The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 67-75

An Improved Quantile-Quantile Plot for Normality Test

Jea-Young Lee[1]   Seong-Won Rhee[2]

Abstract

A new graphical method, named transformed quantile-quantile (TQQ), of a quantile-quantile (Q-Q) plot is developed for the detection of deviations from the normal distribution. It will be shown that TQQ is helpful for detecting patterns of how points depart from normality. TQQ characteristics of the various kinds of representations are illustrated by a generated sample from a composite of a normal distribution and a clinical example for TQQ is constructed and explained.

1. Introduction

A test for normality in the data of a given experiment plays a central role in statistical analysis. A simple graphical method for doing this is a Q-Q, or quantile-quantile plot (Wilk and Gnanadesikan, 1968), in which the cumulative distribution function is compared graphically with a theoretical distribution function F. Its competitor, the so-called probability-probability (P-P) plot, states that the graph is linear even if the two distributions being compared have different scale or location parameters. Dyer (1974) listed seven tests, of which the Shapiro-Wilk (W) and Anderson-Daring (A_n^2) tests generally provide the most powerful one for a reasonable class of alternatives. Large-sample versions of W are given by Shapiro and Francia (1972) and Weisberg and Bingham (1975). Lin and Mudholker (1980) have proposed a test for normality based on the Z statistic. Looney (1995) studied multivariate normality and Holmgren (1995) examined the use of P-P plot as a method for comparing treatment effects.

On the other hand, in clinical experiments, many cases are known bimodal like the distribution of debrisoquin and mephenytoin hydroxylation between Japanese and Caucasians (Nakamura *et al.* 1985). Jackson *et al.* (1989), Nakamura *et al.* (1985), Miller *et al.* (1985) and Lee *et al.* (1997, 1998) have called attention both to the importance of identifying the possible multiplicity of population distributions and to the related difficulties. We want to describe a new graphical method for the identification of deviations from the normal distribution. It is

[1] Assistant Professor, Department of Statistics, Yeungnam University, Gyongsan, 712-749 Korea
[2] International Research and Consulting Institute, 255-5, Bongsan-Dong, Jung-Ku, Taegu, 700-400, Korea

simple and very sensitive to detect outlying point and the multiplicity of distribution.

2. Transformed Quantile-Quantile Plot Method

Let $y=F(x)=pr[X\leq x]$ be the graph of the distribution function for a random variable X and let μ and σ be the mean and the standard deviation of the distribution. We can compare their x quantile values for a set of common y values using the inverse relationship $x=F^{-1}(y)$, which is known as Q-Q plot. If the two random variables to be compared are X_1 and $X_2=(X_1-\mu)/\sigma$, then

$$y=pr[X_1\leq x_1]=pr\left[X_2\leq \frac{x_1-\mu}{\sigma}\right] \quad (pr[X_2\leq x_2],\text{ say}),$$

so that for a common y, the x-values satisfy the linear relation $x_1=\sigma x_2+\mu$. Therefore, if the distributions are identical, $x_1=x_2$ and the Q-Q plot is a straight line through the origin with a unit slope.

But, the one difficulty of Q-Q plot for normality test is to check the linearity line at the diagonal side from the origin. To solve this kind of problem, we want to obtain a new improved Q-Q plot which is named of Transformed Quantile-Quantile (TQQ) plot. The TQQ is a just technical transformation of Q-Q plot and so its effectiveness has not been changed, but the main credit of TQQ is to check the linearity easily or correctly from the $y=0$ horizontal line directly.

Now, we consider the problem about the test for normality. Let Φ be the distribution function for the standard normal distribution and let $x_{(1)}\leq x_{(2)}\leq\cdots\leq x_{(n)}$ be the ordered statistics of $x_1, x_2, \ldots, x_n$. Given the Q-Q coordinate $\left(\Phi^{-1}\left(\frac{i-c}{n-2c+1}\right), x_{(i)}\right)$ where $c\in[0,1)$, we define the TQQ plot based on the following schemes : Based on the null hypothesis whose random samples are come from normally distribution and calculate the standardized value $y_{(i)}$, $y_{(i)}=\frac{x_{(i)}-\bar{x}}{s_x}$ where $\bar{x}$ is the sample mean and s_x is the sample standard deviation, then we have the standardized Q-Q coordinate resulting in $\left(\Phi^{-1}\left(\frac{i-c}{n-2c+1}\right), y_{(i)}\right)$. The next, interchange the values of the x and y coordinates for inverse functions. *i. e.*, interchange the x with the y coordinate value and we get $\left(y_{(i)}, \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right)\right)$.

