

PERFORMANCE ANALYSIS FOR SKEWED DATA

¹S. Ahmad, ²M. Abdollahian, ³P. Zeephongsekul, ⁴B. Abbasi

^{1,2,3}Department of Statistics and Operations Research, RMIT University, Melbourne, Australia,
shafiq.ahmad@rmit.edu.au, mali.abdollahian@rmit.edu.au, panlop.zeephongsekul@rmit.edu.au,
⁴Department of Industrial Engineering, Sharif University of Technology, Tehran, Iran
b.abbasi@gmail.com

ABSTRACT

Information technology and the media have changed the face of business practices today. Customers now play the key role for the success of any business and non-compliance to their specifications will simply lead to failure of the business. Researchers across many disciplines have worked out several modifications of the traditional process capability measures to obtain better estimates of the products capability to meet market specifications. However, these conventional capability measures heavily depend upon the theory of normality. In this paper, we compare and contrast the Cumulative Distribution Function (CDF) method with the latest proposed process capability evaluation methods such as Burr percentile method and commonly used Clements percentile method when the underlying distribution is non-normal. A simulation study using Gamma, Weibull and Beta distributions is conducted and the comparison of the results is presented. Finally, a case study is presented using actual data from a manufacturing process.

Keywords: Process Capability Index (PCI), Proportion of nonconforming in non-normal process, CDF method, Quantile based capability indices.

1 INTRODUCTION

Process capability indices (PCIs), process yield and process expected losses are three basic means that have been widely used in measuring process performance. Of the three, PCI is least complex to understand and deploy to any process. The larger PCI value implies the higher process yield, and the larger PCI also indicates the lower process expected loss. Therefore, the PCI can be viewed as an effective and excellent means of measuring product quality and process performance [1].

The conventional process capability index C_p is defined as:

$$C_p = \frac{usl - lsl}{6\sigma} \quad (1)$$

where, usl and lsl (upper and lower specification limits) are the design tolerance limits also called customer specifications.

The process ratio for off-center process C_{pk} is defined as:

$$C_{pk} = \min\{C_{pu}, C_{pl}\} \quad (2)$$

where C_{pu} and C_{pl} :

$$C_{pu} = \frac{\text{upper specification} - \text{process mean}}{3\sigma} \quad (3)$$

$$C_{pl} = \frac{\text{process mean} - \text{lower specification}}{3\sigma} \quad (4)$$

where C_{pu} and C_{pl} refer to as upper and lower one sided capability indices μ and σ are the process mean and standard deviation respectively.

Process capability index C_p defined here is heavily based on certain assumptions such as collected data is from an in-control process, independent and identically distributed and follows normal distribution. However, most of the processes in the real world produce non-normal data and the quality practitioners need to consider the basic assumptions before deploying any conventional process capability index.

The calculation of the conventional PCI measure requires the values of three points within the process distribution; the upper tail, the point of central tendency and the lower tail [2].

In normal distribution, in terms of quantiles; $X_{0.99865} = \mu + 3\sigma$ is the upper tail and $X_{0.5} = \mu$, in

general; is the median and $X_{0.00135} = \mu - 3\sigma$ corresponds to lower tail.

In case of normal data, it is easy to estimate quantile points. However, for the non-normal data, it is not easy to estimate them. To deal with non-normality; one approach is to transform the non-normal data to approximately normal data using mathematical functions. Johnson [3] proposed a system of distributions based on the moment method called the Johnson transformation system. Box and Cox [4] also used transformation method for non-normal data by presenting family of power transformations. Somerville and Montgomery [5] proposed using a square-root transformation to transform a skewed distribution into a normal one. The main objective of all these transformations is that one can apply conventional PCIs once the data is transformed to normal data.

Clements [6] proposed a percentile method to calculate C_p and C_{pk} indices for the non-normal data using the Pearson family of curves. Liu and Chen [7] proposed a modified Clements PCI percentile method using Burr XII distribution. Ahmad et al. [8] compared Liu and Chen's method with the commonly used Box-Cox method and concluded that Burr method provides slightly better estimates of PCI for the non-normal data.

In this paper, we will review and compare CDF, Clements and Burr methods which are commonly used to evaluate the PCIs for the non-normal data. This paper is organized in the following manner. PCI methods for the comparison study are discussed in section 2. For illustrational purposes, a simulation study using Weibull, Gamma and Beta distributions is presented in section 3 & 4, an application example with real world data is presented in section 5 and the conclusion is given in section 6.

