both charts exhibit no out-of-control points. The estimates of the repeatability, reproducibility and total gauge variability are given as follows:

$$\hat{\sigma}_{\text{repeatability}} = \frac{\bar{\bar{S}}}{c_4} = \frac{1.07}{0.7979} = 1.34$$

$$\hat{\sigma}_{\text{reproducibility}} = \frac{X_{\text{diff}}}{d_2} = \frac{0.5}{1.693} = 0.30$$

$$\hat{\sigma}_{\text{gauge}} = \sqrt{\hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}^2_{\text{reproducibility}}} = \sqrt{1.34^2 + 0.30^2} = 1.37$$

Table 5. *Analysis of variance for the gauge capability studies using Table 4 data*

Source of variability	Degree of freedom	Sum of squares	Mean square		P-value
Parts	24	1360.76	56.70	40.64	0.001
Operator	2	6.88	3.44	2.47	0.089
Repeatability	123	171.62	1.40		
Total	149	1539.26			

Expected mean squares:	Estimated variance components:
$E(MS_P) = \sigma_R^2 + 6\sigma_O^2$	$\hat{\sigma}^2_{\text{repeatability}} = 1.40$
$E(MS_O) = \sigma_R^2 + 50\sigma_O^2$	$\hat{\sigma}^2_{\text{reproducibility}} = \sigma_O^2 = \dfrac{3.44 - 1.40}{50} = 0.04$
$E(MS_R) = \sigma_R^2$	$\hat{\sigma}^2_{\text{gauge}} = \hat{\sigma}^2_{\text{repeatability}} + \hat{\sigma}^2_{\text{reproducibility}} = 1.44$

Table 6. *Comparison of repeatability, reproducibility and total gauge variability by different methods using Table 4 data*

	$\sigma_{\text{repeatability}}$		$\sigma_{\text{reproducibility}}$		σ_{gauge}	
Method	Point estimate	95% confidence interval	Point estimate	95% confidence interval	Point estimate	95% confidence interval
MLE	1.18	(1.00, 1.33)	0.20	**(0.00, 0.40)**	1.20	**(1.01, 1.35)**
RMLE	1.18	(1.00, 1.34)	0.20	(0.00, 0.41)	1.20	(1.00, 1.36)
MLS	1.18	**(1.05, 1.35)**	0.20	*(0.20, 1.52)*	1.20	*(1.07, 2.03)*
CCWB	1.34	*(1.00, 1.54)*	0.30	(0.00, 0.49)	1.37	(1.00, 1.61)

MLE = the maximum likelihood method.
RMLE = the restricted maximum likelihood method.
MLS = the modified large sample method.
CCWB = the control chart with Bootstrapping method.
Italic interval indicates the longest interval among the four methods.
Bold interval indicates the shortest interval among the four methods.

A factorial experiment design was used to analyse this study. The Part by Operator interaction resulted in an estimate of $\hat{\sigma}^2_{PO} \cong 0$ indicating that the interaction is found to be statistically insignificant, while the main effects of Parts and Operators are statistically significant. Table 5 demonstrates the analysis of variance and the estimates of variance components of the data. The normal probability plotting of residuals and plots of residuals versus the predicted responses do not violate the assumptions. Thus, this model is adequate for summarizing the measurement data.

For this example, the approximate 95% confidence intervals for repeatability, reproducibility and total gauge variability by different methods are listed in Table 6. The estimates of repeatability and reproducibility provided by the ANOVA approach and the control chart method are similar. The confidence intervals on the variance components of reproducibility and total gauge variability from the MLS method are wider than those of the MLE, the RMLE, and the CCWB procedures. The interval widths for the repeatability variance component from all methods are similar. The analysis in Table 6 indicates that the major variation is from the component of repeatability. Thus, the action should focus on the inspection device.

Conclusions and recommendations

The adequacy of a gauge capability study is critical to studying the statistical-control characteristics of a production process. This study reviews the methods of gauge capability study and proposes a combined method for estimating point estimates and confidence intervals of the gauge variability. This proposed method adopts the control chart method for calculating the point estimates and the Bootstrap method for obtaining the confidence intervals for the variance components of the gauge capability study. With the control chart method, the R chart or the S chart can be used for detecting whether the operator is having difficulty using the instrument or not. This study demonstrates the use of the combined method along with three types of ANOVA method: the maximum likelihood, the restricted maximum likelihood, and the modified large sample method. The results show that the point estimates and the confidence intervals obtained by the combined method are very close to those from the ANOVA methods. An increase in the number of samples can result in more precise point estimates on the variance components of the gauge capability. Nevertheless, the control chart method is very applicable for a quick R&R study given its simplicity. Moreover, the Bootstrap method is likely to out-perform the ANOVA methods in obtaining shorter confidence intervals, as it did in one of the two examples examined in this study. In fact, it out-performed the MLS method in both examples, indicating the viability of using the Bootstrap method in place of the ANOVA methods. However, more studies are needed in order to draw a generalized conclusion.

