초기 데이터 분석 로드맵을 적용한 사례 연구

최은향, 이상복
이레테크 소프트사업부, 서경대학교 경영대학원 6시그마 전공

The Study on Application of Data Gathering for the site and Statistical analysis process

Choi, Eun Hyang & Sangbok Ree
Eretec Co.
6 sigma Major Dept. Business Administration, Seokyeong University Business Graduate School

ABSTRACT

In this thesis, we present process that remove mistake of data before statistical analysis. If field data which is not simple examination about validity of data, we cannot believe analyzed statistics information.

As statistical analysis information is produced based on data to be input in statistical analysis process, the data to be input should be free of error.

In this paper, we study the application of statistical analysis road map that can enhance application on site by organizing basic theory and approaching on initial data exploratory phase , essential step before conducting statistical analysis. Therefore, access to statistical analysis can be enhanced and reliability on result of analysis can be secured by conducting correct statistical analysis.

Key Words: statistical analysis road map , initial data exploratory

1. 서론

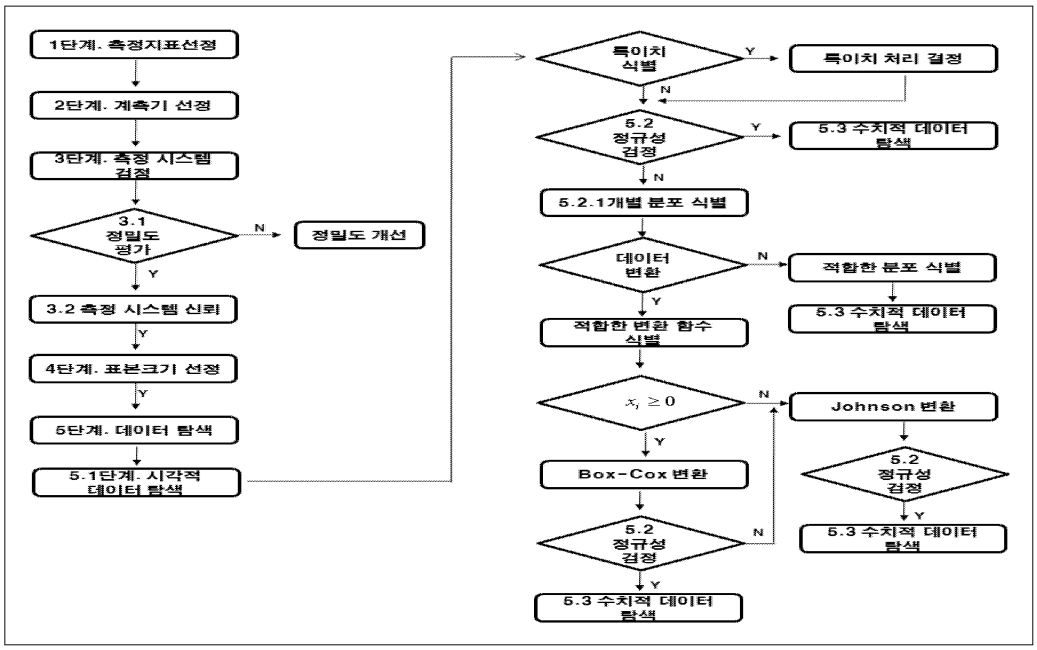
1990년대 말 이후 우리나라 기업들은 6시그마의 도입으로 통계 활용이 과거에 비해 월등히 높아졌다. 통계처리는 입력되는 데이터에 따라 통계분석정보를 제시함으로 입력되는 데이터에 오류가 있어서는 안 된다. 하지만 분석자는 통계분석 시 수집된 데이터에 오류가 있는지 등에 대한 검토를 하고 있는지에 대한 확신이 없으며, 기존의 연구논문이나 기업 컨설턴트들의 지도 자료에도 최초 수집된 데이터에 대한 검증이 언급되었으나 현장에서 활용할 수 있도록 체계적으로 정리 된 것을 찾을 수 없었다. 또한 그동안 보고된 6시그마 사례집 등을 검토하였을

때, 초기 데이터를 분석한 증거를 찾을 수 없었다(서경대학교 경영대학원 6시그마전공 MBA 학생들 중 6시그마 개선 사례를 수행한 기업의 사례집 분석과 6시그마 강사들과의 면담을 통하여 확인하였다).