2. PCI FOR NON-NORMAL DATA

In this section a brief review of the three different methods that are used in this paper is presented.

2.1 Clements Percentile PCI Method

Clements method is popular among quality practitioners in industry. Clements [6] proposed that 6σ in equation (1) be replaced by the length of the interval between the upper and lower 0.135 percentage points of the distribution of X. Therefore, the denominator in equation (1) can be replaced by $(U_p - L_p)$, i.e.

$$C_p = (usl - lsl) / (U_p - L_p) \quad (5)$$

where U_p is the upper percentile i.e. 99.865

percentile of observations and L_p is the lower percentile i.e. 0.135 percentile of observations. Since the median "M" is the preferred central value for a skewed distribution, so he defined C_{pu} and C_{pl} as follows:

$$C_{pu} = (usl - M) / (U_p - M) \quad (6)$$

$$C_{pl} = (M - lsl) / (M - L_p) \quad (7)$$

$$\text{and } C_{pk} = \min\{C_{pu}, C_{pl}\} \quad (8)$$

Clements approach uses the standard estimators of skewness and kurtosis that are based on 3rd and 4th moments respectively, and may not be reliable for very small sample sizes [7]. Wu et al [9] have conducted a research study indicating that the Clements method cannot accurately measure the capability indices, especially when the underlying data distribution is skewed.

2.2 Burr Percentile PCI Method

Burr [11] proposed a distribution called Burr XII distribution, whose probability density function is defined by:

$$f(x) = \begin{cases} \frac{ckx^{c-1}}{(1+x^c)^{k+1}} & \text{if } x \geq 0; c, k \geq 1 \\ 0 & \text{if } y < 0 \end{cases} \quad (9)$$

Cumulative distribution function is defined by:

$$F(x) = 1 - \frac{1}{(1+x^c)^k} \quad \text{if } x \geq 0; c, k \geq 1 \quad (10)$$

where c and k represent the skewness and kurtosis coefficients of the Burr distribution respectively.

Liu and Chen [7] introduced a modification based on the Clements method, whereby instead of using Pearson curve percentiles, they replaced them with percentiles from an appropriate Burr distribution. The proposed modified method is as follows

- Estimate the sample mean, sample standard deviation, skewness and kurtosis of the original sample data.
- Calculate standardized moments of skewness (α_3) and kurtosis (α_4) for the

given sample size n (see Appendix I for details)

- Use the values of α_3 and α_4 to select the appropriate Burr parameters c and k , Burr IW [11]. Then use the standardized tails of the Burr distribution XII to determine standardized 0.135, 0.5, 99.865 percentiles (X).
- Calculate estimated percentiles using Burr table for lower, median, and upper percentiles as follows:
- Calculate estimated percentiles using Burr table for lower, median, and upper percentiles as follows:

$$L_p = \bar{x} + (X_{0.00135} \times s) \quad (11)$$

$$U_p = \bar{x} + (X_{0.99865} \times s) \quad (12)$$

$$M = \bar{x} + (X_{0.50} \times s) \quad (13)$$

- Calculate process capability indices using equations 5-8.

2.3 CDF PCI METHOD

Wierda [12] introduced a new approach to evaluate process capability for a non-normal data using Cumulative Distribution Function (CDF). Castagliola [13] used CDF approach to compute proportion of non-conforming items and then estimate the capability index using this proportion. Castagliola showed the relationship between process capability and proportion of non-conforming items and used CDF method to evaluate PCI for non normal data by fitting a Burr distribution to the process data. He used a polynomial approximation to replace empirical function in the Burr distribution, and then used the proposed method given by equation (14). To calculate C_p we give a short proof of this well known result in Appendix II.

Using CDF method C_p and C_{pk} are defined by;

$$C_p = \frac{\Phi^{-1}(0.5 + 0.5 \int_{Lsl}^{usl} f(x) dx)}{3} \quad (14)$$

$$C_{pk} = \min(C_{pu}, C_{pl}) \quad (15)$$

$$\text{where } C_{pl} = \frac{\Phi^{-1}(0.5 + \int_{Lsl}^T f(x) dx)}{3} \quad (16)$$

$$C_{pu} = \frac{\Phi^{-1}(0.5 + \int_T^{usl} f(x) dx)}{3} \quad (17)$$

where $f(x)$ represents the probability density function of the process and T represents the process mean for normal data and process median for non-normal data. In this paper $f(x)$ in Equation (14) is replaced by Equation (9) i.e. Burr density function (see details in Appendix III).

