CHAPTER 3

PROCESS CAPABILITY INDICES FOR NON-NORMAL PROCESSES

3.1 INTRODUCTION

The process capability indices discussed in the previous chapters are designed under the assumption that the quality characteristic under investigation is normally distributed. All the statistical properties of the PCIs discussed earlier are dependent on this assumption. Especially crucial is the tie to process yield. If the data are even slightly non-normal, the tie to process yield is broken. For example, suppose the process is perfectly symmetric with $C_p = C_{pk} = 1.0$, but the process is actually t-distributed with six degrees of freedom then instead of 2700 parts per million out of specification for a normal process, there will be 10000 parts per million out of specification. Further if the data are actually logistically distributed, there will be 82840 parts per million out of specification. Hence before estimating the process capability of a certain process we must first assess the normality of the process measurements. Though process is in statistical control but the output does not give normally distributed data, then we can not use any of the indices C_p, C_{pk} or C_{pm} etc. Situations of these kind usually occur in practice. See, for example, Sarkar and Pal (1997). In such situations there are several indices available in the literature. In this Chapter we discuss some of them.

In section 3.2, we discuss some wellknown tests for testing normality of process measurements. Section 3.3 describes Clements' technique for process capability estimation of non-normal processes. Section 3.4 discusses Munechika's approach to assess process capability of such processes. In Section 3.5 the two approaches are compared. Further approaches are given in Section 3.6.

3.2 TESTS FOR TESTING NORMALITY

In this sectiom we first discuss the graphical method of assessing the normality of process measurements and then the wellknown chi-square test of goodness of fit. One can get an idea about non-normality by plotting the histogram of the data, when it does not come from symmetric distribution. However, if the data comes from symmetric distribution histogram fails to conclude about non-normality.

3.2.1 Graphical Method

Graphically assessing normality of process data is usually done by means of a so called Q-Q plot. This is in effect a plot of the sample quantile verses the quantile one would expect to observe if the observations actually were normally distributed. Data normality is suspected if the plotted points deviate from a straight line. In addition, the pattern of the deviation provide vital clues as to the nature of non-normality.

Let $x_1, x_2, \ldots, x_n$ represent n observations on any single characteristic X. Let $x_{(1)}, x_{(2)}, \ldots x_{(n)}$ represent these observations after they are ordered according to magnitude. The $x_{(j)}$'s are the sample quantiles. The proportion j/n of the sample to the left of $x_{(j)}$ is often approximated by (j - 1/2)/n for analytical convenience. For standard normal distribution the population quantiles $q_{(j)}$ are defined by the relation

$$P[X \le q_{(j)}] = \int_{-\infty}^{q_{(j)}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{j - 1/2}{n}.$$

Look at the pairs of quantiles $(q_{(j)}, x_{(j)})$ with the same associated cumulative probability, $(j - 1/2)/n$. The first member of the pair is the population or theoretical quantile $q_{(j)}$ and the second is the sample or observed quantile $x_{(j)}$. If the data arises from a normal population, the pair $(q_{(j)}, x_{(j)})$ will be approximately linearly related; That is, if we plot $x_{(j)}$'s against $q_{(j)}$'s, the plotted points will fall along a straight line.

To clarify the ideas, we provide an example in the following. Suppose that a sample of size, $n = 10$ observations is available. These observations are written in ascending order and are given in Table 3.1.

We construct the Q-Q plot for the given data of $x_{(j)}$. The Q-Q plot is as shown in Figure 3.1. The points $(q_{(j)}, x_{(j)})$ lie very nearly along a straight line and we would not reject the notion that these data are normally distributed.