이에 본 논문에서는 현장에서 수집된 데이터를 통계분석하기 위하여 데이터 수집 초기부터 통계 패키지 사용까지 검토해야 할 내용을 현장에서 사용하기 쉽게 정리하여 제시하고자 한다.

2. 초기데이터 탐색 로드맵

본 연구에서는 데이터수집 단계인 측정지표 선정 문제부터 데이터의 정규분포 변환까지 연속형 데이터의 초기데이터 분석에 대하여 5 단계를 제시하였다. 1단계는 분석목적에 부합하는 측정지표 선정, 2단계는 측정지표에 부합하는 계측기 선정, 3단계는 선택된 계측기를 포함한 측정시스템의 정밀도 평가, 4단계는 모집단을 대변할 적정한 표본크기 선정, 5단계 데이터 탑색으로 구성하였다. 특히, 5단계 데이터 탐색은 시각적 데이터 탐색을 통하여 특이치를 식별하고, 데이터의 분포를 식별한 후, 데이터 특성에 적합한 수치적 데이터 탐색을 실행하도록 로드맵을 구성하였다. [그림 1]은 초기 데이터분석 로드맵이다.



[그림 1] 초기데이터분석 로드맵

[표 1]은 초기데이터분석 로드맵 각 단계에서 사용되는 기법들을 정리한 것이다.

[표 1] 초기데이터분석 단계별 분석 내용

분석 단계		분석 내용
[1단계] 측정 지표 선정		• 연구목적에 부합하는 측정지표 선정-타당성 고려 • 측정이 용이한 특성 선정
[2단계] 계측기 선정		• 선정한 측정지표 데이터의 자릿수보다 1/10단계 정밀한 계측기 선정 • 정확도 검정 - 검·교정 받은 계측기 선정 • Minitab 통계분석 > 품질도구 > Gage 연구 > Gage 선형성 및 치우침 연구 / 유형 1 Gage 연구
[3단계] 측정시스템 검정	3.1 정밀도 검정	• 측정시스템의 변동성 분석-Gage R&R • 부품의 반복측정이 가능: 교차연구 • 부품의 반복측정이 불가능한 경우: 내포 연구 • Minitab 통계분석 > 품질도구 > Gage 연구 > Gage R&R(교차)/Gage R&R(내포)
[4단계] 표본크기 선정		• 시간·비용·모집단에 대한 대표성을 고려하여 선정
[5단계] 데이터 탐색	5.1 시각적 데이터 탐색	• 특이치·데이터의 치우침 탐색 • 히스토그램, 상자그림 • Minitab 그래프 > 히스토그램 / 상자그림
	5.2 정규성 검정	• 데이터가 정규분포를 따르는 지 검정 • Minitab 통계분석 > 기초통계학 > 정규성 검정
	5.2.1 개별 분포 식별	• 데이터에 적합한 고유한 분포 식별 • 정규 분포 적합을 위한 적합한 데이터 변환식 식별 • Minitab 통계분석 > 품질도구 > 개별분포 식별
	Box-Cox 변환	• x^{λ}또는$\ln(x)$ 변환식 적용 • 양의 값을 가지는 데이터에만 적용 가능
	Johnson 변환	• Box-Cox 변환에서 적절한 λ값을 찾지 못했을 경우 적용
	5.3 수치적 데이터 탐색	• 데이터의 특성에 맞는 통계량 선정 • 중심위치척도와 산포 위치 척도에 대한 접근 • Minitab 통계분석 > 기초통계학 > 기술통계량 표시

3. 적용사례

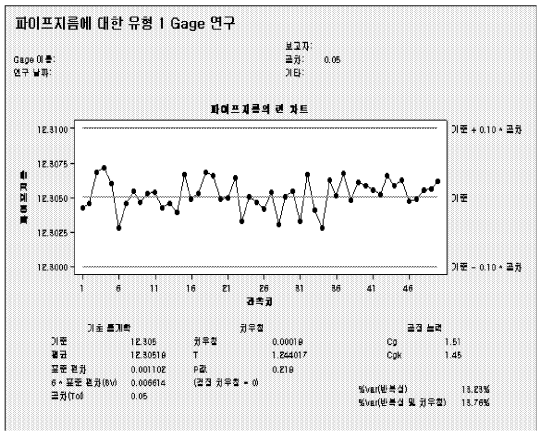
금속 파이프를 생산하는 회사에서 생산되는 파이프의 지름의 일관성을 유지하는데 관심이 있으며 2500개/Lot를 생산하고 있다.