3. SIMULATION STUDY

Three non-normal distributions; Gamma, Weibull and Beta have been used to generate random data in this simulation. These distributions are used to investigate the effects of non-normal data on the process capability index. These distributions are known to have the parameter values that can represent mild to severe departures from normality. These parameters are selected so that we can compare our simulation results with existing results using the same parameters in the literature. The probability density function of Gamma distribution, with parameters α and β , is given by

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-\frac{x}{\beta}}, \alpha, \beta > 0, x \geq 0 \quad (18)$$

The parameters used in this simulation are shape=4.0 and scale=0.5

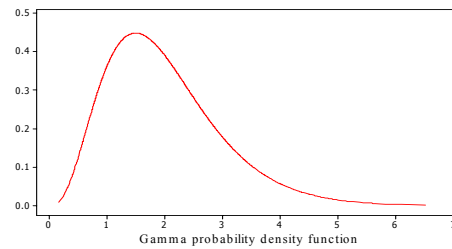


Figure 1: pdf of Gamma distribution with parameters (shape= 4.0, scale= 0.5)

The probability density function of Weibull distribution with shape (α) and scale (β) is given by

$$f(x) = \frac{\alpha}{\beta} (x^{\alpha-1}) e^{-x^{\frac{\alpha}{\beta}}}, \alpha, \beta > 0, x \geq 0 \quad (19)$$

The parameters used in this simulation are: $\alpha = 1.0$ and $\beta = 1.2$

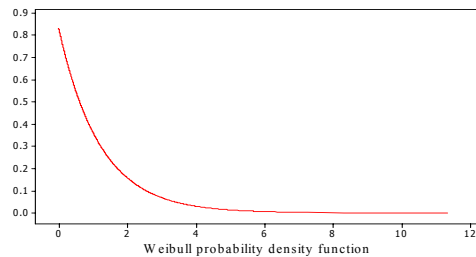


Figure 2: pdf of Weibull distribution with parameters ($\alpha = 1.0, \beta = 1.2$)

The probability distribution function of Beta distribution with shape 1 (α) and shape 2 (β) is given by

$$f(x) = \left[\frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \right] x^{\alpha-1} (1-x)^{\beta-1}, < 0x < 1 \quad (20)$$

The parameters used in this simulation are: $\alpha = 4.4$ and $\beta = 13.3$

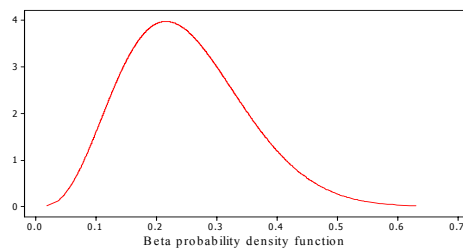


Figure 3: pdf of Beta distribution with parameters ($\alpha = 4.4, \beta = 13.3$)

3.1 Comparison Criteria

The criterion for comparison in this simulation study is based on proportion of non-conformances (PNC). The proportion of non-conforming units for a normal distribution can be determined by [2]

$$PNC = \Phi(-3C_{pu}^*) \quad (21)$$

The C_{pu}^* values in table (1) are computed using equation (17) where $f(x)$ is replaced by the

corresponding distributions (i.e. Gamma, Weibull and Beta). Probability of non-conforming items (PNC) is calculated using equation (21) as suggested by Castagliola [13] for all three methods (e.g. for Gamma distribution with C_{pu} value 0.8698, corresponding PNC value using equation (21) will be 0.0045351).

Figure 4 presents flowchart of estimating PNC and PCI's using different methods and different non-normal distributions. The exact PNC value (p) in this flow chart is obtained using following equation.

$$PNC = 1 - \int_0^{usl} f(x) dx \quad (22)$$

where $f(x)$ represents the corresponding distribution function of Gamma, Weibull and Beta distributions.

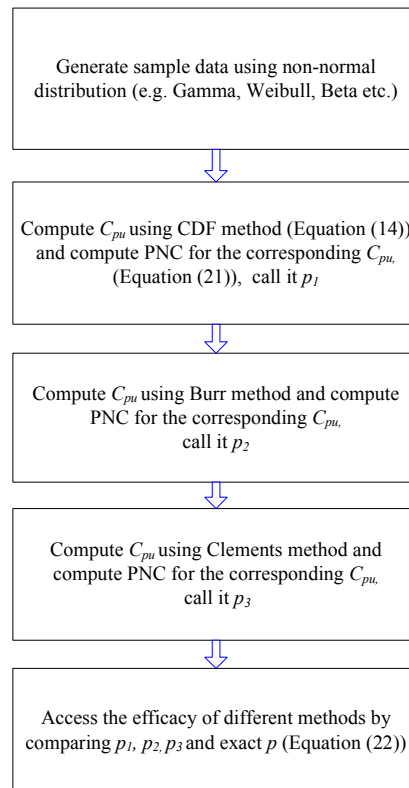


Figure 4: Simulation methodology flowchart

3.2 Simulation Results

These C_{pu}^* values in table (1) are used to access the efficacy of the three method in estimating process capability index for non-normal data. Table (1) shows the results of this comparison.

