J. Soc. Korea Ind. Syst. Eng Vol. 42, No. 2 : 18-27, June 2019
https://doi.org/10.11627/jkise.2019.42.2.018
ISSN : 2005-0461(print)
ISSN : 2287-7975(online)

Analysis of Multivariate Process Capability Using Box-Cox Transformation

Hye-Jin Moon · Young-Bae Chung†

Department of Industrial and Management Engineering, Incheon National University

Box-Cox변환을 이용한 다변량 공정능력 분석

문혜진 · 정영배†

인천대학교 산업경영공학과

The process control methods based on the statistical analysis apply the analysis method or mathematical model under the assumption that the process characteristic is normally distributed. However, the distribution of data collected by the automatic measurement system in real time is often not followed by normal distribution. As the statistical analysis tools, the process capability index (PCI) has been used a lot as a measure of process capability analysis in the production site. However, PCI has been usually used without checking the normality test for the process data. Even though the normality assumption is violated, if the analysis method under the assumption of the normal distribution is performed, this will be an incorrect result and take a wrong action. When the normality assumption is violated, we can transform the non-normal data into the normal data by using an appropriate normal transformation method. There are various methods of the normal transformation. In this paper, we consider the Box-Cox transformation among them. Hence, the purpose of the study is to expand the analysis method for the multivariate process capability index using Box-Cox transformation. This study proposes the multivariate process capability index to be able to use according to both methodologies whether data is normally distributed or not. Through the computational examples, we compare and discuss the multivariate process capability index between before and after Box-Cox transformation when the process data is not normally distributed.

Keywords : Box-Cox Transformation, Non-Normal Distribution, Multivariate Process Capability Index

1. 서 론

생산 현장에서 사용되는 통계적 공정관리 분석은 대부분 공정 데이터들이 정규분포를 따른다는 가정하에 개발된 수리적 척도를 사용한다. 하지만 자동화된 측정시스템을 통해 실시간으로 집계되는 실측 데이터들의 분포를 확인해 보면 정규분포를 따르지 않는 경우도 빈번하다. 표본의 수가 충분히 많으면 중심극한정리에 따라 정규분포에 근사한다는 사실을 기반으로 통계적 분석을 수행할 수는 있겠으나 정확하고 신뢰성 있는 공정분석 결과라고 판단할 수는 없을 것이다. 따라서 다양한 통계적 분석을 적용함에 있어 공정으로부터 수집된 데이터들에 대한 정규성을 우선적으로 확인할 필요성이 있으며 그 결과 정규분포를 따르지 않는다면 이에 대한 적합한 공정분석 방법의 적용이 요구될 것이다.

특히 생산 현장에서는 통계 패키지를 이용하여 공정능력지수를 산출한다. 그러나 공정 데이터들에 대한 정규성 여부를 확인하지 않은 채 분석된 공정능력지수 결과를

Received 22 March 2019; Finally Revised 13 June 2019;
Accepted 14 June 2019
† Corresponding Author : ybchung@inu.ac.kr

토대로 공정상태를 판단하고 있다. 데이터에 대한 정규성을 만족하지 않음에도 불구하고 정규분포를 가정한 분석법에 의해 얻어진 그릇된 평가 결과는 결국 왜곡된 해석과 그로 인한 잘못된 조치를 유발할 것이다.

본 연구는 최근 연구되었던 역정규 손실함수를 이용한 다변량 공정능력지수인 MC_{pI}모형을 대상으로 연구의 적용범위를 확장하고자 한다. 즉, 분석대상의 모든 변수들이 다변량 정규분포를 따른다는 가정을 전제로 제안된 기존 연구모형 MC_{pI}의 한계를 극복할 수 있는 공정능력의 분석을 제안하고자 한다. 따라서 본 연구에서는 다변량의 공정 데이터들이 정규분포뿐 만 아니라 비정규분포에서도 적용 가능한 MC_{pI}의 공정능력 분석 프로세스를 수립할 것이다.

정규성 검정을 위한 방법은 히스토그램, Q-Q(Quantile-Quantile) plot, P-P(Probability - Probability) plot과 같은 그래프를 이용한 시각적 분석방법과 통계량을 이용한 적합도 검정이 있다. 통계량에 의한 일변량 정규성(UVN : Univariate Normality) 검정법은 왜도, 첨도, 카이제곱 검정, Kolmogorov-Smirnov 검정, Shapiro-Wilk 검정, Anderson- Darling 검정, Lilliefors 검정, Jarque-Bera 검정 등 다양한 방법들이 제안되었다. 그리고 이러한 UVN 검정법들은 다변량으로의 확장이 가능하다[5]. 대다수의 통계 소프트웨어 패키지들은 다양한 UVN 검정법의 기능들을 포함하고 있다. 따라서 본 연구에서는 사용자들이 용이하게 다변량 공정능력지수 MC_{pI} 분석을 실시할 수 있도록 통계 패키지의 UVN 검정을 이용하여 다변량으로 확장한 정규성(MVN : Multivariate Normality) 검정을 수행하는 프로세스를 제안하고자 한다.

많은 연구자들이 추천하고 있는 MVN 검정을 위한 첫 번째 단계는 모든 변수들에 대하여 개별적으로 UVN 검정을 수행하는 것이다. 이는 MVN의 필요 충분 조건으로써 각각의 변수들의 주변분포(Marginal Distribution)가 정규분포를 만족해야 한다는 것을 의미한다. 그리고 UVN 검정 결과 변수들 중 하나라도 비정규성을 나타내면 MVN을 만족할 수 없다. 또한 모든 변수들이 주변정규분포를 따른다고 해도 이것이 반드시 다변량 정규분포를 의미하지는 않기 때문에 MVN에 대한 검증 절차가 추가적으로 필요하다. 따라서 본 연구에서 MVN 검정을 수행하기 위한 프로세스의 첫번째 단계로 모든 변수들에 대하여 정규성 그래프와 UVN를 수행한다. 그 다음 단계로 왜도와 첨도를 확인한다. 그런 후에 Shapiro-Wilk검정을 다변량으로 확장한 Royston 검정과 다변량의 왜도와 첨도 접근법인 Small의 Q1, Q2를 실시한다[5].

비정규분포에서의 공정능력지수에 관한 기존 연구에는 정규변환(Normalizing Transformation)을 이용하여 비대칭분포를 정규분포로 변환한 후 전통적인 공정능력지수를 이용하는 방법, 경험분포(Empirical Distribution) 혹은 다양한 상황을 묘사할 수 있는 3 또는 4-모수 분포를 활용하여 분포의 0.135백분위수와 99.865백분위수를 추정한 후 분포의 산포를 구하는 방법, 공정능력지수가 분포의 형태에 강건(Robust)하도록 공정능력지수의 모수를 변화시키는 방법, 공정의 불량률을 추정한 후 추정된 불량률을 공정능력지수로 역산하는 방법, 분포의 산포를 분할하여 분포의 위쪽 부분과 아래쪽 부분에 서로 다른 산포를 활용하는 휴리스틱(Heuristic) 방법 등 다양한 방법들이 제안되었다[3]. 따라서 본 연구에서는 공정의 변수들이 비정규분포를 따른다면 정규변환을 이용하여 정규분포에 근사하도록 변환시킨 후 MVN을 만족하는 조건하에서 다변량 공정능력지수 MC_{pI}를 이용하여 공정능력을 평가하고자 한다. 그리고 MVN을 만족시키기 위해 정규변환을 수행한 경우와 그렇지 않은 경우에서의 MC_{pI} 결과값을 비교하여 공정상태에 관한 평가에 미치는 영향을 비교하고자 한다.

