Journal of Korean
Data & Information Science Society
2006, Vol. 17, No. 1, pp. 41 ~ 52

Goodness-of-Fit-Test from Censored Samples[1)]

Young-Suk Cho[2)]

Abstract

Because most common assumption is normality in statistical analysis, testing normality is very important. The Q-Q plot is a powerful tool to test normality with full samples in statistical package. But the plot can't test normality in type-II censored samples.

This paper proposed the modified the Q-Q plot and the modified normalized sample Lorenz curve(NSLC) for normality test in the type-II censored samples. Using the two Hodgkin's disease data sets and the type-II censored samples, we picture the modified Q-Q plot and the modified normalized sample Lorenz curve.

Keywords : Lorenz curve, Normality, Q-Q plot, Type-II censored samples

1. 서론

일반적으로 데이터의 통계적 분석에서 데이터의 정규성에 관한 검정은 매우 중요한 가정이며 매우 일반화된 가정이라고 할 수 있다. 따라서 지금껏 수많은 학자들이 데이터의 정규성검정에 관하여 지속적인 연구가 이루어지고 있다. 특히 데이터의 통계적 분포와 형태에 관한 추정으로 히스토그램이나 Q-Q 플롯, P-P 플롯, CDF 플롯과 같은 그래프를 이용하여 접근하기도 하는데, 이들 연구는 Jackson 등(1989), Endrenyi와 Patel(1991), Holmgren(1995), Cho 등(1999), Kang과 Cho(2001), 그리고 Peternelli와 Osorio Silva(2003)등에 의해 연구되었다.

완전 표본에서의 정규성검정에 사용하는 Q-Q 플롯을 간단히 소개 하면 다음과 같다. 확률표본 X_1, X_2, …, X_n 의 순서통계량을 $X_{1:n}$, $X_{2:n}$, …, $X_{n:n}$ 이라 하고,

1) This work was supported by Korea Research Foundation Grant. (KRF-2004-003-C00036)

2) Assistant professor, School of Free Major, Miryang National University 50, Cheonghak-ri, Samnangjin-eup, Miryang-si, Gyeongsangnam-do 627-706, Korea
E-mail : choys@mnu.ac.kr

이 확률변수 X가 표준정규분포를 따를 때, 이 확률분표의 누적분포함수(cdf)를 $\Phi(x)$라 하자. 그러면 Q-Q 플롯은 (x, y) 좌표 평면상에 $(x_{(i)},\ \Phi^{-1}(i/n))$를 표시하는 그림을 나타낸다. 따라서 데이터가 정규분포를 따른다면, 이 Q-Q 플롯에서의 기대되는 직선은 $y=\sigma x+\mu$상에 나타나는 경향이 있고, 그 직선의 y절편은 모평균 μ의 추정값, 기울기는 모표준편차 σ의 추정값으로 사용될 수 있다. 우리는 이 직선으로부터 떨어진 정도로 데이터의 정규성을 판단한다.

다른 방법으로 Lorenz curve는 경제학 분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 곡선으로 사람들을 소득의 크기대로 순서를 정한 다음에 낮은 소득을 가진 사람부터 시작해서 수평축에 총인구에 대한 인구의 누적비를 수직축에서는 총소득에 대한 그들의 소득 누적비를 그린 하나의 곡선이다. 이 Lorenz curve를 수학적으로 표시하면

$$L(y)=\int_0^y x\ dF(x)/E(Y)$$

이고, 여기서 Y는 기대값 $E(Y)$가 존재하는 음이 아닌 소득변수이며, $F(y)$는 전체 소득수입자의 누적분포함수이다. 이와 같이 정의된 변수를 이용하여, $F(y)$를 수평축에 표시하고 $L(y)$를 수직축에 표시하여 Lorenz curve를 그릴 수 있다.

다른 정의로 $F^{-1}(p)=\ \inf_x\{x : F(x)\geq p\}$로 정의하면, Lorenz curve (Gastwirth (1971))는 다음과 같이 정의할 수 있다.

$$L(p)=\int_0^p F^{-1}(x)dx/E(Y).$$

일반적으로 확률변수의 누적본포함수가 특정함수로 표시되는 경우에 Lorenz curve를 정확히 계산할 수 없으므로 추정해야한다. 확률표본 $X_1, X_2, \cdots, X_n$의 순서통계량을 $X_{1:n}, X_{2:n}, \cdots, X_{n:n}$이라 하면, 다음과 같이 추정할 수 있다.

$$\widehat{L}(p)=\frac{\sum_{j=1}^{i} X_{j:n}}{\sum_{j=1}^{n} X_{j:n}},\ p=i/n,\ i=1,\ 2,\ \cdots,\ n.$$