1단계: 측정지표 선정

파이프 간의 결합을 위해서는 파이프 지름이 중요한 요소이다. 따라서 측정지표로 파이프 지름을 선정하였으며 관리규격은 12.2″±0.2″이다.

2단계: 계측기 선정

데이터를 측정하기 위해 계측기를 선정하기 위해 소수점 4자리까지 읽을 수 있는 계측기를 선정하였다. 규격을 관리하는 단위가 소수점 1자리이기에 규격의 1/10단위까지 측정할 수 있는 계측기를 선정하였으며, 선정된 계측기는 계측기 공차는 ±0.05″이다. 계측기에 대한 검교정 인증서는 첨부되어 있지만, 계측기 성능을 검증하기 위하여 유형 1 Gage 연구를 시행한 결과 $C_g = 1.51$, $C_{gk} = 1.45$로서 모두 1.33보다 큰 값으로서, 선정된 계측기는 정밀도와 정확도 모두 검증되었다.



[그림 2] 유형 1Gage 연구 실행결과

3단계: 측정시스템 검정

측정데이터의 신뢰성 확보를 위하여 계측기, 검사원(측정자)의 정밀도를 고려하여 Gage R&R (분산분석)을 실행하여 측정시스템을 검정하였다. Gage R&R분석 결과 %기여도=6.83%로서 9%미만이며, %연구변동=26.13%로서 30%미만으로 분석 되었다. 따라서 측정시스템은 정밀한 것으로 검증되었기 때문에, 측정시스템(검사원, 계측기)로 부터 얻은 데이터는 신뢰할 수 있다.

Gage R&R

출처	분산 성분	%기여(분산 성분)
총 Gage R&R	0.0028392	6.83
반복성	0.0013745	3.30
재현성	0.0014647	3.52
측정자	0.0004442	1.07
측정자*부품	0.0010205	2.45
부품-대-부품	0.0387528	93.17
총 변동	0.0415920	100.00

출처	표준 편차(SD)	연구 변동(6 * SD)	%연구 변동(%SV)
총 Gage R&R	0.053284	0.31970	26.13
반복성	0.037074	0.22244	18.18
재현성	0.038271	0.22963	18.77
측정자	0.021075	0.12645	10.33
측정자*부품	0.031946	0.19167	15.66
부품-대-부품	0.196857	1.18114	96.53
총 변동	0.203941	1.22365	100.00

고유한 범주의 수 = 5

[그림 3] Minitab Gage R&R(교차) 결과

4단계: 표본 추출

데이터 수집을 위하여 데이터 추출을 Lot 단위로 선정하였다. Lot단위로 선정한 이유는 Lot 내의 제품은 우연요인으로 인한 발생되는 변동만이 존재하는 상태이기에 Lot 단위는 합리적 부분군으로 정의할 수 있기 때문이다. 알려진 모집단의 표준편차(σ)=0.2이며, 모수에 대한 허용오차는 0.01을 넘지 않는다고 95% 신뢰할 수 있는 표본크기를 산출하고자 하며, 표본추출방법은 단순랜덤 샘플링방법으로 결정하였다. 계산결과 Lot당 8개의 샘플을 랜덤하게 추출하여 데이터를 구상할 경우 모집단(2,500개)을 추정하기에 적합한 표본크기로 볼 수 있다.

$$n = \frac{\sigma^2}{\left(\frac{\epsilon}{Z_{\alpha/2}} + \frac{\sigma^2}{N}\right)} = \left(\frac{0.2^2}{\frac{0.01}{1.96} + \frac{0.2^2}{2500}}\right) = 7.839 \fallingdotseq 8$$

