Estimating and Testing Quantile-based Process Capability Indices for Processes with Skewed Distributions

Cheng Peng
University of Southern Maine

Abstract: This article extends the recent work of Vännman and Albing (2007) regarding the new family of quantile based process capability indices (qPCI) $C_{MA}(\tau, v)$. We develop both asymptotic parametric and nonparametric confidence limits and testing procedures of $C_{MA}(\tau, v)$. The kernel density estimator of process was proposed to find the consistent estimator of the variance of the nonparametric consistent estimator of $C_{MA}(\tau, v)$. Therefore, the proposed procedure is ready for practical implementation to any processes. Illustrative examples are also provided to show the steps of implementing the proposed methods directly on the real-life problems. We also present a simulation study on the sample size required for using asymptotic results.

Key words: Confidence limits, kernel density estimation, process capability index, one-sided specification interval, skewed distribution, testing hypothesis.

1. Introduction

As a numerical measure, the process capability index (PCI) uses both the process variability and the process specifications to determine whether the process is capable. It plays an important role in monitoring and analyzing process quality and productivity. Many PCIs have been proposed since Juran *et al.* (1974) proposed the first PCI C_p. Let USL and LSL be the upper and lower specification limits, $d = (USL - LSL)/2$, $m = (LSL + USL)/2$ and T be the target value. The process mean and standard deviation are denoted by μ and σ. Vännman (1995) proposed the superstructure which unifies the four basic PCIs, namely, C_p, C_{pk}, C_{pm} and C_{pmk}, as follows

$$C_p(u, v) = \frac{d - u|\mu - m|}{3\sqrt{\sigma^2 + v(\mu - T)^2}}. \tag{1.1}$$

where u and v are non-negative constants. We can see from (1.1) that $C_p(0, 0) = C_p$, $C_p(1, 0) = C_{pk}$, $C_p(0, 1) = C_{pm}$ and $C_p(1, 1) = C_{pmk}$. Since the process

mean and variance based process capability indices (mvPCI) implicitly assume the normality of the underlying process, they are sensitive to skewed processes.

Recently many published works (see Spiring *et al.*, 2003, for more references) try to address this important non-normal issue through modifying the existing popular PCIs. Some researchers use different parametric models to deal with non-normal processes. To name a few, Kotz and Lovelace (1998, page 174) and Lin (2004) use the folded normal distribution and Lin (2005) uses the generalized folded normal distribution to model the underlying process and use the special feature of the parametric models to modify PCIs. Chang and Bai (2001) and Chang *et al.* (2002) model the process density with a weighted average of two normal densities (mixture of two normal distributions with known mixing proportions) according to the skewness of the underlying process. Most recently, Chao and Lin (2005) proposed a very general process yield-based PCI as follows

$$C_y = \frac{1}{3}\Phi^{-1}\left[\frac{1}{2}\left(F_\theta(USL) - F_\theta(LSL) + 1\right)\right] \tag{1.2}$$

where Φ is the CDF of standard normal distribution, $F(\cdot)$ and θ are the CDF and the vector of parameters of the underlying process distribution. PCI (1.2) has an elegant analytical expression involving only two tail probabilities of the underlying process and is easy to interpret. The formulation of C_y does not implicitly assume the normality/sysmetry of the underlying process since the CDF of the underlying process $F(\cdot)$ is not specified. From this perspective, C_y has a structural difference from many existing PCIs in handling non-normal processes.

Another direction to relaxing the implicit normality/symmetry assumption is to introduce process quantiles to the definition of PCIs. Motivated from Clements' (1989) quantile idea, Chen and Pearn (1997) modified Vännman's (1995) $C_p(u, v)$ and proposed a quantile-based PCI superstructure without implicitly assuming normality of the underlying process as follows

$$C_{Np}(u; v) = \frac{d - u|\xi_{p_2} - m|}{3\sqrt{(\frac{\xi_{p_3} - \xi_{p_1}}{6})^2 + v(\xi_{p_2} - T)^2}}. \tag{1.3}$$

where ξ_α is the α-th quantile of the process, i.e., $P(X < \xi_\alpha) = \alpha$, and $p_1 = 0.00135$, $p_2 = 0.5$, and $p_3 = 0.99865$. The formulation of the above quantile-based PCI is intended to yield the nonconformity proportion about 0.27% with $C_{Np} = 1$ if the process is approximately normally distributed and on target.

Note that (1.3) is essentially designed for processes with two specification limits. In practice, many processes only have one-sided specification limit, such as zero-bound processes in which zero is the natural bound and measurements with value zero are desirable. The existing standard two-sided and one-sided indices do

not work well on this special type of processes (see Vännman and Albing, 2007). There are only a few discussions on this topic available in the literature. Using the parametric approach, Lovelace *et al.* (1997) studied this process based on the log-normal distribution and proposed an index $C_{pu(ln)}$ (see Kotz and Lovelace (1998), page 167-174) and its modification allowing measurements with zero values. Kotz and Lovelace (1998, page 174) also proposed to use the folded normal distribution on index C_{pu} for zero bound data and estimated the process mean and standard deviation by sample mean and sample standard deviation. Lin (2004) uses the folded normal distribution on zero-bound process and proposed a new estimator for process mean and standard deviation.

