한국통계학회논문집
2011, 18권, 6호, 707–716

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.6.707

표본분위수와 표본분위의 관계

안성진 [1,a]

[a]경상대학교 기초과학연구소 및 컴퓨터정보통신연구소 연구원, 정보통계 학과

요 약

분위수와 분위(또는 타점위치)는 학계에서나 산업계에서 널리 사용되고 있다. 그런데 통계 소프트웨어에 구현되어 있는 표본 분위수 계산방법들과 표본 분위 계산방법들은 각각 적어도 일곱 가지가 있다. 분위수들이나 분위들을 정의하는 방법들 간의 사소해 보이는 차이가 그 값을 토대로 이루어지는 결정의 큰 차이를 가져올 수 있다. 이 논문에서는 경험적 누적확률을 사용한 기본 타점위치 방법과 Blom (1958)의 제안을 토대로 파생된 여섯 가지 타점위치 방법의 특징과 차이점을 논의하였다. 또 통계소프트웨어에 구현되어 있는 일곱 가지 표본분위수 계산방법들의 특징과 차이점들을 논의한 후 이들을 망라하는 하나의 일반식을 제시하였다. 이 논문에서는 이 일반식으로부터 얻어지는 통찰을 토대로 표본분위수에 대응되는 표본분위를 구하는 방법을 제안하였고, 이 제안을 각 표본분위수 방법에 적용하여 대응되는 표본분위 방법을 도출하였다. 이런 대응관계는 표본분위수와 표본분위에 대한 종합적 이해와 적용에 도움을 줄 수 있을 것이다.

주요용어: 분위수, 분위, 타점위치.

1. 서론

분위수(quantiles)와 분위(quantile ranks)는 학계에서나 산업계에서 널리 사용되고 있다. 이들은 한 집단이나 데이터 셋에서 특정 개체나 값 또는 점수의 위치에 대한 정보를 제공한다. 이런 종류의 상대적인 위치 측도들은 통계적 기법들을 사용하는 연구자들에게 매우 유용하다. 예컨대, 국가가 주관하는 대부분의 시험들의 결과는 백분위들로 보고된다. 대학입시에서 많은 대학들은 수능점수를 백분위 점수로 반영한다. 시민들은 어떤 의원들이 활동 평가에서 상위 또는 하위 10%에 드는지를 알고 싶어 한다. 소득분배의 측도로 사용되는 10분위 분배율이나 5분위 배율을 계산하려면 각각 먼저 십분위수들과 오분위수들을 구해야 한다.

데이터에 대한 탐색 단계와 분석 단계에서 널리 사용되는 상자그림(box-and-whisker plot)을 작성할 때 대부분의 소프트웨어에서는 튜키(1977)의 경첩들(hinges) 대신에 사분위수들을 사용하여 상자를 그린다. 회귀분석에서는 최소제곱 회귀선 즉 평균 회귀선에 대한 대안으로 분위수 회귀선이 사용되고 있다 (Koenker과 Bassette, 1978).

또 확률도들(probability plots)은 분위수나 분위를 토대로 작성된다. Q-Q 확률도는 경험적 분위수와 이론적 분위수간의 산점도이고, P-P 확률도는 경험적 분위와 이론적 분위간의 산점도이다 (안성진, 2010). 확률도에서는 분위라는 용어보다는 타점위치(plotting position)라는 용어가 널리 사용된다. 이와같이 실용적으로나 학술적으로 표본분위수와 표본분위가 널리 사용되고 있기 때문에 학생들은 이런 측도들에 대해 정확히 이해할 필요가 있다.

그런데 통계 소프트웨어에 구현되어 있는 표본분위수 계산방법들과 표본분위 계산방법들은 각각 적어도 일곱 가지가 있고 통계학 교과서나 학술논문들에 제시되어 있는 방법들까지 고려하면 그 수는

[1] (660-701) 경남 진주시 가좌동 900, 경상대학교 기초과학연구소 및 컴퓨터정보통신연구소 연구원, 정보통계 학과, 교수. E-mail: ahnsj@gnu.ac.kr