Table (1) shows the results of this comparison

Distribution	USL	C_{pu}^*	C_{pu} Clements	C_{pu} Burr	C_{pu} CDF
Gamma(4,0.5)	6.3405	1.000	0.8698	0.9069	1.0000
Weibull(1,1.2)	5.0	1.043	0.9694	0.9738	1.0292
Beta(4.4,13.3)	0.5954	1.002	0.7434	0.7965	1.0028

*Computed from Equation (21) – percentile and exact distribution

The simulation results given in Table (1) show that C_{pu} values obtained using Clements method are worse than those obtained using Burr and CDF methods. The C_{pu} values obtained using the CDF method are the closest to those C_{pu} values obtained using direct distribution percentiles in the conventional approach; thus, leading to better estimates of the PCIs compare with the Burr method.

Our comparison criteria is that the method which yields expected proportion of non-conformities closest to that obtained using exact distribution would be the most superior method.

Table (2) – proportion of nonconformance

Distribution	Table (2) – proportion of nonconformance (PNC)			
	Comparison of expected proportion of nonconformance (PNC) with exact PNC			
	Clements p_3	Burr p_2	CDF p_1	Exact p
Gamma	0.00454	0.00326	0.00135	0.0013
Weibull	0.00182	0.00170	0.00101	0.0010
Beta	0.01287	0.00844	0.00131	0.0013

Results in table 2 show that PNC values obtained using Clements method are worse than the other 2 methods. In this table PNC values using CDF method are close to the PNC values obtained using exact distribution. Thus the later method is giving better estimates of non-conformances as compared to the commonly used Clements and Burr methods.

4. DISCUSSION

Simulation study shows that both Burr and PNC methods are estimating C_{pu} values more accurately than commonly used Clements method. Looking at the results as depicted in tables 1 & 2, we conclude that:

- CDF method is superior to both percentile methods (Burr & Clements)
- Burr method is still performing better than the commonly used Clements method.

- CDF method is the one for which the estimated C_{pu} value deviates least from the target C_{pu} value.
- For the given sample size, PNC value obtained using CDF method is comparable with the targeted PNC value obtained from exact distribution.

During simulation, we have observed that data having moderate departure from normality provides better estimates of capability indices compared with data having severe departures from normality.

5. REAL DATA EXAMPLE

A case study using data from a manufacturing industry is conducted. All three methods have been deployed to estimate the non-normal process capability for the experimental data. Data has been collected from an in-control manufacturing process. The data is the measurements of bonding area between two surfaces with upper specification limit (USL = 24). The summary statistics of the process data is:

μ (mean) = 23.4809, σ = 0.5650, $\tilde{\mu}$ (median) = 23.3963, μ_3 (skewness) = 1.1098, μ_4 (kurtosis) = 4.9740.

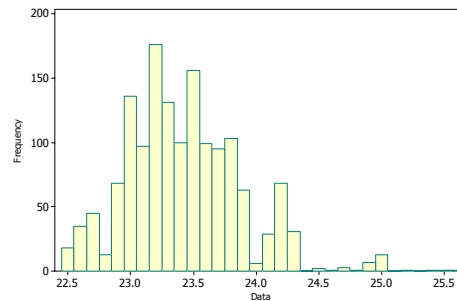


Figure 5: Histogram of the real data

We have selected 30 samples of size 50 from these data points. For each sample; we computed the process capability index C_{pu} and proportion of non-conforming PNC by using Clements, Burr and the CDF method. The mean and standard deviation of the estimated C_{pu} values are given in table (3).