Vännman and Albing (2007) recently defined a promising quantile based index for measuring capability of processes (particularly for skewed distribution) with upper specification limit such as zero-bound processes as follows

$$C_{MA}(\nu; \xi) = \frac{USL}{\sqrt{\xi_{p_3}^2 + \nu \xi_{p_2}^2}} \qquad (1.4)$$

where $p_2 = 0.5$ and $p_3 = 0.9973$. The process yield at $C_{MA} = 1$ is given by

$$P(X < LSL) = P\left(X < \sqrt{\xi_{p_3}^2 + \xi_{p_2}^2}\right) > P(X < \xi_{p_3}) = p_3 = 99.73\%.$$

Vännman and Albing's (2007) work is essentially within nonparametric frame work in which they proposed sample quantile and interpolation based quantile estimators of $C_{MA}(u, v)$, denoted by $\hat{C}_{MA}(u, v)$, and proved that $\hat{C}_{MA}(u, v)$ are asymptotically normally distributed. Since the asymptotic variance of $\hat{C}_{MA}(u, v)$ is dependent on the explicit expression of the density function of the underlying process, their results of asymptotic normality cannot be used directly to construct confidence intervals or test hypotheses for the qPCI unless the distribution of the underlying process is completely specified.

In this paper, we systematically develop both parametric and non-parametric procedures for constructing asymptotic confidence limits and testing hypotheses based on Vännman and Albing's (2007) qPCI $C_{MA}(\nu; \xi)$. The parametric method will be discussed in Section 2. In Section 3, we discuss a nonparametric approach and use the kernel density estimator to estimate the underlying process density function, hence, to obtain the consistent estimate of the variance of nonparametric estimate of $C_{MA}(\nu; \xi)$. Case studies and illustrative examples are given in Sections 4 and 5. We also conduct a simulation study on sample size determination in Section 6. Finally, we make some remarks on using the proposed procedures.

2. Parametric Confidence Limit and Testing Procedure

Let $\{y_1, \cdots, y_n\}$ be an i.i.d. random sample selected from the process with density $f(y; \theta)$ with $\theta = (\theta_1, \cdots, \theta_k)^\tau$, the transpose of the column vector of process parameters. The likelihood and log likelihood functions of θ are given by

$$L(\theta) = \prod_{i=1}^{n} f(y_i; \theta)) \quad \text{and} \quad l(\theta) = \sum_{i=1}^{n} \ln f(y_i; \theta) \tag{2.1}$$

respectively. The αth quantile (ξ_α) of the process distribution is defined implicitly by function

$$\alpha = F(\xi_\alpha; \theta) = \int_{-\infty}^{\xi_\alpha} f(y; \theta) dy \tag{2.2}$$

Let $\tilde{\theta} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_k)^\tau$ be the MLE of θ. By the invariance property of MLE, $\tilde{\xi}_\alpha = \xi_\alpha(\tilde{\theta})$ is the maximum likelihood estimator of quantile ξ_α. Therefore, we propose the parametric maximum likelihood estimators of the $C_{MA}(\nu; \tilde{\theta})$ as follows

$$\tilde{C}_{MA}(\nu; \theta) = C_{MA}(\nu; \tilde{\theta}) = \frac{USL}{\sqrt{\xi_{p_3}^2(\tilde{\theta}) + \nu \xi_{p_2}^2(\tilde{\theta})}} \tag{2.3}$$

Note that $C_{MA}(\nu; \theta)$ is a real valued function of quantiles ξ_{p_3} and ξ_{p_2} which are continuous real functions of the vector of parameters θ. Note also that $\tilde{\theta}$ is a consistent MLE of θ, therefore, $\tilde{C}_{MA}(\nu; \tilde{\theta})$ is consistent MLEs of $C_{MA}(\nu; \theta)$ (see Serfling 1980, page 24). Since $C_{MA}(\nu; \theta)$ is expressed as functions of θ through quantiles $\xi_\alpha(\theta)$, we will use notations $C_{MA}(\nu; \theta)$ interchangeably with $C_{MA}(\nu; \xi)$.

Let $I(\theta)$ be the information matrix of the numeric characteristic of the process corresponding the parameter θ. Under some regularity conditions (Serfling, 1980, page 144-145), the MLE possesses the following asymptotic normality

$$\sqrt{n}(\tilde{\theta} - \theta) \rightarrow_d N(0, I^{-1}(\theta)) \tag{2.4}$$

A consistent estimator of the information matrix is

$$\hat{I}(\theta) = I_n(\hat{\theta}) = -\frac{1}{n} \frac{\partial^2 \ln L(\theta)}{\partial \theta^\tau \partial \theta}\Big|_{\theta = \tilde{\theta}} \tag{2.5}$$

with $\theta = (\theta_1, \cdots, \theta_k)^\tau$ and $\tilde{\theta} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_k)^\tau$. Furthermore, define

$$U_1(\nu, \theta) = \frac{\partial C_{MA}(\nu; \theta)}{\partial \xi_{p_2}} = -\frac{USL \cdot \nu \cdot \xi_{p2}}{\left(\xi_{p_3}^2 + \nu \xi_{p_2}^2\right)^{3/2}} = -C_{MA}(\nu; \theta) \times \frac{\nu \xi_{p_2}}{\xi_{p_3}^2 + \nu \xi_{p_2}^2} \tag{2.6}$$

$$U_2(\nu, \theta) = \frac{\partial C_{MA}(\nu; \theta)}{\partial \xi_{p_3}} = -\frac{USL \cdot \xi_{p3}}{\left(\xi_{p_3}^2 + \nu \xi_{p_2}^2\right)^{3/2}} = -C_{MA}(\nu; \theta) \times \frac{\xi_{p_3}}{\xi_{p_3}^2 + \nu \xi_{p_2}^2} \tag{2.7}$$

U_1 and U_2 are only dependent on the form of $C_{MA}(\nu; \theta)$. Define

$$U(\nu, \theta) = \big(U_1(\nu, \theta), U_2(\nu, \theta)\big) \tag{2.8}$$

and

$$R(\theta) = \begin{pmatrix} \dfrac{\partial \xi_{p_2}(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial \xi_{p_2}(\theta)}{\partial \theta_k} \\ \dfrac{\partial \xi_{p_3}(\theta)}{\partial \theta_1} & \cdots & \dfrac{\partial \xi_{p_3}(\theta)}{\partial \theta_k} \end{pmatrix}. \tag{2.9}$$

