SQC, DOE 및 RE에서 확증적 데이터 분석 (CDA)과 탐색적 데이터 분석(EDA)의 고찰

Review of Confirmatoty Data Analysis and Exploratory Data Analysis in Statistical Quality Control, Design of Experiment and Reliability Engineering

최 성 운*
Sung-Woon Choi*

Abstract

The paper reviews the methodologies of confirmatory data analysis(CDA) and exploratory data analysis(EDA) in statistical quality control(SQC), design of experiment(DOE) and reliability engineering(RE). The study discusses the properties of flexibility, openness, resistance and reexpression for EDA.

Keywords : CDA, EDA, SQC, DOE, RE, Flexibility, Openness, Resistance, Reexpression

1. 서 론

대량의 데이터를 가공처리하는 통계학은 광범위하게 사용하는 효율적인 분석 방법이다. 확증적 데이터 분석(CDA: Confirmatory Data Analysis)은 수학 함수인 확률분포 모형의 모집단의 모수에 대한 변화와 크기에 대해 샘플 데이터의 요약 정리된 기술통계량으로 증거를 평가하는 것이다. 즉 CDA는 모집단의 모수의 변화를 샘플링 오차인 유의수준(α)내에서 변화가 없다는 귀무가설에 반하는 유의확률 P-Value가 작을 경우 유의적인 판정을 하는 가설검정과 신뢰수준의 구간추정에 해당된다.

* 경원대학교 산업공학과

CDA에서 표본의 데이터 종류는 사회과학 분야에서 명목데이터, 순서데이터, 구간데이터, 비율데이터로 구분되어 명목, 순서 데이터는 정성적 데이터로, 구간, 비율 데이터는 정량적 데이터로 재분류된다. 품질 및 신뢰성 분야에서는 소수점 처리가 가능한 계측기를 이용한 정밀, 정확한 계량연속형 데이터와 부적합, 부적합품을 정수로 세는 계수이산형 데이터로 구분된다. 계량연속형 데이터에 대한 CDA로는 평균, 중앙값, 최빈값, 미드레인지 등의 정확도(위치, 축, 편의, 치우침, 중심) 검추정과 편차 제곱합, 분산, 표준편차, 변동계수, 범위 등의 정밀도(넓이, 폭, 산포) 검추정이 있다. 계수이산형 데이터에 대한 CDA로는 부적합, 부적합품의 검추정이 있다.

CDA의 다변량 분석을 위해서 행은 관측치(Observation, Case, Experimental Unit), 열은 변수(Variable, Characteristics, Attributes)로 하는 데이터 행렬(Matrix) 또는 셋(Set)을 구성할 필요가 있다. 변수간의 상관관계를통하여 유사변수를 발견하는 기법으로는 PCA(차원 축소를 통한 데이터 요약), 인자분석(Factor Analysis, 공통인자를 통한 자료요약), 정준(Canonical)상관분석(두 변수 집단 사이의 상관관계), 중회귀분석(변수를 변수로 예측), MANOVA(두개의 특성값에 대한 인자수준간의 차이 영향)등이 있다. 또한 관측치간의 거리를 통하여 유사한 집단으로 분류하는 기법으로는 판별분석(Discrimination Analysis, 탐색적 특성으로 집단별 분리), 분류분석(Classification Analysis, 잘 정의된 분류규칙, 군집분석(Cluster Analysis, 유사성, 비유사성 거리의 계층적, 비계층적 방법), 대응분석(Correspondence Analysis, 분할표 행렬 범주를 저차원 공간점으로 동시표현), 다차원 척도 분석(Multidimensional Scaling Analysis, 선호도, 유사성으로 공간배치) 등이 있다. 위와 같은 다양한 CDA기법들은 모집단을 확률분포 모형(Model)으로 가정하여 표본의 기술 통계량으로 증거를 판정하는 방법이다.