Table (3) – result of the real example based on 30 samples of size n=50

$C_{pu} \rightarrow$	Mean	Standard deviation	Expected PNC using Eq (21)
CDF	0.313277	0.023811	0.17365
Burr	0.347917	0.065859	0.14830
Clements	0.360691	0.080264	0.13961

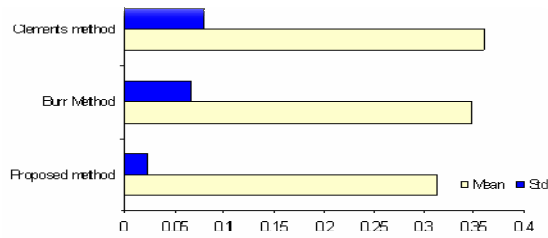


Figure 6: Comparison of three methods

For CDF method; we have replaced the corresponding $f(x)$ by the Burr distribution. The Burr parameters for each sample have been estimated using maximum likelihood estimation. The exact PNC for experimental data is 0.168. This PNC value is obtained by using upper specification limit $USL=24$ and calculating the proportion of data that falls outside the specification limit.

The results presented in table 3, indicates that the expected PNC based on 30 samples of size 50, using CDF method is the closest estimate to the exact PNC. Table 3 also indicates that CDF method has the least variability as compared to the other two methods.

6. CONCLUSIONS

In this paper a comparison between three methods of estimating the process capability and the proportion of non-conformance in the manufacturing industry is presented. The CDF method is not sensitive to distribution of the process data and therefore can be applied to any real set of data as long as a suitable distribution can be fitted to it. However, to apply the CDF method, one must identify the corresponding distribution. One of the significant characteristics of Burr XII distribution is that, when mean, variance, skewness and kurtosis of the process data are obtained; using Burr tables (Liu and Chen [7]) we can fit a suitable Burr distribution. Therefore we can conclude that by replacing the probability density function $f(x)$ in the CDF method with the appropriate Burr density function would lead to a better estimate for PCI and PNC of non-normal data.

Simulation studies for different non-normal distributions show that the CDF method using Burr distribution produces better estimates of PCI.

This paper strongly recommends further research to extend the CDF method to non-normal multivariate PCI studies in this area.

APPENDIX I:

Standardized moments of skewness (α_3) and kurtosis (α_4) for the given sample size n can be computed as follows:

$$\alpha_3 = \frac{(n-2)}{\sqrt{n(n-1)}} * Skewness \quad (1)$$

where

$$Skewness = \frac{n}{(n-1)(n-2)} \sum \left\{ \frac{(x_j - \bar{x})}{s} \right\}^3 \quad (2)$$

where \bar{x} is mean of the observations and s is the standard deviation.

$$\alpha_4 = \frac{(n-2)(n-3)}{(n+1)(n-1)} * (Kurtosis + 3) * \frac{(n-1)}{(n+1)} \quad (3)$$

where

$$Kurtosis = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_j - \bar{x}}{s} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)} \quad (4)$$

APPENDIX II:

Conventionally capability index C_p is defined as:

$$C_p = \frac{usl - lsl}{6\sigma} \quad (1)$$

If the process X is normally distributed with mean μ and standard deviation σ , i.e.

$X \sim N(\mu, \sigma^2)$, then

$$C_p = \frac{1}{3} \Phi_2^{-1} \left(\frac{1}{2} + \frac{B}{2} \right) \quad (2)$$

And $B = F(LSL < X < USL)$. On face value, it is not obvious that (1) and (2) are equal. Here is the proof:

- We first note that $P(X < USL) = \frac{1}{2} + \frac{B}{2}$.
(Draw a normal graph and you will see this!)
- Since $Z = \frac{X-\mu}{\sigma}$, we must also have that

$$P\left(Z < \frac{USL - \mu}{\sigma}\right) = \frac{1}{2} + \frac{B}{2} \quad (3)$$

which is equivalent to:

$$\varphi_Z^{-1}\left(\frac{1}{2} + \frac{B}{2}\right) = \frac{USL - \mu}{\sigma} \quad (4)$$

1. Because the *pdf* of Z is symmetric about the origin,

$$\frac{LSL - \mu}{\sigma} = -\frac{USL - \mu}{\sigma} = \varphi_Z^{-1}\left(\frac{1}{2} + \frac{B}{2}\right) \quad (5)$$

2. By equation (3). Finally,

$$\begin{aligned} C_p &= \frac{USL - LSL}{6\sigma} = \frac{1}{6} \left(\frac{USL - \mu}{\sigma} - \frac{LSL - \mu}{\sigma} \right) \\ &= \frac{1}{6} \left(\varphi_Z^{-1}\left(\frac{1}{2} + \frac{B}{2}\right) - \varphi_Z^{-1}\left(\frac{1}{2} + \frac{B}{2}\right) \right) \quad (**) \\ &= \frac{1}{3} \varphi_Z^{-1}\left(\frac{1}{2} + \frac{B}{2}\right) \end{aligned}$$

where, we have used (3) and (5), which concludes the proof.

