The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.037
26(1), 37–47

A Graphical Method to Assess Goodness-of-Fit for Inverse Gaussian Distribution

Byungjin Choi[a,1]

[a]Department of Applied Information Statistics, Kyonggi University

(Received November 14, 2012; Revised November 27, 2012; Accepted December 10, 2012)

Abstract

A Q-Q plot is an effective and convenient graphical method to assess a distributional assumption of data. The primary step in the construction of a Q-Q plot is to obtain a closed-form expression to represent the relation between observed quantiles and theoretical quantiles to be plotted in order that the points fall near the line $y = a + bx$. In this paper, we introduce a Q-Q plot to assess goodness-of-fit for inverse Gaussian distribution. The procedure is based on the distributional result that a transformed random variable $Y = |\sqrt{\lambda}(X - \mu)/\mu\sqrt{X}|$ follows a half-normal distribution with mean 0 and variance 1 when a random variable X has an inverse Gaussian distribution with location parameter μ and scale parameter λ. Simulations are performed to provide a guideline to interpret the pattern of points on the proposed inverse Gaussian Q-Q plot. An illustrative example is provided to show the usefulness of the inverse Gaussian Q-Q plot.

Keywords: Inverse Gaussian distribution, standard half-normal distribution, Q-Q plot, quantile.

1. 서론

Q-Q 플롯은 자료에 대한 분포적 가정을 평가하기 위해서 사용되는 편리하고 효과적인 그래프 방법이다. Q-Q 플롯은 자료의 분포와 이론적 분포를 비교하기 위한 확률플롯으로 자료에서의 분위수와 이에 대응하는 이론적 분위수를 각각 수직축과 수평축으로 해서 그린 산점도의 형태를 취한다. 플롯상에 나타난 점들이 직선의 형태를 보이면 비교하고자 하는 두 분포는 동일하다고 판단하게 된다. 그러나, 작성된 Q-Q 플롯상의 점들의 형태가 직선을 보이는지를 판단하는 것은 다분히 주관적일 수 밖에 없다. 이런 단점에도 불구하고 Q-Q 플롯이 자료분석에서 유용한 주된 이유는 다음과 같이 들 수 있다.

자료의 분포와 이론적 분포의 일치 여부를 객관적으로 판단하려면 분포에 대한 가설검정을 수행해 보는 것이다. 그러나, 검정결과는 영가설의 채택 유무만을 제공할 뿐이어서 영가설이 기각이 되어졌다면 자료가 어떤 분포를 하는지를 알 수가 없다. 이에 반해 Q-Q 플롯에서는 특이값 존재, 분포의 꼬리형태, 비대칭성 여부 등의 유용한 정보를 점들의 형태를 통해 파악할 수 있기 때문에 자료의 분포에 대한 유추가 어느 정도 가능하다. Q-Q 플롯은 또한 가정된 자료의 분포에서 최대가능도법과 같은 통상적인 방식의 적용에 의한 모수의 추정이 여의치 않는 경우에 추정된 모수를 얻기 위한 도식수단으로 사용할 수가

[1]Associate Professor, Department of Applied Information Statistics, Kyonggi University, Iui-Dong, Yeongtong-Gu, Suwon-Si, Gyeonggi-Do 443-760, Korea. E-mail: bjchoi92@kyonggi.ac.kr

있다. 정규분포를 포함한 다양한 확률분포에 대한 Q-Q 플롯의 작성절차와 해석방법은 Chambers 등 (1983)에 상세히 소개되어 있으므로 참고하기 바란다

역가우스분포 IG(μ, λ)를 따르는 확률변수 X는 확률밀도함수로

$$f_X(x;\mu,\lambda) = \sqrt{\frac{\lambda}{2\pi x^3}} \exp\left\{-\frac{\lambda(x-\mu)^2}{2\mu^2 x}\right\}, \quad x>0,\, \mu>0,\, \lambda>0 \tag{1.1}$$

을 가지게 된다. 여기서, μ와 λ는 각각 위치와 척도를 나타내는 모수들이다. 브라운운동에서 첫 통과 시간의 분포로 Schrödinger (1915)와 Smoluchowsky (1915)에 의해 독립적으로 처음 제시된 역가우스분포는 Tweedie (1957a, 1957b)의 초기 연구 이후에 자료분석을 위한 확률모형으로 많은 관심을 받아왔다. Chhikara와 Folks (1989), Seshadri (1999)는 오른쪽으로 긴 꼬리를 보이는 자료를 분석하기 위한 확률모형으로 역가우스분포의 유용성을 생물학, 생태학, 환경연구, 공학, 경영과학, 약물동태학, 공학, 품질관리, 신뢰성과 생존분석 등 다양한 분야에서의 사례를 통해 소개했다. 분석에 앞서서 확률모형으로 사용할 역가우스분포의 적절성을 이론적 또는 응용적 측면에서 점검하는 것은 매우 중요하다. 분석할 자료가 역가우스분포를 따르는 지를 알아보기 위해서는 분포적 가설검정을 수행하거나 Q-Q 플롯을 작성해 보면 된다. 가설검정의 경우는 Edgeman 등 (1988), Edgeman (1990), Pavur 등 (1992), Mudholkar와 Tian (2002) 등이 제안한 적합도 검정을 이용하면 된다. 하지만 Q-Q 플롯의 경우는 구축과 관련된 방법론이 제시되어 있지 않다. Q-Q 플롯의 구축에서 주된 단계는 플롯상의 점들이 직선의 형태로 나타나게끔 자료에서의 분위수와 비교하고자 하는 확률분포의 누적분포함수의 역함수로부터 계산된 이론적 분위수와의 표현식으로부터 수평축과 수직축을 설정하는 것이다. 그런데, 역가우스분포의 경우, 좌표축의 설정을 위해 요구되는 폐쇄형의 표현식을 얻기가 가능하지 않다. 아마도 이런 이유로 인해 Q-Q 플롯을 구축할 수가 없었된 것으로 판단된다.