Define

$$x_T = y_{(i)},$$

$$y_T = \Phi^{-1}\left(\frac{i-c}{n-2c+1}\right) - y_{(i)}.$$

We obtain the new TQQ coordinate (x_T, y_T), and plotting x_T against y_T, we define the transformed Q-Q plot.

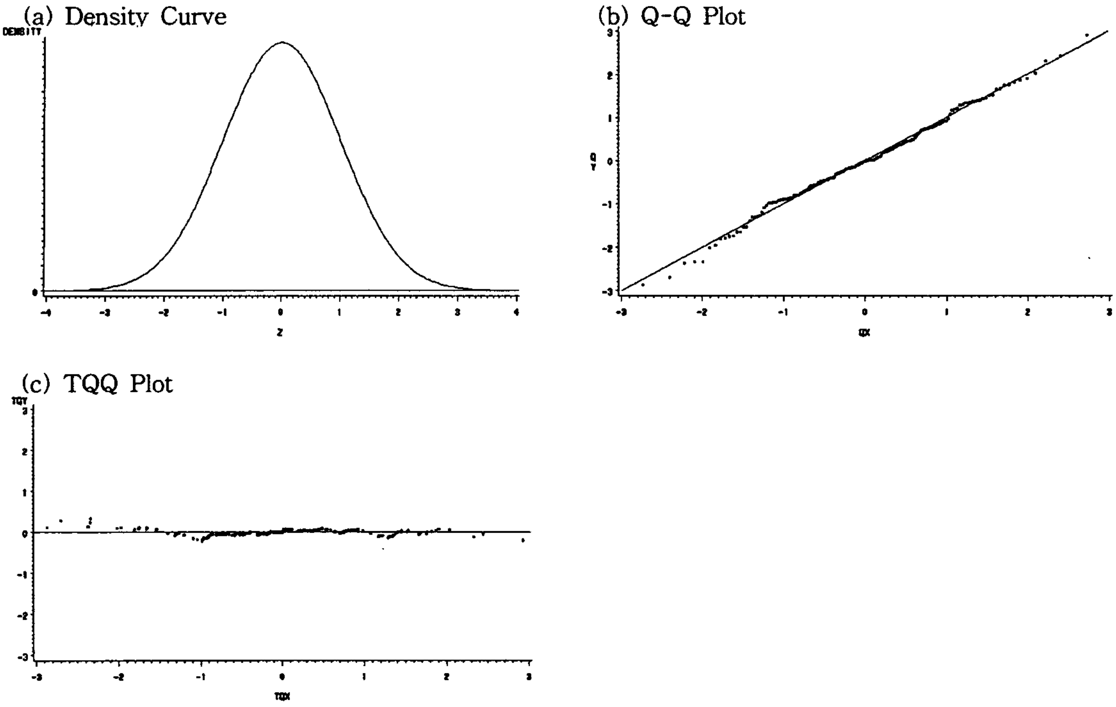
Therefore, the vertical deviation of the TQQ plot presents the difference in the expected value $\Phi^{-1}\left(\frac{i-c}{n-2c+1}\right)$ and ordered $y_{(i)}$. The typical TQQ plot then keeps constant or parallel with the x-axis and so normality is followed of the deviation proportion from $y=0$ horizontal line of TQQ. In the next section, we will illustrate how TQQ is helpful for detecting patterns of departure from normality and bimodality.

3. Effectiveness of TQQ plots

We know if the experimental sample is drawn from a normal distribution the Q-Q plot should be a straight line with a unit slope. The sensitivity of the Q-Q plot has been discussed by Jackson *et al.* (1989) and Nakamura *et al.* (1985). In particular, they showed that the presence of sharp breakpoint in the Q-Q plot have been used to detect bimodality. Now we want to discuss TQQ plots by comparing them with density curves and Q-Q plots.

