The Korean Communications
in Statistics Vol. 6, No. 2, 1999
pp. 645-658

The Limits of Bivariate Q-Q Plots Based on Matching that Minimizes a Distance[1)]

Namhyun Kim[2)]

Abstract

One of the most popular graphical techniques for goodness of fit problems is the quantile-quantile plot (Q-Q plot). Easton and McCulloch(1990) suggested a way of generalizing Q-Q plots to multivariate cases based on finding a matching between the points of the data set whose shape is being examined and a reference sample. In this paper, we investigated the asymptotic behavior of the generalized Q-Q plot for bivariate cases. As a result, we concluded that the standard univariate Q-Q plot and the generalized Q-Q plot have the same limit if two variables are independent.

1. 서론

적합도 검정을 위하여 그래프를 이용하는 방법은 통계학에서 매우 유용하게 사용되고 있다. 그러한 방법들은 자료에 내재해 있는 사실들을 밝혀 내는데 시사하는 바가 크지만 통계량을 이용해 검정하는 방법들 보다 주관적인 것 또한 사실이다. 따라서 그래프를 이용하는 방법들은 자료에 포함되어 있는 여러 가지 관계를 이해하고, 조사하려 하는 현상을 간파하기 위한 보조적인 수단으로써 통계량을 이용한 검정법과 병행하여 이용하는 것이 바람직할 것이다. 그래프를 이용한 방법들로는 경험적 누적분포함수(empirical cumulative distribution functions), Q-Q 플롯(Q-Q plots), 퍼센트 플롯(percent plots) 등이 있다. Wilk와 Gnanadesikan(1968)은 이러한 방법들에 대해서 자세히 설명하고 그것들의 성질과 장점들에 대해서 논의하고 또한 여러 가지 실제 자료들을 통하여 그 방법들의 실용례를 보여주고 있다.

일변량 자료의 경우 적합도 검정을 위해서 가장 널리 쓰이는 그래프를 이용한 방법은 Q-Q 플롯이다. 이에 관련된 이론을 간단히 살펴보자. $X_1, \ldots, X_n$이 미지의 연속확률분포 F에서 얻어진 자료라고 가정하고 귀무가설을

$$H_0 : F(x) = F_0\left(\frac{x-\mu}{\sigma}\right) = F_0(z)$$

라고 하자. 여기서 F_0는 기지의 확률 분포이고 위치모수(location parameter) μ와 척도모수(scale parameter) σ는 미지이며 $z = \frac{x-\mu}{\sigma}$ 이다. 만일 가정된 분포가 사실이라면

1) The author wishes to acknowledge the financial support of the Korea Research Foundation made in the program year of 1997. (971030131)

2) Assistant Professor, Department of Science, Hongik University, Seoul, 121-791, Korea

$$X_i = \mu + \sigma Z_i, \quad i = 1, \ldots, n,$$

이고

$$E(X_{(i)}) = \mu + \sigma E(Z_{(i)}), \quad i = 1, \ldots, n, \tag{1.1}$$

이다. 여기에서 E는 기대치를 나타내고 $Z_1, \ldots, Z_n$은 F_0에서 얻어진 표본이고 $X_{(i)}$와 $Z_{(i)}$는 각각의 표본에서의 i번째 순서통계량을 나타낸다. 따라서 귀무가설 H_0가 사실인 경우 $(X_{(i)}, E(Z_{(i)}))$의 그래프는 근사적으로 직선이 될 것이다. 일반적으로 순서통계량의 기대치 $E(Z_{(i)})$는 대부분의 분포에서 계산하기 힘들다. 그러나 식(1.1)과 유사하게

$$F^{-1}(p_i) = \mu + \sigma F_0^{-1}(p_i)$$

이 성립한다. 여기서 F^{-1}, F_0^{-1}는 각각 F와 F_0의 역함수를 의미한다. 순서 통계량 $X_{(i)}$는 그것의 기대치 $E(X_{(i)})$의 추정치가 될 뿐 아니라 적당한 p_i값, 예를 들면 $p_i = \frac{i}{n+1}$에 대해서 p_i번째 분위수 $F^{-1}(p_i)$의 추정치로 고려될 수 있다. 따라서 표본 분위수 $X_{(i)}$와 이론적인 분위수 $F_0^{-1}(p_i)$의 그래프를 Q-Q 플롯(Quantile-Quantile plot)이라고 한다. p_i에 대한 일반적인 공식은

$$p_i = \frac{i-c}{n-2c+1}$$

로 주어진다. Blom(1958)을 보라. 여기에서 c는 $0 \le c \le 1$인 상수이다. 주로 많이 이용되는 p_i로는

$$p_i = \frac{i-0.5}{n} \quad (c = 0.5\text{일 때, Hazen(1914)})$$

$$p_i = \frac{i}{n+1} \quad (c = 0\text{일 때, Weibull(1939)})$$

$$p_i = \frac{i-0.375}{n+0.25} \quad (c = \frac{3}{8}\text{ 일 때, Blom(1958)})$$

를 들 수 있다. 적당한 p_i의 선택에 대해서는 과거 수십년 동안 많은 연구가 이루어져 왔다. 이에 관해서는 Barnett(1975), Chernoff와 Liberman(1954, 1956) Filliben(1975), Harter(1984), Kimball(1960), Looney와 Gulledge(1985) 등을 보라. 간단히 말해서 Q-Q 플롯은 표본에서의 분위수(quantiles)와 거기에 해당하는 귀무가설 하에서의 이론적인 분위수를 표시하는 것으로 만일 가정된 분포가 사실이라면 그 플롯은 직선으로 나타날 것이고 직선에서 벗어난 정도에 따라 귀무가설을 기각 또는 채택할 것이다. 이와 같은 확률플롯(probability plot)에 관한 일반적인 개념과 예제는 D'Agostino와 Stephens(1986), Wilk와 Gnanadesikan(1968)에서 자세히 다루고 있다.