이 추정된 Lorenz curve를 이용하여 그래프적인 측면에서 특정분포의 좌우 치우침을 보다 잘 파악하기 위하여 Cho 등(1999)은 변환된 Lorenz curve를 $TL(p)=1+L(p)-p$로 계산하고, 데이터가 음수인 경우에도 이 곡선을 추정하기 위해서 다음 Transformed Sample Lorenz Curve를 제시하였다.

$$TSL(p)=\frac{\sum_{j=1}^{i}(X_{j:n}-X_{1:n})}{\sum_{j=1}^{n}(X_{j:n}-X_{1:n})}-p+1,\ p=i/n,\ i=1,\ 2,\ \cdots,\ n.$$

그 후 플롯을 통하여 귀무가설 $H_0 : X \sim F(x)$에 대한 검정을 위하여 Kang과 Cho(2001)은 다음과 같은 NSLC를 제시하였다.

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)}, \quad p = i/n, \quad i = 1, 2, \cdots, n$$

여기서

$$TSL(p) = \frac{\sum_{j=1}^{i}(X_{j:n} - X_{1:n})}{\sum_{j=1}^{n}(X_{j:n} - X_{1:n})} - p + 1,$$

$$TSL_F(p) = \frac{\sum_{j=1}^{i}(F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))}{\sum_{j=1}^{n}(F^{-1}(j/(n+1)) - F^{-1}(1/(n+1)))} - p + 1.$$

이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하는 플롯을 제시하였다. 따라서 데이터가 귀무가설 $H_0 : X \sim F(x)$를 따른다면, 플롯 NSLC는 직선 $y = 0$에 일치하게 나타나고, 이 직선으로부터 떨어진 정도로 귀무가설 $H_0 : X \sim F(x)$를 판단한다.

데이터의 정규성검정을 생각한다면, 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$에 대한 NSLC는 다음과 같다.

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)}, \quad p = i/n, \quad i = 1, 2, \cdots, n$$

여기서

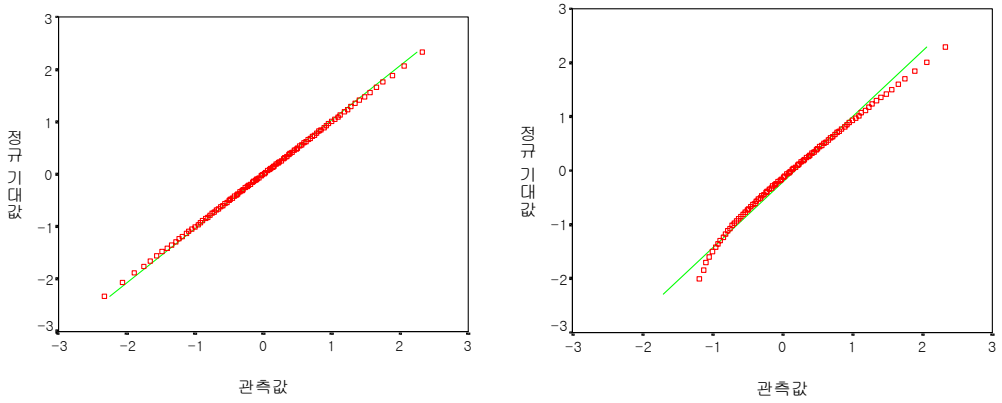
$$TSL_F(p) = \frac{\sum_{j=1}^{i}(\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1)))}{\sum_{j=1}^{n}(\Phi^{-1}(j/(n+1)) - \Phi^{-1}(1/(n+1)))} - p + 1.$$

이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하여 정규성검정을 위한 플롯으로 제시하였다.

본 논문에서는 완전 표본에서의 정규성검정에 사용하는 Q-Q 플롯을 제2종 중도절단표본의 정규성검정에 사용하게 되면 어떤 문제점이 있는가는 알아보고 이를 보완하기 위해서 새로운 수정된 Q-Q 플롯을 제시하고, 다른 방법으로 경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 Lorenz curve을 변환하여 새로운 플롯을 제시한다. 예제로 Hodgkin's disease 데이터와 여러 통계분포를 따르는 데이터를 제2종 중도절단표본에서 새로 수정된 Q-Q 플롯과 Normalized Sample Lorenz Curve(NSLC)을 그려 비교분석하였다.