더욱 불어난다 (Langford, 2006; Hyndamen과 Fan, 1996; Joarder과 Firozzaman, 2001). 물론 각 방법들 간의 차이는 크지 않다. 때로 표본분위수들이나 표본분위들을 정의하는 방법들 간의 차이가 사소하므로 그 차이를 중요하지 않게 여기려는 이들에게 Freund와 Perles (1987)은 1절의 마지막 문단에서 표본분위수나 표본분위에 대한 계산 방법의 작은 차이가 그 값을 토대로 이루어지는 결정의 큰 차이를 가져올 수 있음을 잘 논증하고 있다. 이 뿐 아니라 그 계산방법이 과연 정당한가하는 문제도 제기될 수 있으며, 어떤 의사결정 상황에서는 특정 분위수 계산방법에 대응되는 분위 계산방법이 요구되기도 한다.

타점위치와 관련해서, Gumbel (1958)은 타점위치를 선택하기 위한 다섯 가지 조건을 제시하였고, Kimbal (1960)은 세 가지 기준을 제시하였다. Blom (1958)은 "Blom 타점위치"라고 불리는 타점위치를 제안하였고, Chernoff와 Lieberman (1954)과 Harter (1984)는 최적 타점위치들에 대해 논의하였다. Hyndman과 Fan (1996)은 표본 분위수가 만족시켜야 할 바람직한 여섯 가지 성질을 제시하고, Hazen 타점위치가 이런 모든 성질을 충족한다고 보고하였다. 그들은 타점위치들을 선형 보간함으로써 각 타점위치를 정의하는 방법에 대응되는 분위수 방법으로 정의하였다. Langford (2006)는 교과서나 통계 소프트웨어에서 채택되고 있는 분위수를 구하는 15가지의 서로 다른 방법들을 수집해서 비교하고, SAS의 기본방법(SAS Method 5)만이 분위수에 대한 보편적인 정의(정의 1)을 만족한다고 보고하였다.

한편, Kim과 Kim (2003)과 Kim 등 (2004)은 분위수에 대한 비모수적 검정을 다루었고, Kim과 Oh (2010)는 분위수 회귀분석을 이용하여 고객의 서비스 품질의 효과를 추정하였다. 그리고 김혁주와 유지선 (2008)는 집단화된 데이터의 분위수를 계산하는 방법에 관하여 논의하였다.

표본 분위수 계산과 관련하여, 국내의 기초 통계학 교과서들은 대개 한 가지나 두 가지 계산방법을 제시하고 있다. 특정 소프트웨어로 실습을 하도록 고안된 통계학 교과서의 경우에 그 소프트웨어에서 채택하고 있는 분위수 계산방법을 언급하지 아니하고 이와는 다른 계산방법을 소개하고 있기도 하다.

우리는 다음 절에서 통계 소프트웨어에 구현되어 있는 표본분위수 계산 방법들과 표본분위 계산 방법들을 소개하고 이들 간의 대응관계를 밝히고자 한다. 우리는 많은 동점들(ties)을 가진 데이터나 집단화된 데이터를 고려하지는 않을 것이다.

2. 표본분위수와 표본분위의 대응관계

2.1. 표본분위수와 표본분위에 대한 정의

한 분포의 p-분위수 $Q(p)$는 다음과 같이 정의된다.

$$Q(p) = F^{-1}(p) = \inf\{x|F(x) \geq p\}, \quad 0 < p < 1, \tag{2.1}$$

여기서 $F(x)$는 분포함수이다. 표본 분위수 $\hat{Q}(p)$는 분포 F로부터의 랜덤표본 $\{X_1, \ldots, X_n\}$을 토대로 계산되는 분포의 분위수에 대한 비모수적 추정량이다. $\{X_1, \ldots, X_n\}$의 순서통계량들을 $\{X_{(1)}, \ldots, X_{(n)}\}$로 나타내자. 우리는 방법 k에 의해 계산된 표본 p-분위수를 $\hat{Q}_k(p)$로 나타낼 것이다. 표본 분위수에 대해 널리 통용되는 정의는 다음과 같다 (Langford, 2006).

정리 1. (표본 분위수) 표본 제 p-분위수 $\hat{Q}(p)$는 순서화된 데이터에서 그 값 이하에 $100p\%$이상이 있고, 그 값 이상에 $100(1-p)\%$이상이 있는 그러한 값으로 정의된다.