2. 이론적 배경

2.1 기호

$\mathbf{X}$	: 원본 데이터 벡터($\mathbf{x}' = [x_i] = [x_1 x_2 \cdots x_n]$)
$\mathbf{Y}$	: 변환 데이터 벡터($\mathbf{y}' = [y_i] = [y_1 y_2 \cdots y_n]$)
$\boldsymbol{\mu}$	: 공정 평균 벡터($\boldsymbol{\mu}' = [\mu_i] = [\mu_1 \mu_2 \cdots \mu_n]$)
$\mathbf{T}$	: 목표치 벡터($\mathbf{T}' = [T_i] = [T_1 T_2 \cdots T_n]$)
n	: 품질특성치(변수)의 수
m	: 변수 i에 대한 측정 데이터 수
i	: 품질특성치(i = 1, 2, …, n)
λ_i	: MLE 방법에 의한 Box-Cox 변환 파라미터
A_j	: 목표치 T_i 중심으로 특성치 x_i의 비대칭 구간에서 최대손실(j = 1, 2, …, 2^n)
Σ	: 분산-공분산 매트릭스
Σ^{-1}	: 분산-공분산 매트릭스 역행렬
Σ_T	: 평균제곱오차(MES) 매트릭스
Λ	: 척도모수(scaling parameter) 매트릭스
$K(n)$	: $\chi^2(n, 0.9973)$의 값
$\Gamma(\alpha)$	: 감마함수
$L_I(\mathbf{X},\mathbf{T})$	: 목표치 T에 대한 표본벡터 $\mathbf{X}_i$의 역정규 손실함수
$E[L_I(\mathbf{X},\mathbf{T})]$	: $L_I(\mathbf{X},\mathbf{T})$의 기대손실

2.2 정규성 검정 및 Box-Cox 변환

UVN 검정 가운데 모든 분포에 대하여 검정력이 가장

우수한 검정법은 Shapiro-Wilk 검정이지만 표본의 수가 50 이상일 경우에는 검정력이 떨어지는 한계점을 가지고 있다. 따라서 본 연구에서는 표본의 수에 대한 제약을 받지 않으면서 검정력이 우수한 Anderson-Darling(AD) 검정을 채택하고자 한다. Anderson, Darling에 의해 제안된 AD 검정은 적합도 검정을 기반으로 분석하며, AD 검정 통계량 A^2는 다음과 같다[1].

$$A^2 = -n - \frac{1}{n}\sum_{i=1}^{n}\{(2i-1)\log(F(x_i) + \log(1-F(x_{n+1-i})\} \quad (1)$$

여기서 $F(x_i)$는 순서통계량 x_i의 누적분포함수를 의미한다. MVN 검정을 위해 본 연구에서 채택하고자 하는 검정법은 UVN의 Shapiro-Wilk 검정을 n ≤ 2000인 표본에서 사용 가능하도록 다변량으로 확장한 Royston검정이며 검정 통계량 H는 다음과 같다[8].

$$H = eG \quad (2)$$

여기서 $G = \sum_{j=p}^{p} K_j$이고 $e = p/[1+(p-1)\bar{c}]$이다.
그리고 $K_j = (F^{-1}[F(-Z_j)/2])^2$, $\bar{c} = \sum_{i,j} c_{ij}/(p^2-p)$, $c_{ij} = corr(K_i, K_j)$이다.

Royston 검정과 더불어 MVN 검정을 위해 다변량의 왜도와 첨도도 함께 확인하고자 한다. Small은 변환된 개별 변수의 왜도와 첨도의 이차 조합을 기반으로 다변량에 대한 왜도(Q_1)와 첨도(Q_2)를 다음과 같이 제안하였다[9].

$$Q_1 = \mathrm{y}_1'\mathrm{U}_1^{-1}\mathrm{y}_1, \quad Q_2 = \mathrm{y}_2'\mathrm{U}_2^{-1}\mathrm{y}_2, \quad (3)$$

여기서 y_1과 y_2는 각각 표본의 왜도와 첨도 벡터를 Johnson 변환에 의해 변환된 벡터이고, U_1과 U_2는 벡터 y_1과 y_2의 상관행렬을 의미한다. 그리고 $Q_3 = Q_1 + Q_2$를 이용하여 다변량 정규성의 옴니버스 검정(Omnibus Test)을 실시할 수 있다.

UVN 결과 비정규분포를 따르는 데이터들은 Box-Cox의 정규변환을 이용하여 분석하고자 한다. Box, Cox에 의해 소개된 'Box-Cox 변환(Box-Cox Transformation)'은 멱변환(Power Transformation) 방법의 일종으로 랜덤 변수가 정규분포에 근사하도록 변환하는 방법이다.

Box-Cox 변환 함수의 식은 다음의 식과 같다[2, 4].

$$\mathrm{Y} = \mathrm{X}^{(\lambda)} = \begin{cases} \dfrac{\mathrm{X}^{\lambda}-1}{\lambda}, & (\lambda \neq 0) \\ \log(\mathrm{X}), & (\lambda \neq 0) \end{cases} \quad (4)$$

여기서 변환 전의 원본 데이터 집합은 X이고, 변환 후의 데이터 집합은 Y이다. 그리고 상기 Box-Cox 변환 식(4)는 X가 양수(X > 0)일 때 대해서만 성립한다. 하지만 X가 음수(X < 0)일 경우에는 일정 상수 값을 추가적으로 더하여 양수로 변환한 후에 분석할 수 있다. Box-Cox 변환 과정에서 중요한 파라미터는 λ(Lambda) 값의 결정이다. 원본 데이터를 최대한 정규성에 근사한 데이터로 변환하기 위해서는 λ의 가장 적절한 값을 선택하는 것이 중요하다. 따라서 Box-Cox 변환 방법에서는 최대 우도 추정(MLE : Maximum Likelihood Estimation) 방법에 의하여 최적의 λ (Optimal Lambda)값을 결정한다. 최적의 λ를 구하기 위해서는 우선 초기값으로 λ의 범위를 설정한다. 일반적인 통계프로그램에서는 [-5, 5] 또는 [-2, 2]의 범위에서 설정할 수 있다. 그런 다음 최대 우도 측정법(MLE)에 의한 $L_{\max}$ 값을 다음의 식 (5)와 같이 구한다[2, 4].

$$L_{\max}(\lambda) = -\frac{n}{2}\log\hat{\sigma}_{\lambda}^2 + (\lambda-1)\sum_{i=1}^{m}\log(y_i) \quad (5)$$

여기서 $\hat{\sigma}_{\lambda}^2$값은 다음의 식 (6)과 식 (7)와 같이 구한다.