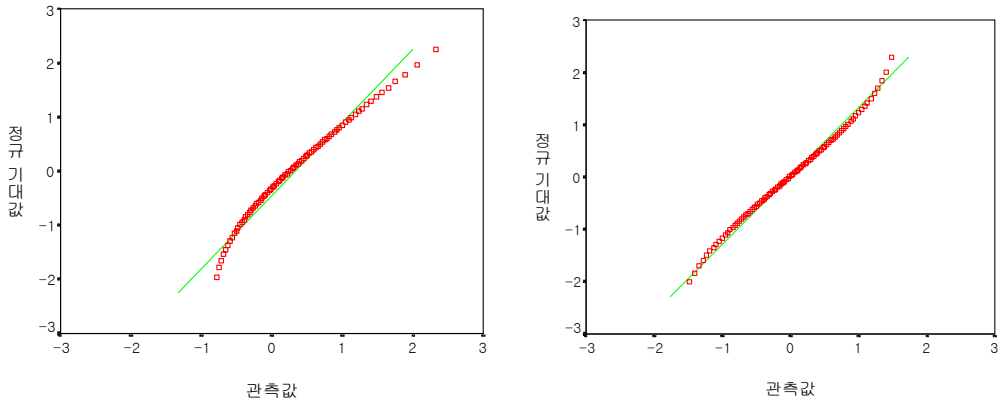
2. 중도 절단표본에서의 정규성검정을 위한 플롯

순서 데이터 $X_{1:n}$, $X_{2:n}$, …, $X_{n:n}$ 에서 처음 r개와 마지막 s개의 데이터가 중도 절단된 제2종 중도절단표본의 정규성검정을 위해서 SPSS11.0 통계패키지를 사용하였다. 정규분포를 따르는 데이터 100개의 Shapiro-Wilk 검정통계량의 P-값은 1.000이며, Q-Q 플롯은 그림 2.1(a)와 같고, 이 정규분포 데이터에서 왼쪽 10개(왼쪽 10% 절단)를 절단한 데이터의 Shapiro-Wilk 검정통계량의 P-값은 .173이며, Q-Q 플롯은 그림 2.1(b)와 같이 나타났다.



(a) (b)

<그림 2.1> 완전 정규분포 데이터와 왼쪽 10% 중도 절단표본의 Q-Q 플롯



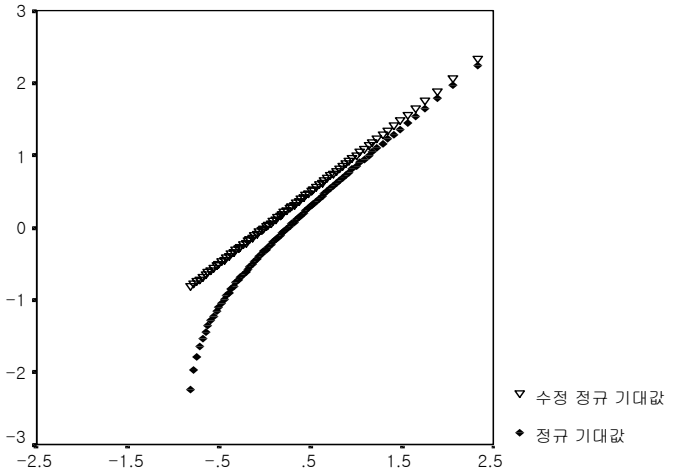
(a) (b)

<그림 2.2> 왼쪽 20% 중도 절단표본과 양쪽 10% 중도 절단표본의 Q-Q 플롯

정규분포 데이터에서 왼쪽 20개(왼쪽 20% 절단)를 중도 절단된 데이터의

Shapiro-Wilk 검정통계량의 P-값은 .034로 계산이 되었으며, Q-Q 플롯은 그림 2.2(a)이며, 정규분포 데이터에서 양쪽 5개(왼쪽 5% 절단, 오른쪽 5% 절단)씩 절단된 데이터의 Shapiro-Wilk 검정통계량의 P-값은 .281이며, Q-Q 플롯은 그림 2.2(b)와 같이 나타났다.