확률변수 X가 위치모수 μ와 척도모수 λ를 가지는 역가우스분포 IG(μ, λ)를 따르면, 변환된 확률변수 $Y = |\sqrt{\lambda}(X-\mu)/\mu\sqrt{X}|$는 평균이 0, 분산이 1인 표준반접정규분포(standard half-normal distribution) HN$(0,1)$를 하게 된다 (Chhikara와 Folks, 1989). 본 논문에서는 이 분포적 결과를 활용하여 역가우스분포에 대한 적합을 알아보기 위한 Q-Q 플롯(이하 역가우스분포 Q-Q 플롯)의 구축방법을 제안한다. 2장에서는 역가우스분포 Q-Q 플롯을 얻기 위한 절차를 다룬다. 3장에서는 역가우스분포와 다른 분포를 따르는 자료를 대상으로 그린 Q-Q 플롯에서 나타나는 점들의 형태를 알아보고자 모의실험을 수행하고 그 결과를 제시한다. 4장에서는 실제 자료에 대한 사례분석을 통해 제안한 Q-Q 플롯의 유용성을 보인다. 끝으로 5장에서는 결론을 내린다.

2. 역가우스분포 Q-Q 플롯

IG(μ, λ)를 따르는 확률변수 X의 누적분포함수는

$$F_X(x;\mu,\lambda) = \Phi\left[\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}-1\right)\right] + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\frac{\lambda}{x}}\left(\frac{x}{\mu}+1\right)\right] \tag{2.1}$$

로 주어진다 (Shuster, 1968). 여기서, Φ는 표준정규분포의 누적분포함수이다. 이 분포함수에 기초한 역가우스분포의 p 분위수는

$$p = F_X(x_p;\mu,\lambda) = \Phi\left[\sqrt{\frac{\lambda}{x_p}}\left(\frac{x_p}{\mu}-1\right)\right] + \exp\left(\frac{2\lambda}{\mu}\right)\Phi\left[-\sqrt{\frac{\lambda}{x_p}}\left(\frac{x_p}{\mu}+1\right)\right] \tag{2.2}$$

를 만족하는 x_p가 된다. x_p를 찾기 위해서는 식 (2.2)로부터 p에 대한 표현식을 도출해야만 한다. 그런데 표현식을 폐쇄형으로 얻는 것이 가능하지가 않기 때문에 Chhikara와 Folks (1989)에 의해 제시된 분포적 결과를 활용하여 역가우스분포 Q-Q 플롯의 구축을 시도하기로 한다.

$X \sim \mathrm{IG}(\mu, \lambda)$일 때 $W = \sqrt{\lambda}(X-\mu)/\mu\sqrt{X}$의 밀도함수는

$$f_W(w;\mu,\lambda) = \left(1 - \frac{w}{\sqrt{4\lambda/\mu + w^2}}\right) \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{w^2}{2}\right), \quad \infty < w < \infty \tag{2.3}$$

가 되고 $Y = |W|$는

$$f_Y(y) = \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \tag{2.4}$$

을 밀도함수로 가지는 평균 0이고 분산 1인 표준반접정규분포 HN(0, 1)를 하게 된다. 이론적 분위수로 사용할 HN(0, 1)의 p 분위수는 Y의 분포함수가 $F_Y(y) = 2\Phi(y) - 1$로 주어지므로

$$p = F_Y(x_p) = 2\Phi(x_p) - 1 \tag{2.5}$$

을 만족하는 x_p가 된다. 식 (2.5)를 p에 대해서 풀게 되면 역가우스분포 Q-Q 플롯의 구축을 위해 필요한 다음의 표현식을 얻게 된다.

$$x_p = \Phi^{-1}\left(\frac{p+1}{2}\right). \tag{2.6}$$

자료에서의 분위수 추정을 위해서 크기 n인 원자료 $x_1, x_2, \ldots, x_n$이 임의의 분포에서 추출되었다고 하자. x_i들을 변환한 $y_1, y_2, \ldots, y_n$을 얻기 위해서 $\mu = \mu_0$와 $\lambda = \lambda_0$로 주어지면

$$y_i = \left|\frac{\sqrt{\lambda_0}(x_i - \mu_0)}{\mu_0\sqrt{x_i}}\right| \tag{2.7}$$

을 이용한다. 그리고 μ와 λ가 알려져 있지 않는 경우에는 각각의 추정치인 $\bar{x}$와 $(n-1)/v$로 대체한

$$y_i = \left|\frac{\sqrt{n-1}(x_i - \bar{x})}{\bar{x}\sqrt{vx_i}}\right| = \left|\sqrt{\frac{n-1}{v}}\left(\frac{\sqrt{x_i}}{\bar{x}} - \frac{1}{\sqrt{x_i}}\right)\right| \tag{2.8}$$

을 적용한다. 여기서, $v = \sum_{i=1}^{n}(1/x_i - 1/\bar{x})$이다. 이렇게 얻은 변환자료를 크기순으로 나열한 순서자료를 $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$으로 표기하기로 한다. 자료값들에 대한 분위수 $p_i = P[Y \le y_{(i)}]$, $i = 1, \ldots, n$은 통상적으로 경험적 분포함수 F_n을 이용하여 추정하게 된다. 제안된 여러 방법들 중에서 Blom (1958)에 의해 정의된

$$\hat{p}_i = F_n\left[y_{(i)}\right] = \frac{i - 3/8}{n + 1/4}, \quad i = 1, \ldots, n \tag{2.9}$$