When density curve and Q-Q plots are constructed for data generated from a single standard normal distribution (Figure 1a, b), they are bell shaped and linear respectively. But the corresponding y term variance of the TQQ plot keeps constant and it is very sensitive to detect outlying points. However, when the sample is taken from a skewed distribution (Exponential(1)), the Q-Q plot becomes markedly nonlinear and the y term variance of the TQQ plot is strongly inconstant (Figure 2b, c). Thus, the inconstancy of the TQQ plot does not conclusively support the presence of bimodality.

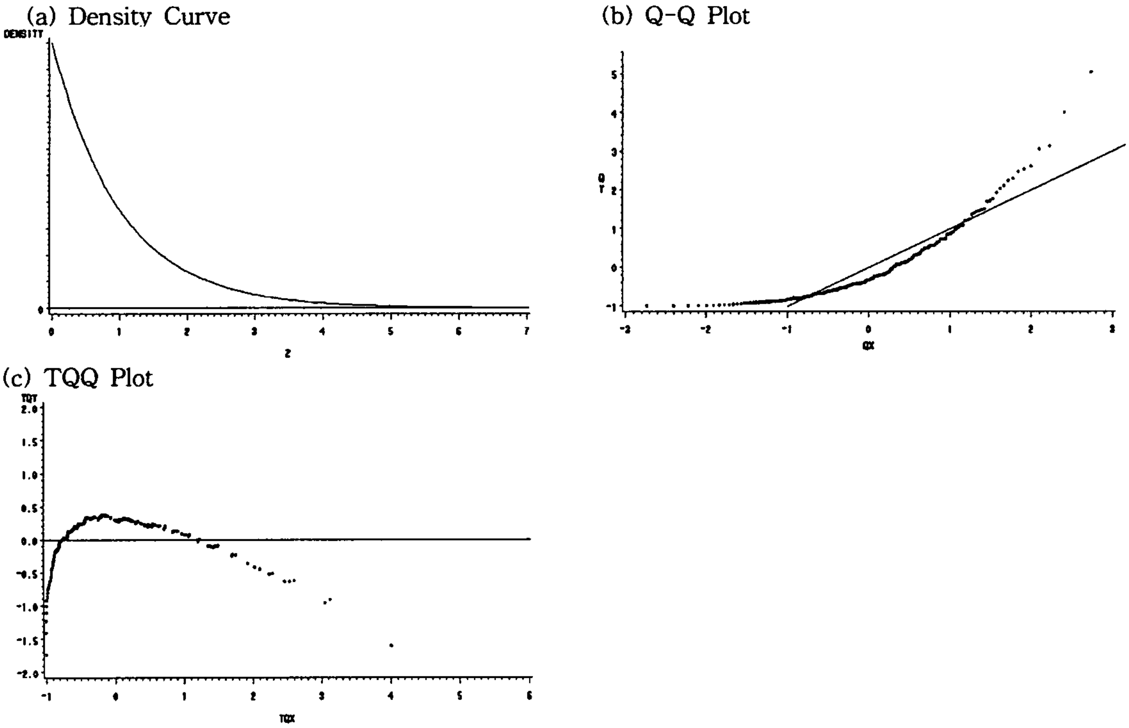
When the random samples are generated from a mixture of two normal distributions based on the same size ratio (Figure 3 and 4), the corresponding Q-Q plots and TQQ plots are both S-shaped. But, in Figure 3b, c, the Q-Q plot has still a theoretical inflection point. In practice, this can not be easy to detect clearly; however, the TQQ plot is much more sensitive to detect normality and bimodality than a Q-Q plot. When the component distributions are of similar variance and not well-separated, there are almost linear (no sharp break points) Q-Q plots (Figure 3b). But, TQQ has a fairly clear inflection point around the zero x-coordinate value(Figure 3c). Of course, when the means are well-separated, then sharp break points are presented in both plots (Figure 4b, c).