Using the first order Taylor expansion and the Slutsky's Theorem (see Casella and Berger, 2002, page 239), we have

$$\sqrt{n}\big(\tilde{C}_{MA}(\nu; \theta) - C_{MA}(\nu; \theta)\big) \to_d N\big(0, \Omega\big) \tag{2.10}$$

where $\Omega = U(\nu; \theta)R(\nu; \theta)I^{-1}(\theta)R^\tau(\nu; \theta)U^\tau(\nu; \theta)$ and $R^\tau(\nu; \theta)$, $U^\tau(\nu; \theta)$ are transpose of $R(\nu; \theta)$ and $U(\nu; \theta)$ respectively. We see from (2.10) that point estimator $\tilde{C}_{MA}(\nu; \theta)$ is an asymptotically unbiased estimator of $C_{MA}(\nu; \theta)$. A consistent estimator of variance Ω is given by

$$\tilde{\Omega} = U(\nu; \tilde{\theta})R(\nu; \tilde{\theta})I^{-1}(\tilde{\theta})R^\tau(\nu; \tilde{\theta})U^\tau(\nu; \tilde{\theta}) \tag{2.11}$$

that is, by replacing the vectors of parameters $\theta = (\theta_1, \cdots, \theta_k)$ with their MLEs $\tilde{\theta} = (\tilde{\theta}_1, \cdots, \tilde{\theta}_k)$, the standard error of $\tilde{C}_{MA}(\nu; \theta)$ is s.e.$\big[\tilde{C}_{MA}(\nu; \theta)\big] = \sqrt{\tilde{\Omega}/n}$. The $100(1 - \alpha)\%$ one-sided confidence interval (with lower limit) of $C_{MA}(\nu; \theta)$ is given by

$$\Big(\tilde{C}_{MA}(\nu; \theta) - z_\alpha \text{s.e.}\big[\tilde{C}_{MA}(\nu; \theta)\big], \infty\Big) \tag{2.12}$$

Since the process is capable if $C_{MA}(\nu, \theta) > 1$, the test statistic for testing $H_0 : C_{MA}(\nu, \theta) \le 1$ vs $H_a : C_{MA}(\nu, \theta) > 1$ is given by

$$Z = \frac{\tilde{C}_{MA}(\nu, \tilde{\theta}) - 1}{\sqrt{\tilde{\Omega}/n}} \sim N(0, 1) \tag{2.13}$$

Constructing the confidence interval using (2.12) and testing process capability using (2.13) with illustrative examples will be discussed in Section 5.

3. Non-parametric Confidence Limits and Testing Procedures

The procedure we discussed in Section 2 is dependent on the assumption that the density function of the underlying process is completely specified. If the density function of the process is unknown and we are not sure which distribution should be used to fit the model, nonparametric procedure methods should be

used. We focus our discussion on constructing confidence intervals and developing testing hypotheses using the distribution free approach in this section.

Let $\hat{\xi}_\alpha$ be the sample αth quantile. That is,

$$\hat{\xi}_\alpha = \max\{y : F_n(y) < \alpha\} \tag{3.1}$$

where $F_n(y)$ is the empirical CDF defined based on the sample data. It is well known that $\hat{\xi}_\alpha$ is a consistent estimator of ξ_α. Furthermore for p_2th and p_3th sample quantiles, we have the following variance and covariance matrix (see Serfling, 1980, page 80.)

$$\frac{\Gamma(f;\xi)}{n} = \text{cov}\begin{pmatrix} \hat{\xi}_{p_2} \\ \hat{\xi}_{p_3} \end{pmatrix} = \begin{pmatrix} \dfrac{p_2(1-p_2)}{nf^2(\xi_{p_2})} & \dfrac{p_3(1-p_2)}{nf(\xi_{p_2})f(\xi_{p_3})} \\ \dfrac{p_3(1-p_2)}{nf(\xi_{p_2})f(\xi_{p_3})} & \dfrac{p_3(1-p_3)}{nf^2(\xi_{p_3})} \end{pmatrix}$$

where f is the density function of the underlying process. Therefore, a consistent estimator of $C_{MA}(\nu;\xi)$ using sample quantiles is given by

$$\hat{C}_{MA}(\nu;\xi) = C_{MA}(\nu;\hat{\xi}) = \frac{USL}{\sqrt{\hat{\xi}_{p_3}^2 + \nu\hat{\xi}_{p_2}^2}} \tag{3.2}$$

Again using the first order Taylor expansion on $\hat{C}_{MA}(\nu;\xi)$ at the true value ξ and Slutsky's Theorem, we have

$$\sqrt{n}(\hat{C}_{MA}(\nu;\xi) - C_{MA}(\nu;\xi)) \to N(0, \sigma_C^2) \tag{3.3}$$

where

$$\sigma_C^2 = U(\nu;\theta)\Gamma(f;\xi)U^\tau(\nu;\theta)$$

$$= \frac{C_{MA}^2(\nu;\xi)}{(\xi_{p_3}^2 + \nu\xi_{p_2}^2)^2}\left[\frac{\nu^2\xi_{p_2}^2}{4f^2(\xi_{p_2})} + \frac{\nu(1-p_3)\xi_{p_2}\xi_{p_3}}{f(\xi_{p_2})f(\xi_{p_3})} + \frac{p_3(1-p_3)\xi_{p_3}^2}{f^2(\xi_{p_3})}\right]$$

which is exactly the same as the one obtained in Vännman and Albing (2007).