그러나 모형은 현실을 간략화, 추상화한 실체로, 여기서 나온 최적해를 그대로 현실에 반영해서는 안된다. 수학 통계 모형인 확률분포는 효율적이긴 하나 효과적이지 못할 수 있기 때문에 통계모형에서 나온 유의성의 결과는 실무적, 기술적인 판단과 연계하여 종합 결론을 내려야 한다.

이렇듯 CDA는 수학 통계 모형에 대한 증거 평가 방법으로 재현성을 판정할 시 대표적인 기술 통계량의 비저항성(이상데이터에 대한 일탈)으로 인해 오도된 의사결정을 내릴 수 있으며 다변량 분석에서는 고등 수준의 수학지식과 과도한 계산량이 요구된다. 따라서 데이터의 수(Number)집단에 대해 융통성있고 개방적인 자세로 데이터의 패턴과 특징(Feature), 징후(Symptom)를 탐지, 탐색하여 CDA모형의 적용기법의 실마리와 분석 능력을 함양케 하는 방법인 탐색적 데이터 분석(EDA, Exploratory Data Analysis)을 병용 사용하여야 한다.

EDA에 대한 많은 연구가 이루어졌으며 [1,10-13], 활용분야는 MINITAB적용[2], R적용[3,7], 품질 분야의 적용[8,9]등이 있다. 특히 EDA의 품질분야의 적용은 층별, 히스토그램, 파레토그림을 포함한 EDA의 기본 원리 및 개괄적인 기법을 제시하고 있다.

따라서 본 연구에서는 SQC(Statistical Quality Control), DOE(Design of Experiment) 및 RE(Reliability Engineering)에서[4-6] CDA의 보완적 관계로 적용할 수 있는 EDA의 기법을 고찰한다. EDA기법은 경직된 모형의 재현성과 평균의 비저항성

기술통계량을 사용하는 CDA와는 다르게 이상치, 영향치의 일탈되지 않는 중앙값같은 저항성 관점에서 표현된다. 또한 숫자의 기술 통계량으로 요약 정리하는 CDA와 달리 그래프 현시화, 가시화에 의한 EDA기법을 살펴보고 평활, 적합값, 신호, 이미지 모형의 CDA와는 다르게 데이터를 제외한 거침(Roughness), 잔차(Residual), 노이즈(Noise), 오차(Error)등을 대상으로 CDA가정, 모델선택, 진단을 수행할 수 있는 EDA기법과 CDA변수 변환과 같은 관점에서 데이터를 재표현하는 EDA기법을 고찰한다.

2. SQC에서 CDA와 EDA

샘플의 계량연속형 데이터를 공식으로 요약정리하는 통계량(Statistics)은 CDA의 대표적인 수치표현 방법이다. 중심의 치우침 또는 정확도를 나타내는 평균, 중앙값, 최빈값과, 미드레인지와 산포, 흩어짐 또는 정밀도로 나타내는 분산, 표준편차, 범위, 변동계수가 있다. 시각적 파악을 위해 분포(Distribution)를 사용하는데 데이터의 불규칙한 실제의 산포를 이론적으로 가정한 수학함수인 분포에 적합한가를 알아보기 위해 히스토그램을 작성한다. 히스토그램에 의한 분포는 기술통계량의 정확도와 정밀도를 그림의 위치 축과 폭 넓이로 나타낸다. 히스토그램의 기둥의 수를 결정하는 공식으로는

Sturges $=[1+\log n/\log 2]$, Velleman $=[2\sqrt{n}]$, Dixon- Kronmal $=[10\log_{10} n]$, Larson=
$[1+2.2\log_{10} n]$이 있으며 n은 샘플 데이터의 크기, $[x]$는 x를 넘지 않는 큰 정수를 나타낸다. Histogram 대신 Dot Diagram, Stem and Leaf Diagram도 사용된다. 평활히스토그램(Smoothed Histogram)에서 사용되는 커늘(Kernel)에는 직사각형 (Rectangular), 삼각형(Triangular), 가우스(Gaussian), 에파네친코프(Epanechinkov), 이중제곱(Biweight) 등이 있다.[3] Sliding Window의 Band Width가 작을 경우 정확도는 좋아지지만 정밀도가 나빠지며 Band Width가 큰 경우 정확도는 나빠지지만 정밀도는 좋아진다.