(a) Density Curve

(b) Q-Q Plot

(c) TQQ Plot

Figure 1. Generated sample distribution from Normal(0,1) with sample 200


(a) Density Curve

(b) Q-Q Plot

(c) TQQ Plot

Figure 2. Generated sample distribution from Exp(1) with sample 200

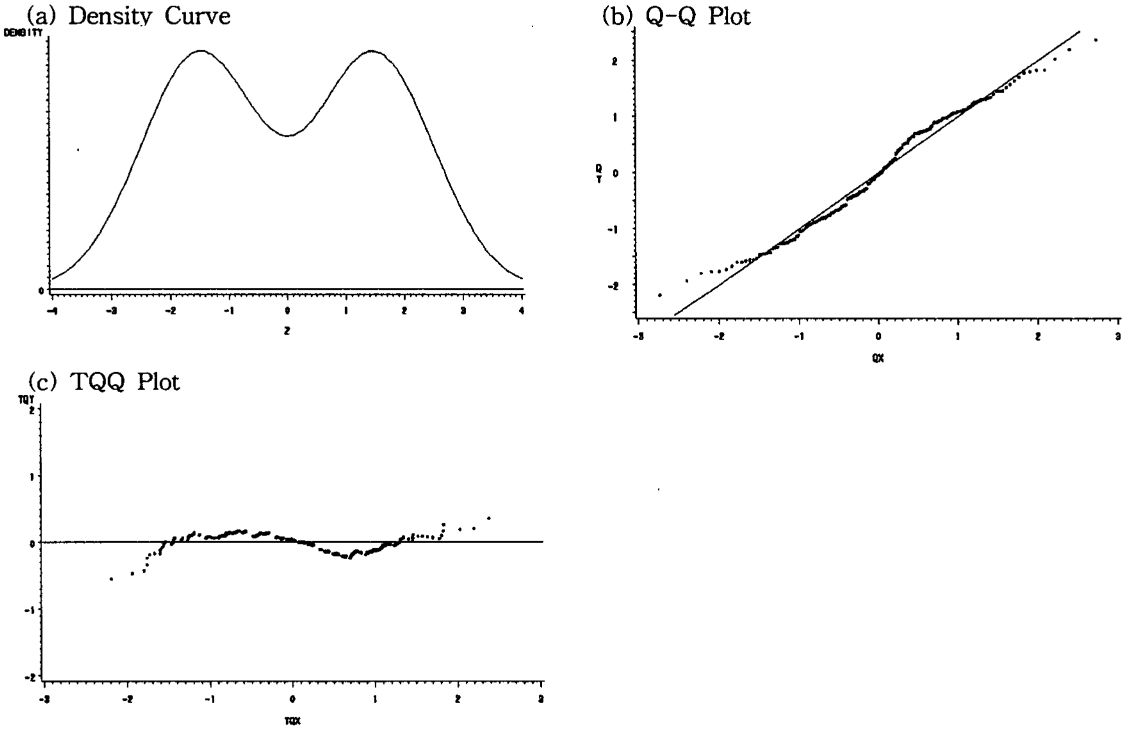


Figure 3. Generated sample distribution from a mixture distribution 1/2*N(-1.5,1)+1/2*N(1.5,1) with sample 200