In order to use the asymptotic result (3.3) to construct the confidence interval of $C_{MA}(\nu;\xi)$ and test hypothesis of process capability, we need a consistent estimator of variance σ_C^2 in (3.3). For consistent quantile estimators, we only use the sample quantiles $\xi_{p_2} = \xi_{0.5}$ and $\xi_{p_3} = \xi_{0.9973}$ in this paper. We can also use sample quantiles or interpolation based quantiles discussed in Hyndman and Fan (1996), Pearn and Chen (1997) generalized Chang and Lu (1994) with a minor modification on σ_C^2.

For the process density, we choose the following nonparametric kernel density estimator. Suppose that $\{y_1, \cdots, y_n\}$ is a random sample collected from the underlying process with density $f(y)$. The kernel density estimator of $f(y)$, denoted by $\hat{f}(y)$, is given by

$$\hat{f}_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - y_i}{h}\right) \tag{3.4}$$

where $K(\cdot)$ is the kernel function which is nonnegative, unimodal and symmetric with respective to the vertical axis and integrates to unity, and h is the bandwidth which controls the degree of smoothing of the estimation. Among several commonly used kernel functions, Gaussian kernels are the most often used. In this paper, we will use Gaussian kernel and the rule-of-thumb of Silverman (1986) for choosing the bandwidth of a Gaussian kernel density estimator.

Using the kernel estimator of the density and the sample quantile estimators, we propose the following variance estimators

$$\hat{\sigma}_C^2 = \frac{C_{MA}^2(\nu; \hat{\xi})}{(\hat{\xi}_{p_3}^2 + \nu \hat{\xi}_{p_2}^2)^2} \left[\frac{\nu^2 \hat{\xi}_{p_2}^2}{4 \hat{f}^2(\hat{\xi}_{p_2})} + \frac{\nu(1 - p_3)\hat{\xi}_{p_2}\hat{\xi}_{p_3}}{\hat{f}(\hat{\xi}_{p_2})\hat{f}(\hat{\xi}_{p_3})} + \frac{p_3(1 - p_3)\hat{\xi}_{p_3}^2}{\hat{f}^2(\hat{\xi}_{p_3})} \right] \tag{3.5}$$

Since the variance is a continuous function of the estimated density function $f(x)$ and the estimated process quantiles, the consistency of estimator $\hat{\sigma}_C^2$ follows immediately from the fact that $\hat{f}(x)$ is (strongly) consistent with $f(x)$.

The $100(1 - \alpha)\%$ one-sided confidence limit for $C_{MA}(\nu; \xi)$ based on sample quantiles is given by

$$\left(\hat{C}_{MA}(\nu; \xi) - z_\alpha \text{s.e.}\left[\hat{C}_{MA}(\nu; \xi)\right], \infty \right) \tag{3.6}$$

where s.e.$\left[\hat{C}_{MA}(\nu; \xi)\right] = \sqrt{\hat{\sigma}_C^2/n}$. The test statistic for testing $H_0 : C_{MA}(\nu, \xi) \leq 1$ vs $H_a : C_{MA}(\nu, \xi) > 1$ is given by

$$Z = \frac{\hat{C}_{MA}(\nu, \hat{\xi}) - 1}{\sqrt{\hat{\sigma}_C^2/n}} \sim N(0, 1) \tag{3.7}$$

Since the above standard errors are explicitly expressed in data measurements and the estimated density function, the confidence limits can be calculated through simple programming.

4. Parametric Case Studies: Lognormal and Weibull Processes

We are interested in the investigating the performance of C_{MA} under skewed process with upper specification. Since lognormal and Weibull distributions are

non-normal and widely used in engineering, particularly in engineering reliability modeling, we choose to use these two distributions to illustrate how to construct parametric asymptotic confidence intervals for C_{MA}. We will provide the explicit expression of the lower confidence limit and the test statistic in terms of sample data values.

4.1 Lognormal process

Recall that the two parameter lognormal distribution has density function

$$f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma x} \exp\left[-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right] \tag{4.1}$$

The α-th quantile of the lognormal process with parameters $\theta = (\mu, \sigma^2)$ is defined by

$$\xi_\alpha = \exp[\mu + \Phi^{-1}(\alpha)\sigma] \tag{4.2}$$

where $\Phi(\cdot)$ is the CDF of standard normal distribution as usual and $\Phi^{-1}(\cdot)$ is the inverse of $\Phi(\cdot)$. Let $x_1, \cdots, x_n$ be a random sample collected from a lognormal process. The log-likelihood function of μ and σ is

$$\ln L(\theta) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 - \sum_{i=1}^n x_i - \frac{1}{2}\sum_{i=1}^n \left(\frac{\ln x_i - \mu}{\sigma}\right)^2 \tag{4.3}$$

One can easily find the consistent MLE of the parameters μ and σ^2 are

$$\tilde{\mu} = \frac{\sum_{i=1}^n \ln(x_i)}{n} \quad \text{and} \quad \tilde{\sigma}^2 = \frac{\sum_{i=1}^n \left[\ln(x_i) - \tilde{\mu}\right]^2}{n} \tag{4.4}$$

The consistent estimator of the information matrix is given by

$$I_n(\tilde{\theta}) = -\frac{1}{n}\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta^\tau}\bigg|_{\theta=\tilde{\theta}} = \begin{pmatrix} \frac{1}{\tilde{\sigma}^2} & 0 \\ 0 & \frac{1}{2\tilde{\sigma}^4} \end{pmatrix} \tag{4.5}$$

where $\tilde{\theta} = (\tilde{\mu}, \tilde{\sigma}^2)$. Note also that the MLE of matrix $R(\theta)$ defined in (2.9) by using the the MLEs of the derivatives of quantiles

$$\frac{\partial\xi_\alpha}{\partial\mu}\bigg|_{\theta=\tilde{\theta}} = \tilde{\xi}_\alpha \quad \text{and} \quad \frac{\partial\xi_\alpha}{\partial\sigma^2}\bigg|_{\theta=\tilde{\theta}} = \tilde{\xi}_\alpha\left(\frac{\ln\tilde{\xi}_\alpha - \tilde{\mu}}{2\tilde{\sigma}^2}\right) = \frac{\tilde{\xi}_\alpha\Phi^{-1}(\alpha)}{2\tilde{\sigma}} \tag{4.6}$$