우리는 모든 p에 대해 정의되는 분위수 방법들만을 고려할 것이다. 따라서 Tukey (1977)의 경첩(hinges)이나 Hoaglin (1983)의 문자값들(letter-values)은 고려에서 제외할 것이다 (이들은, 예컨대, Minitab의 문자값 전시 명령(Lvals)과 텍스트 모드의 상자그림 전시 명령(Boxplot)에 구현되어 있다).

한편, 한 분포에서 값 x의 분위 $\mathrm{QR}(x)$는 다음과 같이 정의된다.

$$\mathrm{QR}(x) = F(x). \tag{2.2}$$

표본 분위 $\widehat{\mathrm{QR}}(x)$는 분포 F로부터의 랜덤표본 $\{X_1, \ldots, X_n\}$을 토대로 계산되는 분포의 분위에 대한 비모수적 추정량이다.

우리는 이 분위수와 분위에 대한 정의들로부터 표본 분위수와 표본 분위가 서로 밀접하게 연관되어 있다는 것을 알 수 있다. 즉 표본 분위수의 정의와 표본 분위의 정의 간에는 서로 대응관계가 있고, 이들은 서로를 정의하기 위해 필요한 존재들이다. 많은 연구들에서 이 두 통계량의 관계가 암시되고 있고 전제되고 있긴 하지만, 이 두 통계량의 관계를 명시적으로 보여주는 문헌은 찾아보기 어렵다.

2.2. 표본 분위와 표본 분위수를 구하는 방법들

우리는 주요 통계소프트웨어에서 채택하고 있는 표본 분위수 계산 방법들에 대응하는 표본 분위 계산방법을 소개하고자 한다. 2.1절의 정의에서도 보았듯이 p-분위수는 분위 p에 대해 정의되고 있으므로 우리는 표본 분위수 계산방법에 대응되는 표본 분위 계산방법을 확정할 필요가 있다.

구체적으로 우리는 표본 분위수를 구하는 일곱 가지 방법과 이들에 대응되는 표본 분위 계산방법을 제안할 것이다. 먼저 표본 분위 즉 타점위치를 구하는 방법들을 살펴보자.

2.2.1. 표본 분위(타점위치)를 구하는 방법

랜덤표본의 순서통계량들 $\{X_{(1)}, \ldots, X_{(n)}\}$에서 $X_{(i)}$까지의 경험적 누적확률 p_i는 다음과 같이 정의된다.

$$p_i = \frac{i}{n}. \tag{2.3}$$

그런데 i/n은 $(-\infty, \infty)$를 치역으로 가지는 분포에 대해서 가장 큰 값을 타점할 수 없기 때문에 타점위치로는 바람직하지 않다. $(i-1)/n$도 가장 작은 값을 타점할 수 없으므로 타점위치로는 바람직하지 않다.

이런 이유 때문에 주로 사용되는 타점위치는 Blom (1958)이 제안한바 다음과 같은 형태를 띤다.

$$p_i = \frac{i-\alpha}{n+1-2\alpha}, \tag{2.4}$$

여기서 α는 상수로서 대개 $0 \le \alpha \le 1$인 값이 사용된다. 이러한 타점위치는 모두 대칭성을 가진다. α값으로 추천되는 값들은 다음과 같다.

① $\alpha = 0$. 이 때 $p_i = i/(n+1)$는 와이블(Weibull) 타점위치라 불린다 (Weibull, 1939). 이는 평균 위치(mean position)를 나타낸다.

$$E[F(X_{(i)})] = \frac{i}{n+1}. \tag{2.5}$$

② $\alpha = 0.3$. 이 때 $p_i = (i-0.3)/(n+0.4)$는 B-B 타점위치라 불린다 (Benard와 Bos-Levenbach, 1953). 이는 Johnson (1951)이 제안한 중위수 위치(median position)에 대한 좋은 근사를 제공한다.

$$\mathrm{Median}[F(X_{(i)})] = F\left(\mathrm{Median}[X_{(i)}]\right) \doteq \frac{i-0.3}{n+0.4}. \tag{2.6}$$

③ $\alpha = 1/3$. 이 때 $p_i = (i-1/3)/(n+1/3)$는 튜키(Tukey) 타점위치라 불린다 (Tukey, 1962).