정규분포에서 대칭성(Symmetry)일 경우 최빈값은 하나의 값을 가지며 평균, 중앙값과 동일한 값을 가진다. 그러나 오른쪽으로 꼬리가 길게 있어 양의 왜도(Positively Skewed)인 경우 최빈값, 중앙값, 평균의 크기순이며 음의 왜도(Negatively Skewed)인 경우 반대의 크기로 위치한다. 대칭성을 검토하는 방법으로 Pearson 왜도 계수= (평균-중앙값)/표준편차로 양의 값인 경우 평균>중앙값으로 오른쪽으로 치우친 경우이고 음의 값인 경우 평균<중앙값으로 왼쪽으로 치우친 경우이다. Fisher는 왜도를 3차 적률로, 첨도(Kurtosis)를 4차 적률로 표현하였다. 첨도는 정규분포의 종모양이 납작한가(Platykurtic), 뾰족한가(Leptokurtic)를 측정하는 것으로 각각 양의 값과 음의 값을 갖는다. 정규분포가 대칭이 아니거나 분포를 가정하지 않는 경우 백분위수 개념을 이용하는 해석이 용이한 비모수 통계방법을 사용한다. 중앙값의 순서를 100점 만점으로 환산한 점수값이 백분위수(Percentile)이다. 백분위수 P에 대한 중앙값은 P_{50}으로 표시되며 P_{50}=60점은 분포값의 50%가 60점이하라는 의미이다. 25백분위수를 1사분위수

(First Quantile) Q_1, 50 백분위수는 2사분위수(Second Quantile) Q_2, 75백분위수는 3사분위수(Third Quantile) Q_3이며 사분위수 범위 IQR(Interquantile Range)= $Q_3 - Q_1$이고 상자도표(Box-and-Whisker's Plot)의 상자를 표시한다. 수염(Whisker)은 $\pm 1.5IQR$로 극단(Extreme) 이상치를 나타낸다. 이상값의 처리방법으로 분포의 양끝 극단값의 일정 퍼센트를 제외하고 계산된 절사 평균(Trimmed Mean)과 가장 크고 작은 값들을 차순위의 크고 작은 값으로 대치하는 윈저화 평균(Winsorized Mean)이 있다. 평균은 무게 중심점(Balance Point)으로 이상치에 대한 비저항성으로 영향을 받으나 중앙값은 동일면적점(Equal-Area Point)으로 이상치에 저항하여 영향을 받지 않는다.

정규분포가 아닌 경우 공정능력지수를 구하는 경우 Box-Cox 변환, Johnson변환을 사용하여 $x^3, x^2, x, x^{1/2}, \log x, -x^{-1/2}, -x^{-1}, -x^{-2}, -x^{-3}$으로 재표현된다.[3]

개선전후의 효과, 시제품테스트의 효과등을 파악하기 위한 검정(Test), 추정(Estimation)은 Z, t, χ^2, F 통계량으로 가설검정 기각치를 판정하는 CDA방법이다. 이에 대한 EDA방법은 신뢰구간 그림, 상자그림에 의한 평균 정확도의 중심위치와 분산 정밀도의 폭넓이의 특징을 발견해 낸다.

단기간의 개선효과를 알아 보는 검정, 추정과 다르게 장기간의 개선효과를 파악하기 위해 합리적인 군구분(Rational Subgrouping)에 의해 관리한계가 기입된 꺾은선 그래프가 관리도(Control Chart)이다. 검정에서 유의수준 α=0.27%, 추정에서 신뢰수준 $1-\alpha = 99.73\%$에 의한 CDA방법을, 3σ원칙에 의해 그림으로 작성하여 플롯된 점으로 공정의 이상원인과 정상요인의 패턴과 징후를 알아보는 EDA의 Shewhart관리도가 있다.