Table 3.1

Observed observations $x_{(j)}$	cumulative probabilities $(j - 1/2)/n$	standard normal quantiles $q_{(j)}$
2.00	0.05	- 1.645
2.90	0.15	- 1.036
3.16	0.25	- 0.674
3.41	0.35	- 0.385
3.62	0.45	- 0.125
3.80	0.55	0.125
4.26	0.65	0.385
4.54	0.75	0.674
4.71	0.85	1.036
5.30	0.95	1.645

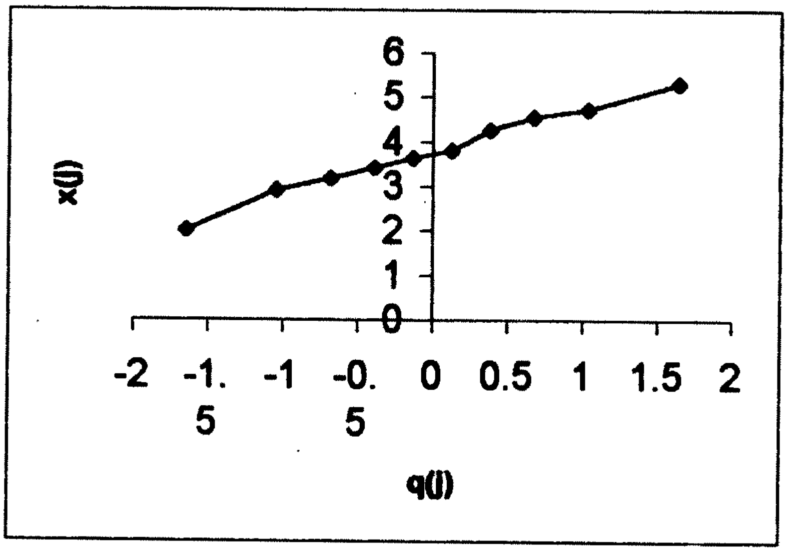


Figure 3.1

Another wellknown test for testing normality is chi-square test of goodness of fit, which is discussed below.

3.2.2 Chi-Square Test of Goodness of Fit

Let the random variable X denotes a quality characteristic with unknown mean μ and unknowm variance σ^2. Let $X_1, X_2, \ldots, X_n$ be independent observations on X. Let $A_1, A_2, \ldots, A_k$ be a collection of disjoint intervals that cover the real line and let

$$\hat{p}_i = P_{(\hat{\mu},\hat{\sigma}^2)}\left\{ X \in A_i \right\} > 0 \quad ; i = 1,2,\ldots, k,$$

where $P_{(\mu,\sigma^2)}$ is the probability distribution associated with $N(\mu, \sigma^2)$ and $\hat{\mu}$ and $\hat{\sigma}^2$ are maximum likelihood estimates of μ and σ^2 respectively. Let O_i denote the observed frequency of A_i and E_i be the expected frequency of A_i, when X is normally distributed. Clearly $E_i = n\hat{p}_i$. Then if X is normally distributed, the statistic

$$V_k = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

is asymptotically distributed as chi-square random variable with k - 3 d.f. We can use the statistic V_k to reject the hypothesis that X is normally distributed at α level of significance, if

$$V_k = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} > \chi^2_{k-3,1-\alpha} .$$

To use this test procedure for testing normality of X, first we have to choose the sets $A_1, A_2, \ldots, A_k$. Frequentely these are chosen to be disjoint intervals. As a rule of thumb, we choose the length of each interval in such a way that the probaility $P_{(\hat{\mu},\hat{\sigma}^2)}(X \in A_i)$, is approximately 1/k. Moreover, it is desirable to have $n/k \geq 5$ or rather $E_i \geq 5$ for each i. If any of the E_i's is < 5, the corresponding interval is pooled with one or more adjoining intervals to make the cell frequency atleast 5.

To illustrate the procedure we test the normality of the following data (the data is generated from N(0, 1) using MINITAB).

-2.884	-0.500	-0.158	0.151	-1.478	-0.366	0.578
-1.379	-0.879	-1.251	-0.582	1.293	0.038	0.586
0.478	-0.152	-0.270	0.625	0.619	0.747	0.955
-1.812	-0.125	2.259	-0.020	-0.327	0.730	-0.733
-1.047	-1.114	0.684	-0.868	0.446	0.389	-0.307
2.452	-1.526	0.758	0.042	-0.692		

Here we have n = 40, $\hat{\mu} = -0.116$, $\hat{\sigma} = 1.0358$. The procedure is summerized in Table 3.2:

Table 3.2

Interval	O_i	E_i
$(-\infty, -1.0)$	8	7.8682
$(-1.0, -0.4)$	6	7.8106
$(-0.4, 0.0)$	8	6.1045
$(0.0, 0.4)$	4	5.8492
$(0.4, 1.0)$	11	6.7416
$(1.0, \infty)$	3	5.6258

We have

$$V_k = 5.7995 \quad \text{and} \quad \chi^2_{3,0.05} = 7.8147$$

Since $V_k < \chi^2_{3,0.05}$, we can not reject the normality of the data.