그러나 이러한 방법의 다변량 자료로의 일반화에 대해서는 그리 많은 연구가 행해지지 않은 실정이다. 그 이유는 분위수의 개념을 다변량으로 간단히 확장하기 힘들기 때문이라고 생각된다. Easton과 McCulloch(1990)는 Q-Q 플롯을 다변량으로 확장하는 한가지 방법을 제안하였다. 그들의 생각은 조사 대상이 되는 자료와 가정된 분포에서 모의실험(simulation)을 통해 얻어진 표본간의 거리의 합을 최소화하는 최적 순열(permutation)을 찾는 것이다. $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$이 미지의 분포

F에서 얻어진 표본이고 $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n$은 가정된 분포 G에서 얻어진 표본이라고 하자. 우리는 F와 G가 과연 같은 분포인가 아닌가를 알고 싶은 것이다. Π_n을 $\{1, \ldots, n\}$의 모든 가능한 순열의 집합이라고 하고 $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{pi})$, $\boldsymbol{y}_j = (y_{1j}, \ldots, y_{pj})$, $\|\boldsymbol{x}_i\| = \sqrt{x_{1i}^2 + \cdots + x_{pi}^2}$이라고 하자. 즉 $\boldsymbol{x}_i$, $\boldsymbol{y}_j$는 p-변량 자료이다. Easton과 McCulloch(1990)의 방법은

$$\min_{\pi_n \in \Pi_n} \frac{1}{n} \sum_{i=1}^{n} \| \boldsymbol{y}_i - \boldsymbol{x}_{\pi_n(i)} \|^2 \tag{1.2}$$

을 만족하는 최적 순열 π_n^*를 찾아내어 $(\boldsymbol{x}_{\pi_n^*(i)}, \boldsymbol{y}_i)$을 각각의 좌표에 대해서 플롯하는 것이다. 위의 문제를 일변량의 경우에 대해서 생각해 보면 최적 순열 π_n^*는 표본 $x_1, \ldots, x_n$의 순서 통계량과 표본 $y_1, \ldots, y_n$의 순서 통계량을 대응시키는 순열이 될 것이고 이것을 플롯하면 우리는 기존의 일변량 Q-Q 플롯을 얻게 된다. 따라서 그들의 방법은 일변량 Q-Q 플롯을 다변량으로 자연스럽게 확장한 것이라고 생각될 수 있다.

위의 방법을 예를 통해서 살펴보기 위하여, 이변량 정규분포 $N_2(\boldsymbol{0}, I)$에서 표본 $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{70}$과 표본 $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{70}$을 추출하여 식(1.2)을 만족하는 최적 순열 π_n^*를 찾아보자. 여기에서 $\boldsymbol{X}_i = (X_{1i}, X_{2i})$, $\boldsymbol{Y}_i = (Y_{1i}, Y_{2i})$이다. 그림 1(a)는 X_{1i}와 Y_{1i}에 대한 일변량 Q-Q 플롯이고 그림 1(b)는 $\boldsymbol{X}$와 $\boldsymbol{Y}$의 최적 순열 π_n^*를 찾아 $(x_{1\pi_n^*(i)}, y_{1i})$를 플롯한 것이다. 기대했던 바와 같이 두 플롯 모두 직선의 형태를 보여주나 그림 1(a)에서 점들이 좀 더 직선을 중심으로 밀집되어 있고, 그림 1(b)는 점들이 직선을 따라 나타나기는 하나, 그림 1(a)보다는 직선을 중심으로 퍼져 있음을 볼 수 있다. 이와 같은 현상 때문에 Easton과 McCulloch(1990)는 그들의 플롯을 "Fuzzy Q-Q plot"이라고 부르고 있다. 두 번째 예로 $\boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_{70}$은 이변량 정규분포 $N_2(\boldsymbol{0}, I)$에서 생성하고, $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{70}$은

$$\begin{bmatrix} X_{1,1} & X_{2,1} \\ \vdots & \vdots \\ X_{1,35} & X_{2,35} \\ X_{1,36} & X_{2,36} \\ \vdots & \vdots \\ X_{1,70} & X_{2,70} \end{bmatrix} = \begin{bmatrix} Z_{1,1} & Z_{1,1} \\ \vdots & \vdots \\ Z_{1,35} & Z_{1,35} \\ -Z_{2,1} & Z_{2,1} \\ \vdots & \vdots \\ -Z_{2,35} & Z_{2,35} \end{bmatrix}$$

과 같이 생성하자. 여기에서 Z_{ij}, $i=1,2$, $j=1, \ldots, 35$, 는 정규분포 $N(0,1)$에서의 표본이다. 표본 $\boldsymbol{X}$의 주변확률분포는 역시 $N(0,1)$이므로 그림 2(a)에서 볼 수 있듯이 (X_{1i}, Y_{1i})의 일변량 Q-Q 플롯을 통해서는 정규성에 대한 가정을 기각할 수 없다. 그러나 최적 순열을 통해서 얻어진 그림 2(b)는 양쪽 끝으로 갈수록 점들이 넓게 퍼져 있음을 볼 수 있고 이를 통하여 정규성의 가정을 의심하게 된다. 이에 대한 좀 더 자세한 내용과 예제는 Eason과 McCulloch(1990)을 보라.

본 논문에서는 미지의 분포 F와 가정된 분포 G가 모두 이변량분포일 때 식(1.2)의 극한에 대해서 생각해 보고자 한다. 이를 위해서 식(1.2)의 다른 표현을 생각해 보자. 각각의 순열 $\pi_n \in \Pi_n$에 대하여

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\{(y_{1i}-x_{1\pi_n(i)})^2+(y_{2i}-x_{2\pi_n(i)})^2\} \\ &= \frac{1}{n}\sum_{i=1}^{n}\{(y_{1(i)}-x_{1(\sigma'_n(i))})^2+(y_{2[i]}-x_{2[\sigma'_n(i)]})^2\} \end{aligned} \tag{1.3}$$