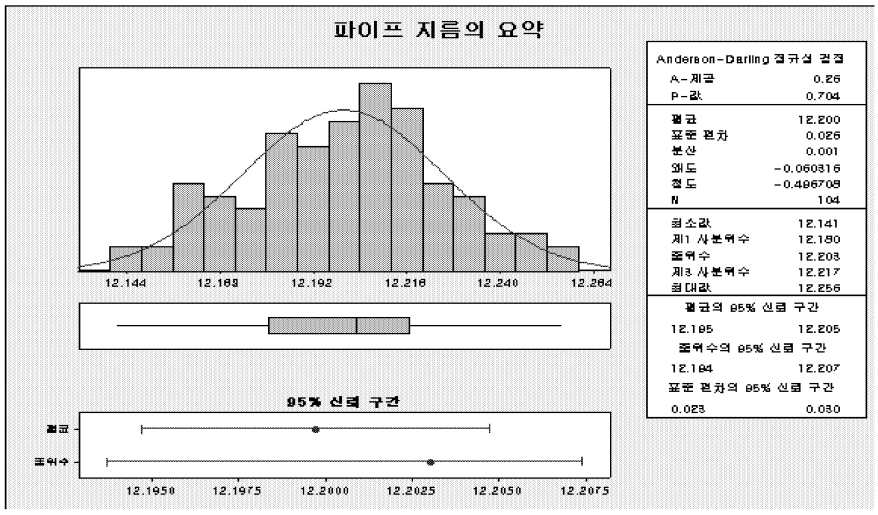
각 Lot당 8개 씩 샘플링하여 측정된 데이터는 [표2]와 같다.

[표 2] 13개의 Lot로부터 수집된 데이터

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
LOT 1	12.1735	12.181	12.2543	12.1803	12.2028	12.2001	12.1931	12.1585
LOT 2	12.1974	12.1926	12.2086	12.256	12.1861	12.2261	12.1821	12.2124
LOT 3	12.2489	12.1612	12.2175	12.2046	12.2331	12.2135	12.1894	12.216
LOT 4	12.2143	12.1593	12.1564	12.1892	12.1903	12.2366	12.1898	12.207
LOT 5	12.2135	12.2146	12.2357	12.1817	12.1786	12.2112	12.1984	12.1518
LOT 6	12.2039	12.2034	12.1882	12.2054	12.1968	12.1942	12.2296	12.164
LOT 7	12.2195	12.2159	12.2028	12.2303	12.1997	12.1689	12.2097	12.175
LOT 8	12.18	12.1636	12.226	12.2107	12.171	12.2238	12.2187	12.1694
LOT 9	12.18	12.2311	12.2061	12.1804	12.2189	12.1414	12.2064	12.2405
LOT 10	12.1994	12.1708	12.1867	12.1593	12.2095	12.1917	12.1807	12.1797
LOT 11	12.2466	12.2267	12.2046	12.2423	12.1507	12.2167	12.2073	12.2087
LOT 12	12.1973	12.2201	12.1675	12.2309	12.2238	12.189	12.1805	12.1627
LOT 13	12.2135	12.2033	12.1732	12.2083	12.2075	12.2277	12.1463	12.2471

5단계: 데이터 탐색

수집된 데이터에 대하여 초기데이터 탐색 단계를 통하여 특이치 제거, 치우침 정도, 분포 탐색을 통하여 데이터를 Cleaning하고 요약하기 위해서, 상자그림과 히스토그램을 통하여 데이터의 특이치와 치우침을 검토하고, AD검정을 통하여 데이터의 정규성 여부를 판정하고, 적합한 통계량을 선택하고자 하였다. 분석결과 작성된 상자그림에서 특이치 기호인 '*'가 발생되지 않음에 따라 수집된 데이터에 대하여 특이치는 존재하지 않으며, 히스토그램에서도 외딴섬과 같은 막대가 존재하지 않는다. 따라서 수집된 데이터에 대하여 특이치는 존재하지 않는 것으로 판단하였다. 정규성 검정결과 P-Value=0.704 > 유의수준(α)=0.05 이기에 데이터는 정규 분포를 따르는 것으로 판단하였다. 따라서 표본 데이터에 대한 요약값으로 평균과 표준편차를 적용할 수 있으며, 시그마수준을 위한 공정능력 분석은 정규분포를 가정하고 분석을 진행할 수 있다. 13개의 Lot로 부터 수집된 데이터는 평균이 12.20″이고 표준편차는 0.020″, 평균에 대한 95%신뢰구간은 (12.195,12.205)인 것으로 분석되었다. 만약 수집된 데이터에서 측정자의 측정값 기입 실수로 인하여 [표3]과 같은 데이터가 얻어 졌다고 가정하자.