If applying these tests of normality it is found that the process output is not normal, then we have to adopt some non-normal process capability estimation technique. In practice there are two commonly being used techniques which are given by Clements (1989) and Munechika (1986). In the following we discuss Clements' technique.

3.3 CLEMENTS' TECHNIQUE

Clements (1989) has proposed a simple technique for calculating C_p and C_{pk} for any shape of distributions, using the pearson family of curves. Consider the normal curve shown in Figure 3.2. For normal distribution with mean μ and standard

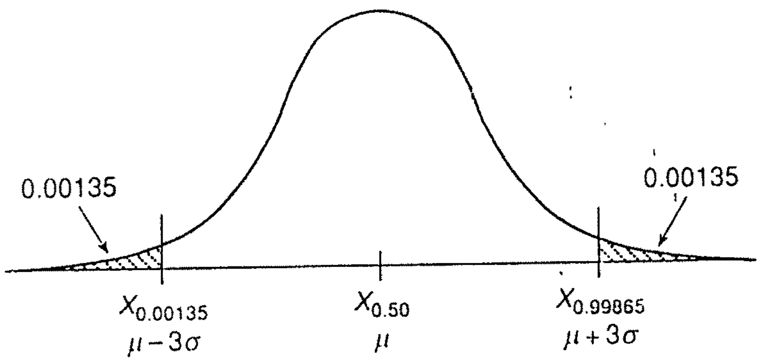


Figure 3.2

deviation σ, 0.135% of the process data lie below $\mu - 3\sigma$, 50% of the data lie below μ and 99.865% lie below $\mu + 3\sigma$, leaving 99.73% of the data lying between the limits $\mu \pm 3\sigma$. In quantile notation, we define x_p as

$$P [X < x_p] = p,$$

where $X \sim N(\mu, \sigma^2)$, so that

$$x_{0.00135} = \mu - 3\sigma$$

$$x_{0.5} = \mu$$

$$x_{0.99865} = \mu + 3\sigma$$

For standard normal case, with variable Z,

$$P[Z < z_{0.00135}] = 0.00135$$

$$P[Z < z_{0.5}] = 0.5$$

$$P[Z < z_{0.99865}] = 0.99865$$

The process capability indices can be redefined in terms of the quantiles of the quality characteristic as

$$C_p = \frac{USL - LSL}{x_{0.99865} - x_{0.00135}} \qquad (3.3.1)$$

$$C_{pk} = \min\left\{ \frac{USL - x_{0.5}}{x_{0.99865} - x_{0.5}}, \frac{x_{0.5} - LSL}{x_{0.5} - x_{0.00135}} \right\} \qquad (3.3.2)$$

In non-normal case, if we can find a better distributional form for the data, one that provides a very good fit, we can obtain more accurate measures of the three required quantiles.

Clements' technique consists of the following steps for estimating non-normal C_p and C_{pk}:

1. Estimate the mean ($\overline{X}$), standard deviation (S), skewness (S_k) and kurtosis (k_u) from the process data. Also note the values of USL and LSL for the process.
2. Based on the estimates of skewness and kurtosis obtained in (1), use either Table 1a or Table 1b (in

Clements'(1989)) to determine $z'_{0.00135}$ and $z'_{0.99865}$, where z'_p is the adjusted standardized normal variate for the non-normal data. For $z'_{0.00135}$ use Table 1a for positive skewness and table 1b for negative skewness. Similarly for $z'_{0.99865}$ use Table 1b for positive skewness and table 1a for negative skewness.

3. Calculate $z'_{0.5}$ using Table 2 [in Clements (1989)] and the estimates of skewness and kurtosis derived in (1):

   (a) for positive skewness reverse sign;

   (b) for negative skewness leave positive.