을 만족하는 순열 $\sigma'_n \in \Pi_n$ 이 유일하게 존재한다. 여기에서 $y_{1(i)}, x_{1(i)}$ 는 각각 표본의 i번째 순서 통계량을 말하고, $y_{2[i]}, x_{2[i]}$ 는 i번째 순서 통계량의 concomitant(또는 induced order statistics)를 말한다. 즉, $y_{2[i]}$ 는 $y_{1(i)}$ 와 짝이 되는 값을 말한다. concomitant의 성질과 응용에 대해서는 David와 Galambos(1974), David(1982)를 보라. 식(1.3)은 i번째 관측치 (y_{1i}, y_{2i}) 를 $(x_{1\pi_n(i)}, x_{2\pi_n(i)})$ 에 대응시키는 순열 π_n 에 대하여 $(y_{1(i)}, y_{2[i]})$ 를 $(x_{1(\sigma'_n(i))}, x_{2[\sigma'_n(i)]})$ 에 대응시키는 유일한 순열 σ'_n 이 존재한다는 것을 의미한다. 주어진 순열 π_n 에 대해서, 식(1.3)을 만족하는 σ'_n 을 찾기 위해서는 우선 자료 $\boldsymbol{y}$ 를 첫 번째 변량의 순서통계량에 따라 배열한 후, $(y_{1(i)}, y_{2[i]})$ 에 대응되는 $\boldsymbol{x}$ 의 첫 번째 변량이 몇 번째 순서통계량인지를 조사하면 된다. π_n 과 σ'_n 의 일대일 대응관계는

$$R_{x_1}(\pi_n(i)) = \sigma'_n(R_{y_1}(i))$$

로부터 쉽게 보일 수 있다. 여기에서 $R_{x_1}(i), R_{y_1}(i)$ 는 각각 x_{1i}, y_{1i} 의 순위(rank)를 말한다. 따라서 식(1.2)는 $p=2$ 일 때

$$\min_{\sigma'_n \in \Pi_n} \frac{1}{n}\sum_{i=1}^{n}\{(y_{1(i)}-x_{1(\sigma'_n(i))})^2+(y_{2[i]}-x_{2[\sigma'_n(i)]})^2\} \tag{1.4}$$

이 된다. 이 경우 $\sigma'_n(i)=i$ 인 항등 순열(identity permutation)은 $(y_{1(i)}, y_{2[i]})$ 을 $(x_{1(i)}, x_{2[i]})$ 에 대응시키는 순열로, $\boldsymbol{x}$ 와 $\boldsymbol{y}$ 의 첫 번째 변량을 모두 오름차순으로 나열하는 순열임에 주의하라.

2. 주요 결과

$(\Omega, \boldsymbol{F}, \Pr)$ 을 확률공간(probability space)이라고 하고, $\boldsymbol{X}_i=(X_{1i}, X_{2i})$, $i=1,\ldots,n$, $\boldsymbol{Y}_j=(Y_{1j}, Y_{2j})$, $j=1,\ldots,n$, 을 각각 결합확률분포함수 $F(x_1, x_2)$, $G(y_1, y_2)$ 에서의 확률표본이라고 하자. 또한 F, G 는 각각 연속함수 F_1, F_2, G_1, G_2 를 주변확률분포함수(marginal distribution function)로 갖는다. $n \geq 1$ 일 때 $X_{1(1)} \leq X_{1(2)} \leq \cdots \leq X_{1(n)}$ 을 $X_{11}, X_{12}, \ldots, X_{1n}$ 의 순서통계량이라고 하고, $F_{1n}(x)$, $x \in \boldsymbol{R}$, 을

$$F_{1n}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_{1i} \in (-\infty, x]), \quad x \in \boldsymbol{R},$$

로 정의되는 $X_{11}, X_{12}, \ldots, X_{1n}$ 의 경험적 분포함수(empirical distribution function)라고 하자. 여기서 $I(A)$ 는 집합 A 의 indicator 함수이다. 또한 $F_{1n}^{-1}(y)$, $0 \leq y \leq 1$, 을

$$F_{1n}^{-1}(y) = \inf\{x: F_{1n}(x) \geq y\}$$

$$= X_{1(i)} \quad \text{만일} \quad \frac{i-1}{n} < y \leq \frac{i}{n}, \quad i=1, \ldots, n,$$

으로 정의되는 표본 분위함수(sample quantile function)라고 하자. 마찬가지로 분위함수(quantile funciton) $F_1^{-1}(y)$, $0 \leq y \leq 1$, 은

$$F_1^{-1}(y) = \inf\{x: F_1(x) \geq y\} \tag{2.1}$$

로 정의하고, 확률표본 $Y_{11}, Y_{12}, \ldots, Y_{1n}$ 에 대해서 $G_{1n}(\cdot)$, $G_{1n}^{-1}(\cdot)$, $G_1^{-1}(\cdot)$ 도 같은 식으로 정의된다.

또한 P를

$$m\{x: \sigma(x) \leq y\} = y, \quad 0 \leq y \leq 1,$$

를 만족하는 모든 측도보존변환(measure preserving transformations) σ의 집합이라고 하자. 여기서 m은 $[0,1]$에서의 Lebesgue measure이다. 1절에서의 $\sigma'_n \in \Pi_n$을 $(0,1]$에서의 함수로 나타내기 위해서, 각각의 σ'_n에 대응해서 σ_n을

$$\sigma_n(x) = x - \frac{i}{n} + \frac{1}{n}\sigma'_n(i), \quad \frac{i-1}{n} < x \leq \frac{i}{n}, \quad i=1, \ldots, n,$$

와 같이 정의하자. 즉, $\sigma_n(x)$는 각각의 구간 $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, $i=1, \ldots, n$, 에서 기울기가 1인 구간별 선형함수(piecewise linear function)이고,

$$\sigma_n\left(\frac{i}{n}\right) = \frac{\sigma'_n(i)}{n}, \quad \sigma_n\left(\frac{i-1}{n}\right) = \frac{\sigma'_n(i)}{n} - \frac{1}{n}$$

을 만족한다. 정의에 의해서, 각각의 σ'_n에 해당하는 σ_n은 $(0,1]$에서의 측도보존변환이고, 모든 σ_n의 집합 $\{\sigma_n\}$을 P_n이라고 하면, 모든 n에 대해서

$$P_n \subset P$$

를 만족한다. $\sigma'_n(i)$ 번째 순서통계량과 concomitant를 나타내기 위해서, 편의상 $X_{1(\sigma'_n(i))}$ 대신 $X_{1(\sigma_n(i))}$을, $X_{2[\sigma'_n(i)]}$ 대신 $X_{2[\sigma_n(i)]}$를 쓰기로 하자.

