The Korean Communications
in Statistics Vol. 8, No. 1, 2001
pp. 185-192

A Study on Distribution Based on the Normalized Sample Lorenz Curve

Suk-Bok Kang[1] and Young-Suk Cho[2]

Abstract

Using the Lorenz curve that is proved to be a powerful tool to measure the income inequality within a population of income receivers, we propose the normalized sample Lorenz curve for the goodness-of-fit test that is very important test in statistical analysis. For two Hodgkin's disease data sets, we compare the Q-Q plot and the proposed normalized sample Lorenz curve.

Keywords : Goodness-of-fit test, Lorenz curve, Normality

1. 서론

주어진 데이터의 통계적 분포에 관한 추정은 매우 중요하다. 특히 분포의 형태에 관한 추정은 히스토그램이나 Q-Q 플롯과 같은 그래프를 이용하여 접근하기도 하는데, 이들 연구는 Jackson et al. (1989), Endrenyi와 Patel (1991), Holmgren (1995), Lee et al. (1998), 그리고 Cho et al. (1999) 등에 의해 연구되었다. 그래프를 이용한 방법 외에도 검정통계량을 이용한 대표적인 방법으로는 Kolmogorov-Smirnov 검정, Shapiro와 Wilk (1965)의 W 검정 통계량, Shapiro와 Francia (1972)의 W' 검정 통계량 등에 의해서 계속 연구되어 왔으며, Gail and Gastwirth (1978)는 Lorenz Curve를 이용하여 데이터의 지수성(exponentiality)검정에 대해 연구하였고, Looney (1995)는 다중정규성 (multivariate normality)에 대한 연구를 하였으며 그리고 Kang과 Cho (1999, 2000)는 변환된 Lorenz Curve를 이용하여 정규성(normality)과 지수성(exponentiality)검정에 대해 연구하였다.

본 논문의 2절에서는 귀무가설 $H_0 : X \sim N(\mu, \sigma^2)$, $H_0 : X \sim \text{UNIF}(\theta_1, \theta_2)$, $H_0 : X \sim \text{EXP}(\mu, \sigma)$에 대한 새로운 normalized sample Lorenz curve (*NSLC*)를 제시하고, 모의실험을 통하여 몇 몇 특정분포에 대해 일반적으로 정규성 검정에 사용되는 Q-Q 플롯과 *NSLC*를 비교하였다. 3절에서

1) Professor, Department of Statistics and Institute of Natural Sciences, Yeungnam University, Kyongsan, Kyongbuk 712-749, Korea.
E-mail : sbkang@yu.ac.kr
2) Adjunct Assistant Professor, Department of Statistics, Yeungnam University, Kyongsan, Kyongbuk 712-749, Korea.
E-mail : choys@yu.ac.kr

는 예제로 Hodgkin's disease 데이터 (Alterman(1992))를 이용하여 정규성 검정에서 사용되는 Q-Q 플롯과 새로 제시한 *NSLC*를 비교하였다.

2. 정규성 검정을 위한 플롯

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 확률변수의 누적분포함수(cdf)를 $\Phi(x)$라 하자. 그러면 Q-Q 플롯은 (x, y) 좌표 평면상에, $(\Phi^{-1}(i/n),\ X_{(i)})$를 표시하는 그림을 나타낸다. 이때, y좌표의 기대값을 구하면 다음과 같이 근사적으로 계산된다. $E[X_{(i)}] \approx \Phi^{-1}[(i-c)/(n-2c+1)]$. 이 때 사용되는 c값은 0과 1사이의 상수이다. 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.