정규분포 데이터에서 왼쪽 20개를 중도 절단한 데이터의 Shapiro-Wilk 검정통계량의 P-값은 .034이므로 유의수준 5%에서 정규성을 기각한다. 데이터의 정규성검정을 하기 위해서 그래프 측면에서 Q-Q 플롯은 완전표본에서 정규성검정을 검정할 수 있으나 제2종 중도절단표본인 경우에서는 앞에서 제시한 Q-Q 플롯의 예시에서 보았듯이 정규성이 기각되는 오류를 범하여 옳은 결정을 하지 못함을 알 수 있다. 이는 현재 사용하고 있는 모든 통계패키지가 Q-Q 플롯을 (x, y) 좌표 평면상에 절단된 데이터에 대한 프로그림의 작성 없이 좌표평면에 $(X_{(i)},\ \Phi^{-1}(i/(n-r-s)))$로 표시하기 때문에 발생하는 문제이다. 따라서 제2종 중도절단표본에서 절단된 표본의 수 $(r+s)$를 포함하여 Q-Q 플롯을 (x, y) 좌표 평면상에 $(X_{(i)},\ \Phi^{-1}(i/n))$, $i = r+1,\ 2,\ \cdots,\ n-s$로 표시 할 수 있도록 새로운 프로그램을 작성하여야 한다. 정규분포 데이터에서 왼쪽 20% 중도 절단표본에 대한 기존의 Q-Q 플롯과 새로 수정된 Q-Q 플롯은 그림 2.3과 같이 나타났다. 정규성검정을 위해 현재 사용하고 있는 Q-Q 플롯은 완전표본에서는 가능하나 제2종 중도절단 데이터에서는 문제점을 가지고 있으므로 프로그램을 수정하여 사용하여야 한다.



<그림 2.3> 완쪽 20% 절단표본의 Q-Q 플롯과 수정된 Q-Q 플롯

데이터의 정규성검정을 위한 다른 그래프적인 방법으로 Lorenz curve를 변환하여 귀무가설 $H_0 : X \sim F(x)$에 대한 검정방법으로 Kang과 Cho(2001)은 Normalized Sample Lorenz Curve($NSLC$)를 제시하였다. 완전표본에서의 정규성검정을 위한 NSLC 블롯을 제2종 중도절단표본에서 데이터의 적합성검정을 위한 새로운 플롯으로 다음과 같은 플롯을 제시한다.

우리는 이 플롯을 통하여 귀무가설 $H_0 : X \sim F(x)$에 대한 검정을 위하여 다음과 같은 새로운 수정된 Normalized Sample Lorenz Curve를 제시한다.

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)}, \quad p = i/n, \quad i = r+1, \ 2, \ \cdots, \ n-s$$

여기서

$$TSL_F(p) = \frac{\sum_{j=r+1}^{i} (F^{-1}(j/(n+1)) - F^{-1}((r+1)/(n+1)))}{\sum_{j=r+1}^{n-s} (F^{-1}(j/(n+1)) - F^{-1}((r+1)/(n+1)))} - p + 1.$$

이 곡선을 (x, y) 좌표 평면상에 $(1-p, \ 1-NSLC(p))$를 표시하는 새로운 플롯을 제시한다. 따라서 데이터가 귀무가설 $H_0 : X \sim F(x)$를 따른다면, 이 수정된 NSLC은 x축에 일치 하게 나타나고, 플롯이 x축으로부터 떨어진 정도로 귀무가설 $H_0 : X \sim F(x)$를 판단한다.

데이터의 정규성을 생각한다면, 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$에 대한 수정된 NSLC는 다음과 같다.

$$NSLC(p) = \frac{TSL(p)}{TSL_F(p)}, \quad p = i/n, \quad i = r+1, \ 2, \ \cdots, \ n-s$$

여기서

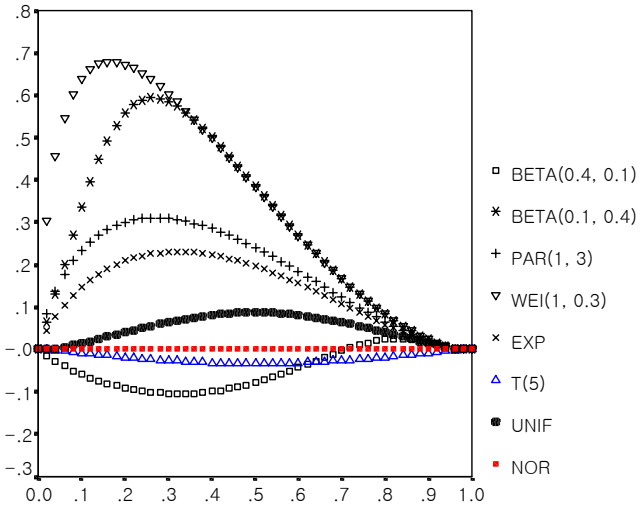
$$TSL(p) = \frac{\sum_{j=r+1}^{i} (X_{j:n} - X_{r+1:n})}{\sum_{j=r+1}^{n-s} (X_{j:n} - X_{r+1:n})} - p + 1,$$

$$TSL_F(p) = \frac{\sum_{j=r+1}^{i} (\Phi^{-1}(j/(n+1)) - \Phi^{-1}(r+1/(n+1)))}{\sum_{j=r+1}^{n-s} (\Phi^{-1}(j/(n+1)) - \Phi^{-1}(r+1/(n+1)))} - p + 1.$$