$$\hat{\sigma}_{\lambda}^2 = \frac{S(\lambda)}{n} \quad (6)$$

$$S(\lambda) = \mathrm{I} - \mathrm{X}(\mathrm{X}'\mathrm{X})^{-1}\mathrm{X}' \quad (7)$$

2.3 다변량 공정능력 지수 MC_{pI}

Moon, Chung에 의해 제안된 다변량 공정능력지수 MC_{pI}는 Tamm, Subbaiah, Liddy에 의해 제안된 MC_{pm}모형을 응용한 다변량 공정능력지수이다[3, 4]. 즉, 공정영역의 산포(Σ_T) 대신에 목표치로부터 공정평균이 멀어짐에 따른 손실함수의 기대손실을 적용하였다. 그리고 MC_{pI} 모형에 적용된 손실함수는 역정규 손실함수(Inverted Normal Loss Function)로써 Spiring에 의해 제안되었다. 이 손실함수는 품질변동에 따른 손실을 정규분포의 p.d.f의 역함수를 이용함으로써 손실에 대하여 보다 합리적으로 설명한다는 장점과 목표치 T_i를 중심으로 최대손실 A_j에 대하여 손실함수의 형태가 대칭과 비대칭 모두일 때 제약없이 적용이 가능하다는 장점을 가진 모형이다[6].

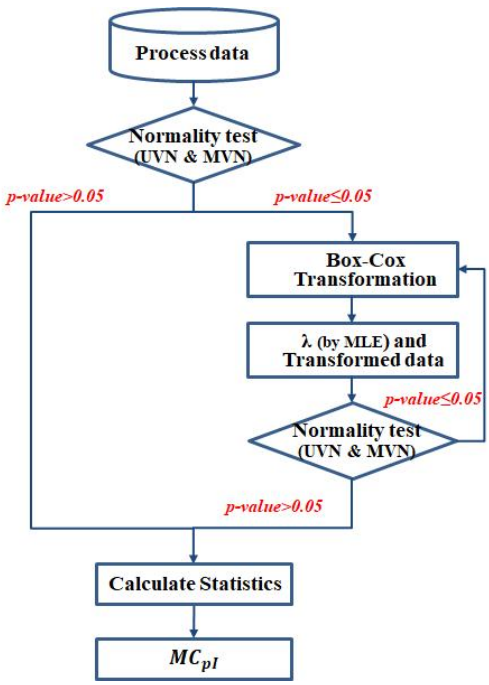
$$MC_{pI} = \frac{\text{수정규격허용오차영역}(R1)\text{의면적또는부피}}{\text{공정영역의면적또는부피}} \quad (8)$$

$$= \frac{\text{수정규격허용오차영역}(R1)\text{의면적또는부피}}{|E[L_I(\mathbf{X},\mathbf{T})]|^{\frac{1}{2}}(\pi K)^{\frac{n}{2}}[\Gamma(\frac{n}{2}+1)]^{-1}}$$

3. Box-Cox를 이용한 다변량 공정능력분석

기존 연구에서 제안되었던 다변량 공정능력지수 MC_{pI}는 모든 변수들이 다변량 정규분포를 따를 때 적합한 척도 모형이다. 그러기에 본 연구에서는 이러한 한계점을 벗어나 비정규분포일 경우에도 MC_{pI}를 이용하여 공정능력을 분석할 수 있도록 확장하는 방법을 제안한다. 즉, 공정으로부터 수집된 원본 데이터에 대하여 정규성을 검정하고 그 결과가 정규분포를 따르지 않는다고 판정될 때 이에 대한 적절한 조치를 취함으로써 올바른 공정능력 평가를 수행하고자 한다.

<Figure 1>의 Flow Chart에서 보듯이 다변량 공정능력 분석과정에서 가장 먼저 선행하고자 하는 분석이 정규성 검정(Normality Test)이다. 우선적으로 개별 변수들에 대하여 그래프(히스토그램, Q-Q plot, P-P plot)와 UVN 검정들(AD 검정, 왜도, 첨도)을 실시한다. 매트랩(MATLAB) 프로그램을 이용하여 Anderson-Darling 검정을 수행하게 되면 h값과 p-value의 결과에 따라 정규분포에 대한 가설검정을 판정한다. h값(Hypothesis Test Result)의 판정기준은 h = 1이면 H0(귀무가설 : 정규분포를 따른다)를 기각하고 h = 0이면 H0를 채택한다. 또한 p-value ≤ 0.05이면 H0를 기각하고 p-value > 0.05이면 H0를 채택한다. 또한 각각의 변수들에 대한 왜도와 첨도를 분석하여 AD 검정 결과와 함께 검토한다. 이들 중 하나라도 정규성을 벗어



<Figure 1> Flow Chart for Evaluating Multivariate Process Capability Using Box-Cox Transformation Under Non-Normal Distribution

난다면 Box-Cox변환 방법에 의해 원본 데이터를 정규분포에 근사하도록 변환시킨다. 만약 UVN 검정이 모두 만족되더라도 Royston 검정과 다변량 왜도와 첨도를 분석하여 다변량 정규분포에 대한 검증을 실시한다.

UVN 검정 결과 비정규분포를 따르는 변수들은 Box-Cox 변환을 수행한다. 이때 MLE 방법에 의해 각각의 변수들에 대한 최적의 λ_i을 구한다. 이때 λ_i의 값에 대한 95% 신뢰구간과 최적의 λ_i값을 나타내는 그래프로 같이 도출하여 확인한다. 그리고 선택된 최적의 λ_i값을 가지고 Box-Cox 변환 함수에 의해 변환된 데이터 집합을 구한다. 이때 원본 데이터에 대한 규격의 상한과 하한, 목표치에 대한 수치들도 최적의 λ_i에 의해 동일하게 변환시킨다. 정규 변환된 데이터에 대해서는 변환 전과 같은 UVN의 AD 검정법에 의해 정규성 검정을 실시하고 히스토그램, Q-Q plot, P-P plot 을 통해 변환된 데이터들이 정규분포에 근사하도록 바뀌었는지 확인한다. 그리고 MVN 검증을 위해 Royston 검정과 다변량 왜도와 첨도를 분석한다.

변환된 데이터가 다변량 정규분포를 만족하면 해당 통계량을 산출하여 다변량 공정능력지수 MC_{pI}에 대입하기 위한 입력값들을 구한다.

식 (8)에서 분자식의 '수정 규격허용오차영역의 면적 또는 부피 [Vol.($R1$)]'를 구한다. 품질특성치(변수)의 수(n)에 대한 면적 또는 부피를 구하는 일반적인 식은 다음과 같다.

$$\mathrm{Vol.}(R1) = \frac{2\prod_{i=1}^{n} a_i}{n} \times \frac{\pi^{n/2}}{\Gamma(\frac{n}{2})} \tag{9}$$

여기서 a_i는 각 품질특성(변수)별 '규격범위(USL_i-LSL_i)의 1/2'이다.