대수의 강법칙(strong law of large numbers)에 의해서

$$\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} \xrightarrow{a.s.} E(X_1)$$

등등이 성립하므로 식(1.4)의 극한을 찾는 것은

$$\sup_{\sigma_n \in P_n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2[\sigma_n(i)]} Y_{2[i]}\right)$$

의 극한을 찾는 것과 동일하다. 이를 위하여 $\psi_n(\sigma_n)$, $\psi(\sigma)$을 각각

$$\int h_1(F_{1n}^{-1}(\sigma_n(u))) \cdot h_2(G_{1n}^{-1}(u))du \overset{let}{=\!=} \psi_n(\sigma_n)$$

$$\int h_1(F_1^{-1}(\sigma(u))) \cdot h_2(G_1^{-1}(u))du \overset{let}{=\!=} \psi(\sigma)$$

와 같이 정의하자. 적분의 범위가 생략되어 있는 경우는 $\int_0^1$ 로 간주하기로 하자.

정리 1. h_1, h_2 가 Lebesgue measure 0인 점을 제외하고 연속인 함수이고 $E(h_1^2(X_1))<\infty$, $E(h_2^2(Y_1))<\infty$ 일 때

$$\overline{\lim_{n\to\infty}} \sup_{\sigma_n\in P_n} \psi_n(\sigma_n) \le \sup_{\sigma\in P} \psi(\sigma) \quad a.s.$$

이 성립한다.

증명 위의 정리를 증명하기 위하여, 우선

$$\left|\int h_1(F_{1n}^{-1}(\sigma_n(u)))\cdot h_2(G_{1n}^{-1}(u))du - \int h_1(F_1^{-1}(\sigma_n(u)))\cdot h_2(G_1^{-1}(u))du\right| \to 0 \text{ as } n\to\infty \tag{2.2}$$

이 됨을 보이자. Cauchy-Schwartz 부등식과 $\sigma_n\in P_n\subset P$ 로 부터

$$\begin{aligned}
&\left|\int h_1(F_{1n}^{-1}(\sigma_n(u)))\cdot h_2(G_{1n}^{-1}(u))du - \int h_1(F_1^{-1}(\sigma_n(u)))\cdot h_2(G_1^{-1}(u))du\right| \\
&\le \left|\int h_1(F_{1n}^{-1}(\sigma_n(u)))\cdot h_2(G_{1n}^{-1}(u))du - \int h_1(F_{1n}^{-1}(\sigma_n(u)))\cdot h_2(G_1^{-1}(u))du\right| \\
&\quad + \left|\int h_1(F_{1n}^{-1}(\sigma_n(u)))\cdot h_2(G_1^{-1}(u))du - \int h_1(F_1^{-1}(\sigma_n(u)))\cdot h_2(G_1^{-1}(u))du\right| \\
&\le \left[\int h_1^2(F_{1n}^{-1}(u))du\right]^{1/2}\left[\int (h_2(G_{1n}^{-1}(u)) - h_2(G_1^{-1}(u)))^2 du\right]^{1/2} \\
&\quad + \left[\int (h_1(F_{1n}^{-1}(u)) - h_1(F_1^{-1}(u)))^2 du\right]^{1/2}\left[\int h_2^2(G_1^{-1}(u))du\right]^{1/2}
\end{aligned} \tag{2.3}$$

이 성립한다. 또한 Glivenko-Cantelli 정리로부터

$$\sup_{-\infty<x<\infty} |F_{1n}(x) - F_1(x)| \xrightarrow{a.s.} 0 \text{ as } n\to\infty \tag{2.4}$$

이고 F_1 이 연속함수이므로, 모든 $\varepsilon>0$ 에 대해서

$$\sup_{\varepsilon<u<1-\varepsilon} |F_{1n}^{-1}(u) - F_1^{-1}(u)| \xrightarrow{a.s.} 0 \text{ as } n\to\infty \tag{2.5}$$

이 성립한다. 식(2.5)와 h_1 의 연속성으로부터, 모든 $\varepsilon>0$ 에 대해서

$$\int_{\{u:\varepsilon<u<1-\varepsilon\}} (h_1(F_{1n}^{-1}(u)) - h_1(F_1^{-1}(u)))^2 du \xrightarrow{a.s.} 0 \text{ as } n\to\infty \tag{2.6}$$

이 된다.

$$S_\varepsilon = \{u\in[0,1] : u<\varepsilon \text{ or } u>1-\varepsilon\}$$

$$S_\varepsilon' = \{v\in[0,1] : |u-v|<\varepsilon \text{ for some } u\in S_\varepsilon\}, \text{ i.e.,}$$

$$S_\varepsilon' = \{v\in[0,1] : v<2\varepsilon \text{ or } v>1-2\varepsilon\}$$

라고 하면, $E(h_1^2(X_1))<\infty$ 로 부터

$$\int_{S_\varepsilon} h_1^2(F_1^{-1}(u))du \to 0 \text{ as } \varepsilon\to 0 \tag{2.7}$$

이 된다. 또한 식(2.4)로부터, 어떤 자연수 n_0 가 존재하여, $n \geq n_0$ 이면, 모든 $x \in \boldsymbol{R}$ 에 대해서

$$F_{1n}(x) - \varepsilon \leq F_1(x) \leq F_{1n}(x) + \varepsilon$$

이 성립한다. 따라서 $n \geq n_0$ 인 n 에 대하여

$$\begin{aligned}
\int_{S_\varepsilon} h_1^2(F_{1n}^{-1}(u))du &= \frac{1}{n}\sum h_1^2(X_{1(i)}) \cdot I(F_{1n}(X_{1(i)}) \in S_\varepsilon) \\
&\leq \frac{1}{n}\sum h_1^2(X_{1(i)}) \cdot I(F_1(X_{1(i)}) \in S_\varepsilon') \\
&= \frac{1}{n}\sum h_1^2(X_{1i}) \cdot I(F_1(X_{1i}) \in S_\varepsilon') \qquad (2.8) \\
&\xrightarrow{a.s.} E(h_1^2(X_1) \cdot I(F_1(X_1) \in S_\varepsilon')) \text{ as } n \to \infty \\
&\to 0 \text{ as } \varepsilon \to 0
\end{aligned}$$