이 곡선을 (x, y) 좌표 평면상에 $(1-p, \ 1-NSLC(p))$를 표시하는 정규성검정을 위한 플롯으로 제시한다. 제2종 중도절단표본에서 데이터가 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$를 따른다면, 수정된 $NSLC$는 x축에 일치하게 나타나고, 플롯이 x축으로부터 떨어진 정도로 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$를 판단하는 그래프적인 방법을 제안한다.

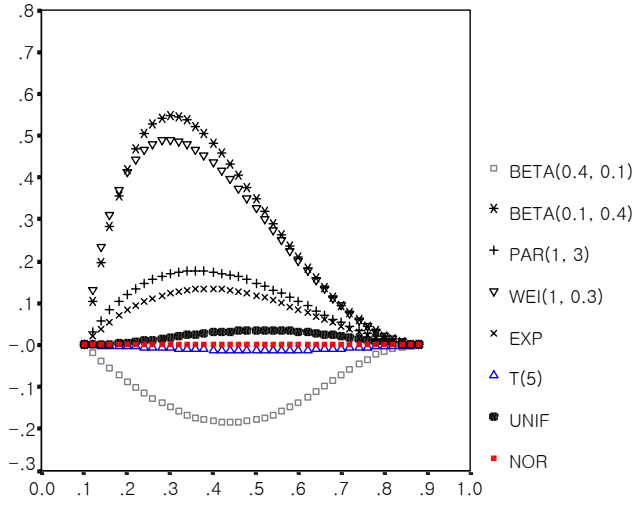
표준정규분포(NOR), 자유도 3인 T분포(T(3)), 균일분포(UNIF), 지수분포(EXP(0, 1)), 와이블분포(WEI(1, 0.3)), 파레토분포(PAR(1, 3)), 베타분포(BETA(0.1, 0.4)), 베타분포(BETA(0.4, 0.1))에서 각각 50개의 데이터를 생성하여 그린 각각의 NSLC들을 그림 2.4에 나타내었다. 이 플롯 그림 2.4로부터 정확히 표준정규분포 데이터의 NSLC는 x축에 일치하고, 와이블분포(WEI(1, 0.3)), 베타분포(BETA(0.1, 0.4), 파레토분포(PAR(1, 3)), 지수분포(EXP(0, 1)), 균일분포(UNIF), T분포(T(3))의 순으로 x축으로

부터 많이 떨어져 있으며, 베타분포(BETA(0.4, 0.1))의 NSLC는 x축 아래에서 위로 곡선을 그리고 있다.



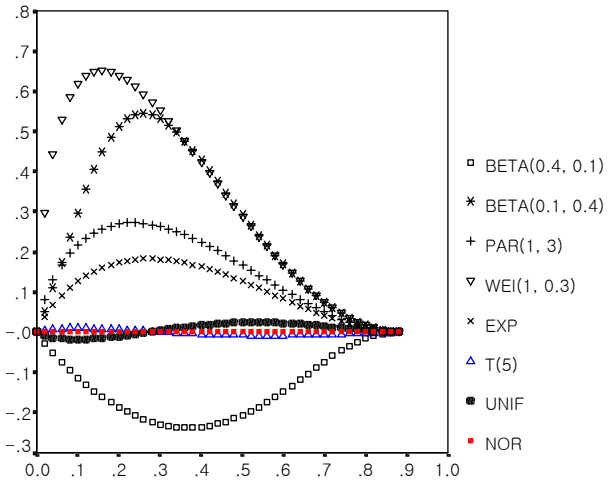
<그림 2.4> 특정분포 데이터의 NSLC

제2종 중도절단표본에서 정규성검정의 가능성을 알아보고자 특정분포의 50개 데이터를 양쪽 5개(왼쪽 10% 절단, 오른쪽 10% 절단) 씩 절단한 데이터의 수정된 NSLC들은 그림 2.5이고, 이 데이터를 왼쪽 5개(왼쪽 10% 절단) 절단표본의 수정된 NSLC들은 그림 2.6이며, 오른쪽 5개(오른쪽 10% 절단) 절단표본의 수정된 NSLC들은 그림 2.7이다.