를 추정분위수로 사용하기로 한다. 원자료 $x_1, x_2, \ldots, x_n$이 IG(μ, λ)에서 추출되었다면 변환된 순서자료 $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$는 HN(0, 1)를 따르게 되고 $\hat{p}_i$ 분위수에 해당하는 이론적인 값은 식 (2.6)에서 p를 $\hat{p}_i$으로 대체한

$$\widehat{x_{p,i}} = \Phi^{-1}\left(\frac{\hat{p}_i + 1}{2}\right), \quad i = 1, \ldots, n \tag{2.10}$$

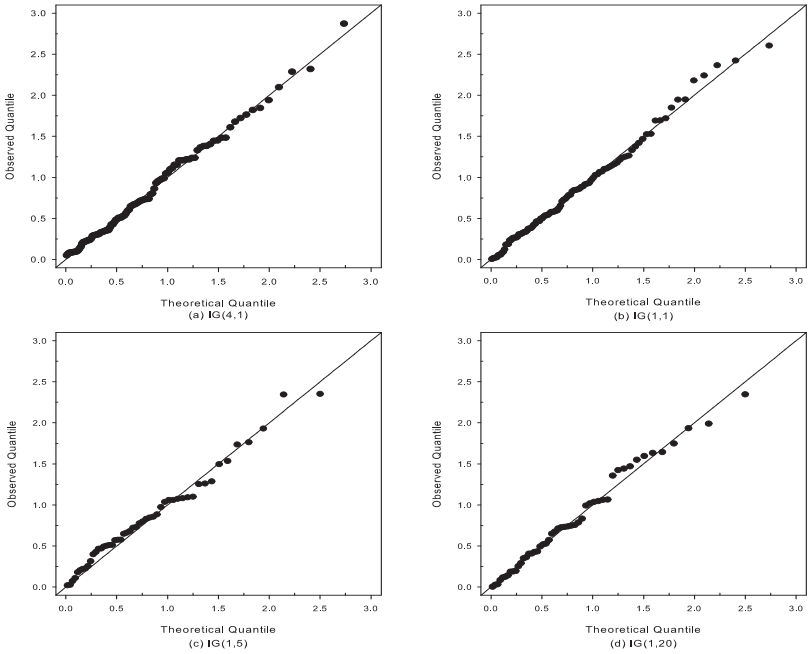

Figure 3.1. Q-Q plots of simulated data from inverse Gaussian distributions with $\phi = \lambda/\mu = 0.25, 1, 5, 20$. (a) and (b) were drawn based on data with $n = 100$. (c) and (d) were drawn based on data with $n = 50$.

가 된다. 따라서, x_i들의 확률분포가 IG(μ, λ)를 하게 된다면 $y_{(i)} \cong \widehat{x_{p,i}}$가 예상되므로 관계식

$$y_{(i)} \cong \Phi^{-1}\left(\frac{\hat{p}_i + 1}{2}\right), \quad i = 1, \ldots, n \tag{2.11}$$

이 성립하게 될 것이다. 그러므로, 수평축과 수직축을 각각 $\Phi^{-1}(\hat{p}_i/2 + 1/2)$와 $y_{(i)}$로 하여 n개의 좌표점 $(\widehat{x_{p,i}}, y_{(i)})$를 찍은 역가우스분포 Q-Q 플롯은 원점을 지나는 기울기 1인 직선의 경향선을 보이게 된다.

3. 모의실험 결과

주어진 자료가 역가우스분포를 충실하게 따른다면 제안한 Q-Q 플롯에서 점들은 직선의 형태로 나타날 것이다. 만일 플롯상의 점들이 직선의 경향선으로부터 이탈되어 있는 형태를 보인다면 자료가 역가우스분포가 아닌 다른 분포를 따른다고 판단할 것이다. 그렇다면, 후자의 경우에, 자료가 어떠한 분포에서 추출되었다고 할 수 있는가? 이 물음에 답하기 위해서는 역가우스분포와 다른 특정분포를 따르는 자료를 대상으로 그린 Q-Q 플롯에서 나타나는 점들의 형태를 토대로 자료의 분포를 유추하기 위한 역가우스분포 Q-Q 플롯의 해석지침이 필요하다. 이런 목적으로 여러 분포들에서 생성한 자료를 이용하여 그린 역가우스분포 Q-Q 플롯에서 점들이 어떠한 형태로 나타나는지 알아보기 위해 모의실험을 수행했다.

Figure 3.1은 Michael 등 (1976)이 제시한 알고리즘을 이용하여 역가우스분포로부터 생성한 모의자료를 대상으로 그린 Q-Q 플롯이다. Figure 3.1(a)와 (b)는 IG$(4, 1)$과 IG$(1, 1)$로부터 각각 독립적으로 생성하여 얻은 $n = 100$인 자료를 가지고 그린 Q-Q 플롯이고 Figure 3.1(c)와 (d)는 IG$(1, 5)$와

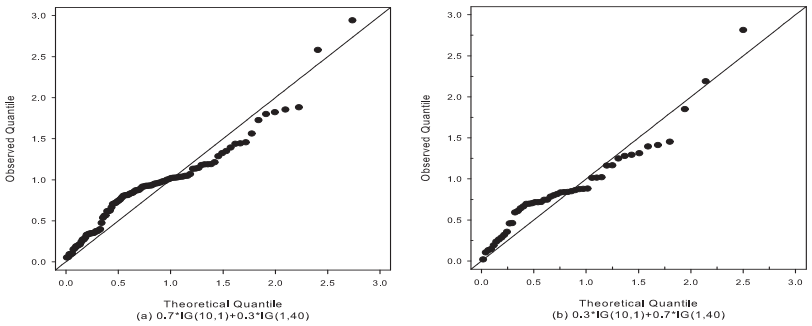

Figure 3.2. Q-Q plots of simulated data from mixed inverse Gaussian distributions. Left figure was drawn based on data with $n = 100$. Right figure was drawn based on data with $n = 50$.