이 된다. 식(2.7), (2.8)과 Cauchy-Schwartz 부등식으로부터,

$$\int_{S_\varepsilon} (h_1(F_{1n}^{-1}(u)) - h_1(F_1^{-1}(u)))^2 du \xrightarrow{a.s.} 0 \text{ as } n \to \infty \text{ and } \varepsilon \to 0 \qquad (2.9)$$

이 성립하고 식(2.6)과 식(2.9)로 부터

$$\int (h_1(F_{1n}^{-1}(u)) - h_1(F_1^{-1}(u)))^2 du \xrightarrow{a.s.} 0 \text{ as } n \to \infty \qquad (2.10)$$

이 성립한다. 마찬가지로,

$$\int (h_2(G_{1n}^{-1}(u)) - h_2(G_1^{-1}(u)))^2 du \xrightarrow{a.s.} 0 \text{ as } n \to \infty \qquad (2.11)$$

도 성립한다. 그리고 식(2.2)는 식(2.3), (2.10), (2.11)과 $E(h_1^2(X_1)) < \infty$, $E(h_2^2(Y_1)) < \infty$ 로 부터 성립한다. 한편, 모든 $n \geq 1$ 에 대해서

$$\begin{aligned}
&\sup_{\sigma_n \in P_n} \int h_1(F_{1n}^{-1}(\sigma_n(u))) \cdot h_2(G_{1n}^{-1}(u))\, du \\
&\overset{let}{=} \int h_1(F_{1n}^{-1}(\sigma_n^*(u))) \cdot h_2(G_{1n}^{-1}(u))\, du \\
&\leq \left| \int h_1(F_{1n}^{-1}(\sigma_n^*(u))) \cdot h_2(G_{1n}^{-1}(u))\, du - \int h_1(F_1^{-1}(\sigma_n^*(u))) \cdot h_2(G_1^{-1}(u))\, du \right| \qquad (2.12) \\
&\quad + \sup_{\sigma \in P} \int h_1(F_1^{-1}(\sigma(u))) \cdot h_2(G_1^{-1}(u))\, du
\end{aligned}$$

이므로, 식(2.12)의 양변에 $\overline{\lim}$ 을 취하고, 식(2.2)를 이용하면 정리가 증명된다.

□

보조정리 1. 모든 $\sigma \in P$ 에 대해서

$$\sigma_n(x) \to \sigma(x) \quad \text{for almost everywhere(a.e.)} \quad x \in (0,1]$$

을 만족하는 $\{\sigma_n\ ;\ n = 1, 2, \cdots\}$ 이 존재한다.

증명 주어진 측도보존변환 $\sigma \in P$ 와 각각의 $n \in N$ 에 대해서, $A_{n,k}$ 를

$$A_{n,k}=\left\{x\in(0,1]:\frac{k-1}{n}<\sigma(x)\le\frac{k}{n}\right\},\quad k=1,\ldots,n,$$

라고 하면 $m(A_{n,k})=\frac{1}{n}$ 이다. 여기서 m 은 Lebesgue measure이다. 일반적인 measure theory로부터 (Royden(1968)을 보라)

$$m(A_{n,k}\,\Delta\,B_{n,k})<\frac{\varepsilon}{n},\quad m(B_{n,k})=\frac{1}{n}$$

을 만족하고, 각 $B_{n,k}$는 유리수를 끝점으로 하는 구간들의 합이며, 또한 $B_{n,k}$, $k=1,\ldots,n$, 간에 교집합이 없는 $\{B_{n,k}\,;\,k=1,\ldots,n\}$ 이 존재한다. 여기서 $A\Delta B$는

$$A\,\Delta\,B=(A-B)\cup(B-A)$$

으로 대칭 차집합(symmetric difference)을 의미한다. 또한 $B_{n,k}$, $k=1,\ldots,n$, 간의 양 끝점의 분모의 최소공배수를 n_c라고 하면 $B_{n,k}$는

$$B_{n,k}=\bigcup_{j=1}^{n_c/n}\left(\frac{d_{n,k,j}-1}{n_c},\frac{d_{n,k,j}}{n_c}\right]$$

와 같이 나타낼 수 있다. 다음으로 σ_{n_c}를

$$\sigma_{n_c}\left(\frac{d_{n,k,j}}{n_c}\right)=\frac{k-1}{n}+\frac{j}{n_c},\quad j=1,\ldots,\frac{n_c}{n},\quad k=1,\ldots,n,$$

$$\sigma_{n_c}(x)=x-\frac{i}{n_c}+\sigma_{n_c}\left(\frac{i}{n_c}\right)\quad \text{만일}\quad \frac{i-1}{n_c}<x\le\frac{i}{n_c},\quad i=1,\ldots,n_c,$$

와 같이 정의하자. 그러면 $\sigma_{n_c}\in P_{n_c}$ 이고, $x\in A_{n,k}\cap B_{n,k}$인 경우 $|\sigma_{n_c}(x)-\sigma(x)|\le\frac{1}{n}$ 을 만족하므로

$$m\left\{x:|\sigma_{n_c}(x)-\sigma(x)|\le\frac{1}{n}\right\}\ge 1-\varepsilon$$

이 된다. 따라서 $\frac{1}{N}<\varepsilon$인 $N\in\mathbf{N}$을 택하면, $n\ge N$인 n에 대해서

$$m\{x:|\sigma_{n_c}(x)-\sigma(x)|\le\varepsilon\}\ge 1-\varepsilon$$

이므로

$$\sigma_{n'_c}(x)\to\sigma(x)\quad \text{for almost everywhere}\quad x\in(0,1]$$

을 만족하는 $\{\sigma_{n_c}\}$의 부분열(subsequence) $\{\sigma_{n'_c}\}$가 존재한다.

□

정리 2. 정리 1의 가정하에서

$$\varliminf_{n\to\infty}\ \sup_{\sigma_n\in P_n}\psi_n(\sigma_n)\ge\sup_{\sigma\in P}\psi(\sigma)\quad a.s.$$

이 성립한다.