Lorenz curve는 경제학분야에서 소득분배의 불균형 정도에 대한 척도로 널리 이용되는 곡선으로 사람들을 소득의 크기대로 순서를 정한 뒤, 낮은 소득을 가진 사람부터 시작해서 수평축에 총 인구에 대한 인구의 누적비를, 수직축에서는 총소득에 대한 그들의 소득 누적비를 그린 하나의 곡선이다. 이 Lorenz curve를 수학적으로 표시하면

$$L(y) = \int_0^y x\, dF(x)/E(Y) \tag{2.1}$$

이고, 여기서 Y는 기대값 $E(Y)$가 존재하는 음이 아닌 소득변수이며, $F(y)$는 전체 소득수입자의 누적분포함수(cdf)이다. 이와 같이 정의된 변수를 이용하여, $F(y)$를 수평축에 표시하고 $L(y)$를 수직축에 표시하여 Lorenz curve를 그릴 수 있다. $F^{-1}(p) = \inf_x\{x : F(x) \geq p\}$로 정의하면, Lorenz curve (Gastwirth (1971))는 다음과 같이 정의할 수 있다.

$$L(p) = \int_0^p F^{-1}(x)dx/E(Y) \tag{2.2}$$

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

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

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

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

여기서

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

이다. 이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하는 새로운 플롯을 제시한다. 따라서 데이터가 귀무가설 $H_0:X\sim F(x)$를 따른다면, 이 $NSLC$에서의 기대되는 직선은 $y=0$상에 나타난다. 우리는 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 표시하는 정규성 검정을 위한 새로운 플롯으로 제시한다.

다음은 확률밀도함수가 $f(x)=1/\theta,\ 0<x<\theta$인 균일분포(UNIF(0, θ))에 대해 생각해보면, 이 분포의 Lorenz Curve는 $L(p)=p^2,\ 0\le p\le 1$이고, $TL(p)=p^2-p+1$이다. 이 결과를 이용하여 일반적인 귀무가설 $H_0:X\sim$ UNIF(θ_1,θ_2)에 대한 $NSLC$는 다음과 같다.

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

이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하는 균일성 검정을 위한 새로운 플롯으로 제시한다.

그리고 확률밀도함수가 $f(x)=\frac{1}{\sigma}\exp^{-1/\sigma},\ x>0$ 인 지수분포(EXP(0, σ))에 대하여 생각해보면, 이 분포의 Lorenz Curve는 $L(p)=p+(1-p)\ln(1-p),\ 0\le p\le 1$이고, $TL(p)=1+(1-p)\ln(1-p)$이다. 이 결과로부터 두 모수를 가지는 일반적인 귀무가설 $H_0:X\sim$ EXP(μ,σ)에 대한 $NSLC$는 다음과 같다.

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

이 곡선을 (x, y) 좌표 평면상에, $(1-p,\ 1-NSLC(p))$를 표시하는 지수성 검정을 위한 새로운 플롯으로 제시한다. 우리는 몇 몇 특정분포에서 Q-Q플롯과 새로 제시한 $NSLC$를 비교하고자 표준정규 (NOR), t (STT(3)), 균일 (UNIF), 지수 (EXP(0, 1)), 베타분포 (BETA(0.2, 0.1))에서 각각 난수를 100개 발생하여 그들의 히스토그램과 Q-Q플롯을 그림 2.1, 2.2, 2.3, 2.4, 2.5에 제시하고 그 들의 $NSLC$는 그림 2.6에 제시한다. 이들 그림들로부터 여러 Q-Q플롯에서 각 분포를 비교하는 것 보다 새로 제시한 $NSLC$에서의 비교가 더욱더 용이함을 알 수 있다. Cho et al. (1999)가