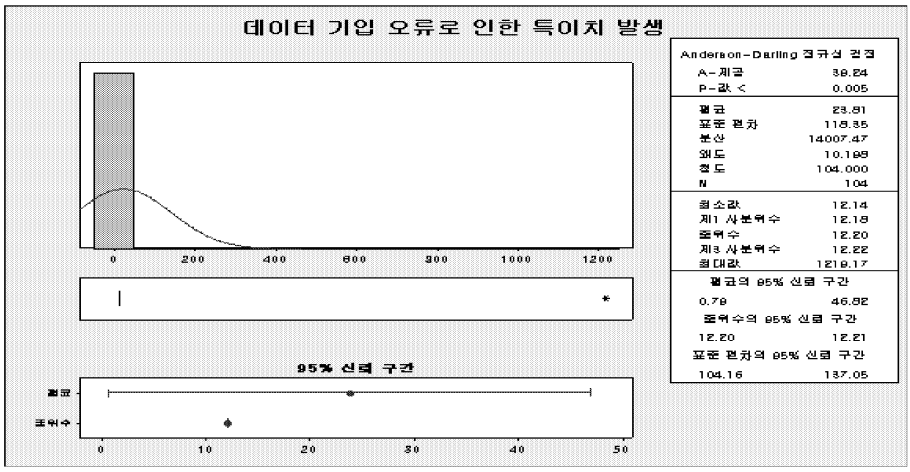
[그림 4] Minitab 그래픽 요약 분석 결과

[표3] 측정값 입력 오류를 포함한 데이터

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
LOT 1	12.1735	12.181	12.2543	12.1803	12.2028	12.2001	12.1931	12.1585
LOT 2	12.1974	12.1926	12.2086	12.256	12.1861	12.2261	12.1821	12.2124
LOT 3	12.2489	12.1612	12.2175	12.2046	12.2331	12.2135	12.1894	12.216
LOT 4	12.2143	12.1593	12.1564	12.1892	12.1903	12.2366	12.1898	12.207
LOT 5	12.2135	12.2146	12.2357	12.1817	12.1786	12.2112	12.1984	12.1518
LOT 6	12.2039	12.2034	12.1882	12.2054	12.1968	12.1942	12.2296	12.164
LOT 7	12.2195	12.2159	12.2028	12.2303	12.1997	12.1689	12.2097	12.175
LOT 8	12.18	12.1636	12.226	12.2107	12.171	12.2238	12.2187	12.1694
LOT 9	12.18	12.2311	12.2061	12.1804	12.2189	12.1414	12.2064	12.2405
LOT 10	12.1994	12.1708	12.1867	12.1593	12.2095	121.917	12.1807	12.1797
LOT 11	12.2466	12.2267	12.2046	12.2423	12.1507	12.2167	12.2073	12.2087
LOT 12	12.1973	12.2201	12.1675	12.2309	12.2238	12.189	12.1805	12.1627
LOT 13	12.2135	12.2033	12.1732	12.2083	12.2075	12.2277	12.1463	12.2471

특이치를 식별하지 않고, 평균과 표준편차를 산출 한다면 분석 결과 평균=23.81, 표준편차 =118.35, 평균에 대한 95%신뢰구간 =(0.79,46.82)로서 수집된 데이터가 공정을 많이 벗어나며, 산포가 매우 큰 것으로 유추될 수 있을 것이다. 따라서 특이치 식별은 전체 데이터 탐색에 있어서 매우 중요한 단계이다. 또 다른 경우로서, 데이터가 정규 분포를 따르지 않는 상황을 가정하자.