증명 $\sup_{\sigma\in P}\psi(\sigma)=\psi(\sigma^*)$ 라고 하면, 보조정리 1에 의해서 $\sigma_n^* \xrightarrow{a.e.} \sigma^*$ 이 성립하는 $\{\sigma_n^*\}$가 존재하고

$$\lim_{n\to\infty}\psi_n(\sigma_n^*) = \psi(\sigma^*) \quad a.s.$$

이 성립한다. 따라서

$$\sup_{\sigma\in P}\psi(\sigma)=\psi(\sigma^*)=\lim_{n\to\infty}\psi_n(\sigma_n^*) \le \underline{\lim}_{n\to\infty} \sup_{\sigma_n\in P_n}\psi_n(\sigma_n) \quad a.s.$$

으로, 정리는 성립한다.

□

정리 3. 정리 1의 가정 하에서

$$\lim_{n\to\infty}\sup_{\sigma_n\in P_n}\psi_n(\sigma_n) = \sup_{\sigma\in P}\psi(\sigma) \quad a.s.$$

이다.

증명 정리 1과 정리 2로부터 자명하다.

□

정리 4. (Hardy, Littlewood와 Pólya 1952, p.278) ϕ_1, ϕ_2 가 $(0,1)$ 에서 적분가능한 함수이고, F_{ϕ_1}, F_{ϕ_2} 를 각각 $\phi_1(U), \phi_2(U)$의 분포함수라고 할 때

$$E(\phi_1(U)\cdot\phi_2(U)) \le E(F_{\phi_1}^{-1}(U)\cdot F_{\phi_2}^{-1}(U))$$

이다. 단, U는 $(0,1)$ 에서의 균일분포이고, $F_{\phi_i}^{-1}$, $i=1,2$, 는 식(2.1)과 같이 정의된다.

□

따름정리 1. $E(X_1^2)<\infty$, $E(Y_1^2)<\infty$ 일 때

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}X_{1(i)}Y_{1(i)}=\int F_1^{-1}(u)\,G_1^{-1}(u)\,du \quad a.s.$$

이다.

증명 정리 3과 정리 4로부터 자명하다.

□

정리 5. 두 변량 X_{1i}, X_{2i}, $i=1,\ldots,n$, 이 서로 독립이고, 마찬가지로 두 변량 Y_{1j}, Y_{2j}, $j=1,\ldots,n$, 도 서로 독립일 때,

$$\lim_{n\to\infty}\sup_{\sigma_n\in P_n}\left(\frac{1}{n}\sum_{i=1}^{n}X_{1(\sigma_n(i))}Y_{1(i)}+\frac{1}{n}\sum_{i=1}^{n}X_{2[\sigma_n(i)]}Y_{2[i]}\right)$$
$$=\int F_1^{-1}(u)\cdot G_1^{-1}(u)\,du+\int F_2^{-1}(u)\cdot G_2^{-1}(u)\,du \quad a.s.$$

이다.

증명 $X_{1i}, X_{2i}, Y_{1i}, Y_{2i}$, $i=1,\ldots,n$, 이 서로 독립이므로,

$$X_{2[\sigma_n(i)]} \overset{d}{=} X_{2\sigma_n(i)}$$

$$Y_{2[i]} \overset{d}{=} Y_{2i}$$

이고, 따라서 concomitants $X_{2[1]},\ldots,X_{2[n]}$, $Y_{2[1]},\ldots,Y_{2[n]}$ 은 각각 $i.i.d.$ (independent and identically distributed) 표본이다. 여기서 $\overset{d}{=}$는 두 확률변수가 같은 분포를 가짐을 의미한다.

그러므로, 모든 σ_n, π_n에 대해서

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))}Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2[\sigma_n(i)]}Y_{2[i]} \\ \overset{d}{=}\ &\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))}Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2\pi_n(i)}Y_{2i} \end{aligned} \tag{2.13}$$

이고, 식(2.13)의 양변에 $\sup_{\sigma_n,\pi_n}$ 을 취하면 따름정리 1에 의해서,

$$\begin{aligned} &\lim_{n\to\infty}\sup_{\sigma_n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))}Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2[\sigma_n(i)]}Y_{2[i]}\right) \\ \overset{d}{=}\ &\lim_{n\to\infty}\sup_{\sigma_n,\pi_n}\left(\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))}Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2\pi_n(i)}Y_{2i}\right) \\ =\ &\int F_1^{-1}(u)\cdot G_1^{-1}(u)\,du + \int F_2^{-1}(u)\cdot G_2^{-1}(u)\,du \quad a.s. \end{aligned}$$

이 성립한다.

□

정리 6. 정리 5의 가정 하에서

$$\begin{aligned} &\lim_{n\to\infty}\inf_{\sigma_n\in P_n}\left(\frac{1}{n}\sum_{i=1}^{n}(X_{1(\sigma_n(i))}-Y_{1(i)})^2 + \frac{1}{n}\sum_{i=1}^{n}(X_{2[\sigma_n(i)]}-Y_{2[i]})^2\right) \\ =\ &\int (F_1^{-1}(u)-G_1^{-1}(u))^2\,du + \int (F_2^{-1}(u)-G_2^{-1}(u))^2\,du \end{aligned}$$

이다.

증명 대수의 강법칙(SLLN)에 의해서

$$\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} = \frac{1}{n}\sum_{i=1}^{n} X_{1i} \xrightarrow{a.s.} E(X_1) = \int \{F_1^{-1}(u)\}^2 du$$

등등이 성립하므로, 정리 5에 의해서 결과가 성립한다.