④ $\alpha = 3/8$. 이 때 $p_i = (i-3/8)/(n+1/4)$는 블롬(Blom) 타점위치라 불린다 (Blom, 1958). 이는 정규분포의 $F(E[X_{(i)}])$에 대한 좋은 근사를 제공하므로 정규확률도를 작성할 때 널리 사용된다.

표 1: Plotting positions and their characteristics

p_i	α	Characteristics	Name(Softwares)
i/n	-	empirical cumulative probability	Parzen (Minitab: Kaplan-Meier)
$i/(n+1)$	0	mean position	Weibull (Minitab: Herd-Johnson; SAS: VW)
$(i-0.3)/(n+0.4)$	0.3	(approximate) median position	Benard & Bos-Levenbach (Minitab: Bernard)
$(i-1/3)/(n+1/3)$	1/3		Tukey (SAS: Tukey)
$(i-3/8)/(n+1/4)$	3/8	approximate $F(E[X_{(i)}])$ for the normal distribution	Blom (Minitab: NORMTEST, NSCORE; SAS: Blom)
$(i-0.5)/n$	0.5		Hazen (Minitab: Modified Kaplan-Meier)
$(i-1)/(n-1)$	1	modal position	Gumbel (Excel: PERCENTRANK)

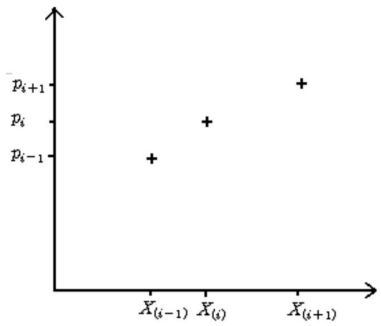

그림 1: *Plot resulting from a plotting-position method*

⑤ $\alpha = 1/2$. 이 때 $p_i = (i-1/2)/n$는 하젠(Hazen) 타점위치라 불린다 (Hazen, 1914). 이는 i/n과 $(i-1)/n$의 중간 위치(midpoint position)를 나타낸다.

⑥ $\alpha = 1$. 이 때 $p_i = (i-1)/(n-1)$는 검벨(Gumbel) 타점위치라 불린다 (Gumbel, 1939). 이는 미국의 NIST가 권장하는 타점위치이다. 이는 또 최빈 위치(modal position)를 나타낸다.

$$\text{Mode}\left[F\left(X_{(i)}\right)\right] = \frac{i-1}{n-1}. \tag{2.7}$$

이상의 결과들은 표 1에 요약되어 있다. 그림 1은 각 값과 이에 대응되는 타점위치에 대한 산점도이다.

어떤 타점 위치가 적절한지는 확률도의 사용 용도와 비교하려는 분포 등에 좌우될 것이다.

첫째, 단순히 데이터를 가정하고 있는 분포와 비교하기 위한 경우에는 타점위치 선택이 그다지 중요하지 않다.

둘째, 특정한 누적확률값에 대응되는 변수값(즉 특정 분위수)을 추정하기 위한 경우에는 타점위치로 $F(E[X_{(i)}])$가 적절하다.

셋째, 특정한 측정값에 대응되는 누적확률(즉 특정 분위)을 추정하기 위한 경우에는 타점위치로 평균위치나 중위수 위치가 적절하다 (각각 최소제곱(합)회귀와 최소절대편차(합)회귀를 사용하여 모수들을 추정한다).

2.2.2. 표본 분위수를 구하는 방법

여러 통계 소프트웨어에 구현되어 있는 표본 분위수 계산방법들을 살펴보자.

① 방법 1(SAS의 Method 1)

$(n)p = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_1(p) = (1-g)X_{(j)} + gX_{(j+1)}. \tag{2.8}$$

② 방법 2(SAS의 Method 2)

$(n)p = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_2(p) = \begin{cases} x_{(j)}, & g < 0.5, \\ x_{(j)}, & g = 0.5\text{이고 } j\text{가 짝수}, \\ x_{(j+1)}, & g = 0.5\text{이고 } j\text{가 홀수}, \\ x_{(j+1)}, & g > 0.5. \end{cases} \tag{2.9}$$

이 방법은 가장 가까운 정수 순서통계량을 돌려준다. 동점으로 맞서는 경우에는 짝수 순서통계량을 돌려준다.

