플롯을 이용한 중도절단표본에서의 정규성 검정[1)]

조영석[2)] 강석복[3)]

요 약

통계학의 주요 관심인 표본의 정규성 검정을 위해 통계패키지에서 사용하고 있는 Q-Q(quantile-quantile) 플롯을 중도절단표본에서 사용함으로 발생하는 문제점을 알아보고 이를 보완하여 수정된 Q-Q플롯과 수정된 Normalized Sample Lorenz Curve(*NSLC*)을 제시한다. 예제로 Hodgkin's disease 데이터를 중도절단하여 새로 제시한 Normalized Sample Lorenz Curve을 그려보았다.

주요용어 : 중도절단표본, 정규성 검정, Q-Q 플롯

1. 서론

일반적으로 자료의 통계적 분석에서 데이터의 정규성에 관한 검정은 매우 중요한 가정이며 매우 일반화된 가정이라고 할 수 있다. 따라서 지금껏 수많은 학자들이 이 연구에 심혈을 기울여 왔다. 특히 분포의 형태에 관한 추정으로 히스토그램이나 Q-Q 플롯과 같은 그래프를 이용하여 접근하기도 하는데, 이들 연구는 Jackson et al. (1989), Endrenyi와 Patel (1991), Holmgren (1995), Cho et al. (1999), Kang et al. (2001), 그리고 Peternelli와 Osorio Silva (2003)등에 의해 연구되었다.

완전 표본에서의 정규성 검정에 사용하는 Q-Q 플롯을 간단히 소개 하면 다음과 같다. 확률표본 $X_1, X_2, \cdots, X_n$ 의 순서통계량을 $X_{1:n}, X_{2:n}, \cdots, X_{n:n}$ 이라 하고, 이 확률변수 X가 표준정규분포를 따를 때 이 확률분수의 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, 기울기 모표준편차 σ의 추정값으로 사용될 수 있다. 우리는 이 직선으로부터 떨어진 정도로 데이터의 정규성을 판단한다.

본 논문에서는 완전 표본에서의 정규성 검정에 사용하는 Q-Q 플롯을 중도절단표본의 정규성 검정에 사용하게 되면 어떤 문제점이 있는가는 알아보고 이를 보완하기 위한 새로운 플롯을 제시한다.

2. 정규성 검정을 위한 플롯

1) 이 논문은 2004년도 학술진흥재단의 지원에 의하여 연구되었음. (KRF-2004-003-C00036)
2) 627-706 경상남도 밀양시 삼랑진읍 청학리 50번지, 밀양대학교 자율전공학부 조교수. E-mail: choys@mnu.ac.kr
3) 712-749 경상북도 경산시 대동 214-1, 영남대학교 통계학과 교수. E-mail: sbkang@yu.ac.kr

플롯을 통하여 귀무가설 $H_0 : X \sim F(x)$에 대한 검정을 위하여 Kang et al. (2001)은 다음과 같은 Normalized Sample Lorenz Curve를 제시하였다.

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

여기서

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

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

이다. 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, 이 $NSLC$에서의 기대되는 직선은 $y = 0$ 상에 나타난다. 이 직선으로부터 떨어진 정도로 귀무가설 $H_0 : X \sim F(x)$를 판단한다.

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

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

여기서

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

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

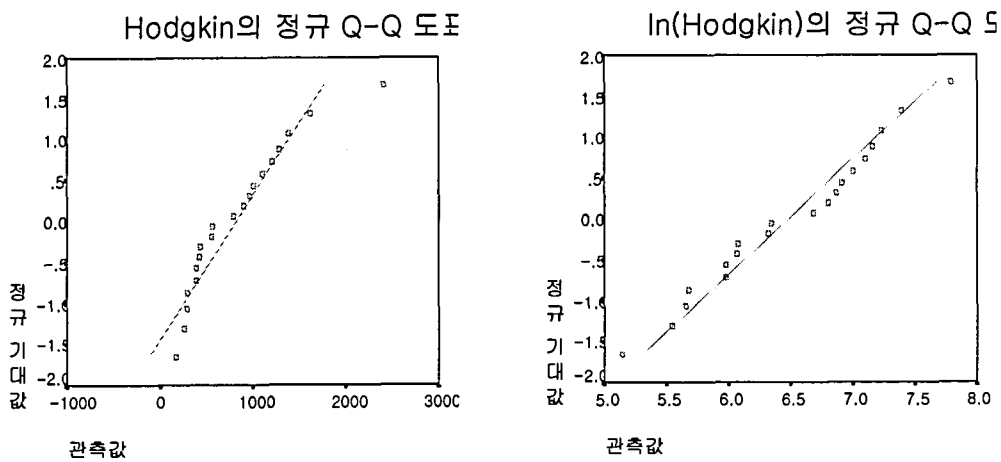


그림 1. Hodgkin's disease 데이터의 Q-Q 플롯

예로서 Hodgkin's disease 데이터 (Alterman(1992))에서 회복된 20명의 환자 혈액샘플에서 mm^3당 세포의 개수를 조사한 자료의 Q-Q 플롯은 그림 1 왼쪽과 같이 나타났고, Shapiro-Wilk 검정통계량의 P-값은 0.031이므로 정규성을 따른다는 가설을 기각한다. 이 데이터를 자연 Log변환하여 Q-Q 플롯를 나타낸 결과 그림 1 오른쪽과 같이 나타났고, Shapiro-Wilk 검정통계량의 P-값은 0.772이므로 정규성이라고 판단한다. 한편, Hodgkin's disease 데이터와 자연로그 변환된 Hodgkin's disease 데이터의 *NSLC*는 그림 2와 같이 나타났다.