□

정리 7. (X_{1i}, X_{2i}), $i=1,\ldots,n$, 이 이변량 정규분포 $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1, \rho_1\\ \rho_1, 1\end{pmatrix}\right)$ 을 따르고,

(Y_{1i}, Y_{2i}), $i=1,\ldots,n$, 이 이변량 정규분포 $N_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1, \rho_0\\ \rho_0, 1\end{pmatrix}\right)$을 따를 때

$$\lim_{n\to\infty} \sup_{\sigma_n \in P_n} \left(\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2[\sigma_n(i)]} Y_{2[i]}\right) = (1+\rho_0\rho_1) + \sqrt{1-\rho_0^2}\sqrt{1-\rho_1^2}$$

이고

$$\lim_{n\to\infty} \inf_{\sigma_n \in P_n} \left(\frac{1}{n}\sum_{i=1}^{n} (X_{1(\sigma_n(i))} - Y_{1(i)})^2 + \frac{1}{n}\sum_{i=1}^{n} (X_{2[\sigma_n(i)]} - Y_{2[i]})^2\right)$$
$$= 4 - 2\left[(1+\rho_0\rho_1) + \sqrt{1-\rho_0^2}\sqrt{1-\rho_1^2}\right]$$

이다.

증명 이변량 정규분포의 성질에 의해서,

$$X_{2[i]} = \rho_1 X_{1(i)} + Z_{[i]},$$
$$Y_{2[i]} = \rho_0 Y_{1(i)} + W_{[i]}$$

이다. 여기서 $Z_{[i]}$, $i=1,\ldots,n$, 는 $i.i.d.$ $N(0,(1-\rho_1^2))$ 이고 $X_{1(i)}$ 에 독립이다. 마찬가지로 $W_{[i]}$, $i=1,\ldots,n$, 는 $i.i.d.$ $N(0,(1-\rho_0^2))$ 이고 $Y_{1(i)}$ 에 독립이다. David와 Galambos(1974)를 보라.

따라서 정리 5에 의해서

$$\lim_{n\to\infty} \sup_{\sigma_n \in P_n} \left(\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} X_{2[\sigma_n(i)]} Y_{2[i]}\right)$$
$$= \lim_{n\to\infty} \sup_{\sigma_n \in P_n} \left\{(1+\rho_1\rho_0)\frac{1}{n}\sum_{i=1}^{n} X_{1(\sigma_n(i))} Y_{1(i)} + \frac{1}{n}\sum_{i=1}^{n} Z_{[\sigma_n(i)]} W_{[i]}\right)$$
$$= (1+\rho_1\rho_0)\int (\Phi^{-1}(u))^2 du + \sqrt{1-\rho_1^2}\sqrt{1-\rho_0^2}\int (\Phi^{-1}(u))^2 du$$

으로 정리가 성립한다. 여기서 Φ^{-1}는 정규분포의 확률분포함수 Φ의 역함수이다.

□

3. 토의

본 논문에서는 Easton과 McCulloch(1990)가 제안한 거리를 최소화하는 순열을 이용하여 다변량으로 일반화한 Q-Q 플롯의 극한에 대해서 연구하였다. 2절의 정리 5 또는 정리 6으로부터 두 변량 X_{1i}, X_{2i}가 서로 독립일 경우에는 각각의 변량에 대해서 일변량 Q-Q 플롯을 적용하던지, 아니면 1절에서 제안한 "fuzzy Q-Q 플롯"을 적용하던지 그 극한은 같다는 것을 알 수 있다. 이는 직관적으로 우리가 기대할 수 있는 결과이며, 이변량 Q-Q 플롯이 유용한 방법이 될 수 있음을 의미하기도 한다. 그러나 그림 1(a),(b)에서 짐작할 수 있듯이 극한으로의 수렴속도는 당연히 일변량 Q-Q 플롯에서 더욱 빠를 것이다. 두 변량이 서로 독립일 경우에는 $X_{1(i)}$가 $X_{2[i]}$에 아무런 정보를 주지 못하므로 사실상 각 변량에 대해서 일변량 Q-Q 플롯을 적용하는 것이 보다 효과적일 것이다.

또한 분포 F와 G가 각각 이변량 정규분포일 경우, 이변량 Q-Q 플롯의 극한에 대해서 고려해 보았다. 정리 7에서 보는 바와 같이 $F=G$일 경우 이 극한은 0으로 수렴한다는 것을 알 수 있고 이는 이변량 Q-Q 플롯이 정규분포의 가정을 검토하는 데 유용하게 쓰일 수 있음을 암시한다. 그러나 실제로 자료가 이변량 정규분포를 따르는지 아닌지를 검정하기 위해서는 식(1.2)의 최적 순열을 찾는 것으로는 불충분하다. $\boldsymbol{X}=(X_1, X_2)$가 이변량 정규분포를 따른다면 $\boldsymbol{X}$의 affine 변환 $A\boldsymbol{X}+\boldsymbol{k}$도 역시 이변량 정규분포를 따르므로, $\boldsymbol{Y}$가 $N_2(\boldsymbol{0}, I)$를 따르는 확률변수일 때,

$$A\boldsymbol{X}+\boldsymbol{k} \overset{d}{=} \boldsymbol{Y}$$

인 행렬 A와 벡터 $\boldsymbol{k}$가 존재한다. 따라서 이 경우에는 식(1.2)의 해를 찾는 대신

$$\min_{A, \boldsymbol{k}, \pi_n} \frac{1}{n} \sum_{i=1}^{n} \| \boldsymbol{y}_i - (A\boldsymbol{x}+\boldsymbol{k})_{\pi_n(i)} \|^2 \tag{3.1}$$

을 만족하는 최적 순열을 찾아야 한다. 식(3.1)의 최적 순열은 일변량일 경우에는 A와 k에 관계없이 $x_{(i)}$와 $y_{(i)}$를 대응시키는 순열이지만, 다변량에서는 행렬 A와 벡터 $\boldsymbol{k}$에 의존하므로 구하기가 수월하지 않다. 이에 대한 좀 더 자세한 설명과 최적순열을 구하기 위한 알고리즘은 Easton과 McCulloch(1990)에 설명되어 있다.

위와 같이 affine 변환을 고려했을 경우에도 이변량 Q-Q 플롯의 극한에 대해서 생각해 볼 필요가 있다. 또한 두 분포 F와 G가 2절에서 다룬 독립인 경우나 정규분포인 경우 외에 좀 더 일반적인 분포일 경우에도 유사한 문제를 고려할 수 있고 이에 관해서는 본 논문에서의 결과를 토대로 하여 좀 더 깊이 있는 연구가 필요하다고 생각된다.