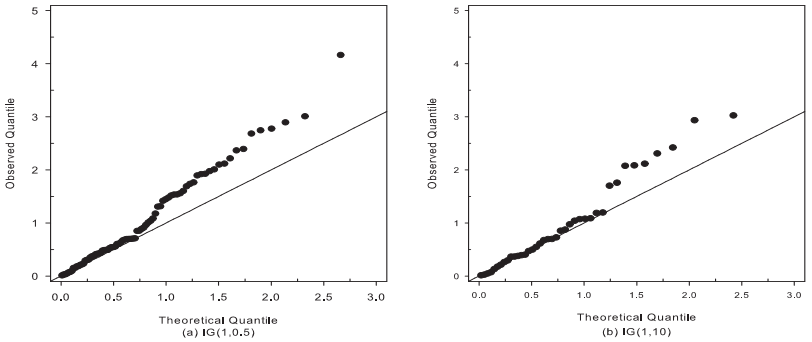

Figure 3.3. Q-Q plots of simulated data from inverse Gaussian distribution truncated at both tails. Left figure was drawn based on data with $n = 100$. Right figure was drawn based on data with $n = 50$.

IG$(1, 20)$로부터 추출한 $n = 50$인 자료에 대한 것이다. 각 그림에서의 실선은 자료의 적합 여부의 판단에 참조하기 위한 경향선을 나타낸 것이다. 모든 플롯에서 거의 대부분의 점들은 표본크기에 상관없이 일직선위에 놓여 있음을 볼 수 있다.

Figure 3.2는 혼합역가우스분포로부터 온 자료의 Q-Q 플롯을 보여준다. 플롯의 작성을 위해 사용할 모의자료는 IG$(10, 1)$와 IG$(1, 40)$에서 크기가 각각 $(n_1, n_2) = (70, 30), (15, 35)$인 자료를 Michael 등 (1976)이 제시한 알고리즘을 이용하여 생성한 다음 이것을 혼합시켜서 만들었다. 왼쪽 Figure 3.2 (a)는 $(n_1, n_2) = (70, 30)$인 혼합자료에 대한 플롯으로 점들이 직선의 경향선에서 많이 이탈되어져 있는 것을 볼 수 있다. 플롯에서 왼쪽 하단 부분의 점들은 경향선의 위에 놓여 있어 위로 볼록한 모양으로 나타나는 반면에 중간 부분의 점들은 경향선의 아래 쪽에 있어 아래로 오목한 모양을, 오른쪽 상단 부분의 점들은 직선과 유사한 모양을 보인다. 플롯은 전반적으로 오른쪽으로 비스듬히 놓여 있는 N자를 양 끝에서 잡아 당겨서 늘려놓은 형태로 나타나고 있어 세 개의 직선이 존재함을 알 수 있다. 크기 (n_1, n_2)이 $(15, 35)$인 혼합자료의 플롯(Figure 3.2(b)) 또한 Figure 3.2(a)와 비슷한 형태로 나타나고 있어서 Figure 3.2(a)에서와 유사한 해석을 얻게 된다. 혼합분포로부터 생성한 모의자료로부터 그린 Q-Q 플롯 역시 표본크기에 상관없이 거의 유사한 형태를 보이는 것을 알 수 있다.

양쪽 꼬리의 끝이 잘려져 있는 분포를 따르는 자료에서의 플롯 형태를 보기 위하여 역가우스분포 IG$(1, 0.5)$에서 $n = 100$인 모의자료를 Michael 등 (1976)의 알고리즘에 따라 생성했다. 그런 다음

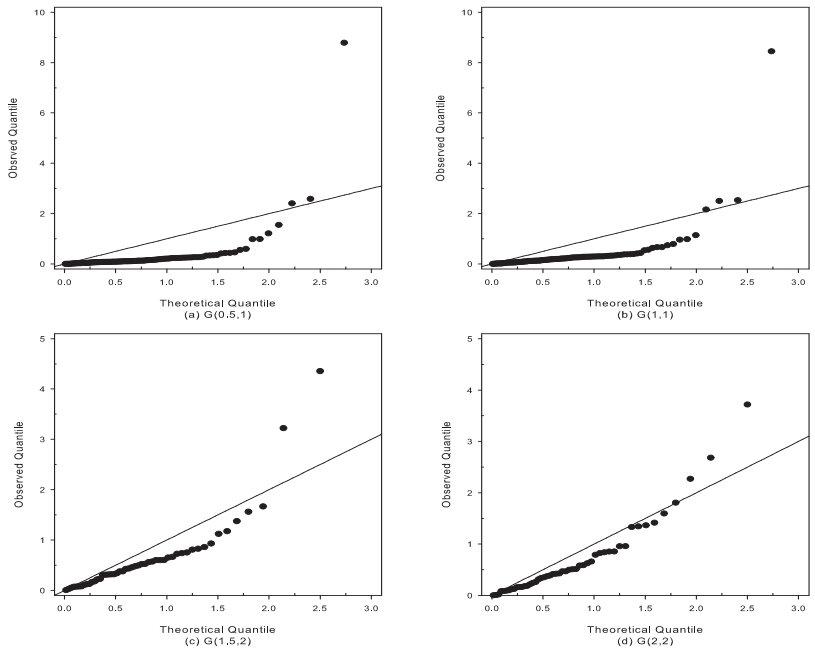

Figure 3.4. Q-Q plots of simulated data from gamma distributions. (a) and (b) were drawn based data with $n = 100$. (c) and (d) were drawn based on data with $n = 50$.