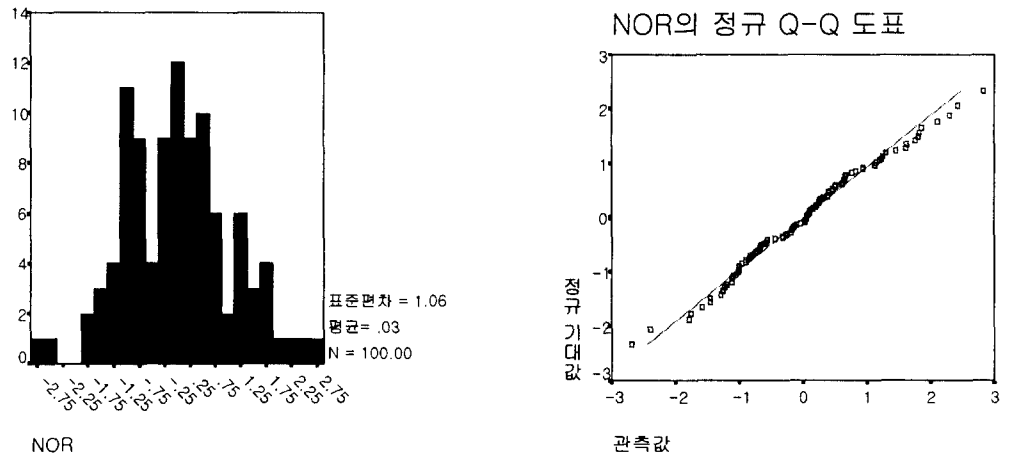

그림 2.1 표준정규분포의 히스토그램과 Q-Q플롯

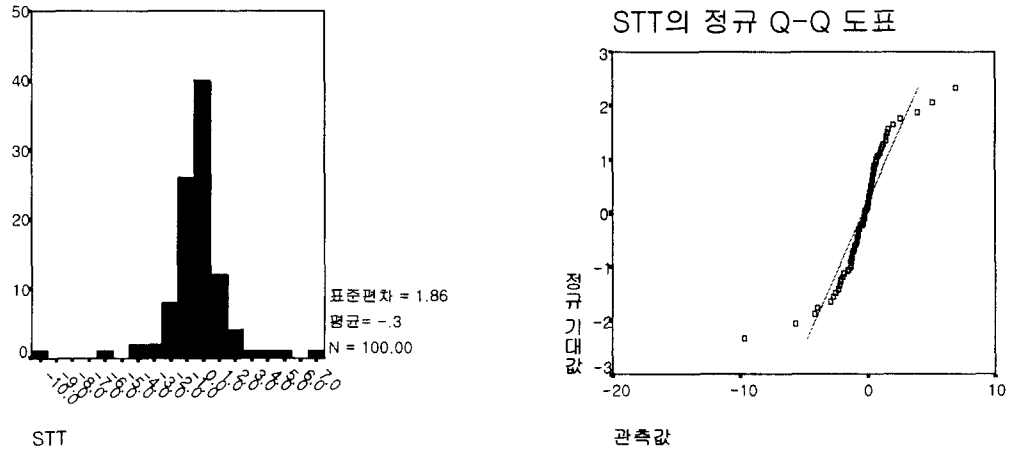

그림 2.2 t-분포의 히스토그램과 Q-Q플롯

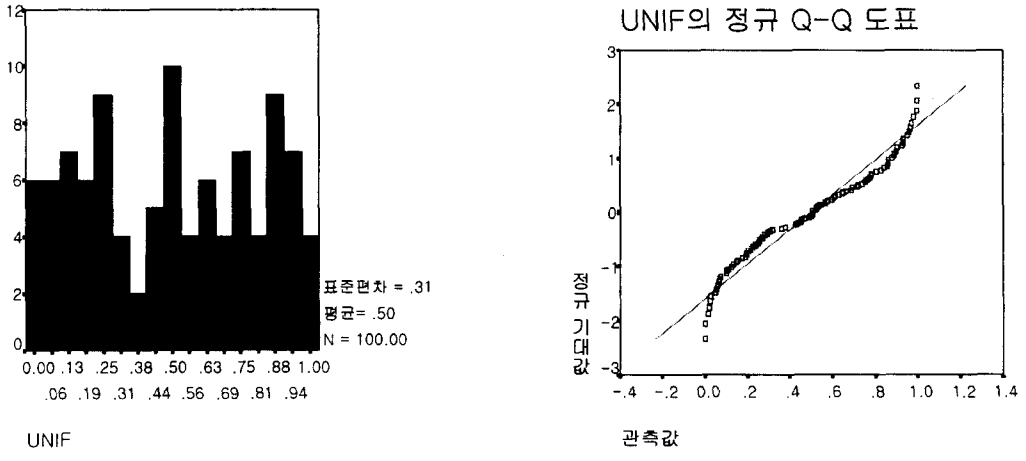

그림 2.3 균일분포의 히스토그램과 Q-Q플롯

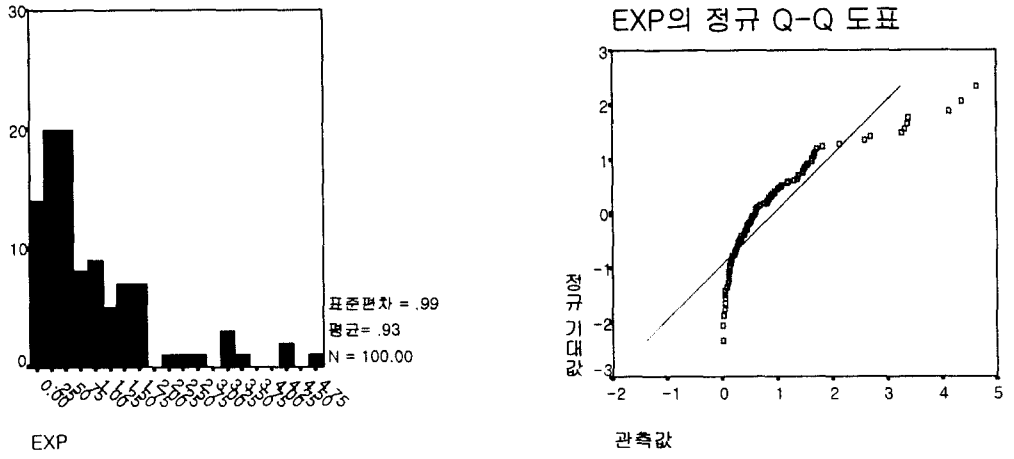

그림 2.4 지수분포의 히스토그램과 Q-Q플롯