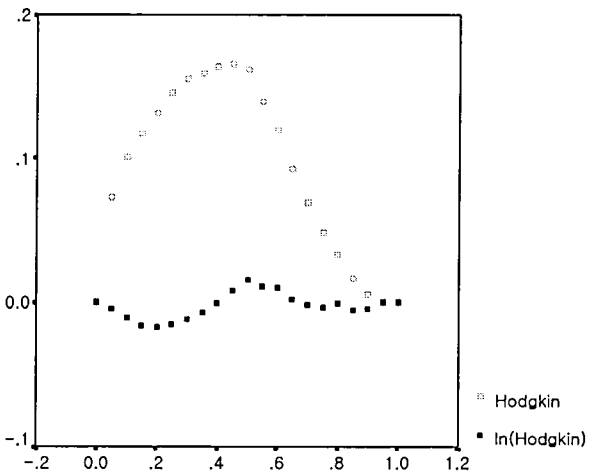
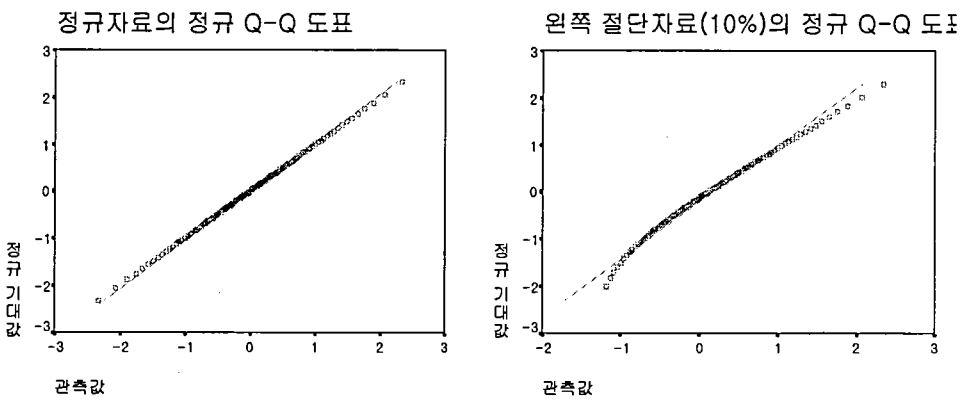


그림 2. Hodgkin's disease 데이터의 *NSLC*

3. 중도 절단표본에서의 정규성 검정을 위한 플롯

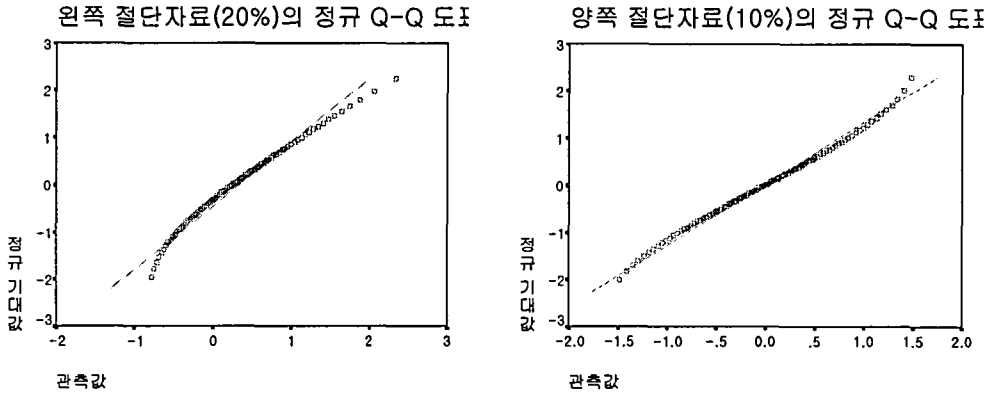
순서 자료 n에서 처음 r와 마지막 s개의 자료가 절단된 자료의 정규성 검정을 위해서 통계 패키지를 사용하였다. 정규자료 100개의 (Shapiro-Wilk 검정통계량의 P-값은 1.000) Q-Q 플롯은 그림3 (a)이고, 정규자료 왼쪽 10개를 절단한 자료의 (Shapiro-Wilk 검정통계량의 P-값은 .173) Q-Q 플롯은 그림3 (b)와 같이 나타났다.