Using MLEs (4.4), (4.5) and (4.6) we can easily evaluate the consistent MLE of the variance Ω in (2.11) by using the MLEs of the α-th quantile and the matrix

of (2.9) expressed below

$$\tilde{\xi}_\alpha = \exp\left[\tilde{\mu} + \Phi^{-1}(\alpha)\tilde{\sigma}\right] \quad \text{and} \quad \tilde{R}(\mu, \sigma) = R(\tilde{\mu}, \tilde{\sigma}) = \begin{pmatrix} \tilde{\xi}_{p_2}, & \Phi^{-1}(p_2)\tilde{\xi}_{p_2}/(2\tilde{\sigma}) \\ \tilde{\xi}_{p_3}, & \Phi^{-1}(p_3)\tilde{\xi}_{p_3}/(2\tilde{\sigma}) \end{pmatrix}.$$

(4.7)

Therefore, the parametric asymptotic confidence interval (2.12) and the test statistics (2.13) can be easily calculated.

4.2 Weibull process

Next we consider the underlying process which follows the Weibull distribution with density function

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \exp\left[-\left(\frac{x}{\theta}\right)^\beta\right], \quad x \geq 0$$

(4.8)

where θ is the scale parameter and β is the shape parameter. The theoretical αth quantile of this Weilbull process, denoted by ξ_α, is determined by

$$\xi_\alpha = \theta\left[-\ln(1-\alpha)\right]^{1/\beta}$$

(4.9)

Let $x_1, \cdots, x_n$ be a random sample collected from a Weibull process. The log-likelihood function of θ and β is

$$l(\theta, \beta) = n\ln\left(\frac{\beta}{\theta}\right) + (\beta-1)\sum_{i=1}^{n}\ln x_i - n(\beta-1)\ln\theta - \sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^\beta$$

(4.10)

Let $\tilde{\theta}$ and $\tilde{\beta}$ be the maximum likelihood of θ and β and

$$\frac{\partial l^2(\theta,\beta)}{\partial\theta^2} = \frac{n}{\beta} - \frac{\beta(\beta+1)}{\theta^2}\sum_{i=1}^{n}\left(\frac{x_i}{\beta}\right)^\beta, \quad \frac{\partial l^2(\theta,\beta)}{\partial\beta^2} = \frac{n}{\beta^2} - \sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^\beta\ln^2\left(\frac{x_i}{\theta}\right)$$

$$\frac{\partial l^2(\theta,\beta)}{\partial\theta\partial\beta} = -\frac{n}{\theta} + \frac{\beta}{\theta}\sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^\beta\ln\left(\frac{x_i}{\theta}\right) + \frac{1}{\theta}\sum_{i=1}^{n}\left(\frac{x_i}{\theta}\right)^\beta = \frac{\partial l^2(\theta,\beta)}{\partial\beta\partial\theta}$$

The observed information matrix is given by

$$\hat{I}(\theta,\beta) = I_n(\tilde{\theta},\tilde{\beta}) = -\frac{1}{n}\begin{pmatrix} \frac{\partial l^2(\theta,\beta)}{\partial\theta^2} & \frac{\partial l^2(\theta,\beta)}{\partial\theta\partial\beta} \\ \frac{\partial l^2(\theta,\beta)}{\partial\beta\partial\theta} & \frac{\partial l^2(\theta,\beta)}{\partial\beta^2} \end{pmatrix}\Bigg|_{(\theta,\beta)=(\tilde{\theta},\tilde{\beta})}$$

(4.11)

The MLE of matrix $Q(\xi,\theta,\beta)$ defined in (2.9) is specified as follows

$$Q(\xi,\tilde{\theta},\tilde{\beta}) = \begin{pmatrix} [-\ln(1-p_2)]^{1/\tilde{\beta}} & -\frac{\tilde{\theta}}{\tilde{\beta}^2}[-\ln(1-p_2)]^{1/\tilde{\beta}}\ln[-\ln(1-p_2)] \\ [-\ln(1-p_3)]^{1/\tilde{\beta}} & -\frac{\tilde{\theta}}{\tilde{\beta}^2}[-\ln(1-p_3)]^{1/\tilde{\beta}}\ln[-\ln(1-p_3)] \end{pmatrix}.$$

(4.12)

Using the MLE of the quantile function (4.9), $\tilde{\xi}_\alpha = \tilde{\theta}\left[-\ln(1-\alpha)\right]^{1/\tilde{\beta}}$, along with the observed information matrix (4.11) and the MLEs of the quantile derivative matrix (4.12), we can calculate the consistent MLE of the variance Ω specified in (2.11). Hence the parametric asymptotic confidence limit (2.12) and the test statistic defined in (2.13) can be easily expressed with sample data values.

5. Illustrative Examples

In this section, we simulate two datasets from the two-parameter lognormal family and Weibull family respectively and use them to illustrate how to construct parametric and nonparametric confidence limits and test hypothesis introduced in Sections 2, 3 and 4. Throughout this section and the next section of simulation, we choose $\nu = 1$.