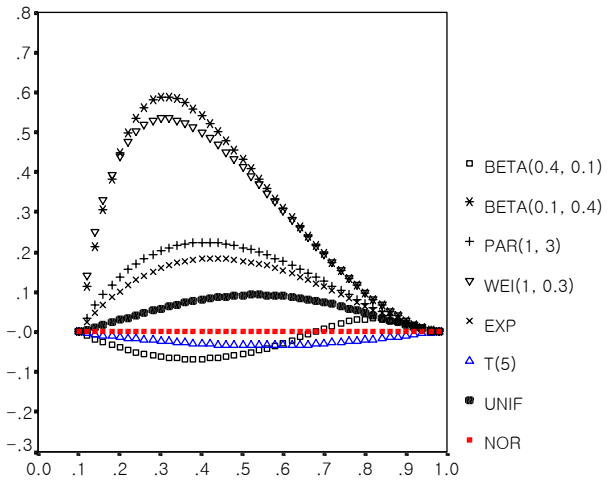
<그림 2.5> 특정분포에서 양쪽 10% 절단표본의 수정된 NSLC

특정분포에서 양쪽 10% 씩 절단된 데이터의 수정된 NSLC 그림 2.5를 보면, 마찬가지로 표준정규분포는 x축에 일치하고, 베타분포(BETA(0.1, 0.4)), 와이블분포(WEI(1, 0.3)), 베타분포(BETA(0.4, 0.1)), 파레토분포(PAR(1, 3)), 지수분포(EXP(0, 1)), 균일분포(UNIF), T분포(T(3))의 순으로 x축으로부터 많이 떨어져 있는 곡선을 그리고 있다.



<그림 2.6> 특정분포에서 왼쪽 10% 절단표본의 수정된 NSLC

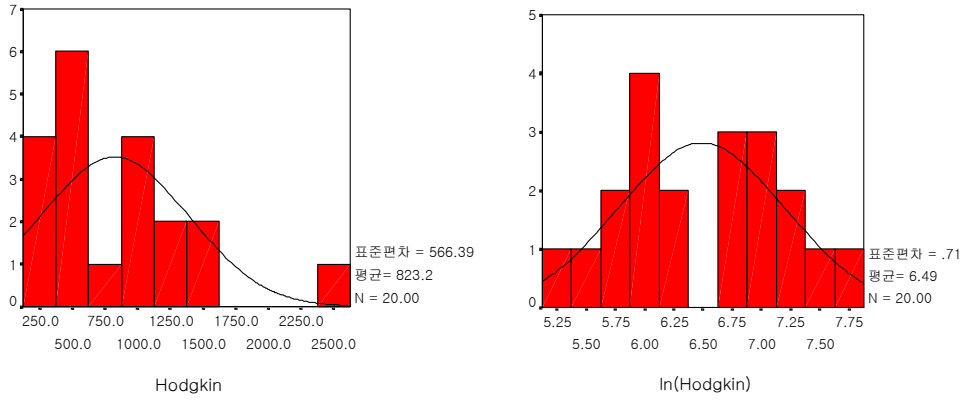
특정분포에서 왼쪽 10% 씩 절단된 데이터의 수정된 NSLC 그림 2.6을 보면, 마찬가지로 표준정규분포는 x축에 일치하고, 와이블분포(WEI(1, 0.3)), 베타분포(BETA(0.1, 0.4)), 파레토분포(PAR(1, 3)), 베타분포(BETA(0.4, 0.1)), 지수분포(EXP(0, 1)), T분포(T(3))의 순으로 x축으로부터 많이 떨어져 있으며, 균일분포(UNIF)의 수정된 NSLC는 x축을 아래에서 위로 곡선을 그리고 있다.



<그림 2.7> 특정분포에서 오른쪽 10% 절단표본의 수정된 *NSLC*

특정분포에서 오른쪽 10% 씩 절단된 데이터의 수정된 NSLC 그림 2.7을 보면, 마찬가지로 표준정규분포는 x축에 일치하고, 베타분포(BETA(0.1, 0.4)), 와이블분포(WEI(1, 0.3)), 파레토분포(PAR(1, 3)), 지수분포(EXP(0, 1)), 균일분포(UNIF), T분포(T(3))의 순으로 x축으로부터 많이 떨어져 있으며, 베타분포(BETA(0.4, 0.1))분포의 수정된 NSLC는 x축을 아래에서 위로 곡선을 그리고 있다.