그리고 식 (8)에서 분모식의 최대손실 A_j 전개방식에 따른 역정규 손실함수 $L_I(\mathbf{X}, \mathbf{T})$와 기대손실 $E[L_I(\mathbf{X}, \mathbf{T})]$는 다음과 같다.

$$L_I(\mathbf{X},\mathbf{T}) = \begin{cases} A_1\left[1-\exp\left\{-\frac{1}{2}(\mathbf{X}-\mathbf{T})^{\mathrm{T}}\Lambda^{-1}(\mathbf{X}-\mathbf{T})\right\}\right], \\ \qquad x_1 < T_1, x_2 < T_2, \cdots, x_n < T_n \\ A_2\left[1-\exp\left\{-\frac{1}{2}(\mathbf{X}-\mathbf{T})^{T}\Lambda^{-1}(\mathbf{X}-\mathbf{T})\right\}\right], \\ \qquad x_1 < T_1, x_2 < T_2, \cdots, x_n \geq T_n \\ \qquad\qquad \vdots \\ A_{2^n-1}\left[1-\exp\left\{-\frac{1}{2}(\mathbf{X}-\mathbf{T})^{T}\Lambda^{-1}(\mathbf{X}-\mathbf{T})\right\}\right], \\ \qquad x_1 \geq T_1, x_2 \geq T_2, \cdots, x_n < T_n \\ A_{2^n}\left[1-\exp\left\{-\frac{1}{2}(\mathbf{X}-\mathbf{T})^{T}\Lambda^{-1}(\mathbf{X}-\mathbf{T})\right\}\right], \\ \qquad x_1 \geq T_1, x_2 \geq T_2, \cdots, x_n \geq T_n \end{cases} \tag{10}$$

$$E[L_I(\mathbf{X},\mathbf{T})] = \tag{11}$$

$$\int_{-\infty}^{T_1}\cdots\int_{-\infty}^{T_n} L_I(\mathbf{X},\mathbf{T})f(\mathbf{X})dX + \int_{-\infty}^{T_1}\cdots\int_{T_n}^{\infty} L_I(\mathbf{X},\mathbf{T})f(\mathbf{X})dX$$

$$\cdots + \int_{T_1}^{\infty}\cdots\int_{-\infty}^{T_n} L_I(\mathbf{X},\mathbf{T})f(\mathbf{X})dX + \int_{T_1}^{\infty}\cdots\int_{T_n}^{\infty} L_I(\mathbf{X},\mathbf{T})f(\mathbf{X})dX$$

따라서 분자의 입력값인 규격영역에 대한 면적 또는 부피인 Vol.($R1$) 값과 분모의 역정규 손실함수의 기대손실 $E[L_I(\mathbf{X},\mathbf{T})]$값을 구하고 나면 다음의 식에 대입하여 MC_{pI}를 구한다.

$$MC_{pI} = \frac{\mathrm{Vol.}(R1)}{|E[L_I(\mathbf{X},\mathbf{T})]|^{\frac{1}{2}}(\pi K)^{\frac{n}{2}}[\Gamma(\frac{n}{2}+1)]^{-1}}$$

4. 수치 예

본 연구에 이용된 시뮬레이션은 다음의 공통적 가정과 범위를 전제로 한다.

(1) 품질특성치는 망목특성이며, 연속적인 값을 가진다.
(2) 척도 매트릭스(Scaling matrix)에서 Λ는 공정 파라미터가 규격한계선에 있을 때 손실이 90%일 때를 기준으로 한다($\gamma_i = 0.233 \times$ 규격공차$_i$).
(3) 구간별 최대손실비용 A_j은 기업으로부터 정해진 상수값이다.
(4) 시뮬레이션 분석에 사용한 프로그램은 매트랩 (MATLAB)이다.

수치 예는 3변량일 때 다음의 규격 조건을 만족해야 하는 경우로 가정한다.

$$USL_1 = 4.0,\ LSL_1 = 2.0,\ T_1 = 3.0$$
$$USL_2 = 3.5,\ LSL_2 = 1.5,\ T_2 = 2.5$$
$$USL_3 = 3.0,\ LSL_3 = 1.0,\ T_3 = 2.0$$

3변량에 대한 다변량 공정능력 분석 프로세스는 <Figure 1>의 Flow Chart에 따라 실시하였다. 그리고 다변량 정규분포를 만족시키기 위해 정규변환을 수행한 경우와 그렇지 않은 경우의 MC_{pI} 결과값을 비교하고자 한다.

4.1 원본 데이터와 통계량(Original Data and Statistics)

3변량에 대한 원본 데이터들(Original Data Set)은 비정규분포 하에서 각각 60개의 난수(Random Number)를 생성하였으며 각 변수에 대한 통계량은 다음과 같다.

<Table 1> Original Data and Statistics

No.	X_1	X_2	X_3
1	2.7434	2.1746	1.7010
2	2.7507	2.1819	1.7083
3	3.3437	2.7749	2.3013
4	2.7633	2.1945	1.7209
5	3.0059	2.4371	1.9635
6	2.7682	2.1994	1.7258
7	2.8974	2.3286	1.8550
8	2.4151	1.8463	1.3727
9	2.2865	1.7177	1.2441
10	2.3941	1.8253	1.3517
11	2.1573	1.5885	1.1149
12	2.1987	1.6299	1.1563
13	2.9225	2.3537	1.8801
14	3.2959	2.7271	2.2535
15	2.7421	2.1733	1.6997
16	2.9157	2.3469	1.8733
17	2.8890	2.3202	1.8466
18	3.2944	2.7256	2.2520
19	2.9750	2.4062	1.4394
20	2.7738	2.2050	1.2382
21	2.9406	2.3718	1.4050
22	3.1457	2.5769	1.6101
23	3.0658	2.4970	1.5302
24	3.0795	2.5107	1.5439
25	2.8618	2.2930	1.3262
26	2.9216	2.3528	1.3860
27	2.7201	2.1513	1.1845
28	3.0388	2.4700	1.5032
29	2.8407	2.2719	1.3051
30	2.6523	2.0835	1.1277
31	3.1389	2.5701	1.6033
32	3.2290	2.6602	1.6934
33	2.7810	2.2122	1.2454
34	3.2130	2.6442	1.6774
35	2.8761	2.3073	1.3405
36	3.0620	2.4932	1.5264
37	2.9182	2.3494	1.3826
38	2.9907	2.4219	1.4551
39	2.9091	2.4403	1.3735
40	3.3164	2.8476	1.7808
41	2.8237	2.3549	1.2881
42	2.7221	2.2533	1.1865
43	2.6633	2.1945	1.1277
44	2.9713	2.5025	1.4357
45	2.9329	2.4641	1.3973
46	3.3759	2.9071	1.8403
47	3.0491	2.5803	1.5135
48	2.8861	2.4173	1.3505
49	2.1597	1.6909	1.2220
50	2.7230	2.2542	1.7853
51	2.2980	1.8292	1.3603
52	2.5529	2.0841	1.6152
53	2.1783	1.7095	1.2406
54	2.3207	1.8519	1.3830
55	2.6194	2.1506	1.6817
56	2.2839	1.8151	1.3462
57	2.4912	2.0224	1.5535
58	2.2196	1.7508	1.2819
59	2.1314	1.6626	1.1937
60	2.3121	1.8433	1.3744
mean	2.782477	2.250343	1.514600
std	0.338272	0.325099	0.282890
var	0.114428	0.105690	0.080027

4.2 원본 데이터에 대한 정규성 검정(Normality Test for the Original Data)