5.1 Example 1: Weibull process

We use the sample process specification limit ($USL = 10$) and the target ($T = 0$) values used in Vännman and Albing (2007) and generate 100 random sample data values from Weibull population with scale parameter $\theta = 2.2$ and shape parameter $\beta = 1.5$ as follows
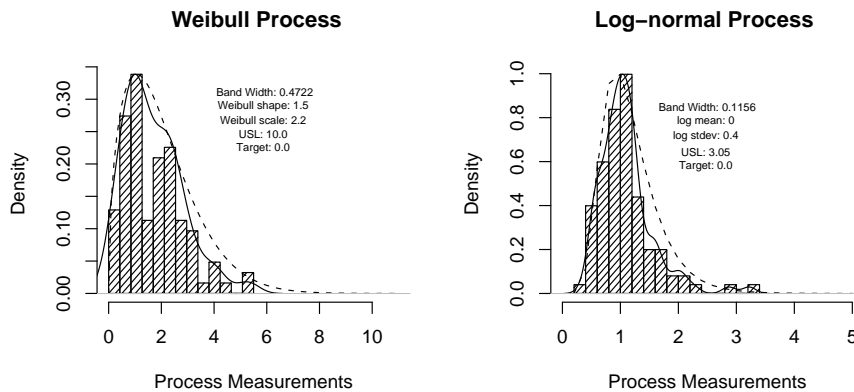

Figure 1: The broken curves represent the estimated parametric density curves and the solid curves represent the nonparametric kernel density curves

The left panel of Figure 1 gives the histogram based on the simulated data along with the true density curve (scale $\theta = 2.2$ and shape $\beta = 1.5$) and the kernel density curve (binwidth = 0.4722).

For the parametric approach, we first make a histogram of the data and then choose appropriate parametric distribution(s) based on the histogram to fit the

data. A goodness-of-fit test is conducted to avoid model misspecification. In this example, we use Kolmogorov-Smirnov goodness-of-fit test based on Weibull and log-normal distributions and found that the two p-values are 0.9683 and 0.1714 respectively. Since the p-value based on Weibull distribution is higher than that based on log-normal distribution, we choose Weibull as the final model (in fact, the data was generated from Weibull distribution). The MLE of θ and β are $\tilde{\theta} = 2.211263$ and $\tilde{\beta} = 1.5141$. The MLE of $C_{MA}(\theta, \beta)$ is $C_{MA}(\tilde{\theta}, \tilde{\beta}) = 1.3587$. The 95% asymptotic confidence interval of C_{MA} is $(1.25, \infty)$. The p-value of the asymptotic parametric normal test for testing $H_0 : C_{MA}(\xi) \leq 1$ versus $H_a : C_{MA}(\xi) > 1$ is approximately 0 which indicates that the underlying process is capable.

For the nonparametric approach, the median and the 99.73-th quantile are $\hat{\xi}_{p_2} = 1.715$ and $\hat{\xi}_{p_3} = 6.305$ respectively. The kernel density estimator evaluated at the two sample quantiles gives $\hat{f}(1.715) = 0.278843$ and $\hat{f}(6.305) = 0.01638945$ respectively. Using these values, we obtain the point estimate $\hat{C}_{MA}(\xi) = 1.53$. The 95% asymptotic nonparametric confidence interval (3.6) and the nonparametric test statistic (3.7) for testing $H_0 : C_{MA}(\theta, \beta) \leq 1$ versus $H_a : C_{MA}(\theta, \beta) > 1$ are given by $(1.382, \infty)$ and 7.312 with p-value 0.

The results obtained based on both parametric and nonparametric approaches agree with that of Vännman and Albing (2007) in which the sample quantiles and the true parametric density were used.

5.2 Example 2: Lognormal process

In the second example, we generate 100 data values from log-normal distribution with geometric mean 0 and geometric standard deviation 0.4. We choose $USL = 3.05$ to guarantee that the proportion of nonconforming is at least 99.73% if the associated C_{MA} is at least 1. The right panel in Figure 1 gives the true density (broken curve) and the kernel density (solid curve) with binwidth 0.1156.

Similar to the steps used in example 1, we performed Kolmogorov-Smirnov test based on Weibull and Lognormal distributions and obtained p-values 0.04242 and 0.61 respectively. Since the test rejects the Weibull distribution and fails to reject lognormal, we choose lognormal distribution (in fact, the dataset was generated from the lognormal distribution).

For the parametric approach, we first find MLEs $\tilde{\mu} = 0.02258$ and $\tilde{\sigma} = 0.3830$. The MLE of $C_{MA}(\mu, \sigma)$ is $C_{MA}(\tilde{\mu}, \tilde{\sigma}) = 0.9712$. The 95% asymptotic confidence interval of C_{MA} is $(0.8075, \infty)$. The p-value of the asymptotic parametric t test for testing $H_0 : C_{MA}(\mu, \sigma) \leq 1$ versus $H_a : C_{MA}(\mu, \sigma) > 1$ is approximately 0.929 indicating that the process is barely capable or incapable.

For the nonparametric approach, the median and the 99.73-th quantile are $\hat{\xi}_{p_2} = 1.03$ and $\hat{\xi}_{p_3} = 3.126$ respectively. The kernel density estimator evaluated at

the two sample quantiles yields $\hat{f}(1.03) = 1.173559$ and $\hat{f}(3.126445) = 0.02786794$ respectively. Using these values, we obtain the point estimate $\hat{C}_{MA}(\xi) = 0.9266$. The 95% asymptotic nonparametric confidence interval (3.6) and the nonparametric test statistic (3.7) for testing $H_0 : C_{MA}(\xi) \leq 1$ versus $H_a : C_{MA}(\xi) > 1$ are given by $(0.8286, \infty)$ and -1.4696 with p-value 0.9292 which matches the parametric result.