플롯의 작성을 위해서 각 자료를 크기순으로 나열하여 아래에서 10개, 위에서 10개를 버린 나머지 80개로 절단된 자료를 만들었다. 또한 IG$(1, 10)$으로부터 생성한 크기 $n = 50$인 모의자료를 기초로 IG$(1, 0.5)$의 경우와 동일한 방식으로 아래에서 5개, 위에서 5개를 버린 40개를 이용하여 플롯 작성에 사용할 자료를 만들었다. Figure 3.3의 왼쪽은 IG$(1, 0.5)$로부터 구성한 절단된 자료를 바탕으로 그린 역가우스분포 Q-Q 플롯을 보여 준다. 플롯에서 왼쪽 하단의 점들은 대체로 경향선 위에 놓여 있게 되고 오른쪽 상단으로 갈수록 경향선 위쪽으로 이탈된 모습으로 나타난다. 플롯의 형태는 전반적으로 비스듬한 S자를 길게 늘려드린 모양을 취하고 있어서 적어도 2개의 직선이 존재함을 알 수 있다. IG$(1, 10)$에서 만든 자료로부터 작성한 Figure 3.3(b)에서도 Figure 3.3(a)와 거의 동일한 현상을 발견하게 된다. 절단자료의 경우 역시 표본크기에 상관없이 플롯상에 나타난 점들은 대체적으로 유사한 형태를 가짐을 볼 수 있다.

오른쪽으로 치우침을 보이는 자료의 분석에서 감마분포, 와이블분포와 로그정규분포는 역가우스분포의 대안으로 많이 활용된다. 이들 분포에서 추출한 자료의 경우에 역가우스분포 Q-Q 플롯상에서 점들이 어떤 형태로 나타나는지를 살펴볼 필요가 있다. Figure 3.4는 감마분포 $G(\alpha, \beta)$에서 $(\alpha, \beta) = (0.5, 1), (1, 1)$로 해서 생성한 $n = 100$인 자료들로부터 얻은 Q-Q 플롯이다. 플롯 작성에 사용한 모의자료의 생성은 IMSL의 DRNGAM 부프로그램을 이용하였다. 제시된 플롯의 위에 있는 Figure 3.4(a)와 (b)는 α와 β가 각각 $(0.5, 1)$과 $(1, 1)$인 감마분포로부터 얻은 $n = 100$인 자료에 대한 것으로 대부분의 점들이 실선의 경향선 아래에 놓여 있는 것을 볼 수 있다. 플롯의 형태는 비스듬히 누워있는 J자 모양으로 나타난다. α와 β가 각각 $(1.5, 2)$과 $(2, 2)$인 감마분포에서 생성한 $n = 50$인 자료에 대한 아래쪽 Figure 3.4(c)와 (d)에서도 윗쪽 그림들에서와 비슷한 현상이 포착된다. 플롯상에 나타난 점들의 형태는 표본크기에 상관없이 거의 유사함을 알 수 있고 α와 β가 커짐에 따라 대부분의 점들에서 보였던 경

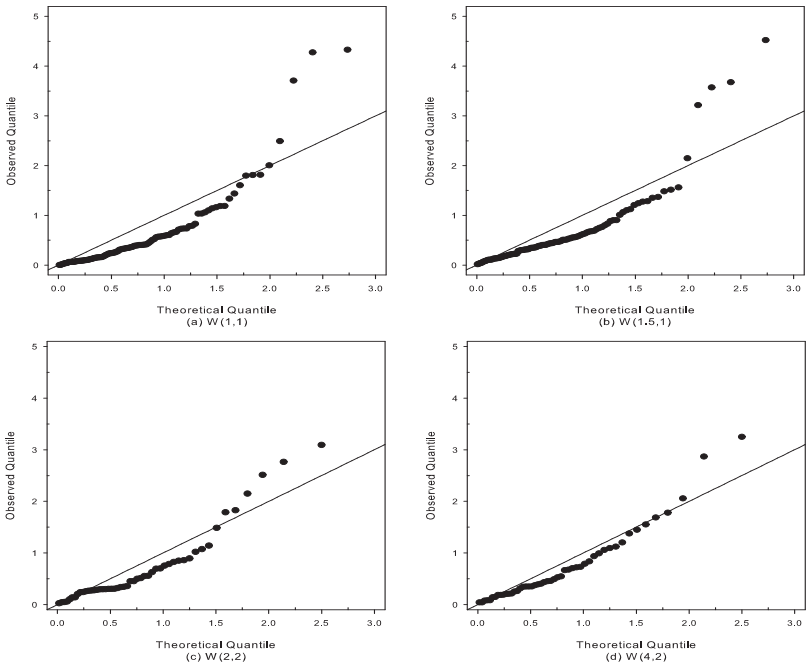

Figure 3.5. Q-Q plots of data generated from Weibull distributions. (a) and (b) were drawn based data with $n = 100$. (c) and (d) were drawn based on data with $n = 50$.

향선으로부터의 이탈 정도가 완화됨을 볼 수 있다.