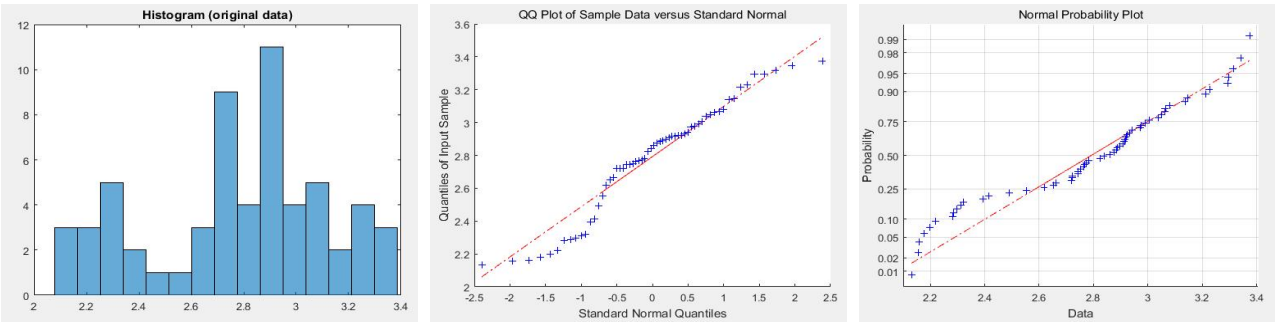
모든 변수에 대하여 원본 데이터의 UVN의 Anderson-Darling(AD) 검정, 왜도, 첨도 분석을 실시한 결과는 다음의 <Table 2>와 같다. 또한 동시에 원본 데이터에 대한 히스토그램을 통해 분포의 형태를 확인하고 Q-Q plot, P-P plot을 통해 UVN 검정을 실시한 결과는 <Figure 2>, <Figure 3>, <Figure 4>와 같다. 왜도와 첨도는 비정규성이 약해 보이긴 하지만 AD 검정 결과는 X_1에 대한 p-value = 0.0096로 H0를 기각, X_2의 p-value = 0.0366로 역시 H0를 기각, 그리고 X_3 역시 p-value = 0.0064로 H0를 기각한다. 또한 MVN 검정을 실시한 결과 <Table 3>에서와 같이 Q_1과 Royston 검정에서 p-value < 0.05로 일변량과 다변량 모두에 대하여 정규분포를 따른다고 할 수 없다.

<Table 2> UVN Tests for Original Data(α = 5%)

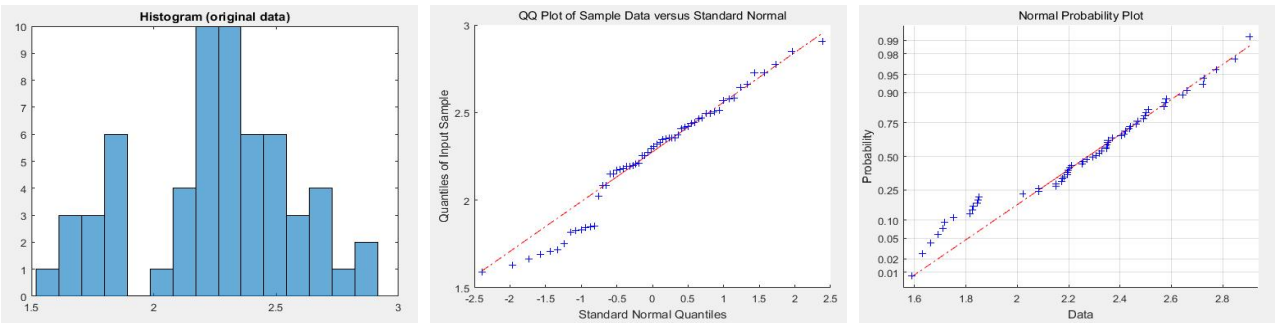
Var.	UVN Test(AD) p-value	Skewness	Kurtosis
X_1	0.0096	-0.3712	2.2998
X_2	0.0366	-0.2926	2.4272
X_3	0.0064	0.8787	3.4305

<Table 3> MVN Tests for Original Data(α = 5%)

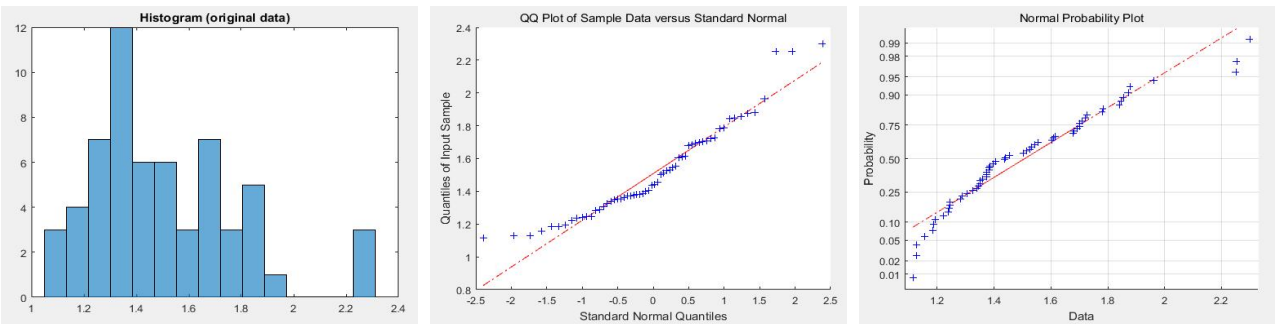
MVN Test	Test Statistics	p-value
Small's skewness(Q_1)	12.5211	0.0058
Small's kurtosis(Q_2)	5.13746	0.1620
Royston	14.6734	0.00099



<Figure 2> Graphs for UVN of Original Data X_1



<Figure 3> Graphs for UVN of Original Data X_2



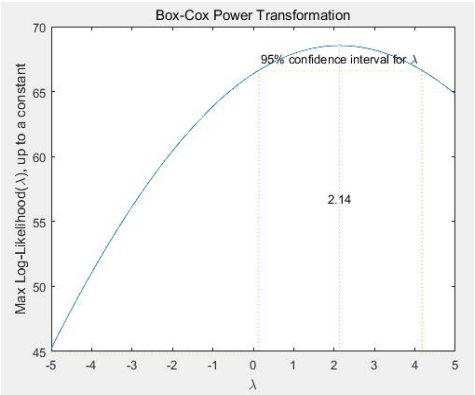
<Figure 4> Graphs for UVN of Original Data X_3

4.3 Box-Cox 변환 및 변환된 데이터(Box-Cox Transformation and Transformed Data)

정규분포를 따르지 않는 원본 데이터들에 대하여 Box-Cox 변환에 의한 데이터를 얻기 위해서는 가장 먼저 MLE (Maximum Likelihood Estimation) 방법에 의한 최적의 λ (Optimal Lambda) 값을 구하는 것이 필요하다. λ값의 범위는 [-5, 5]으로 설정하였고, 초기값 -5에서부터 0.01씩 증가시키며 구해진 $L(\lambda_i)$의 값에 대한 95% 신뢰구간의 λ를 나타내는 그래프와 MLE 방법에 의한 최대값 $[L_{\max}(\lambda)]$을 가지는 최적의 λ_i결과값은 다음과 같다. 최적의 λ 값이 결정되면 3변량에 대한 변환된 데이터들(Transformed Data Set)을 <Table 4>와 같이 얻을 수 있다. 그리고 변환된 데이터들의 통계량을 구하면 다음과 같다.