APPENDIX III:

In this paper we fit Burr distribution function $f(x)$ to process data and then evaluate the PCI using CDF method. To fit the data distribution with Burr distribution, we need to estimate c and k parameters. The likelihood function of univariate Burr is:

$$L(c, k; x_1, \dots, x_n) = \frac{c^n k^n \prod_{i=1}^n (x_i)^{c-1}}{\prod_{i=1}^n (1 + x_i^c)^{k+1}} \quad (1)$$

In univariate Burr distribution there are two parameters c and k ; and to estimate these

parameters the maximum likelihood function with sample size n is:

$$\log L = n \log(c) + \log(k) - (1+k) \sum_{i=1}^n \log(1 + x_i^c) + (c-1) \sum_{i=1}^n \log x_i \quad (2)$$

The differential equations with respect to parameters c and k are:

$$\frac{\partial l}{\partial c} = \frac{n}{c} + \sum_{i=1}^n \log x_i - (k+1) \sum_{i=1}^n \frac{\log x_i \log x_i^c}{1 + x_i^c} \quad (3)$$

$$\frac{\partial l}{\partial k} = \frac{n}{k} - \sum_{i=1}^n \log(1 + x_i^c) \quad (4)$$

In this paper, unknown Burr parameters c and k have been determined by maximizing equation (2) using systematic random search algorithm named "Simulated Annealing".

REFERENCES

- [1] M. Deleryd, K. Vannman 'process capability plots—a quality improvement tool' Qual. Reliab. Engng. Int. 15: 213–227 (1999).
- [2] L C Tang, S E Than (1999) Computing process capability indices for non-normal data : a review and comparative study. Qual. Reliab. Engng. Int. 15: 339-353.
- [3] Johnson NL (1949) System of frequency curves generated by methods of translation. Biometrika 36:149–176
- [4] Box GEP, Cox DR (1964) An analysis of transformation. J Roy Stat Soc B 26:211–243
- [5] Somerville S, Montgomery D (1996) Process capability indices and non-normal distributions. Quality Engineering 19(2):305–316.
- [6] Clements JA (1989) Process capability calculations for non-normal distributions. Quality Progress 22:95–100
- [7] Pei-Hsi Liu, Fei-Long Chen (2006), "Process capability analysis of non-normal process data using the Burr XII distribution", Int J Adv Manuf Technol 27: 975–984
- [8] S. Ahmad, M. Abdollahian, P. Zeepongsekul (2007) Process capability analysis for non-quality characteristics using Gamma distribution. 4th international conference on information technology – new generations, USA, April, 02-04: 425-430

- [9] Wu HH, Wang JS, Liu TL (1998) Discussions of the Clements-based process capability indices. In: Proceedings of the 1998 CIEE National Conference, pp 561–566
- [10] Burr IW (1942) Cumulative frequency distribution. *Ann Math Stat* 13:215–232
- [11] Burr IW (1973) Parameters for a general system of distributions to match a grid of α_3 and α_4 . *Commun Stat* 2:1–21
- [12] Wierda SJ. A multivariate process capability index. ASQC Quality Congress Transactions, Boston, MA, 1993, American Society for Quality Control: Milwaukee, WI, 1993; 342–348.
- [13] Castagliola P (1996) Evaluation of non-normal process capability indices using Burr's distributions. *Qual Eng* 8(4):587–593
- [14] Rodriguez RN (1977) A guide to the Burr type XII distributions. *Biometrika*, 64:129–134
- [15]. Chou CY, Cheng PH (1997) Ranges control chart for non-normal data. *J Chinese Inst Ind Eng* 14(4):401–409
- [16] Hatke M.A. (1949) A certain cumulative probability function. *Ann Math Stat*, Vol. 20, No. 3:461-463.
- [17] C.H. Yeh, F.C. Li, P.K. Wang, Economic design of control charts with Burr distribution for non- normally data under Weibull shock models: 12th international conference on Reliability and Quality in Design, (2006) 323-327.
- [18] V.E. Kane, Process capability indices, *J. Qual. Technol.* 18 (1986) 41–52 .
- [19] Montgomery, D., 'Introduction to Statistical Quality Control 5th edition, Wiley, New York, New York
- [20] Zimmer WJ, Burr IW (1963) Variables sampling plans based on non normal populations. *Ind Qual. Control* July:18–36