와이블분포 $W(\alpha, \beta)$의 경우는 IMSL의 DRNWIB 부프로그램을 사용하여 $(\alpha, \beta) = (1, 1), (1.5, 1)$로 해서 $n = 100$를 모의자료를 생성했다. 또한 $(\alpha, \beta) = (2, 2), (4, 2)$인 와이블분포로부터 $n = 50$인 자료를 추출했다. Figure 3.5는 생성된 이들 모의자료로부터 얻은 Q-Q 플롯이다. 위쪽에 있는 Figure 3.5(a)와 (b)는 α와 β가 각각 $(1, 1)$과 $(1.5, 1)$에 대한 플롯으로 점들의 대부분이 경향선으로 표시한 실선의 직선 밑에 위치하고 있고 아래로 오목한 곡선적인 모습으로 나타난다. 이에 반해 경향선에서 이탈되어 있는 오른쪽 위에 있는 몇 개의 점들은 위로 볼록한 모양을 보인다. α와 β가 각각 $(2, 2)$와 $(4, 2)$인 자료를 이용하여 그린 아래쪽에 전시된 플롯들에서도 정도의 차이가 있지만 위쪽의 그림에서와 유사한 모양을 볼 수 있다. 플롯의 형태를 보면 표본크기와 무관하게 전반적으로 비스듬한 S자 성장곡선으로 나타남을 알 수 있다.

Figure 3.6의 위쪽에 전시된 (a)와 (b)는 μ와 σ^2을 $(0, 0.5), (0, 1)$로 한 로그정규분포에서 IMSL의 DRNLNL 부프로그램을 사용하여 추출한 $n = 100$인 모의자료의 Q-Q 플롯이다. 감마분포와 외이블분포의 자료에 비해, 모든 플롯에 나타난 점들의 경향선으로부터 이탈된 정도는 심하지 않는 것을 볼 수 있다. 점들은 왼쪽 아래와 중간 부분에서는 실선의 경향선에 근접되어 있는 일직선으로 나타나고 오른쪽 위에서는 볼록한 보습을 보인다. Figure 3.6의 아래쪽에 전시된 (c)와 (d)는 μ와 σ^2을 $(0, 0.5), (0, 1)$로 한 로그정규분포에서 동일한 IMSL 부프로그램을 가지고 생성한 $n = 50$인 모의자료의 Q-Q 플롯으로, 점들의 형태는 형태는 (a)와 (b)에서 관측한 것과 동일한 현상을 보여 준다. 플롯의 형태는 표본크기에 무관하게 대체적으로 거의 비슷하게 나타나고 전반적으로 마치 숟가락을 엎어놓은 듯한 모습을 보이고 있음을 알 수 있다.

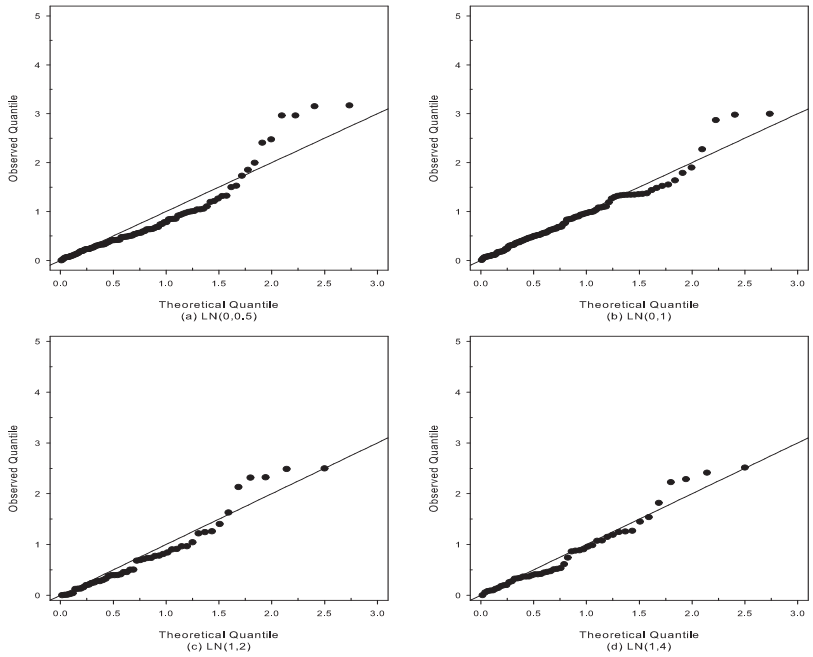

Figure 3.6. Q-Q plots of data generated from lognormal distributions. (a) and (b) are drawn based data with $n = 100$. (c) and (d) are drawn based on data with $n = 50$.

4. 사례분석

제안한 역가우스분포 Q-Q 플롯의 예시를 위해 사용한 $n = 45$인 fluid failure 자료는 전압량 30kV, 32kV와 34kV 하에서 액체 절연재의 고장시간을 분 단위로 측정한 것으로 Nelson (1975)에 실려 있는 자료의 일부분이다. Figure 4.1의 왼쪽은 자료로부터 얻은 역가우스분포 Q-Q 플롯으로 점들은 실선으로 표시된 경향선에서 많이 이탈되어 있는 것을 볼 수 있다. 점들의 형태는 와이블분포를 따르는 자료로부터 작성된 Q-Q 플롯(Figure 3.5) 또는 감마분포를 따르는 자료로부터 작성된 Q-Q 플롯(Figure 3.4)와 유사함을 알 수 있다. Figure 4.1의 오른쪽은 표준반접정규분포 HN$(0, 1)$에 대한 적합을 보기 위해 자료로부터 계산된 $\bar{x} = 38.2633$과 $v = 19.4915$를 식 (2.8)에 대입한 변환을 통해 얻은 자료 $y_1, y_2, \ldots, y_{45}$의 막대그래프와 HN$(0, 1)$의 밀도함수를 겹쳐 그려놓은 것이다. 플롯에서 보듯이 자료가 잘 적합이 되지 않음을 알 수 있다.