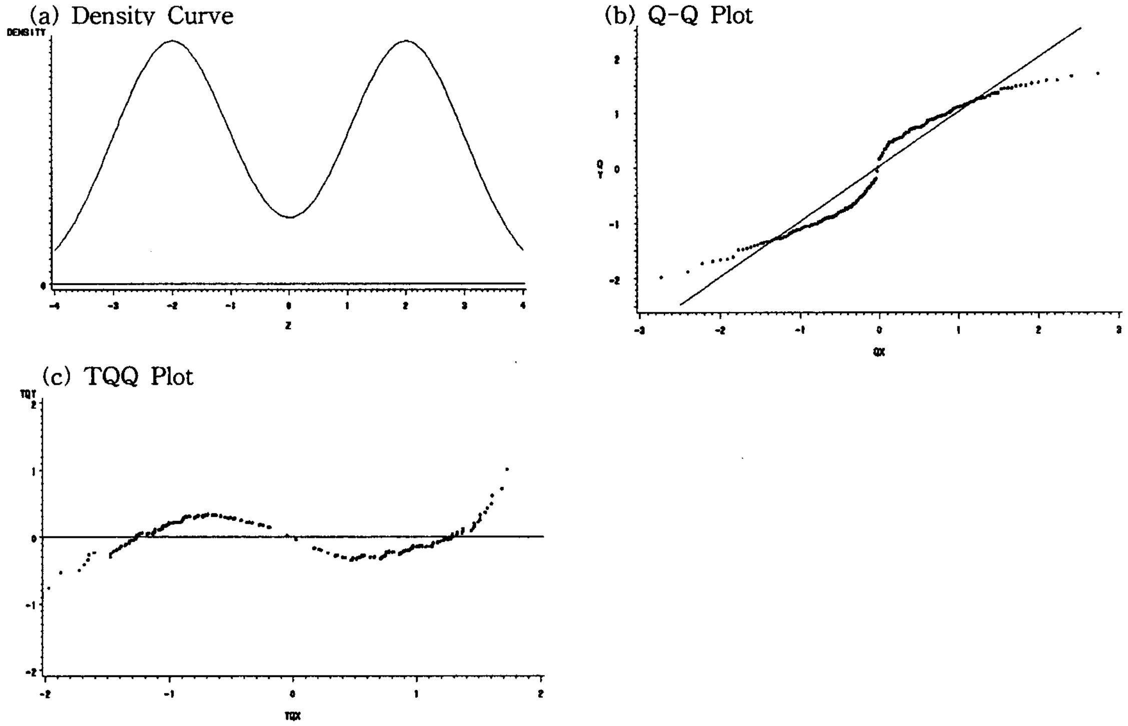
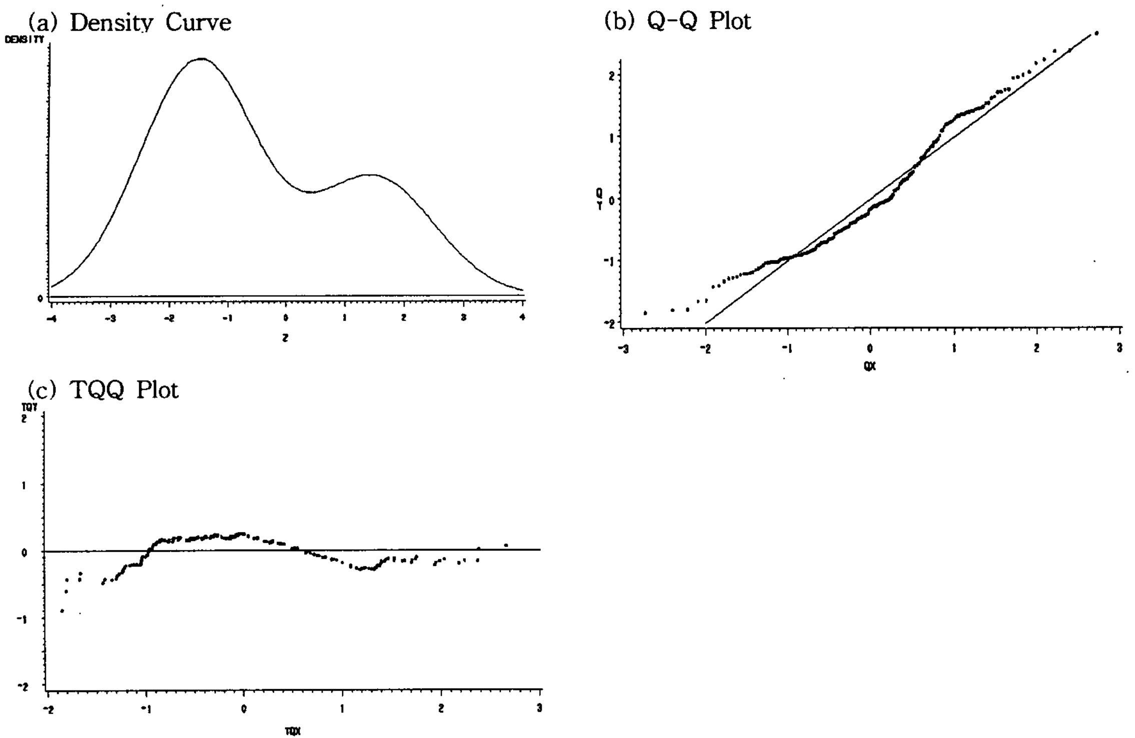