③ 방법 3(SAS의 Method 3)

$$(n)p = j + g\text{라 할 때 } (j\text{: 정수부분}, g\text{: 실수부분}). \tag{2.10}$$

④ 방법 4(SAS의 Method 4)

$(n+1)p = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_4(p) = (1-g)x_{(j)} + gx_{(j+1)}. \tag{2.11}$$

⑤ 방법 5(SAS의 Method 5)

$(n)p = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_5(p) = \begin{cases} \dfrac{x_{(j)} + x_{(j+1)}}{2}, & g = 0, \\ x_{(j+1)}, & g > 0. \end{cases} \tag{2.12}$$

이 방법은 SAS사에 의해 “EDF with averaging”이라고 불리는데, 일반적으로는 CDF(cumulative distribution function) 법이라고 불린다 (Hyndman과 Fan, 1996; Langford, 2006).

⑥ 방법 6(Excel method)

$(n-1)p + 1 = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_6(p) = (1-g)x_{(j)} + gx_{(j+1)}. \tag{2.13}$$

⑦ 방법 7(Hydrologists’s method)

$(n)p + 0.5 = j + g$라 할 때 (j: 정수부분, g: 실수부분)

$$Q_7(p) = (1-g)x_{(j)} + gx_{(j+1)}. \tag{2.14}$$

이 표본 분위수 방법들은 아무 데이터에 대해 중위수(median)만을 동일하게 정의할 뿐 일반적으로 사분위수나 백분위수들을 다르게 산출한다. 이들 가운데 방법 1, 4, 6, 그리고 7은 보간법(interpolation method)으로 분류되고, 방법 5는 평균법(averaging method)으로 분류된다 (Langford, 2006).

표 2: m, l, and γ in expression (2.15) corresponding to each of the quantile rank methods and the softwares adopting that method: the I or A behind the name denotes the interpolation or the averaging, respectively.

Method k (name)	m	l	γ	Softwares
1 (SAS-1)-I	0	0	g	SAS(method 1)
2 (SAS-2)	0	0	$\begin{cases} 1, & \text{if } g > 0.5, \\ 2(0.5j - \lfloor 0.5j \rfloor), & \text{if } g = 0.5, \\ 0, & \text{if } g < 0.5. \end{cases}$	SAS(method 2)
3 (SAS-3)(EDF)	0	0	$\begin{cases} 1, & \text{if } g > 0, \\ 0, & \text{if } g = 0. \end{cases}$	Mathematica(quantiles) SAS(method 3)
4 (SAS-4)(Minitab)-I	0	1	g	Minitab, SPSS, BMDP, JMP, SAS(method 4)
5 (SAS-5)(CDF)-A	0	0	$\begin{cases} 1, & \text{if } g > 0, \\ 1/2, & \text{if } g = 0. \end{cases}$	GLIM, SAS(default: method 5)
6 (Excel)-I	1	−1	g	Excel, R
7 (Hydrologists)-I	0.5	0	g	Mathematica(Interpolated Quantiles)

2.2.3. 표본 분위수 방법과 표본 분위 방법 간의 대응관계

앞 소절(2.2.2절)에서 제시된 분위수 계산 방법들은 모두 한 개 또는 두 개의 순서통계량들을 토대로 정의되고 있으며 다음과 같이 표현될 수 있다. $j = \lfloor np \rfloor$, $g = np - j$, $m \in R$, 그리고 $l \in R$라 하자 (여기서 $\lfloor x \rfloor$는 x를 넘지 않는 가장 큰 정수를 나타낸다). 그러면 $(j-m)/(n+l) \le p < (j-m+1)/(n+l)$에 대해,

$$\hat{Q}_k(p) = (1-\gamma)X_{(j)} + \gamma X_{(j+1)}, \tag{2.15}$$

여기서 γ는 np의 정수부분 j와 분수부분 g의 함수로 주어진다.

표 2는 분위수를 구하는 각 방법에 대해 대응되는 m, l, γ와 이들을 채택하고 있는 통계소프트웨들을 나타내고 있다. Hyndman과 Fan (1996)도 분위수 계산 방법에 대한 한 일반적인 표현을 제시하고 있다.