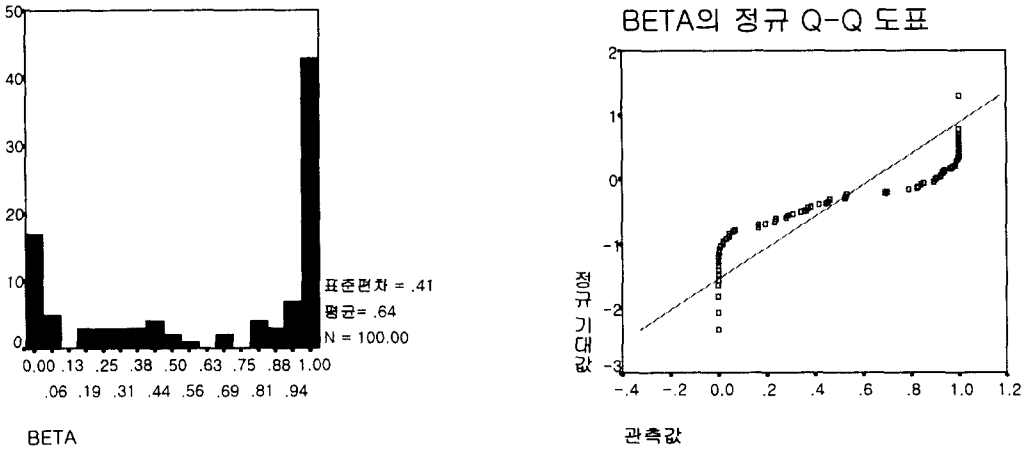

그림 2.5 베타분포의 히스토그램과 Q-Q플롯

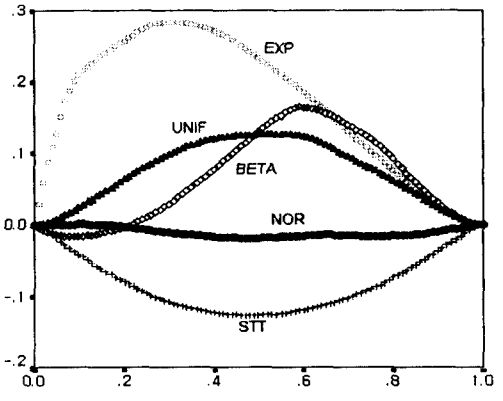

그림 2.6 특정분포의 *NSLC*

제시한 변환된 Lorenz curve는 검정하고자 하는 분포에 대하여 변환된 Lorenz curve를 그린 다음 주어진 데이터의 변환된 Lorenz curve를 비교하여야 하는 번거로움이 있으나 새로 제시한 *NSLC*는 단순히 $y=0$(x축) 직선으로부터 떨어진 정도로 분포를 판단 할 수 있다.