4.3.1 X_1의 최적의 λ_1과 95% 신뢰구간

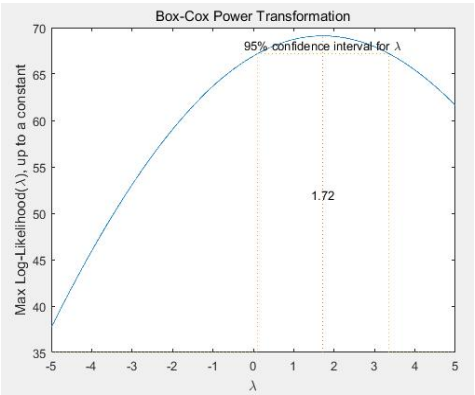
• Optimal Lambda (λ_1) = 2.14



<Figure 5> Optimal Lambda (λ_1) and Lambda in 95% Confidence Interval

4.3.2 X2의 최적의 λ_2과 95% 신뢰구간

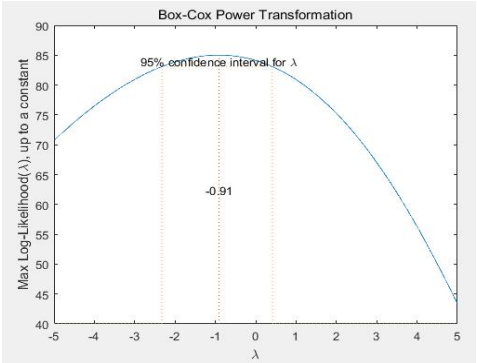
• Optimal Lambda (λ_2) = 1.72



<Figure 6> Optimal Lambda (λ_2) and Lambda in 95% Confidence Interval

4.3.3 X2의 최적의 λ_3과 95% 신뢰구간

• Optimal Lambda (λ_3) = -0.91



<Figure 7> Optimal Lambda (λ_3) and Lambda in 95% Confidence Interval

<Table 4> Transformed Data Via Box-Cox

No.	Y_1	Y_2	Y_3
1	3.5834	1.6305	0.4212
2	3.6065	1.6433	0.4239
3	5.7190	2.7826	0.5842
4	3.6465	1.6654	0.4284
5	4.4582	2.1095	0.5042
6	3.6621	1.6741	0.4301
7	4.0856	1.9067	0.4726
8	2.6164	1.0878	0.2752
9	2.2756	0.8929	0.1981
10	2.5593	1.0553	0.2636
11	1.9546	0.7074	0.1036
12	2.0552	0.7657	0.1360
13	4.1704	1.9531	0.4802
14	5.5313	2.6836	0.5743
15	3.5793	1.6282	0.4208
16	4.1473	1.9405	0.4782
17	4.0574	1.8913	0.4700
18	5.5255	2.6805	0.5739
19	4.3505	2.0511	0.3100
20	3.6800	1.6839	0.1942
21	4.2321	1.9867	0.2925
22	4.9615	2.3804	0.3865
23	4.6707	2.2242	0.3527
24	4.7199	2.2508	0.3588
25	3.9667	1.8417	0.2490
26	4.1673	1.9514	0.2824
27	3.5101	1.5899	0.1569
28	4.5743	2.1723	0.3405
29	3.8970	1.8035	0.2365
30	3.3010	1.4735	0.1138
31	4.9364	2.3670	0.3838
32	5.2738	2.5470	0.4185
33	3.7031	1.6967	0.1989
34	5.2131	2.5147	0.4126
35	4.0142	1.8677	0.2572
36	4.6570	2.2169	0.3510
37	4.1558	1.9451	0.2806
38	4.4051	2.0807	0.3178
39	4.1250	2.1156	0.2756
40	5.6114	2.9356	0.4489
41	3.8413	1.9553	0.2261
42	3.5164	1.7700	0.1584
43	3.3345	1.6654	0.1138
44	4.3377	2.2349	0.3082
45	4.2058	2.1610	0.2884
46	5.8472	3.0630	0.4681
47	4.6110	2.3872	0.3452
48	4.0477	2.0720	0.2629
49	1.9604	0.8535	0.1833
50	3.5192	1.7716	0.4504
51	2.3052	1.0613	0.2684
52	3.0052	1.4745	0.3886
53	2.0053	0.8808	0.1958
54	2.3642	1.0965	0.2808
55	3.2016	1.5887	0.4142
56	2.2690	1.0396	0.2605
57	2.8281	1.3710	0.3629
58	2.1067	0.9421	0.2223
59	1.8928	0.8125	0.1635
60	2.3418	1.0832	0.2761
mean	3.781673	1.794642	0.324916
std	1.062535	0.574216	0.121226
var	1.128982	0.329724	0.014696

그리고 각각의 변수에 대한 최적의 λ_i을 공정의 규격치에도 대입하여 Box-Cox 변환을 실시한 결과 다음과 같다.

$$USL_1' = 8.6108,\ LSL_1' = 1.5923,\ T_1' = 4.4376$$
$$USL_2' = 4.4336,\ LSL_2' = 0.5864,\ T_2' = 2.2300$$
$$USL_3' = 0.6945,\ LSL_3' = 0.0000,\ T_3' = 0.5141$$

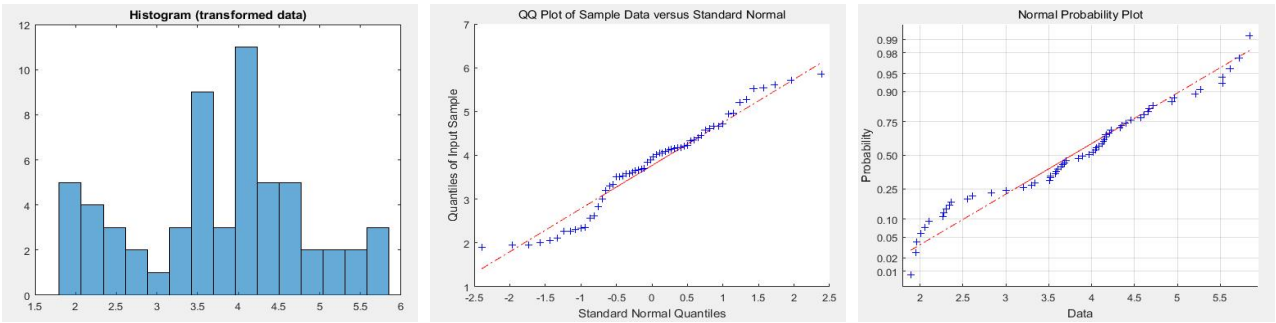
4.4 변환된 데이터에 대한 정규성 검정(Normality Test for the Transformed Data)

Box-Cox 변환에 의해 변환된 데이터들의 UVN 검정을 위해 AD검정, 왜도, 첨도 분석을 실시하고 히스토그램, Q-Q plot, P-P plot을 실시한 결과 <Table 5>, <Figure 8>, <Figure 9>, <Figure 10>과 같다.