6. A Simulation Study on the Sample Size Requirement

The open source statistical package R is used to carry out all data analysis presented in the previous section and the simulation study in this section as well. The program is available from the author upon request.

In this section, we will investigate the sample size needed in asymptotic normal approximation for both parametric and nonparametric confidence limits and tests of the process capability. To be more specific, we first choose different sample sizes and population parameters which generates different skewed populations, then conduct the Shapiro's normality test to see the discrepancy between the normal distribution and the sampling distribution of the consistent estimator of the qPCI. The p-value of Shapiro's test of each of the simulated samples with different sample sizes and population parameters will be reported. The distributions we use in this simulation are Weibull and lognormal distributions.

For the Weibull family, we first fix the scale parameter at $\theta = 2.2$ and choose various values for the shape parameter $\beta = 0.2, 0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$. One can see the two upper panels of Figure 2 that as the the value of β decreases the tail of Weibull distribution gets longer. Then we fix the shape parameter at $\theta = 1.5$ and use different values of the scale parameter $\theta = 1.0, 1.2, 1.4, 1.6, 1.8, 2.0$. The density curves are shown in the bottom left panel of Figure 2.

For the lognormal family, we only choose different values of geometric standard deviation $\sigma = 0.5, 0.7, 0.9, 1.1, 1.3, 1.5$ and fix the location parameter $\mu = 0$ since the location parameter does not affect the shape of the distribution. The corresponding density curves are given in the bottom right panel of Figure 2.

The sample sizes that we used in this simulation are $n = 50, 100, 150, 200, 250, 300, 350, 400, 500, 600, 700, 800, 900$, and 1000. For each combination of population parameters, (θ, β) for the Weibull and (μ, σ) for the lognormal, and the sample size, we generate 1000 samples from the corresponding distribution. We then use each of these 1000 simulated samples to evaluate parametric (MLE) and nonparametric PCIs proposed in (2.3) and (3.2) in Sections 2 and 3 and
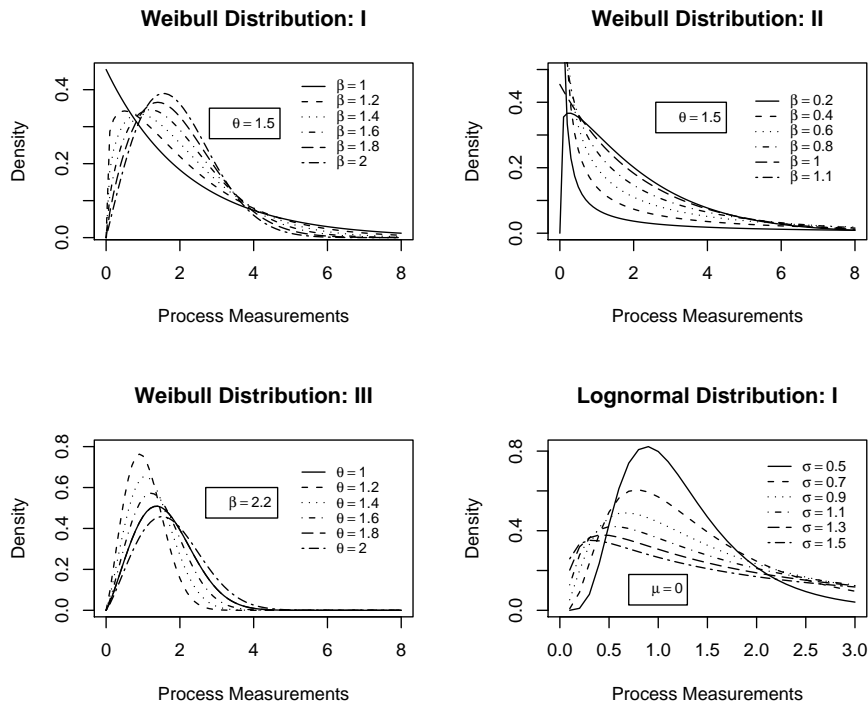
Figure 2: The density curves from lognormal and Weibull families with different values of shape related population parameters.

obtain two sets of 1000 estimated PCIs. Finally we conduct Shiparo's normality test on the two sets of estimated PCIs. The p-values obtained from the simulation indicate that (the detailed numerical results of the simulation is available at http://www.usm.maine.edu/~cpeng/jds582sim.pdf):

1. for both parametric and nonparametric PCIs, the sample size required for asymptotic normal approximation gets larger (in order to achieve an appropiate significant level or p-value) as the shape parameter gets smaller (or equivalently the tail of the distribution gets longer, see also the upper two panels of Figure 2); parametric PCI requires a relatively larger sample size (with the sample cut-off p-value).

2. for the fixed shape parameter $\beta = 1.5$, different values of the scale parameter do not affect the sample very much.

3. as the value of geometric standard deviation increases (the tail of the distribution becomes longer, see the bottom right panel of Figure 2), the sample size required for asymptotic normal approximation increases. Similar to the

Weibull case, the nonparametric PCI requires a relatively smaller sample size than its parametric counterpart.