3. 예제를 통한 비교

간단한 예로서 Hodgkin's disease 데이터 (Alterman(1992))에서 회복된 20명의 환자 혈액샘플에서 mm^3당 세포의 개수를 조사한 자료를 이용하여 히스토그램을 그린 결과는 그림 3.1 (a)에 나타나 있고, 이 자료를 자연로그 변환한 히스토그램은 그림 3.1 (b)에 나타나 있다. 이 자료의 정규 Q-Q 플롯은 그림 3.2 (a)와 같이 정규성에 벗어난 것처럼 보이며 이 데이터의 분포가 정규성을 따른다는 가설을 기각하므로 (Shapiro-Wilk 검정통계량의 p-값은 0.031) 데이터를 자연로그 변환하여 정규 Q-Q 플롯를 나타낸 결과 그림 3.2 (b)와 같이 직선형태의 정규성을 따랐다 (Shapiro-Wilk 검정통계량의 p-값은 0.772).

한편, 우리는 Hodgkin's disease 데이터와 자연로그 변환된 Hodgkin's disease 데이터의 *NSLC*를 구하여 그림 3.3에 제시하였다. 실제로 Hodgkin's disease 데이터의 정규 Q-Q 플롯인 그림 3.2 (a)와 그 자료의 자연로그 변환한 데이터의 정규 Q-Q 플롯인 그림 3.2 (b)의 변화를 비교하는 것보다 실제로 Hodgkin's disease 데이터와 자연로그 변환한 데이터의 정규화한 표본 로렌츠 곡선을 비교하는 것이 변화를 잘 감지할 수 있다고 생각한다. 물론 이 예제는 단편적인 예제에 불과하지만 새로 제시한 정규화한 표본 로렌츠 곡선을 정규성 검정에 적용할 수 있다는 확신을 가진다.