두 변수가 확률변수로 정비례, 반비례 관계를 파악하는 상관(Correlation)과 다르게 회귀(Regression)는 고정 독립변수(Fixed Independent Variable, Input or Explanatory Variable)와 확률 종속변수(Random Dependent Variable, Output Variable or Response Variable)의 함수식(Functional Equation)으로 표현, 예측하는 적극적인 CDA 방법이다. 주어진 Data Set에 대한 적절한 함수식을 시각적으로 파악하기 위한 방법이 산점도(Scatter Diagram)이며 실제데이터와 회귀식의 잔차(Residual, Error)의 그림에 의해 함수식의 차수, 이상치(Outlier), 회귀계수에 변화를 주는 영향치(Influential Points) 등을 파악하는 회귀진단(Regression Diagnostics)이 대표적인 EDA방법이다.

Box-Jenkins 시계열 모형(Time Series Model)은 데이터 관측치간 독립표본(Random Sample, Independently and Identically Distributed)이 아닌 시계열간 상관관계된 표본(Serially Autocorrelated Sample)에 적용된다. ARIMA(Autoregressive Integrated Moving Average)모형은 식별(Identification), 추정(Estimation), 진단(Diagnosis)의 3단계 절차에 의해 분석 평가되는데 식별, 진단단계에서는 ACF(Auto Correlation Function)와 PACF(Partial ACF)의 그림에 의한 EDA방법으로 추정단계에서는 적률추정법, 최소제곱 추정법, 최우추정법 등의 CDA방법을 사용한다.

품질개선에 사용되는 QC 7가지 도구인 파레토차트, 층별, 히스토그램, 관리도, 산점도, 체크시트, 특성요인도 등은 수치 데이터를 그림이나 표로 표현하는 대표적인 EDA

방법이며 언어 데이터를 표로 표현하는 신 QC 7가지 도구 역시 마찬가지이다.

3. DOE에서 CDA와 EDA

3개 이상의 인자 수준간의 특성값에 대한 모평균의 차이를 파악하기 위한 검정방법이 CDA의 분산분석 ANOVA(Analysis of Variance)이다. 1차 회귀분석이 직선 여부를 파악하는 방법이라면 ANOVA는 수평선 여부를 파악하는 방법이다. 인자 수준간의 평균값이 수평선의 총 평균(Grand Mean)과 차이가 없는 경우 귀무가설(H_0)를 채택하고 차이가 있는 경우 인자수준이 특성값에 영향을 준다라는 대립가설(H_1)를 채택(또는 귀무가설을 기각)한다. 이 경우 수준내의 데이터, 수준간의 평균값을 플로트한 그림을 사용하여 ANOVA에 대한 시각적인 특징을 파악할 수 있다. ANOVA 실시 이전에 오차항의 정규성, 등분산성, 독립성, 불편성 등의 가정에 대한 만족여부를 파악하기 위해 잔차의 히스토그램, 정규확률도, 잔차대 적합치, 잔차 대 순서 그림을 이용한다. 평균간의 차이인 효과(Effect, Contrast)를 파악하기 위한 주효과도, 교호작용 효과도, 상자그림과 RSM(Response Surface Methodology)에서 반응 최적화를 위한 2차원의 등고선(Contour Plot), 3차원의 표면도(Surface Plot)등의 EDA방법이 있다.