For the example of the mini-motor manufacturing company, when both repeatability and reproducibility are taken into account, the gauge capability is not as good as we would like. Training the operators to produce more uniform work methods in using the gauge would help reduce $\hat{\sigma}_{\text{reproducibility}}$, but since $\hat{\sigma}_{\text{repeatability}}$ is the largest component of $\hat{\sigma}_{\text{gauge}}$, some effort should also be directed toward finding another inspection device.

Acknowledgements

The authors gratefully acknowledge the referees of this paper who helped to clarify and improve the presentation. In addition, the first author is grateful for partial support by the National Science Council in Taiwan under grant NSC-89-2213-E182-034.

References

BLACK, S.A. & PORTER, L.J. (1996) Identification of the critical factors of TQM, *Decision Sciences*, 27, pp. 1–21.

BORROR, C.M., MONTGOMERY, D.C. & RUNGER, G.C. (1997) Confidence intervals for variance components from gauge capability studies, *Quality and Reliability Engineering International*, 13, pp. 361–369.

BURDICK, R.K. & GRAYBILL, F.A. (1992) *Confidence Intervals on Variance Components* (New York, Marcel Dekker).

BURDICK, R.K. & LARSEN, G.A. (1997) Confidence intervals on measures of variability in R & R studies, *Journal of Quality Technology*, 29, pp. 261–273.

DAYTON, N.A. (2001) Total quality management critical success factors, a comparison: the UK versus the USA, *Total Quality Management*, 12(3), pp. 293–298.

DEUTLER, T. (1991) Grubbs-type estimators for reproducibility variances in an interlaboratory test study, *Journal of Quality Technology*, 23, pp. 324–335.

DOLEZAL, K.K., BURDICK, R.K. & BIRCH, N.J. (1998) Analysis of a two-factor R&R study with fixed operators, *Journal of Quality Technology*, 30, pp. 163–170.

EFRON, B. & TIBSHIRANI, R.J. (1993) *An Introduction to the Bootstrap* (New York, Chapman & Hall).

HOGUET, R.G. (1994) A simplified gage method for test reliability, *Ceramic Engineering & Science Proceeding*, 15, pp. 59–71.

Kanji, G.K. & Barker, R.L. (1990) Implementation of total quality management, *Total Quality Management*, 1, pp. 375–389.

Martin, R. (1997) Do we practice quality principles in the performance measurement of critical success factors?, *Total Quality Management*, 8(6), pp. 429–444.

Montgomery, D.C. & Runger, G.C. (1993) Gauge capability and designed experiments, part II: experimental design models and variance component estimation, *Quality Engineering*, 6, pp. 289–305.

Montgomery, D.C. (2001) *Introduction to Statistical Quality Control*, 4th edn (New York, Wiley).

Prada-Sanchez, J.M. & Cotos-Yanez, T. (1997) Simulation study of iterated and non-iterated bootstrap methods for bias reduction and confidence interval estimation, *Communication in Statistics—Simulation and Computation*, 26, pp. 927–946.

SAS/STAT Software (1995) Release 6.07, SAS Institute, Gary, NC.

Taylor, W.A. (1996) Sectoral differences in total quality management implementation: the influence of management mind-set, *Total Quality Management*, 7(3), pp. 235–248.

Taylor, B.N. & Kuyatt, C.E. (1994) *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*, NIST, Technical Note 1297 (Washington, DC: US Government Printing Office).

Wang, F.K. & Chen, C.Y. (1998) Confidence intervals for multivariate capability indices using bootstrap, *Journal of the Chinese Institute of Industrial Engineers*, 15, pp. 429–438.

Yusof, S.M. & Aspinwall, E. (2000a) Total quality management implementation frameworks: comparison and review, *Total Quality Management*, 11(3), pp. 281–294.

Yusof, S.M. & Aspinwall, E. (2000b) Critical success factors in small and medium enterprises: survey results, *Total Quality Management*, 11(4/5&6), pp. S448–S462.