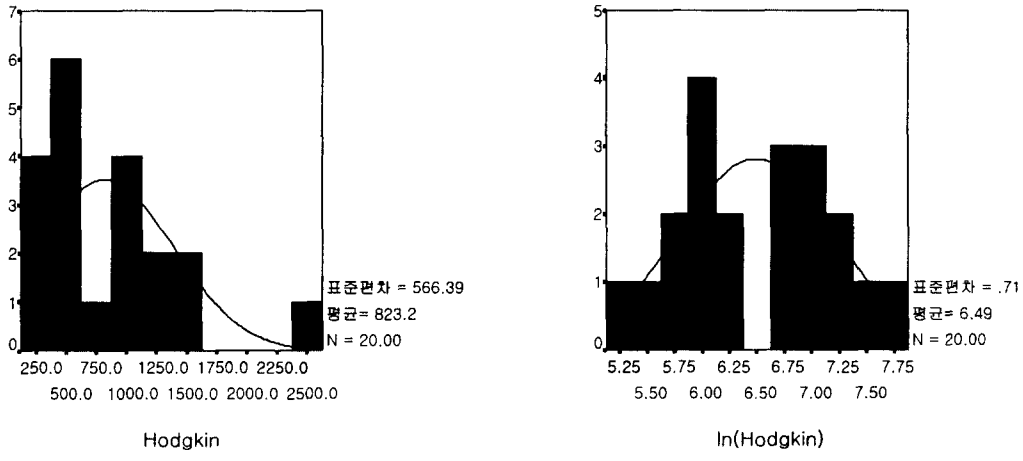

그림 3.1 Hodgkin's disease 데이터의 히스토그램

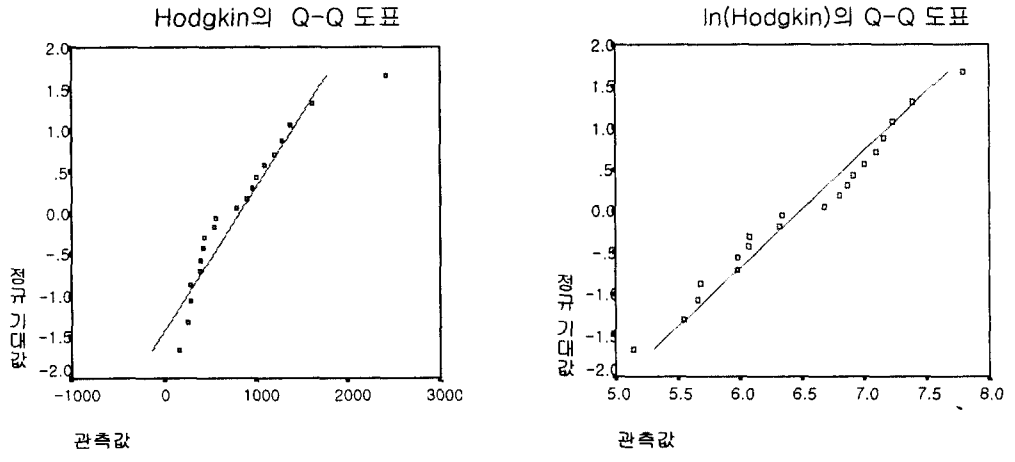

그림 3.2 Hodgkin's disease 데이터의 Q-Q플롯

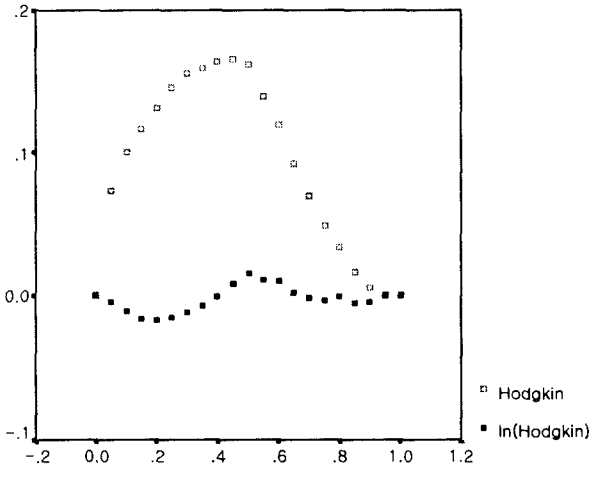

그림 3.3 Hodgkin's disease 데이터의 *NSLC*

References

[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] Gail, M. H. and Gastwirth, J. L. (1978). A Scale-free goodness-of-fit test for the exponential distribution based on Lorenz curve. *Journal of American Statistical Association*, Vol. 73, 787-793.

[5] Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, Vol. 39, 1037-1038.

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

[7] 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.

[8] Kang, S. B. and Cho, Y. S. (1999). Test of normality based on the transformed Lorenz curve. *The Korean Communications in Statistics,* Vol. 6(3), 901-908.

[9] Kang, S. B. and Cho, Y. S. (2000). Goodness-of-fit test for the exponential distribution based on the transformed sample Lorenz curve. *The Korean Communications in Statistics,* Vol. 7(1), 277-283.

[10] Lee, J. Y., Woo, J. S., and Choi, D. W. (1998). Using a normal test variable (NTV) for clinical research, <응용통계연구>, 제11권 1호, 1-12.

[11] Looney, S. W. (1995). How to use tests for univariate normality to assess multivariate normality, *The American Statistician*, Vol. 49, 64-70.

[12] Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples), *Biometrika*, Vol. 52, 591-611.

[13] Shapiro, S. S. and Francia, R. S. (1972). An approximation analysis of variance test for normality, *Journal of American Statistical Association*, Vol. 67, 215-216.