또한, 일변량 Q-Q 플롯에서는 분포의 비대칭성이나 꼬리의 정도 등을 직선에서 플롯이 벗어나는 형태로부터 쉽게 진단할 수 있는 반면, 이변량 Q-Q 플롯에서는 이와 같은 자료의 진단이 그다지 간단하지 않을 것이고, 이에 관해서도 좀 더 연구가 되어야 할 것이다.

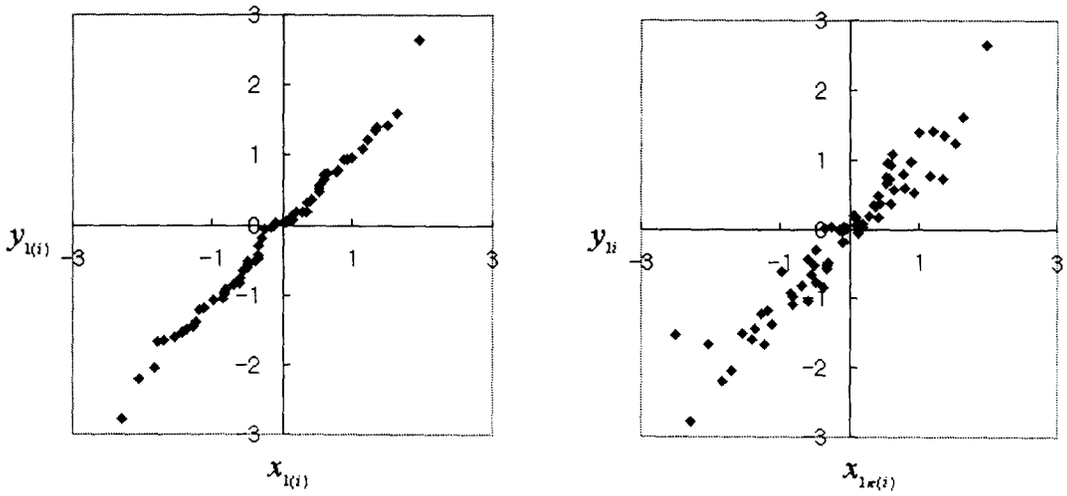

그림 1 (a), (b)

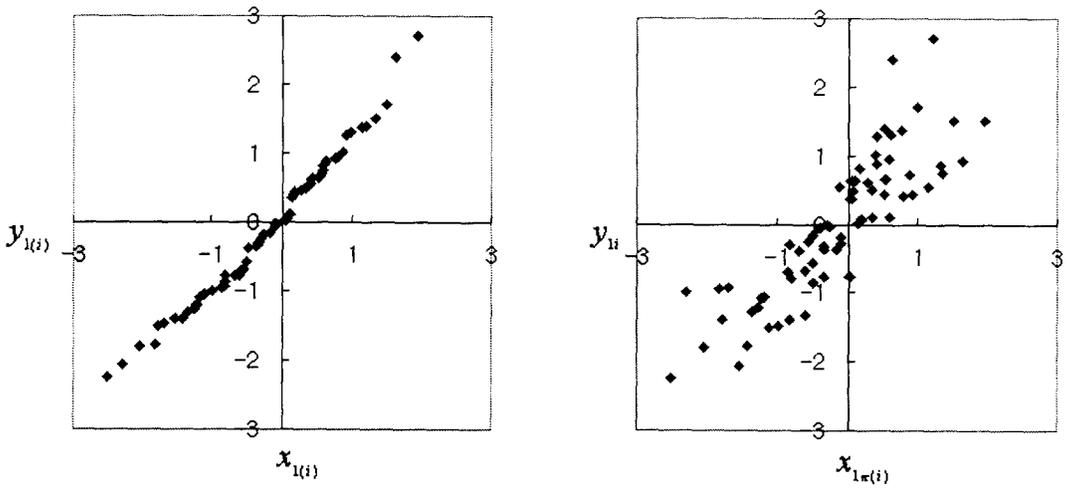

그림 2 (a), (b)

참 고 문 헌

[1] Barnett, V. (1975). Probability plotting methods and order statistics. Applied Statistics, 24, 95-108.

[2] Blom, G. (1958). Statistical estimates and transformed beta-variables. New York: Jorn Wiley.

[3] Chernoff, H. and Liberman, G. J. (1954). Use of normal probability paper. Journal of the American Statistical Association, 49, 778-785.

[4] Chernoff, H. and Liberman, G. J. (1956). The use of generalized probability paper for continuous distributions. Annals of Mathematical Statistics, 27, 806-818.

[5] D'Agostino, R. B. and Stephens M. A. (1986). Goodness-of-fit Techniques. Marcel Dekker, New York.

[6] David, H. A. (1982). Concomitants of order statistics : Theory and applications. In Some recent advances in Statistics (J. Tiago de Oliveira and B. Epstein eds.), Academic, New York.

[7] David, H. A. and Galambos, J. (1974). The asymptotic thoery of concomitants of order statistics. Journal of Applied Probability, 11, 762-770.

[8] Easton, G. S. and McCulloch, R. E. (1990). A multivariate generalization of quantile-quantile plots. Journal of the American Statistical Association, 85, 376-386.

[9] Filliben, J. J. (1975). The probability plot correlation coefficient test for normality. Technometrics, 17, 111-117.

[10] Hardy, G. H., Littlewood, J. E. and Pólya, G. (1952). Inequalities. Cambridge University Press, London.

[11] Harter, H. L. (1984). Another look at plotting positions. Communications in Statistics - Theory and Methods, 13, 1613-1633.

[12] Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. Transactons of the American Society of Civil Engineers, 77, 1529-1669.

[13] Kimball, B. F. (1960). On the choice of plotting positions on probability paper. Journal of the American statistical Association, 55, 546-560.

[14] Looney, S. W. and Gulledge, T. R. Jr. (1985). Probability plotting positions and goodness of fit for the normal distribution. The statistician, 34, 297-303.

[15] Royden, H. L. (1968). Real analysis. Macmillan Publishing Co., New York.

[16] Weibull, W. (1939). The phenomenon of rupture in solids. Ingeniors Vetenskaps Akademien Handlingar (Stockholm), 153, 17.

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