한편 Hyndman과 Fan (1996)은 타점위치 방법들을 분위수 방법들로 간주하고 있다. 타점위치 방법들은 모두 그림 1과 같이 각 $X_{(i)}$에 대해 타점위치 p_i를 제시해 줄뿐이므로 아무 확률 p에 대해 p-분위수를 특정해 주지 못한다. 그들은 $p \ne p_i$에 대해서는 각 점들을 선분으로 연결하여 (즉 선형 보간을 사용하여) 분포함수에 대한 추정량 $\hat{F}$를 구하고 이로부터 p-분위수를 $\hat{F}^{-1}(p)$와 같이 구해줌으로써 이들을 분위수 방법들로 간주하고 있다. 우리는 이들을 분위수 계산 방법으로 포함시키지 않았는데, 그 이유는 다음과 같다. 첫째로, 이들은 통계 소프트웨어에 구현되어 있는 분위수 계산방법들이 아니다. 다음으로, 각 타점 위치들을 선분 대신에 층계함수로 연결하는 것도 분포함수에 대한 추정량 $\hat{F}$를 구하는 또다른 직관적인 방법이 될 수 있다. 끝으로 그리고 가장 중요한 이유로, 아래의 제안들에서 보게 되듯이 각 분위수 계산방법들은 그들에 대응되는 분위 계산방법을 가지고 있기 때문이다.

한편 분위수를 정의하는 각 방법은 주어진 아무 p에 대해 p-분위수를 구해줄 뿐 아니라, 그 방법이 정의하는 분포함수에 대한 추정량 $\hat{F}$로부터 아무 값 x에 대한 분위를 정의해 줄 수 있다. 식 (2.15)에서 $\gamma = 0$이면 우변은 $X_{(j)}$가 되고 따라서 좌변의 인수인 p는 $X_{(j)}$의 분위가 된다. 또 $\gamma = 1$이면 우변은 $X_{(j+1)}$가 되고 따라서 좌변의 인수인 p는 $X_{(j+1)}$의 분위가 된다. 이런 통찰로부터 우리는 아래의 제안들에서 분위수에 대한 정의로부터 분위를 구하는 방법을 제안한다. 값 x의 순위를 $r(x)$로 나타내기로 하자.

표 3: Quantile rank methods corresponding to each of the quantile methods

Quantile Method k (name)	m	l	Quantile Rank Method $\widehat{\mathrm{QR}}_k(x)$	Name of $\widehat{\mathrm{QR}}_k(x)$
1 (SAS-1)	0	0	$\frac{r(x)}{n}$	Kaplan-Meier
2 (SAS-2)	0	0	$\frac{r(x)}{n}$	Kaplan-Meier
3 (SAS-3)(EDF)	0	0	$\frac{r(x)}{n}$	Kaplan-Meier
4 (SAS-4)(Minitab)	0	1	$\frac{r(x)}{n+1}$	Weibull
5 (SAS-5)(CDF)	0	0	$\frac{r(x)-0.5}{n}$	Hazen, Modified Kaplan-Meier
6 (Excel)	1	−1	$\frac{r(x)-1}{n-1}$	Gumbel, NIST-recommended
7 (Hydrologists)	0.5	0	$\frac{r(x)-0.5}{n}$	Hazen, Modified Kaplan-Meier

[제안 1] (보간법으로 분류되는 분위수 방법들- 1, 4, 6, 7 -의 경우) "보간" 분위수 방법들의 경우에 다음 관계가 성립한다: $\gamma = g$. 그런데 실수부분 g는 $0 \le g < 1$의 범위를 가진다. 따라서 γ는 1이 될 수 없고 $g = 0$일 때 0이 될 수 있을 뿐이다. 이로부터 보간법으로 분류되는 각 분위수 방법에서 값 x의 분위 $\widehat{\mathrm{QR}}(x)$는 다음과 같이 정해진다.

$$\widehat{\mathrm{QR}}(x) = \frac{r(x) - m}{n + l}. \tag{2.16}$$

[제안 2] (평균법으로 분류되는 분위수 방법- 5 -의 경우) 이 경우에는 $g > 0$에 대해 γ가 1이 될 수 있을 뿐이다. 따라서 이 "평균" 분위수 방법에서 값 x의 분위 $\widehat{\mathrm{QR}}(x)$는 다음과 같이 정해진다.