4. RE에서 CDA와 EDA

품질이 구매, 생산, 검사, 인도하는 정적인 시점에서 고객이 요구하는 스펙의 만족여부를 판정하는 것이라면 신뢰성은 내용수명동안 스펙의 만족여부를 파악하는 동적인 품질보증(Quality Assurance)의 일환이다. 따라서 신뢰성 분포는 시간, 수명, 주행거리 등을 확률변수 t로 한 PDF(Probability Distribution Function) $f(t)$, CDF(Cumulative Distribution Function) $F(t)$, FR(Function, Hazard or Instantaneous Failure Rate Function) $h(t)$, RF(Survival, Reliability Function) R(t)의 CDA에 의해 신뢰성을 추정, 예측한다. 시간 t와 FRF $h(t)$의 관계를 LCA(Life Cycle Approach)에 의해 그림으로 표현된 것이 BTFR(Bath-Tub Failure Rate)곡선으로 출하시점, 고객의 사용기간, 폐기수리시점에 따른 신뢰성을 가시적으로 파악할 수 있다.

신뢰성 분포를 파악하고 CDA에 의한 방법으로 모수를 추정하기 위해서는 시간과 비용이 많이 소모되므로 가능한한 적은 수의 샘플 데이터를 사용하는 효율적인 통계적인 방법이 요구된다. 이 경우 CDA의 분포를 비모수 순위척도를 이용한 EDA의 정규확률지, 와이블 확률지 등의 그림으로 분포의 적합도, 모수의 추정 및 예측을 실시한다. 비모수 순위 추정량의 종류로는 Jacquelin 추정량, Filliben 추정량, 메디안순위(Median Rank, Bernard and Bos-Levenbach) 추정량, 평균순위(Mean Rank, Herd-Johnson, Weibull) 추정량, IEC 56 추정량, Blom 추정량, Kaplan-Meier 추정량,

Tukey 추정량, 모드순위(Mode Rank, Gumbel) 추정량 등이 있다.[6]

5. 결 론

본 연구에서는 가정된 확률수학 모형에 대한 모수의 재현성에 대한 증거를 엄격하게 파악하는 확증적 데이터 분석(CDA)방법과 CDA적용기법의 실마리를 융통성과 개방성 관점에서 데이터, 잔차, 오차의 패턴과 특징으로 파악하는 탐색적 데이터 분석(EDA) 방법을 비교하였다. 특히 통계적 품질관리(SQC), 품질실험설계(DOE), 신뢰성공학(RE)에서 적용된 그래프와 그림을 이용한 가시적인 방법, 이상치에 대한 적합성, 비모수 순위척도 등의 EDA 특징을 중심으로 고찰하였다.

6. 참 고 문 헌

[1] 김주환외, 탐색적 통계 그래프분석, 자유아카데미, 2006.
[2] 안성진, Minitab 14를 이용한 데이터 탐색법, 자유아카데미, 2007.
[3] 이태림외, 탐색적 자료분석, 한국방송통신대학교 출판부, 2003.
[4] 최성운, “정확도 및 정밀도 관점에서의 통계적 품질기법의 해석”, 대한안전경영과학회지, 9(1) : (2007) 119-131.
[5] 최성운, “안전 및 환경 적용을 위한 최소 실험계획,” 대한안전경영과학회지, 7(5) (2005) : 69-84.
[6] 최성운,“ 소표본인 경우 신뢰성 순위척도의 고찰”, 대한안전경영과학회지, 9(2) (2007) : 161-169.
[7] 허명회, R을 활용한 탐색적 자료분석, 자유아카데미, 2007.
[8] De Mast J., Bergman M.,"Hypothesis Generation in Quality Improvement Projects : Approaches for Exploratory Studies", Quality and Reliability Engineering International, 22(2006) : 839-850.
[9] De Mast J., Trip A.,“Exploratory Data Analysis in Quality-Improvement Projects", Journal of Quality Technology, 39 (4) (2007): 301-311.
[10] Hoaglin D.C., Mosteller F., Tukey J.W., Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, 1983.
[11] Mallows C.,"Tukey's Paper After 40 Years“, Technometrics, 48(3)(2006) : 319-336.
[12] Tukey J.W., Exploratory Data Analysis, Addison-Wesley, 1977.
[13] http://www.itl.nist.go