추가적으로 자료에 대해서 Pavur 등 (1992)과 Mudholkar와 Tian (2002)의 적합도 검정들을 수행해 보았다. Pavur 등 (1992)이 제시한 검정은 $D_n^*(\phi) = A^2(\sqrt{n} + \hat{\beta}_1(\phi)/\sqrt{n} + \hat{\beta}_2(\phi)/n)$를 검정통계량으로 사용한다. 여기서, $\phi = \lambda/\mu$는 역가우스분포의 형상모수이고 A^2은 앤더슨-달링 검정통계량이다. 또한 $\hat{\beta}_1(\phi)$, $\hat{\beta}_2(\phi)$는 검정통계량의 계산에 필요한 회귀계수들이다. 자료로부터 계산한 앤더슨-달링 검정통계량은 $A^2 = 5.4599$이고 기각값 결정을 위해 필요한 형상모수로 $\phi = 0.059$를 얻게 된다.

그런데, Pavur 등 (1992)이 제시한 표에서, 계산된 ϕ값에 해당하는 회귀계수들이 주어져 있지 않아서 검정통계량을 계산할 수 없으므로 보간법을 적용하여 검정통계량의 값을 계산해야만 한다. 가장 가까운 두 값인 $\phi = 0.5, 0.001$에 대한 회귀계수 $\hat{\beta}_1\,(0.5) = 34.2958$, $\hat{\beta}_2\,(0.5) = 17.7135$와 $\hat{\beta}_1\,(0.001) = 31.662$, $\hat{\beta}_2\,(0.001) = 18.9032$를 적용해서 얻은 $D_n^*(0.5) = 62.3906$과 $D_n^*(0.001) = 60.1030$를 이용하면 보간된

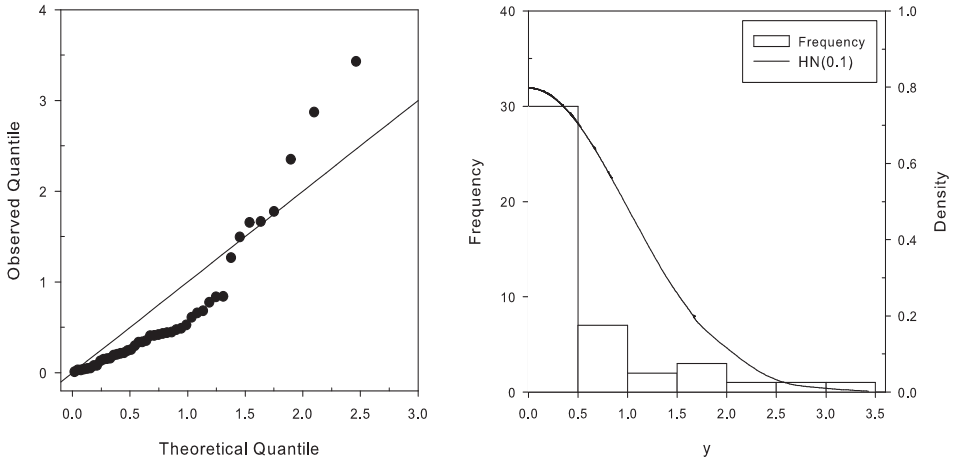

Figure 4.1. Q-Q plot and histogram of fluid failure data.

검정통계량의 값은 $D_n^*(0.059) = 60.3689$가 된다. $\phi = 0.059$일 때의 유의수준 5%에 대한 기각값 또한 표에 제시되어 있지 않으므로 $\phi = 0.5, 0.001$일 때의 기각값 $C(0.5) = 9.9942$와 $C(0.001) = 14.7013$을 사용하여 보간하면 $C(0.059) = 14.1542$가 된다. $D_n^*(0.059) = 60.3689 > C(0.059) = 14.1542$이므로 주어진 자료는 역가우스분포를 하지 않음을 알 수 있다.

Mudholkar와 Tian (2002)이 제안한 검정의 경우, Vasicek (1976)의 권고에 따라 윈도크기를 $m = 4$로 하여 검정통계량을 계산해 보면 $K_{4,45} = 2.8371$가 된다. 검정통계량의 값이 $n = 45$와 $m = 4$일 때 유의수준 5%에 대한 기각값 $K_{4,45}(0.05) = 3.2735$보다 작게 되므로 자료가 역가우스분포를 따른다는 영가설을 기각하게 되므로 Pavur 등 (1992)의 검정과 동일한 결론을 얻게 된다.

이제, Figure 4.1의 Q-Q 플롯에서 관측한 점들의 형태에 관해서 얻은 정보를 바탕으로 와이블분포의 적합을 시도해 보기로 한다. SAS의 UNIVARIATE 프로시저로 분석한 적합도 검정결과에서, 계산된 크래머-미제스와 앤더슨-달링과 크래머-미제스 검정통계량은 각각 $U^2 = 0.0289$와 $A^2 = 0.2402$가 된다. 두 값 모두는 0.25보다 큰 유의확률은 가지게 되므로 유의수준 5%에서 유의하다. 그러므로, 자료는 와이블분포에 잘 적합이 되는 것으로 판단할 수 있다.