Figure 4. Generated sample distribution from a mixture distribution 1/2*N(-2.0,1)+1/2*N(2.0,1) with sample 200

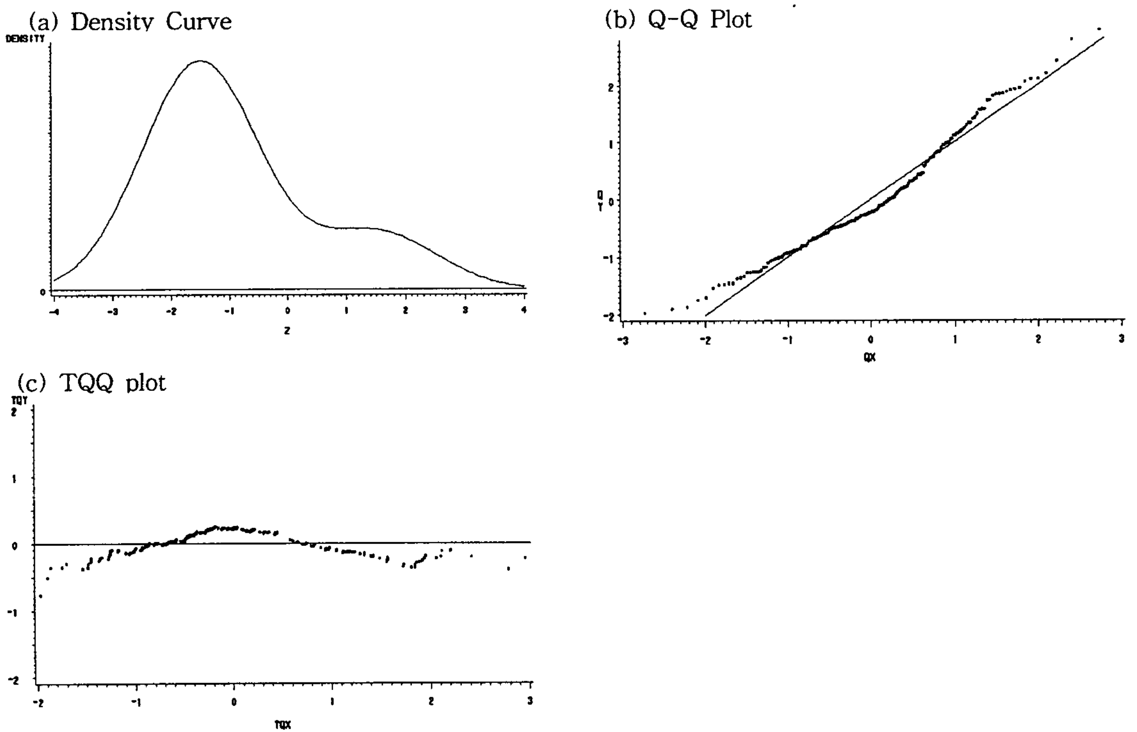


(a) Density Curve

(b) Q-Q Plot

(c) TQQ Plot

Figure 5. Generated sample distribution from a mixture distribution 2/3*N(-1.5,1)+1/3*N(1.5,1) with sample 200



(a) Density Curve

(b) Q-Q Plot

(c) TQQ plot

Figure 6. Generated sample distribution from a mixture distribution 4/5*N(-1.5,1)+1/5*N(1.5,1) with sample 200

Table 1 Numbers of T_4 and $\log_{10} T_4$ cells per mm^3 in their blood samples from 20 patients in remission from hodgkin's disease (Shapiro *et al.*,1986).