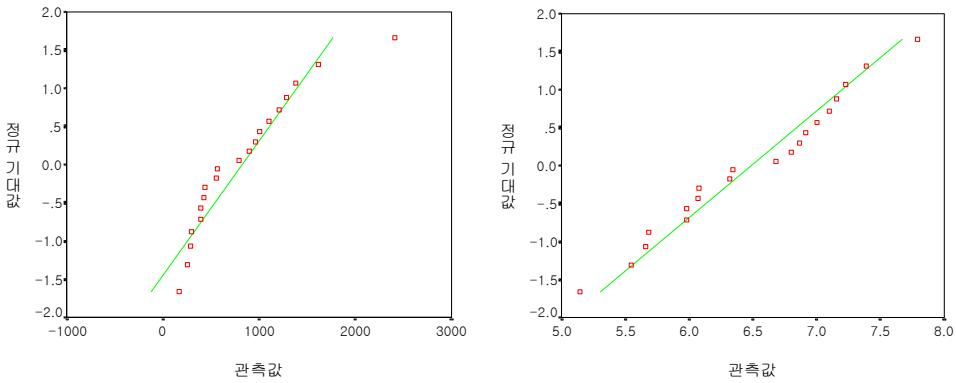
다른 예제로 Hodgkin's disease 데이터(Alterman(1992))에서 회복된 20명의 환자 혈액샘플에서 mm^3당 세포의 개수를 조사한 데이터를 이용하여 히스토그램을 그린 결과는 그림 2.8(a)에 나타나 있고, 이 자료를 자연로그 변환한 히스토그램은 그림 2.8(b)에 나타나 있다.



(a) (b)

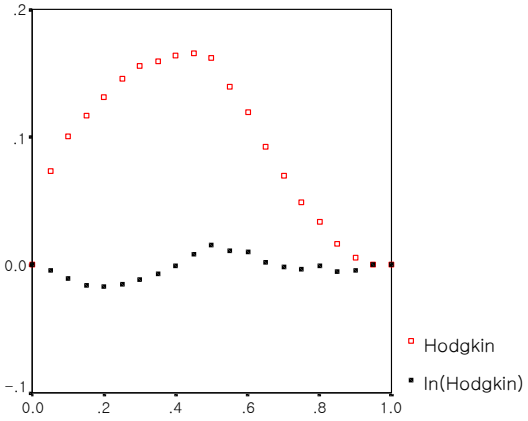
<그림 2.8> Hodgkin's disease 데이터의 히스토그램

이 데이터의 Q-Q 플롯은 그림 2.9(a)와 같이 나타났고, Shapiro-Wilk 검정통계량의 P-값은 0.031이므로 유의수준 5%에서 정규성을 따른다는 가설을 기각한다. 이 데이터를 자연 Log변환하여 Q-Q 플롯를 나타낸 결과 그림 2.9(b) 오른쪽과 같이 나타났고, Shapiro-Wilk 검정통계량의 P-값은 0.772이므로 정규성이라고 판단한다. 한편, Hodgkin's disease 데이터와 자연로그 변환된 Hodgkin's disease 데이터의 NSLC는 그림 2.10과 같이 나타났다.



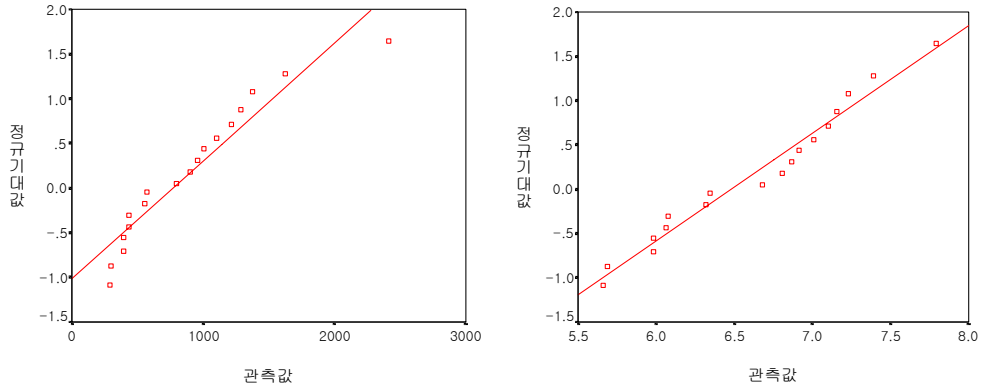
(a) (b)