$$\widehat{\mathrm{QR}}(x) = \frac{r(x) + 1 - m}{n + l}. \tag{2.17}$$

[제안 3] (반올림법으로 분류되는 분위수 방법- 2 -의 경우) $\gamma_{g=0}$이 $g = 0$일 때의 γ의 값을 나타낸다고 하자. 그리고 값 x의 순위를 $r(x)$로 나타내자. 그러면 각 분위수 방법에서 값 x의 분위 $\widehat{\mathrm{QR}}(x)$는 다음과 같이 구해진다.

$$\widehat{\mathrm{QR}}(x) = \frac{r(x) - m - \gamma_{g=0}}{n + l}. \tag{2.18}$$

[제안 4] (EDF법으로 분류되는 분위수 방법- 3 -의 경우) $\gamma_{g=0}$이 $g = 0$일 때의 γ의 값을 나타낸다고 하자. 그리고 값 x의 순위를 $r(x)$로 나타내자. 그러면 각 분위수 방법에서 값 x의 분위 $\widehat{\mathrm{QR}}(x)$는 다음과 같이 구해진다.

$$\widehat{\mathrm{QR}}(x) = \frac{r(x) - m - \gamma_{g=0}}{n + l}. \tag{2.19}$$

표 3은 이 제안들을 이용하여 분위수로부터 도출된 분위들을 나타낸다. 표 3으로부터 각 분위수 방법과 짝을 이루는 분위 방법을 살펴볼 수 있다. Excel은 $\hat{Q}_6$를 함수 PERCENTILE( )로 그리고 이와 짝을 이루는 $\widehat{\mathrm{QR}}_6$를 함수 PERCENTRANK( )로 명시적으로 제공하고 있지만 다른 통계 소프트웨어들은 분위수와 분위(타점위치)를 관련지어서 제공하고 있지 아니한 실정이다.

3. 결론

분위수와 분위는 학계에서나 산업계에서 널리 사용되고 있다. 그런데 통계 소프트웨어에 구현되어 있는 표본 분위수 계산방법들과 표본 분위 계산방법들은 각각 적어도 일곱 가지가 있고 통계학 교과서나 학술논문들에 제시되어 있는 방법들까지 고려하면 그 수는 더욱 불어난다. 분위수들이나 분위들을 정의하는 방법들 간의 사소해 보이는 차이가 그 값을 토대로 이루어지는 결정의 큰 차이를 가져올 수 있다. 이 뿐 아니라 그 계산방법이 과연 정당한가하는 문제도 제기될 수 있으며, 어떤 의사결정 상황에서는 특정 분위수 계산방법에 대응되는 분위 계산방법이 요구되기도 한다.

우리는 먼저 표본분위수와 표본분위에 대한 정의를 하고, 경험적 누적확률을 사용한 기본 타점위치 방법과 Blom (1958)의 제안을 토대로 파생된 여섯 가지 타점위치 방법의 특징과 차이점을 논의하였다. 또 통계소프트웨어에 구현되어 있는 일곱 가지 표본분위수 계산방법들의 특징과 차이점들을 논의한 후 이들을 망라하는 하나의 일반식을 제시하였다. 우리는 이 일반식으로부터 얻어지는 통찰을 토대로 표본분위수에 대응되는 표본분위를 구하는 방법을 제안하였다. 우리는 이 제안을 각 표본분위수 방법에 적용하여 대응되는 표본분위 방법을 도출하였다. 이런 대응관계는 표본분위수와 표본분위에 대한 종합적 이해와 적용에 도움을 줄 수 있을 것이다.

참고 문헌

김혁주, 유지선 (2008). 집단화된 자료의 분위수를 계산하는 방법에 관하여, *Journal of the Korean Data Analysis Society*, **10**, 3453–3464.

안성진 (2010). <미니탭 매크로와 함께 하는 데이터 탐색>, 자유아카데미.

Benard, A. and Bos-Levenbach, E. C. (1953). Het Uitzetten van Waarnemingen op Waarschijnlijkheitspapier, *Statistica*, **7**, 163–173.