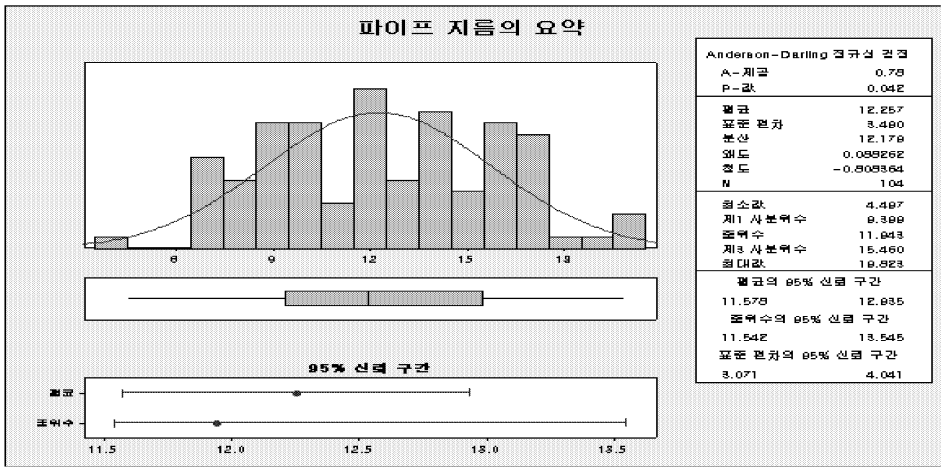
13개의 Lot로 부터 각각 8개씩 단순 랜덤 추출하여 수집된 데이터가 [표4]와 같다. 정규성 검정 결과 P-Value=0.042 < 유의수준(α)=0.05로서 데이터는 정규분포를 따르지 않는다.



[그림 5] 데이터 기입 오류 데이터의 그래픽 요약 결과

[표 4] 13개의 Lot로 부터 수집된 데이터 - 비정규 분포

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8
LOT 1	11.3598	17.7004	13.8897	11.9000	13.8159	16.6301	15.5873	16.7982
LOT 2	18.7117	13.7385	8.7009	8.6117	17.0898	12.3618	8.6104	19.7318
LOT 3	15.5269	15.6315	9.6255	14.6160	16.3202	15.7537	19.7933	16.7868
LOT 4	9.7657	15.6913	13.7983	11.8103	11.5546	13.8818	6.9306	9.1079
LOT 5	8.7705	14.7993	15.3012	9.8049	10.1706	13.9440	9.8931	19.8232
LOT 6	16.5547	12.9424	12.7379	4.4971	6.5748	7.7347	9.8869	13.8459
LOT 7	16.7303	11.8310	8.9964	8.6732	7.2221	8.6208	9.3546	13.1623
LOT 8	12.7283	11.5585	10.9624	9.8033	16.7002	13.0808	14.6792	13.6387
LOT 9	11.0439	15.5298	15.5714	7.7501	15.6982	13.5687	14.0731	7.9770
LOT 10	9.5318	9.7203	11.8174	15.5131	13.5257	14.8726	11.9568	11.7433
LOT 11	15.6134	8.0105	7.9837	6.9118	8.5547	11.5273	7.4691	9.9645
LOT 12	7.8112	6.8257	6.6595	11.6962	14.0103	16.9304	9.9055	12.7512
LOT 13	6.6459	16.6279	11.9300	11.7009	10.6455	8.7043	16.9674	12.0975



[그림 6] 데이터에 대한 그래픽 요약 결과

따라서 현재 수집된 데이터로 부터 평균과 표준편차를 통하여 수치적 데이터를 탐색하는 것은 적합하지 않을 수 있으며, 정규 분포를 가정한 공정능력분석을 수행할 수 없다. 따라서 개별분포 식별을 통하여, 적합한 데이터 변환을 찾기로 하였다.

개별분포 식별을 수행한 결과 [그림 11]과 같이 Box-Cox변환은 적합하지 않으며, Johnson 변환이 적합한 것으로 분석되었다.