4. Estimate $x_{0.00135}$, $x_{0.5}$ and $x_{0.99865}$ as

$$\hat{x}_{0.00135} = \bar{X} - z'_{0.00135}S$$

$$\hat{x}_{0.5} = \bar{X} + z'_{0.5}S$$

$$\hat{x}_{0.99865} = \bar{X} + z'_{0.99865}S$$

5. Estimate process capability indices using (3.3.1) and (3.3.2).

Note that Clements defines kurtosis as $m_4/S^4 - 3$,

where, $$m_4 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^4}{n}$$

To illustrate the technique, we consider the following example.

Example 3.3.1:

The following 50 observations represent the output of some non-normal process. The specifications of the process are LSL = 2.8 and USL = 29.4.

It has been found that the data do not come from normal process, but from a skewed prosses.(see Figure 3.3.)

10.39	8.29	3.84	12.61	7.77	10.80	2.86
8.97	12.06	16.64	10.18	7.71	7.14	11.28
15.44	13.18	2.82	7.40	7.23	5.79	6.28
9.54	10.88	3.85	7.67	9.67	8.75	7.61
20.35	3.57	5.45	9.85	8.35	13.32	5.86
13.44	6.08	13.65	5.20	15.13	10.50	3.95
8.51	11.36	7.73	5.55	29.47	8.48	15.23
9.33						

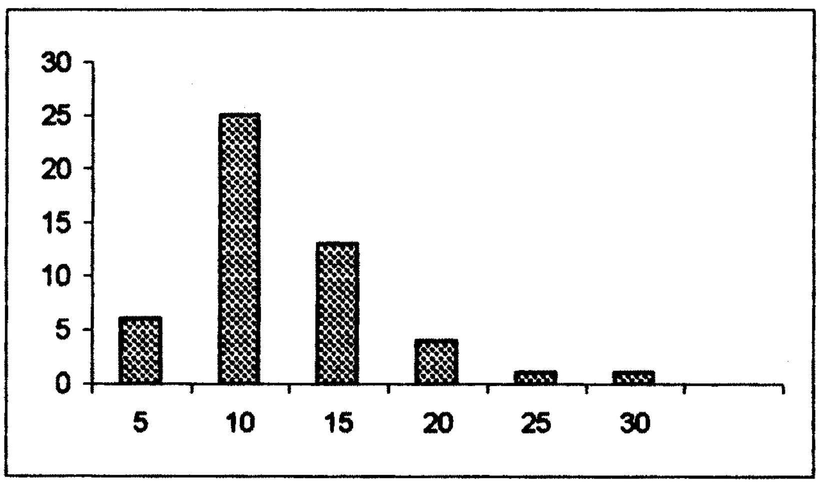


Figure 3.3

We have $\overline{X}$ = 9.54, S = 4.764, skewness = 1.69 and kurtosis = 2.25. From Table 1a, Table 1b and Table 2 of Clements (1989), we have

$$z'_{0.00135} = 0.7235, \quad z'_{0.5} = -0.480 \quad \text{and} \quad z'_{0.99865} = 3.8588$$

Using these values we get the quantile estimates of the PCIs as

$$\hat{C}_p = 1.2185 \quad \text{and} \quad \hat{C}_{pk} = 1.0917.$$

Clements' technique has following merits:

1. When the distribution is normal, the indices are exactly the same as those given by the traditional method.
2. The only difference from the traditional procedure is the method of calculating the width and position of the top and bottom halves of the distribution relative to the tolerance limits.
3. It does not require mathematical transformation of the data.
4. It is easy to visualize graphically and explain to nonstatisticians.
5. It is relatively easy to calculate manually or by using hand-held calculator.
6. It could be readily used as a subroutine in most existing SPC subroutines. (SYSTAT 7.0 does not take an account of non-normality.)

7. The system of Pearson curves on which it is based can be used to provide estimates of percenage out of specification for a wide variety of distributions.

Sundaraiyer (1996) utilized various bootstrap techniques to study point and interval estimates of Clements' C_{pk}, when the data are non-normal and has concluded that, the percentile and biased-corrected percentile bootstrap methods provide the best estimates of Clements' C_{pk}.

In the following we discuss another approach due to Munechika (1986).