<그림 2.9> Hodgkin's disease 데이터의 Q-Q 플롯



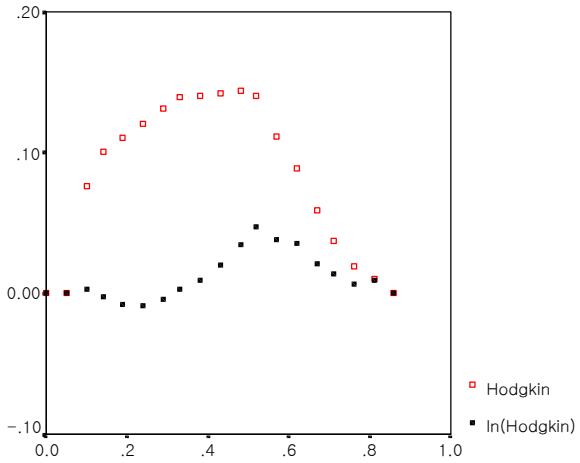
<그림 2.10> Hodgkin's disease 데이터의 NSLC

Hodgkin's disease 데이터에서 왼쪽 10% 절단한 데이터의 수정된 Q-Q 플롯은 그림 2.11(a)와 같이 나타났고, 이 데이터를 자연 Log변환하여 수정된 Q-Q 플롯을 나타낸 결과 그림 2.11(b)와 같이 나타났다. 한편, Hodgkin's disease 데이터와 자연로그 변환된 Hodgkin's disease 데이터의 왼쪽 10% 절단한 데이터의 수정된 NSLC는 그림 2.12와 같이 나타났다.



(a) (b)

<그림 2.11> 왼쪽 10% 절단된 Hodgkin's disease 데이터의 수정된 Q-Q 플롯



<그림 2.12> 왼쪽 10% 절단된 Hodgkin's disease 데이터의 수정된 NSLC

3. 결론

일반적으로 통계패키지에서 데이터의 정규성검정을 위한 그래프적인 방법으로 Q-Q 플롯과 P-P플롯을 많이 사용하고 있으나, 제2종 중도절단표본의 정규성검정에서는 문제점을 가지고 있다는 것을 알 수 있었다. 따라서 Q-Q 플롯을 수정하여 문제점을 개선한 수정된 Q-Q 플롯을 제시하였고, Kang과 Cho(2001)가 데이터의 적합성검정을 위해 제시한 NSLC를 이용하여 제2종 중도절단 데이터의 적합성검정을 위한 새로운 수정된 NSLC를 제시하여 특정분포에서 그려보았다. 그 결과를 보면 제2종 중도절단 표본에서 정규성검정을 위한 플롯으로 가능하다고 판단이 된다. 또한 Hodgkin's

disease 데이터와 자연로그 변환된 Hodgkin's disease 데이터의 수정된 Q-Q 플롯인 그림 2.11(a)와 자연로그 변환한 데이터의 수정된 Q-Q 플롯인 그림 2.11(b)의 변화를 비교하는 것 보다 실제로 Hodgkin's disease 데이터와 자연로그 변환한 데이터의 수정된 NSLC 그림 2.12를 비교하는 것이 변화를 잘 감지할 수 있다고 생각한다. 물론 이 예제는 단편적인 예제에 불과 하지만 새로 제시한 수정된 NSLC을 제2종 중도절단표본의 정규성검정에 적용할 수 있다는 확신을 가진다. 그러나 제2종 중도절단표본에서 그래프적인 비교는 주관적인 판단이므로 새로운 검정통계량을 제시하여 검정력을 비교하는 연구가 필요하다고 생각한다.

참고문헌

1. Alterman, D. G. (1992). *Practical Statistics for Medical Research,* Chapman and Hall, London.
2. Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999). 변환된 Lorenz curve를 이용한 분포 연구, <응용통계연구>, 제12권 1호, 153-163.
3. Endrenyi, L. and Patel, M. (1991). A new, sensitive graphical method for detecting deviations from the normal distribution of drug responses: the NTV plot, *British Journal Clinical Pharmacology*, Vol. 32, 159-166.
4. Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, Vol. 39, 1037-1038.
5. Holmgren, E. B. (1995). The P-P plot as a method for comparing treatment effects, *Journal of American Statistical Association*, Vol. 90, 360-365.
6. Jackson, P. R., Tucker, G. T. and Woods, H. F. (1989). Testing for bimodality in frequency distributions of data suggesting polymorphisms of drug metabolism histograms and probit plots, *British Journal Clinical Pharmacology*, Vol. 28, 647-653.
7. Kang, S. B, and Cho, Y. S. (2001). A study on distribution based on the Normalized Sample Lorenz Curve, *The Korean Communications in Statistics,* Vol. 8, No. 1, 185-192.
8. Peternelli, L. A. and Osorio Silva, C. H. (2003). A simulation study of a proposed graphical diagnostic for assessing goodness-of-fit, *Journal of Statistical Planning and Inference,* Vol. 112, 185-194.

[2005년 12월 접수, 2006년 1월 채택]