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

Chernoff, H. and Lieberman, G. J. (1954). Use of normal probability paper, *Journal of the American Statistical Association*, **49**, 778–785.

Freund, J. E. and Perles, B. M. (1987). A new look at quartiles of ungrouped data, *The American Statistician*, **41**, 200–203.

Gumbel, E. J. (1939). La Probabilité des Hypothèses, *Comptes Rendus de l'Académie des Sciences (Paris)*, **209**, 645–647.

Gumbel, E. J. (1958). *Statistics of Extremes*, Colombia University Press, New York.

Harter, H. L. (1984). Another look at Plotting Positions, *Communications in Statistics-Theory and Methods*, **13**, 1613–1633.

Hazen, A. (1914), Storage to be Provided in Impounding Reservoirs for Municipal Water Supply, (with discussion), *Transactions of the American Society of Civil Engineers*, 77, 1539-1669.

Hoaglin, D. C. (1983). *Letter Values: A set of Selected Order Statistics*, In Hoaglin, D. C., Mosteller, F., and Tukey, J. W. (Editors), Understanding Robust and Exploratory Data Analysis, John Wiley & Sons, New York.

Hyndman, R. J. and Fan, Y. (1996). Sample quantiles in statistical packages, *The American Statistician*, **50**, 361–365.

Joarder, A. H. and Firozzaman, M. (2001). Quartiles for discrete data, *Teaching Statistics*, **23**, 86–89.

Johnson, L. G. (1951). The median ranks of simple values on their population with an application to certain fatigue studies, *Industrial Mathematics*, **2**, 1–9.

Kim, D. H., Kim, D. W. and Kim, H. G. (2004). Nonparametric test on ranked ordering-set sampling for quantiles, *Journal of the Korean Data Analysis Society*, **6**, 407–418.

Kim, D. H. and Kim, H. G. (2003). Wilcoxon signed rank test For quantiles under ranked ordering-set sample, *Journal of the Korean Data Analysis Society*, **5**, 467–477.

Kim, S. H. and Oh, H. C. (2010). The effect of Perceived Service Quality on Customer's Repeat Patronage in the University Training Restaurant: OLS and quantile regression approach, *Journal of the Korean Data Analysis Society*, **12**, 69–81.

Kimball, B. F. (1960). On the choice of plotting positions on probability paper, *Journal of the American Statistical Association*, **55**, 546–560.

Koenker, R. and Bassette, G. W. (1978). Regression quantiles, *Econometrica*, **46**, 33–50.

Langford, E. (2006). Quartiles in elementary statistics, *Journal of Statistical Education*, **14**, www.amstat.org/publications/jse/v14n3/langford.html.

Tukey, J. W. (1962). The future of data analysis, *Annals of Mathematical Statistics*, **22**, 1–67.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Edison-Wesley, Leading, MA, USA.

Weibull, W. (1939). The phenomenon of rupture in solids, *Ingeniörs Vetenskaps Akademien Handlingar*, **153**, 17.

2011년 8월 접수; 2011년 9월 채택

Relationship between the Sample Quantiles and Sample Quantile Ranks

Sung Jin Ahn[1,a]

[a]Department of Information Statistics, RINS & RICIC, Gyeongsang National University

Abstract

Quantiles and quantile ranks(or plotting positions) are widely used in academia and industry. Sample quantile methods and sample quantile methods implemented in some major statistical software are at least seven, respectively. Small looking differences between the methods can make big differences in outcomes that result from decisions based on them. We discussed the characteristics and differences of the basic plotting position using the empirical cumulative probability and the six plotting positions derived from the suggestion of Blom (1958). After discussing the characteristics and differences of seven quantile methods used in the some major statistical software, we suggested a general expression covering all seven quantile methods. Using the insight obtained from the general expression, we proposed four propositions that make it possible to find the plotting position method that correspond to each of the seven quantile methods. These correspondences may help us to understand and apply quantile methodology.

Keywords: Quantile, quantile rank, plotting position.

[1] Professor, Department of Information Statistics, RINS & RICIC, Gyeongsang National University, Jinju 660-701, Korea.
E-mail: ahnsj@gnu.ac.kr