3.4 MUNECHIKA'S APPROACH

Manechika (1986) has provided a non-normal modification of C_{pk} using Cornish-Fisher expansion. Cornish-Fisher expansion is essentially an expansion of any continuous distriution into an infinite series in terms of standerdized normal random variables U with the d.f. $\Phi(\cdot)$. Almost any standardized random variable X (that is, with mean 0 and variance 1) can be expressed as

$$X = U + B_1(U) + B_2(U) + \dots + B_k(U)$$

where $B_1(U) = \frac{1}{6} k_3(U^2 - 1)$

$$B_2(U) = \frac{1}{24} k_4(U^3 - 3U) - \frac{1}{36} k_3^2(2U^3 - 5U)$$

where k_i, $i = 1, 2, \ldots, 6$ are the cumulants related to the moments as follows:

$$k_1 = \mu \quad \text{(the mean)}$$

$$k_2 = \mu_2 \quad \text{(the second central moment)}$$

$$k_3 = \mu_3 \quad \text{(the third central moment)}$$

$$k_4 = \mu_4 - 3\mu_2^2$$

$$k_5 = \mu_5 - 10\mu_3\mu_2$$

Munechika (1986) truncates the expansion after $B_1(U)$ and shows that if X is chi-square random variable with any number of d.f., the error in truncation is negligible.

Denote, therefore,

$$g_1(u) = u + \frac{k_3}{6}(u^2 - 1) \tag{3.4.1}$$

and use g_1 to correct normal C_{pk}. He defines

$$C_{p\lambda} = \lambda C_{pk} \tag{3.4.2}$$

where

$$\lambda = \frac{\sqrt{k_3^2 + 18k_3 C_{pk} + 9} - 3}{3k_3 C_{pk}}, \quad k \neq 0 \tag{3.4.3}$$

(For details refer Munechika (1986).)

Munechika's procedure is thus as follows:

1. Calculate $\overline{X} = \frac{\Sigma x_i}{n}$,

$$S = \sqrt{\frac{\Sigma(X_i - \overline{X})^2}{n - 1}} ,$$

$$\hat{k}_3 = \frac{n\,\Sigma(X_i - \overline{X})^3}{[\Sigma(X_i - \overline{X})^2]^{3/2}} .$$