Subject	T_4	$\log_{10} T_4$
1	396	2.59770
2	568	2.75435
3	1212	3.08350
4	171	2.23300
5	554	2.74351
6	1104	3.04297
7	257	2.40993
8	435	2.63849
9	295	2.46982
10	397	2.59879
11	288	2.45939
12	1004	3.00173
13	431	2.63448
14	795	2.90037
15	1621	3.20978
16	1378	3.13925
17.	902	2.95521
18	958	2.98137
19	1283	3.10823
20	2415	3.38292

On the other hand, random samples are generated from a mixture of two normal distributions based on the various relative contributions of the two components (Figure 5 and 6). In density curves (Figure 5a and 6a), the lesser component became decreasingly noticeable as its contribution is reduced. In the Q-Q plots (Figure 5b and 6b), there is at least one break point and the breadth of the inflection range increased with a growing discrepancy between the two components (Endrenyi and Patel, 1991). But, in general, the discontinuity is not always so sharp and the objective assessment of its presence is difficult (Jackson *et al.*, 1989). However, those kinds of problems have been solved by using TQQ plots (Figure 5c and 6c).

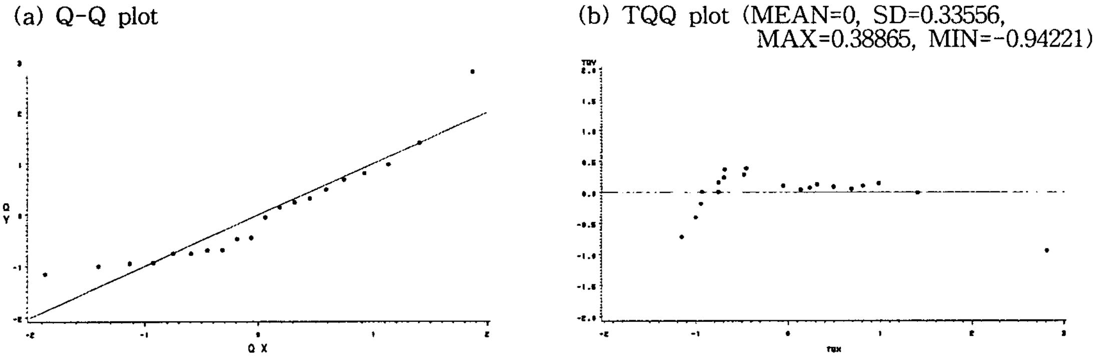
4. Clinical Example (Shapiro *et al.*,1986)

We shall illustrate the paired samples analysis using data from a study of lymphocyte abnormalities in patients in remission from Hodgkin's disease. There were 20 patients. Table 1 shows the numbers of T_4 and $\log_{10} T_4$ cells per mm^3 in their blood. The raw data are showing the skewness and unequal scatter. The result shows the success of the log transformation in producing data that are plausibly normal and have similar standard deviations (Alterman, D. G. 1991).

Q-Q and TQQ plots are constructed from the sample (Figure 7 and 8). In Figure 7a, the Q-Q plot is not linear and in Figure 7b, TQQ doesn't keep constant, also. But TQQ is detecting outlier point, *i. e.,* the minimum value of TQQ is -0.9422 which is detected as outlying point (in Figure 7b, mean=0 and SD=0.3355). After log transformation, Q-Q plot is shaped as linear line but not clear yet. However TQQ plot keeps obviously constant and therefore, it means normally distributed.