(a) (b)

그림 3. 완전 정규자료와 왼쪽10% 절단자료의 Q-Q 플롯

정규자료에서 왼쪽 20개를 절단한 자료의 (Shapiro-Wilk 검정통계량의 P-값은 .034) Q-Q 플롯은 그림4 (a)이고, 정규자료에서 양쪽 5개씩 절단한 자료의 (Shapiro-Wilk 검정통계량의 P-값은 .281) Q-Q 플롯은 그림4 (b)와 같이 나타났다.



(a) (b)

그림 4. 왼쪽20% 절단자료와 양쪽10% 절단자료의 Q-Q 플롯

정규자료에서 왼쪽 20개를 절단한 자료의 Shapiro-Wilk 검정통계량의 P-값은 .034이므로 유의수준 5%에서 정규성을 기각한다. Q-Q 플롯이 완전표본에서는 정규성을 제대로 검정할 수 있다. 중도절단표본인 경우는 그림에서 보았듯이 정규성이 기각되는 오류를 범하여 옳은 결정을 하지 못함을 알 수 있다. 이는 통계패키지가 Q-Q 플롯을 (x, y) 좌표 평면상에 $(X_{(i)}, \Phi^{-1}(i/(n-r-s))$로 표시하기 때문에 발생하는 문제이다. 따라서 중도절단 자료의 수를 포함하여 Q-Q 플롯을 (x, y) 좌표 평면상에 표시 할 수 있도록 새로운 프로그램을 작성하여 그림 5와 같이 나타났다.

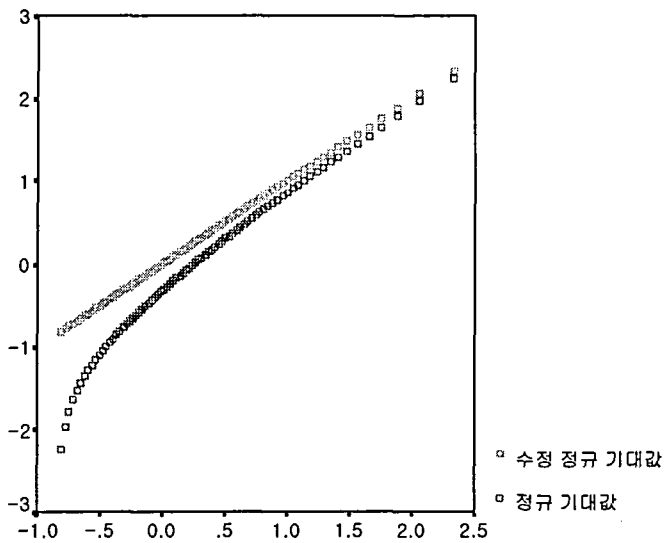


그림 5. 20% 절단자료의 Q-Q 플롯과 수정된 Q-Q 플롯

중도절단표본의 정규성 검정을 위한 새로운 플롯으로 Normalized Sample Lorenz Curve를 이용하여 다음과 같이 제시한다.

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

여기서

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

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

이다. 이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하는 정규성 검정을 위한 플롯으로 제시한다. Hodgkin's disease 데이터에서 왼쪽 10% 절단한 자료의 *NSLC*는 그림 6과 같이 나타났다. 이 예제는 단편적인 예제에 불과 하지만 새로 제시한 *NSLC*을 중도절단표본에서 정규성 검정에 적용할 수 있다는 생각한다.

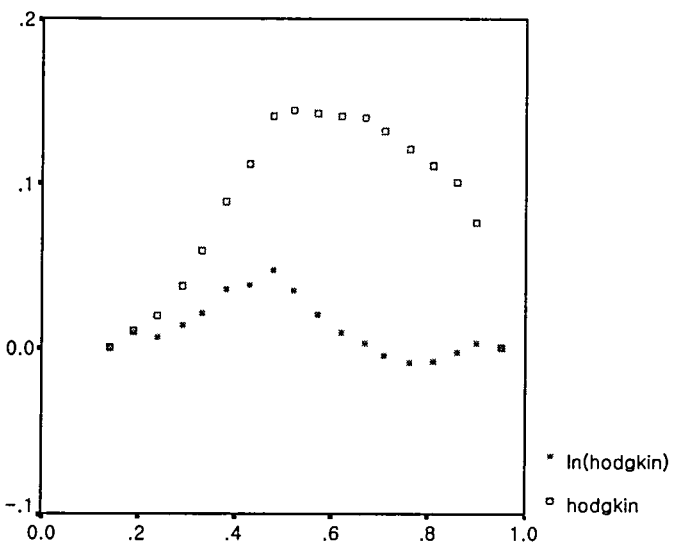


그림 6. Hodgkin's disease 데이터에서 10% 절단자료의 *NSLC*

참고문헌

Alterman, D. G. (1992), *Practical Statistics for Medical Research*, Chapman and Hall, London.

Cho, Y. S., Lee, J. Y., and Kang, S. B. (1999), 변환된 Lorenz curve를 이용한 분포 연구, <응용통계연구>, 제12권 1호, 153-163.

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.

Holmgren, E. B. (1995), The P-P plot as a method for comparing treatment effects, *Journal of American Statistical Association*, Vol. 90, 360-365.

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.

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.

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.