2. Calculate $\hat{C}_{pk}$.

3. Calculate $\hat{\lambda}$ using (3.4.3) with C_{pk} replaced by $\hat{C}_{pk}$ and k_3 by $\hat{k}_3$.

4. Calculate $\hat{C}_{p\lambda} = \hat{\lambda}\hat{C}_{pk}$.

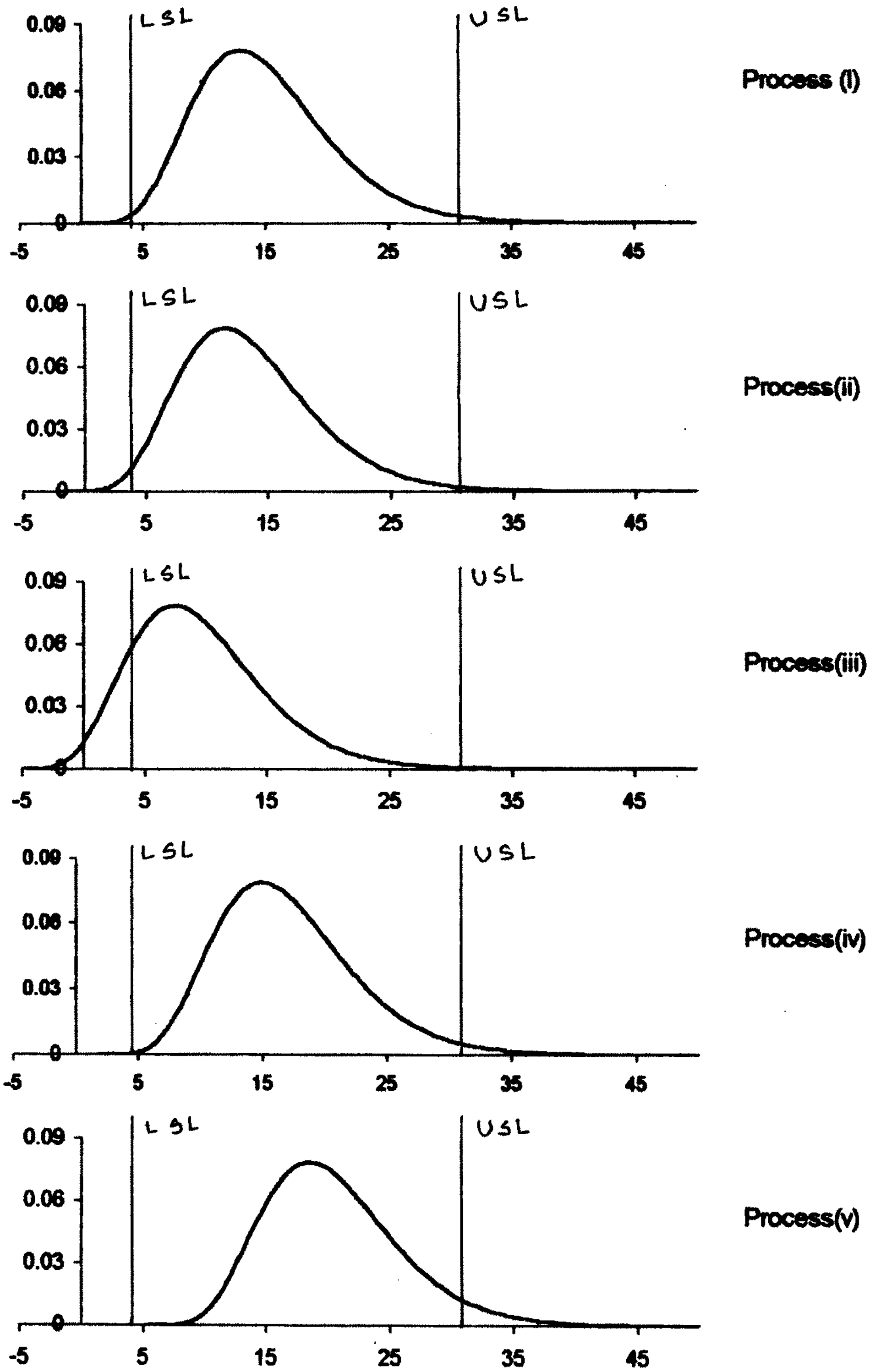
For the measurements in Example (3.3.1), we have

$$\hat{C}_{p\lambda} = 0.4179.$$

It appears that though Clements' technique is being widely used in practice, it has not been compared with any other technique in the non-normal set up. In the following an attempt is made to compare Clements' technique with Munechika's technique.

3.5 COMPARISON OF CLEMENTS' AND MUNECHIKA'S APPROACHES

Consider the 5 processes as shown in Figure 3.4. All the 5 processes are having same set of specification limits as LSL =



LSL
USL
Process (I)
Process(ii)
Process(iii)
Process(iv)
Process(v)
0.09
0.06
0.03
0
-5
5
15
25
35
45

Figure 3.4

3.6624 and USL = 30.8076. Distributions of the processes (i), (ii), (iii), (iv) and (v) are respectively χ^2_{15}, $\chi^2_{15} - 1.5$, $\chi^2_{15} - \text{s.d.}(\chi^2_{15})$, $\chi^2_{15} + 2$, $\chi^2_{15} + \text{s.d.}(\chi^2_{15})$. The Clements' C_{pk}, Munechika's $C_{p\lambda}$, actual probabilities of non-conformance and the bounds of probability of non-conformance associated with C_{pk} given by (1.3.5), for all above processes are tabulated in Table 3.3.

Table 3.3

Process	*Clements'* C_{pk}	*Munichika's* $C_{p\lambda}$	*prob. of non-conformance (p)*	*bounds of p using (1.3.5)*
i	1.0000	0.3334	0.0027	(0.00135, 0.0027)
ii	0.8595	0.3334	0.0102	(0.0050, 0.0099)
iii	0.4870	0.3333	0.1301	(0.0720, 0.144)
iv	0.9110	0.3333	0.0026	(0.00314, 0.0063)
v	0.7562	0.3355	0.0079	(0.0117, 0.0223)

From this example, it is clear that $C_{p\lambda}$ fails to take into account the location of process data while assessing the process capability. On the other hand clements' C_{pk} is a better measure of process capability. As probability of non-conformance associated with the process increases (decreases), Clements'· C_{pk} increases (decreases). However, the bounds on the probability of non-conformance associated with Clements' C_{pk}, obtained using

(1.3.5) are not true. It has been observed that if the process is located left to USL, the actual probability of non-conformance is above or in the vicinity of the upper bound given by (1.3.5) and if the process is located right to LSL, it is below or in the vicinity of the lower bound given by (1.3.5). Perhaps, this happens because the underlying peocess is positively skewed. It will be perhaps possible to correct these bounds for such skewed processes by influncing them by some measure of skewness.