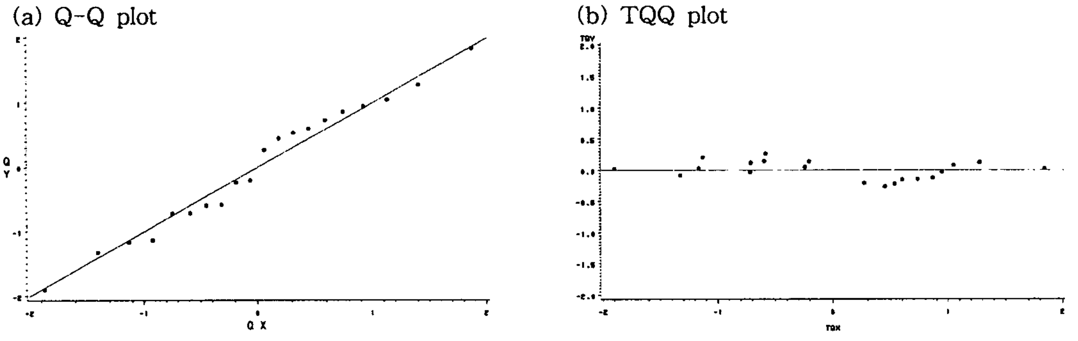
5. Conclusions

A new graphical method, named TQQ, of a Q-Q plot is developed for the detection of deviations from the normal distribution. In Figure 1, the TQQ plot is more clearly detected deviation points outside the normal distribution. The characteristics of the various kinds of representations, which are based on the effect of the separation between distributions (Figure 3 and 4) and the effect of the size ratio of the components (Figure 5 and 6), are illustrated by a generated sample from a composite of normal distribution. It is shown that TQQ is helpful for detecting the patterns of points departing from normality and TQQ also points out break points more clearly. Furthermore, the TQQ plots for bimodal density distributions are constructed and compared with Q-Q plots. We may therefore conclude that TQQ is a more improved plotting system to detect deviations from the normal distribution than Q-Q.


(a) Q-Q plot

(b) TQQ plot (MEAN=0, SD=0.33556, MAX=0.38865, MIN=-0.94221)

Figure 7. Sample in Hodgkin disease T4 case


(a) Q-Q plot

(b) TQQ plot

Figure 8. Sample in Hodgkin disease Log T4 case

6. References

[1] Altman, D. G. (1992). *Practical statistics for medical research,* Chapman and Hall, London.

[2] Dyer, A. R. (1974). Comparison of tests for normality with a cautionary note, *Biometrika,* Vol. 61, 185-189.

[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, *Br. J. clin. pharmac.*, Vol. 32, 159-166.

[4] Holmgren, E. B. (1995). The p-p plot as a method for comparing treatment effects, *J. Am. Stat. Assoc.*, Vol. 90, 360-365.

[5] 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, *Br. J. clin. pharmac.*, Vol. 28, 647-653.

[6] Lee, J.-Y. (1997a). Unidentifiable linear model analysis for medical models using derivative approach model, *J of Information and Optimization Sciences*, Vol. 18-1, 145-156.

[7] Lee, J.-Y. (1997b). Quasi-identifiability analysis in linear compartmental medical models, *J of Information and Optimization Sciences*, Vol. 18-3, 301-309.

[8] Lee, J.-Y. and Rhee, S.-W. (1997). 특정분포에 따른 확률 Plot들의 정규성과 Bimodality 비교, 『한국통계학회논문집』, 4권 1호, 243-254.

[9] Lee, J.-Y., Woo, J. S., and Choi, D. W. (1998). Using a normal test variable (NTV) for clinical research, 『응용통계연구』, 11권 1호, In print.

[10] Lin, C.-C. and Mudholkar, G. (1980). A simple test for normality to against asymmetric alternatives, *Biometrika*, Vol. 67, 455-461.

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

[12] Miller, C. A. *et al.* (1985). Polymorphism of theophylline metabolism in man, *J. Clin. Invest.*, Vol. 75, 1415-1425.

[13] Nakamura, K. *et al.* (1985). Interethnic differences in genetic polymorphism of debrisoquin and mephenytoin hydroxylation between Japanese and Caucasian populations, *Clin. Pharmac. Ther.*, Vol. 38, 402-408.

[14] Shapiro, C., Beckmann, E. and Christiansen, N. (1986). Immunologic status of patients remission from Hodgkin's disease and disseminated malignancies, *Am. J. Med. Sci.*, Vol. 293, 366-370.

[15] Shapiro, S. S. and Francia, R. S. (1972). An approximation analysis of variance test for normality, *J. Am. Stat. Assoc.*, Vol. 67, 215-216.

[16] Weisberg, S. and Bingham, C. (1975). An approximation analysis of variance test for non-normality suitable for machine calculation, *J. Am. Stat. Assoc.*, 17, 133-134.

[17] Wilk, M. B. and Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*. Vol. 55. 1-17.