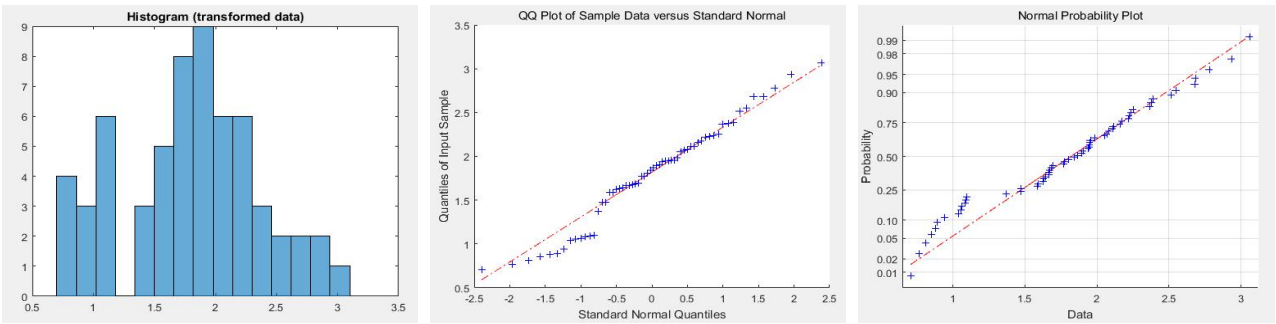
변환 후의 3변량에 대한 AD 검정 결과를 보면, Y_1에 대한 p-value = 0.0727로 H0를 채택, Y_2에 대한 p-value = 0.1465로 H0를 채택, 그리고 Y_3 역시 p-value = 0.3794로 H0를 채택한다. 따라서 3변량 모두에 대한 AD 검정, 왜도, 첨도 모두 UVN을 만족하고 있다. 그리고 MVN 검정의 추가 분석 결과 <Table 6>과 같이 모두 p-value > 0.05로 다변량 정규분포를 따른다고 할 수 있다.

<Table 5> UVN Tests for Transformed Data(α = 5%)

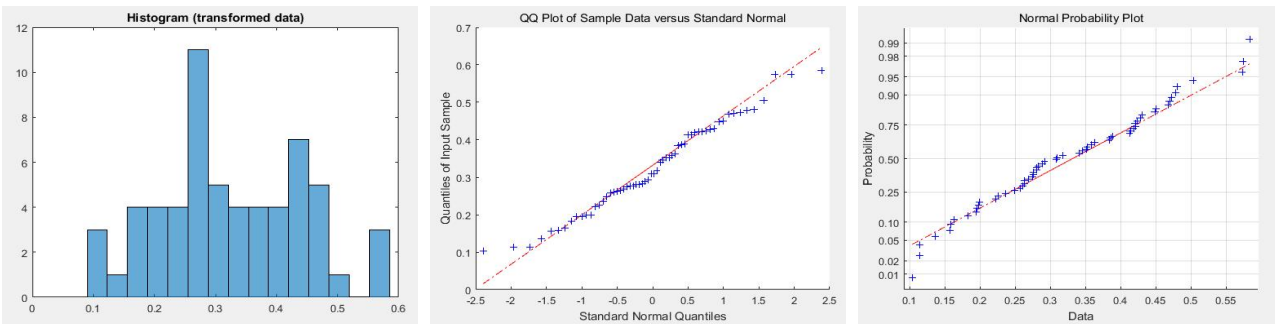
Var.	UVN Test(AD) p-value	Skewness	Kurtosis
Y_1	0.0727	-0.1210	2.2874
Y_2	0.1465	-0.0776	2.4537
Y_3	0.3794	-0.1564	2.2948



<Figure 8> Graphs for UVN of Box-Cox Transformation Data Y_1



<Figure 9> Graphs for UVN of Box-Cox Transformation Data Y_2



<Figure 10> Graphs for UVN of Box-Cox Transformation Data Y_3

<Table 6> MVN Tests for Transformed Data(α = 5%)

MVN Test	Test Statistics	p-value
Small's skewness(Q_1)	0.99756	0.80184
Small's kurtosis(Q_2)	6.00674	0.11128
Royston	5.05459	0.54066

4.5 다변량 공정능력지수 MC_{pI}의 비교(Comparison MC_{pI} between Before and After Box-Cox)

4.5.1 상관계수 변화에 따른 공정능력 비교

3변량($\mathbf{X}' = [x_1\ x_2\ x_3]$)에 대하여 상관계수 ρ가 0, 0.2, 0.4, 0.6, 0.8과 같이 변화할 때 Box-Cox 변환 전과 변환 후의 다변량 공정능력지수 MC_{pI} 결과를 비교하면 다음과 같다. 여기서 상관계수 ρ는 두 쌍의 변수간 상관계수를 의미하며 모든 쌍간의 상관계수들이 모두 동일하다는 조건을 전제로 한다. 그리고 각 변수의 λ_i는 각 변수별 최적의 λ값일 때 변환된 데이터를 기준으로 하며 최대손실 $\mathbf{A}' = [0.01\ 0.01\ 0.01\ 0.01\ 0.01\ 0.01\ 0.01\ 0.01]$이다.

상관계수 ρ의 변화에 따라 분산-공분산 행렬(Σ)값이 변하게 되고 이는 정규확률밀도함수 $f(\mathbf{X})$에 영향을 준다. 그로 인하여 역정규 손실함수의 기대값이 변하기 때문에 결국 다변량 공정능력지수 MC_{pI}의 결과에 대한 변화를 확인하고자 한다.

<Table 7> Comparison MC_{pI} between Before and After Box-Cox Transformation According to Correlation (λ = Optimal Lambda)

		Before Box-Cox (Original Data)		After Box-Cox (Transformed Data)	
No.	ρ	$E[L_I(X,T)]$	MC_{pI}	$E[L_I(Y,T)]$	MC_{pI}
1	0.0	0.00681	0.22849	0.00707	0.52312
2	0.2	0.00659	0.23119	0.00688	0.53053
3	0.4	0.00633	0.23595	0.00664	0.53978
4	0.6	0.00600	0.24237	0.00636	0.55168
5	0.8	0.00556	0.25163	0.00602	0.56711

4.5.2 손실비용 변화에 따른 공정능력 비교

상관계수 ρ = 0.8이고 각각 변수에 대한 λ_i가 최적의 λ(Optimal Lambda) 값일 때를 기준으로 3변량의 최대손실($) $\mathbf{A}' = [A_1\ A_2\ A_3\ A_4\ A_5\ A_6\ A_7\ A_8$ 값이 동일한 대칭일 때를 고려하였다. 최대손실 A($) = 1, 0.5, 0.1, 0.05, 0.01 일 각각의 경우에 따른 Box-Cox 변환 전과 변환 후의 다변량 공정능력지수 MC_{pI} 결과를 비교하면 다음과 같다.