In the following we give some more indices which have been reported in the literature for non-normal set up.

3.6 SOME MORE INDICES FOR NON-NORMAL PROCESSES

3.6.1. The Index C_θ

Pearn et al. (1992) proposed the index C_θ for non-normal process as

$$C_\theta = \frac{USL - LSL}{\theta\sigma} \tag{3.6.1}$$

where θ is such that the probability

$$P_\theta = P[\mu - \theta\sigma < X < \mu + \theta\sigma]$$

is as insensitive as possible to the form of the distribution of X. For $P_\theta = 0.99$ Pearn et al. (1992) recommended the value 5.15

for θ as it is quite robost over a wide range of distributions. (See Table 6 in Pearn et al. (1992)) Note that C_θ does not depend on skewness or kurtosis of the distribution of X.

C_θ can be estimated as

$$\hat{C}_\theta = \frac{USL - LSL}{\theta S} \quad . \tag{3.6.2}$$

3.6.2 The Index C_p^w

Bai and Choi (1997) have proposed the weighted variance method to adjust the capability index value to account for the degree of skewness of non-normal process data. This technique computes the standard deviations above and below the mean seperately. The PCIs based on this method differ from the original indices because the standard deviations above and below the mean are multiplied by two different factors: the upper standard deviation is multiplied by $\sqrt{2P_x}$ and the lower factor by $\sqrt{2(1 - P_x)}$, where $P_x = P(X \le \mu_x)$. C_p^w, the weight variance version of C_p, is defined as

$$C_p^w = \min\left\{ \frac{USL - LSL}{6\sigma_x\sqrt{2P_x}}, \frac{USL - LSL}{6\sigma_x\sqrt{2(1 - P_x)}} \right\} \tag{3.6.3}$$

Similarly, the weight variance versions of C_{pk} and C_{pm} are respectively,

$$\check{C}_{pk} = \min\left\{ \frac{USL - \mu_x}{3\sigma_x\sqrt{2P_x}}, \frac{\mu_x - LSL}{3\sigma_x\sqrt{2(1 - P_x)}} \right\} \qquad (3.6.4)$$

and

$$\check{C}_{pm} = \frac{USL - LSL}{6\sigma_x'} \min\left\{ \frac{1}{\sqrt{2P_T}}, \frac{1}{\sqrt{2(1 - P_T)}} \right\} \qquad (3.6.5)$$

where $P_T = P(X \le T)$ and $\sigma_x' = \sqrt{\sigma_x^2 + (\mu - T)^2}$.

Estimates of $\check{C}_p$, $\check{C}_{pk}$ and $\check{C}_{pm}$ can be obtained by replacing μ_x, σ_x, P_x and P_T in (3.6.3), (3.6.4) and (3.6.5) by their following estimates.

$$\hat{\mu} = \overline{X}$$

$$\hat{\sigma}_x = \overline{R}/d_2'$$

$$\hat{\sigma}_x' = \sqrt{S_x^2 + (\overline{X} - T)^2}$$

$$\hat{P}_x = \text{number of observations} \le \overline{X}$$

and

$$\hat{P}_T = \text{number of observations} \le T$$

where $\overline{R}$ is the mean range obtained from the $\overline{X}$ - R chart and d_2' is the constant for the skew population corresponding to d_2 for the normal equivalent.

3.7 FUTURE IDEAS

After having gone through the latest literature on Process Capability Analysis, we propose to study the following in future.

(1) Comparison of various PCI's for non-normal set up, in the context of probability of non-conformance.

(2) In the current literature only the point estimates of the PCI's for non-normal set up have been suggested. We will try to investigate properties of these estimates as well as confidence bounds for these indices.

(3) Process Capability Analysis by transforming non-normal process data to normality.