Finally, in order to see the pattern of how the length of tail affects the sample size required for asymptotic approximation, we choose three lognormal distributions with geometric standard deviations $\sigma = 0.5, 1.0, 1.5$ (based on the fixed geometric mean 0) and three Weibull distributions with shape parameter $\beta = 0.4, 1.0, 1.6$ (based on the fixed scale parameter 2.2). These density curves are given in Figure 3. Simulation results show that the length of the distribution tail affects the sample size required for asymptotic normality: the longer the tail is, the larger the sample size required.
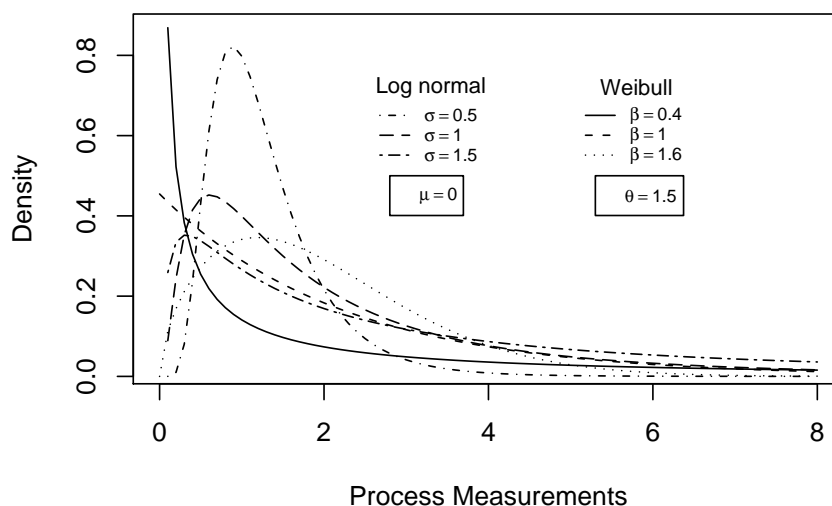

Figure 3: The density curves from both lognormal and Weibull families with different values of shape parameter for Weibull and different geometric standard deviations for lognormal distributions.

7. Summary and Concluding Remarks

We have systematically developed both parametric and nonparametric methods of applying the family of PCIs proposed by Vännman and Albing (2007). The major contributions are

1. We introduce a kernel density estimator for estimating the density of the underlying skewed process and obtain a consistent estimator of the variance

of the proposed qPCI which has NOT been discussed by Vännman and Albing (2007). Therefore, our work makes the proposed qPCI available for practical implementation under a purely nonparametric setting.

2. We also develop a general asymptotic parametric procedure for the proposed qPCI. The general recommendation for using this parametric procedure is to use parametric method if the underling distribution is given or can be easily identified by performing a goodness-of-fit test for the fitted model.

The procedures discussed in this article are based on large sample theory. In practice, the sample size required for both methods depends on the length of the tail of the distribution for the process. The longer the tails, the larger the sample size is reqiured. A general practical recommendation is to use different ways, for example resampling methods, to get the sampling distribution of the estimated PCI and make sure that the sample size is sufficiently large for using the asymptotic results. Similar to Chao and Lin's (2005) C_y using two extreme tail probabilities of the underlying process, the quantile based PCIs discussed in this paper are dependent on the extreme process quantiles. Therefore, Chao and Lin's (2005) recommendation of giving the first priority to the parametric approach whenever possible also applies to our case.

Acknowledgment

References

Casella, G. and Berger, R. L. (2002). *Statistical Inference, 2nd ed.* Duxbury Press.

Chang, P. L. and Lu, K. H. (1994). PCI calculations for any shape of distribution with percentile, *Quality World*, Technical Section. September, 110-114.

Chang, Y. S. and Bai, D. S. (2001). Control charts for positively-skewed populations with weighted standard deviations, *Quality and Reliability Engineering International* **17**, 397-406.

Chang, Y. S., Choi, I. S. and Bai, D. S. (2002). Process capability indices for skewed populations, *Quality and Reliability Engineering International* **18**, 383-393.

Chao, M.-T. and Lin, D. K. J. (2005). Another look at the process capability index, *Quality and Reliability Engineering International* **22**, 153-163.

Chen, K. S. and Pearn, W. L. (1997). An application of non-normal process capability indices, *Quality and Reliability Engineering International* **13**, 355-360.

Chen, S. M. and Hsu, Y. S. (2003). Asymptotic analysis of estimators for C_{Np} based on quantile estimators, *Journal of Nonparametric Statistics* **15**, 137-150.

Clements, J. A. (1989). Process capability calculations for non-normal distributions. *Quality Progress* **22**, 95 - 100.

Hyndman, R. J. and Fan Y. (1996). Sample quantiles in statistical packages *American Statistician* **50**, 361 - 365.

Juran, J.M., Gryna, F. M. and Binghan, R. S. (1974). *Quality Control Handbook*, McGraw-Hill, New York.

Kotz, S. and Lovelace, C. R. (1998). *Process Capability Indices in Theory and Practice.* Arnold.

Lin, H. C. (2004). The measurement of a process capability for folded normal process data. *Journal of Advanced Manufacturing Technology* **24**, 223-228.

Lin, P. C. (2005). Application of the generalized folded-normal distribution to the process capability measures. *Journal of Advanced Manufacturing Technology* **26**, 825-830.

Lovelace, C. R., Swan, J. and Messimer, S. (1997). A modification of C_{pk} for non-normal, zero-bound process data using lognormal quality control techniques, Manuscript, University of Alabama in Huntsville, Huntsville, AL.

Pearn, W. L. and Chen, K. S. (1997). Capability indices for non-normal distributions with an application in electrolytic capacitor manufacturing. *Microelectronics and Reliability* **37**, 1853 -1858.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics.* John Wiley.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman and Hall.

Spiring, F., Leung, B., Cheng, S., and Yeung, A. (2003). A bibliography of process capability papers. *Quality and Reliability Engineering International* **19**, 445-460.

Vännman, K. (1995). A unified approach to process capability indices. *Statistica Sinica* **5**, 805-820.

Vännman, K. and Albing, M. (2007). Process capability indices for one-sided specification itervals and skewed distributions. *Quality and Reliability Engineering International* **23**, 755-765.

Cheng Peng
Department of Mathematics and Statistics
University of Southern Maine
96 Falmouth Street
Portland, Maine 04104, USA
cpeng@usm.maine.edu