Johnson 변환 함수

$$x' = 0.0964845 + 0.986536 \times \ln((x - 3.73657)/(21.2815 - x)) \quad (3.2)$$

Box-Cox 변환: 람다 = 0.5

Johnson 변환 함수:
0.0964845 + 0.986536 * Ln((X - 3.73657) / (21.2815 - X))

적합도 검정

분포	AD	P	LRT P
정규 분포	0.776	0.042	
Box-Cox 변환	0.807	0.036	
로그 정규 분포	1.112	0.006	
3-모수 로그 정규 분포	0.802	*	0.029
지수 분포	24.271	<0.003	
2-모수 지수 분포	13.782	<0.010	0.000
Weibull 분포	0.686	0.074	
3-모수 Weibull 분포	0.659	0.065	0.165
최소극단값 분포	1.361	<0.010	
최대 극단값 분포	1.140	<0.010	
감마 분포	0.894	0.023	
3-모수 감마 분포	0.806	*	0.322
로지스틱 분포	0.993	0.006	
로그 로지스틱 분포	1.223	<0.005	
3-모수 로지스틱 분포	1.000	*	0.119
Johnson 변환	**0.412**	**0.334**	

[그림 7] 개별 분포 식별 분석결과

4. 결론

올바른 통계분석 방법을 적용하기 위해서는 타당성과 신뢰성이 확보된 데이터 수집단계를 거쳐, 수집된 데이터의 특성을 파악하여야 한다. 본 논문에서는 이러한 초기데이터 검증 단계에 대하여 5단계의 로드맵을 제시하였다.

Six-Sigma 로드맵에서 제시되는 대부분의 통계분석 방법이 정규분포를 가정한 통계분석 방법임을 고려하였을 때, 5단계의 분포식별 단계는 매우 중요한 검증 단계일 것이다.

이번 연구는 측정지표 선정 단계에서부터 측정이 가능한 지표를 선정할 것을 제안하면서 5단계의 로드맵이 모두 연속형 데이터일 경우 적용 가능한 로드맵이다.

이산형 자료 또는 질적 척도 데이터에 대한 비중이 높은 비제조 분야에서의 적용에는 한계가 있기에 다양한 데이터 유형에 대한 로드맵 확대 및 사례는 추후 연구과제로 남아있다.

[참고문헌]

(1) 김우철, 김재주, 박병욱, 박성현, 송문섭, 이상열, 이영조, 전종우, 조신섭(1996) “일반 통계학”, 영지문화사

(2) 안성진(2007) “Minitab 14를 이용한 데이터 탐색법”, 자유아카데미

(3) 유성모, 박현주(2006) “Minitab으로 배우는 기초통계”, 이레테크

(4) 이상복, 염경철(2006) “ISO 2859 새로운 규격에 맞춘 샘플링 검사실무”, 이레테크

(5) 이승훈(2008) “Minitab을 이용한 공학통계 자료분석”, 이레테크

(6) 최은향(2010), “현장 데이터 취합과 분석절차에 대한 연구”, 서경대학교 경영대학원 석사학위논문

(7) Automotive Industry Action Group (AIAG) (2002). “Measurement Systems Analysis Reference Manual, 3rd edition.”, Chrysler, Ford, General Motors Supplier Quality Requirements Task Force.

(8) J.J. Filliben (1975). "The Probability Plot Correlation Coefficient Test for Normality," Technometrics, 17, 111.

(9) Keith M. Bower (2003), “Some Misconceptions about the Normal Distribution“Six Sigma Forum - American Society for Quality

(10) Montgomery, D.C., Runger, G.C. (1993). “Gauge Capability and Designed Experiments: Part I:Basic Methods”,Quality Engineering, 6(1), 115-135.

(11) Quinn,D.L., et al(2003) “Ralth&Strong's Six Sigma Leadership Handbook“, John Wilkey&Sons.Inc.

(12) R.B. D'Agostino and M.A. Stephens, Eds. (1986). “Goodness-of-Fit Techniques”, Marcel Dekker.

(13) Y. Chou, A.M. Polansky, and R.L. Mason (1998). "Transforming Nonnormal Data to Normality in Statistical Process Control," Journal of Quality Technology, 133-141