<Table 8> Comparison MC_{pI} between Before and After Box-Cox Transformation According to Maximum Loss (ρ = 0.8, λ = Optimal Lambda)

		Before Box-Cox (Original Data)		After Box-Cox (Transformed Data)	
No.	A($)	$E[L_I(X,T)]$	MC_{pI}	$E[L_I(Y,T)]$	MC_{pI}
1	1.00	0.55642	0.02516	0.60191	0.05671
2	0.50	0.27821	0.03559	0.30096	0.08022
3	0.10	0.05564	0.07958	0.06019	0.17934
4	0.05	0.02782	0.11253	0.03010	0.25362
5	0.01	0.00556	0.25163	0.00602	0.56711

<Table 7>과 <Table 8>의 결과를 비교해 보면 Box-Cox 변환한 후의 공정능력 결과값이 Box-Cox 변환 전과 비교하여 약 2배 정도의 차이를 보이며 높게 평가되었다. 다변량 공정능력지수 MC_{pI}는 다변량 정규분포를 가정하여 개발된 모형이다. 정규분포를 따르지 않는 원본 데이터를 정규분포에 근사하도록 변환 후 MC_{pI}의 평가모형에 적합한 공정능력을 구할 수 있다. 따라서 공정 분석에 사용되는 평가척도의 적합성을 만족하지 않은 채 평가된 결과는 공정상태에 대한 왜곡된 해석과 그로 인한 잘못된 시정조치를 초래할 수 있을 것이다.

5. 결 론

기존 연구의 다변량 공정능력지수 MC_{pI} 모형은 품질특성치들이 다변량 정규분포를 따른다는 가정하에서 제안된 척도이다. 하지만 본 연구에서는 공정 데이터들이 비정규분포를 따를 경우도 포함하여 평가할 수 있는 MC_{pI}의 분석 방법을 확장하여 제안하였다. 그리고 MC_{pI} 공정능력 분석을 위한 가장 선행하는 분석은 개별 변수들의 UVN 검정과 MVN 검정을 실시하여 그 결과에 따라 분석 방법을 이원화하여 수행하도록 제안하였다. 정규성 검정 결과 정규분포를 따르지 않는다면 다양한 정규변환 방법 가운데 Box-Cox 변환에 의해 비정규분포의 데이터를 정규분포에 근사하도록 변환 방법을 공정능력 분석 프로세스에 추가하여 확장하였다.

수치 예에서 3변량의 원본 공정 데이터가 일변량과 다변량 모두 정규분포를 따르지 않음을 확인할 수 있었다. 사용자가 MC_{pI}를 이용하여 다변량 공정능력 분석을 수행할 때 정규성 여부를 무시한 채 그대로 공정능력을 분석한 결과와 Box-Cox 변환을 수행하여 정규분포에 근사하도록 데이터를 변환한 후 공정능력을 분석한 결과를 비교해 보았다. 그리고 상관계수와 손실비용의 변화에

따른 MC_{pI}의 결과값 추이도 Box-Cox 변환 전과 후를 비교하였다. 우선 상관계수와 손실비용의 변화에 따른 양쪽 모두의 MC_{pI} 결과값은 Box-Cox 변환 전보다는 변환 후의 MC_{pI} 값이 높게 나타났다. 그리고 상관계수의 변화에 따른 Box-Cox 변환 전과 변환 후의 MC_{pI} 결과값을 비교해 본 결과 전반적으로 변환 전과 변환 후의 MC_{pI} 값이 약 2배정도 차이를 보이고 있다. 또한 손실비용의 변화에 따른 Box-Cox 변환 전과 변환 후의 MC_{pI} 결과값 역시 전반적으로 변환 전과 변환 후의 차이가 약 2배정도 나타났다. 이는 다변량 정규분포를 따르지 않음에도 불구하고 그대로 MC_{pI}를 이용하여 평가한다면 공정능력지수의 결과값에 대한 정확도가 떨어질 수 있음을 의미한다.

따라서 본 연구는 목표치에 대한 치우침, 경제적 손실, 변수들간의 상관관계, 대칭성, 비대칭성 등 다양한 정보력과 유용성의 장점을 가진 다변량 공정능력 지수 MC_{pI}에 대하여 정규분포를 따를 경우나 따르지 않을 경우 모두를 분석할 수 있는 MC_{pI}의 적용범위를 보다 넓게 확장시키는 제안이 될 것이다.

References

[1] Anderson, T.W. and Darling, D.A., A test of goodness of fit, *Journal of the American Statistical Association*, 1954, Vol. 49, No. 268, pp. 765-769.

[2] Box, G.E.P. and Cox, D.R., An Analysis of Transformations, *Journal of the Royal Statistical Society*, 1964, Series B, Vol. 26, No. 2, pp. 211-252.

[3] Chang, Y.S., Heuristic Process Capability Indices Using Distribution-decomposition Methods, *Journal of the Korean Society for Quality Management*, 2013, Vol. 41, No. 2, pp. 233-248

[4] Hosseinifard, S.Z., Abbasi, B., Ahmad, S., and Abdollahian, M., A transformation technique to estimate the process capability index for non-normal processes, *International Journal of Advanced Manufacturing Technology*, 2009, Vol. 40, pp. 512-517

[5] Looney, S.W., How to Use Tests for Univariate Normality to Assess Multivariate Normality, *The American Statistician*, 1995, Vol. 49, No. 1, pp. 64-70.

[6] Moon, H.J. and Chung, Y.B., Multivariate Process Capability Index Using Inverted Normal Loss Function, *Journal of Society of Korea Industrial and Systems Engineering*, 2018, Vol. 41, No. 2, pp. 174-183.

[7] Razali, N.M. and Wah, Y.B., Power comparisons of shapiro-wilk, kolmogorov-smirnov, lilliefors and anderson-darling tests, *Journal of Statistical Modeling and Analytics*, 2011, Vol. 2, No. 1, pp. 21-33.

[8] Royston, J.P., Some Techniques for Assessing Multivarate Normality Based on the Shapiro-Wilk W, *Journal of the Royal Statistical Society. Series C*, 1983, Vol. 32, No. 2, pp. 121-133.

[9] Small, N.J.H., Marginal Skewness and Kurtosis in Testing Multivariate Normality, *Journal of the Royal Statistical Society. Series C*, 1980, Vol. 29, No. 1, pp. 85-87.

[10] Taam, W., Subbaiah, P., and Liddy, J.W., A Note on Multivariate Capability Indices, *Journal of Applied Statistics*, 1993, Vol. 20, pp. 339-351.

[11] Wu, C.H., Lin, S.J., Yang, D.L. and Pearn, W.L., Box-Cox Transformation Approach for Evaluating Non-Normal Processes Capability Based on the Cpk Index, *Journal of Testing and Evaluation*, 2014, Vol. 42, No. 4, pp. 949-961

[12] Yap, B. W. and Sim, C. H., Comparisons of various types of normality tests, *Journal of Statistical Computation and Simulation*, 2011, Vol. 81, No. 12, pp. 2141-2155.

[13] Zhang, T. and Yang, B., Box-Cox Transformation in Big Data, *Technometrics*, 2017, Vol. 59, No. 2, pp. 189-201

ORCID

Hye-Jin Moon | http://orcid.org/0000-0002-2477-6669

Young-Bae Chung | http://orcid.org/0000-0003-4259-6677