5. 결론

Q-Q 플롯은 자료에 대한 분포적 가정을 평가하기 위해서 사용되는 편리하고 효과적인 그래프 방법이다. Q-Q 플롯의 구축에서 주된 단계는 플롯상의 점들이 직선의 형태로 나타나게끔 자료에서의 분위수와 비교하고자 하는 확률분포의 누적분포함수의 역함수로부터 계산된 이론적 분위수와의 표현식으로부터 수평축과 수직축을 설정하는 것이다. 그런데, 역가우스분포의 경우에는 좌표축의 설정을 위해 요구되는 표현식이 폐쇄형으로 주어지지 않기 때문에 Q-Q 플롯을 구축할 수가 없다.

본 논문에서는 확률변수 X가 위치모수 μ와 척도모수 λ를 가지는 역가우스분포를 따르면, 변환된 확률변수 $Y = |\sqrt{\lambda}(X-\mu)/\mu\sqrt{X}|$는 평균이 0, 분산이 1인 표준반접정규분포를 하게 되는 분포적 결과를 활용하여 역가우스분포 Q-Q 플롯의 구축방법을 소개했다. 이와 함께, 역가우스분포와 다른 특정분포를 따르는 자료를 대상으로 그린 Q-Q 플롯에서 나타나는 점들의 형태를 토대로 자료의 분포를 유추하기

위한 Q-Q 플롯의 해석에 관한 지침을 제공하기 위해 모의실험을 수행했다. 실제 자료에 대한 사례분석을 통해 제안한 Q-Q 플롯의 유용성을 보였다.

References

Blom, G. (1958). *Statistical Estimates and Transformed Beta Variables*, John Wiley, New York.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, Boston.

Chhikara, R. S. and Folks, J. L. (1989). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications*, Marcel Dekker, New York.

Edgeman, R. L. (1990). Assessing the inverse Gaussian distribution assumption, *IEEE Transactions on Reliability*, **39**, 352–355.

Edgeman, R. L., Scott, R. C. and Pavur, R. J. (1988). A modified Kolmogorov-Smirnov test for the inverse density with unknown parameters, *Communications in Statistics-Simulation and Computation*, **17**, 1203–1212.

Michael, J. R., Schucany, W. R. and Haas, R. W. (1976). Generating random variables using transformation with multiple roots, *The American Statistician*, **30**, 88–90.

Mudholkar, G. S. and Tian, L. (2002). An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test, *Journal of Statistical Planning and Inference*, **102**, 211–221.

Nelson, W. B. (1975). Analysis of accelerated life test data-least squares methods for the inverse power law model, *IEEE Transactions on Reliability*, **24**, 103–107.

Pavur, R. J., Edgeman, R. L. and Scott, R. C. (1992). Quadratic statistics for the goodness-of-fit test of the inverse Gaussian distribution, *IEEE Transactions on Reliability*, **41**, 118–123.

Schrödinger, E. (1915). Zur theorie der fall und steigversuche an teilchen mit Brownscher bewegung, *Physikalische Zeitschrift*, **16**, 289–295.

Seshadri, V. (1999). *The Inverse Gaussian Distribution: Statistical Theory and Applications*, Springer, New York.

Shuster, J. J. (1968). On the inverse Gaussian distribution function, *Journal of the American Statistical Association*, **63**, 1514–1516.

Smoluchowsky, M. V. (1915). Notiz über die berechning der Brownschen molkularbewegung bei des ehrenhaft-milikanchen versuchsanordnung, *Physikalische Zeitschrift*, **16**, 318–321.

Tweedie, M. K. (1957a). Statistical properties of inverse Gaussian distributions-I, *Annals of Mathematical Statistics*, **28**, 362–377.

Tweedie, M. K. (1957b). Statistical properties of inverse Gaussian distributions-II, *Annals of Mathematical Statistics*, **28**, 696–705.

Vasicek, O. (1976). A test for normality based on sample entropy, *Journal of the Royal Statistical Society, Series B*, **38**, 54–59.

역가우스분포에 대한 적합도 평가를 위한 그래프 방법

최병진[a,1]

[a]경기대학교 응용정보통계학과

(2012년 11월 14일 접수, 2012년 11월 27일 수정, 2012년 12월 10일 채택)

요 약

Q-Q 플롯은 자료에 대한 분포적 가정을 평가하기 위해서 사용되는 편리하고 효과적인 그래프 방법이다. Q-Q 플롯은 자료의 분포와 이론적 분포를 비교하기 위한 확률플롯으로 자료에서의 분위수와 이에 대응하는 이론적 분위수를 각각 수직축과 수평축으로 해서 그린 산점도의 형태를 취한다. 본 논문에서는 확률변수 X가 위치모수 μ와 척도모수 λ를 가지는 역가우스분포를 따르면, 변환된 확률변수 $Y = |\sqrt{\lambda}(X-\mu)/\mu\sqrt{X}|$는 평균이 0이고 분산이 1인 표준반접정규분포를 하게 되는 분포적 결과를 활용하여 역가우스분포 Q-Q 플롯의 구축방법을 소개한다. 역가우스분포와 다른 분포를 따르는 자료를 대상으로 그린 Q-Q 플롯에서 나타나는 점들의 형태를 알아보고자 모의실험을 수행하고 그 결과를 제시한다. 실제 자료에 대한 사례분석을 통해 제안한 Q-Q 플롯의 유용성을 보인다.

주요용어: 역가우스분포, 표준반접정규분포, Q-Q 플롯, 분위수.

[1]교신저자: (443-760) 경기도 수원시 영통구 이의동 산 94-6, 경기대학교 응용정보통계학과, 부교수.
E-mail: bjchoi92@kyonggi.ac.kr