Journal of the Korean Data & Information Science Society
2017, **28**(2), 383–393

http://dx.doi.org/10.7465/jkdi.2017.28.2.383
한국데이터정보과학회지

집단화된 통계자료의 도수다각형에 근거한 새로운 분위수 계산법[†]

김혁주[1]

[1]원광대학교 수학·정보통계학부 및 기초자연과학연구소

접수 2017년 2월 28일, 수정 2017년 3월 24일, 게재확정 2017년 3월 27일

요 약

집단화되어 있는 통계자료의 통계량을 구하고자 하는 경우 통계량의 참값에 보다 가까운 값을 얻게 해주는 계산 방법을 사용하는 것이 바람직하다. 본 논문에서는 집단화된 자료의 분위수들을 계산하는 새로운 방법을 제시하였다. 제시된 방법의 주된 아이디어는, 히스토그램에 따라 그려지는 도수다각형에서 각 계급구간에 대응하는 오각형의 넓이를 그 계급구간의 도수보다 하나 많은 개수의 부분으로 등분함으로써 자룟값들을 계산하는 것이다. 제시된 방법을 모의실험을 통해 기존의 방법들과 비교하였는데, 통계학개론 교재에 주어져 있는 몇 가지의 자료를 대상으로 하였다. 모의자료의 생성 방법은, 각 계급구간에서 도수다각형에 의해 주어진 모양의 확률밀도함수를 갖는 분포를 찾아낸 뒤 역변환 방법을 이용하여, 이 분포를 따르는 모의자룟값들을 각 계급구간에서 주어진 도수와 같은 개수만큼 발생시키는 방식이다. 모의자료의 분위수와의 차의 제곱합을 기준으로 할 때 제시된 방법이 기존의 방법들보다 거의 모든 사분위수와 십분위수에서 우세한 결과를 주는 것을 볼 수 있었다.

주요용어: 도수다각형, 모의실험, 분위수, 집단화된 자료.

1. 서론

통계자료가 개개의 자룟값들로 표시되어 있지 않고 몇 개의 계급구간으로 묶여서 주어진 경우가 종종 있는데, 이러한 자료를 집단화된 자료라 한다. 집단화된 자료의 경우에는 자료의 특성을 나타내는 통계량 (모집단 자료인 경우에는 모수)을 구할 때도 근삿값을 구할 수밖에 없다. 그러므로 통계량들의 참값에 보다 가까운 값을 얻게 해주는 계산 방법을 사용할 필요가 있다.

통계량 중 중요한 것이 자룟값들의 위치와 관련된 특성을 나타내는 분위수 (quantile)이다. 분위수 중 대표적인 것이 사분위수이며, 세분하면 십분위수와 백분위수도 있다. 자료 (개개의 값들이 주어져 있는)로부터 분위수를 구하는 방법으로 몇 가지가 소개되어 있지만, 세부적 계산의 방법론에서 미세한 차이가 있을 뿐이고 기본적인 개념에서는 차이가 없기 때문에 어느 교재에 정의된 내용을 사용해도 큰 차이가 없을 것이다. 여기서는 Kim 등 (2000)에 정의된 내용을 사용한다. 먼저 백분위수를 정의한다. 자료를 작은 값부터 큰 값까지 순서대로 늘어놓았을 때 적어도 p%의 관측값이 그 값보다 작거나 같으며, 동시에 적어도 $(100-p)$%의 관측값이 그 값보다 크거나 같게 되는 값 (이 값이 유일하게 결정되지 않을 때는 그 값들의 평균을 사용함)을 제p백분위수라고 정의한다. 또한 제1, 제2, 제3사분위수는 각각 제25,

[†] 이 논문은 2016학년도 원광대학교의 교비지원에 의해서 수행됨.
[1] (54538) 전북 익산시 익산대로 460, 원광대학교 수학·정보통계학부 및 기초자연과학연구소, 교수.
E-mail: hjkim@wonkwang.ac.kr

제50, 제75백분위수와 같은 것이며, 제1, 제2,$\cdots$,제9십분위수는 각각 제10, 제20, $\cdots$, 제90백분위수와 같은 것이다. 여기서 제2사분위수는 제5십분위수와 같은 것이며, 중앙값(median)이라고도 한다.

집단화된 자료의 분위수 계산 방법은 대부분의 통계학개론 교재에서 다음과 같이 설명되고 있다. Table 1.1은 29개의 주식에 대하여 주가와 한 주식당 당해 연도 당기순이익의 비율에 관한 자료를 나타낸 도수분포표이다. 이 자료는 Kim 등 (2008)에 나와 있는 자료를 예로 든 것이다.

Table 1.1 Stockprice-earnings ratio data

class interval	frequency	cumulative frequency
7.5 ~ 12.5	7	7
12.5 ~ 17.5	2	9
17.5∼22.5	8	17
22.5∼27.5	4	21
27.5∼32.5	2	23
32.5∼37.5	4	27
37.5∼42.5	2	29
total	29	

i번째 순서통계량 (order statistic)의 값, 즉 자룟값들을 작은 것부터 큰 것까지 순서대로 늘어놓았을 때 i번째 위치에 오는 값을 $x_{(i)}$로 나타내면, 이 자료의 사분위수 3개는 $Q_1 = x_{(8)}$, $Q_2 = x_{(15)}$, $Q_3 = x_{(22)}$이다. 다음과 같이 사분위수들을 계산하는 것이 대부분의 교재에서 설명하는 방식이다 (이것을 방법 1이라 한다).

$$Q_1 = x_{(8)} = 12.5 + 5 \times \frac{8-7}{9-7} = 15.0$$

$$Q_2 = x_{(15)} = 17.5 + 5 \times \frac{15-9}{17-9} = 21.25$$

$$Q_3 = x_{(22)} = 27.5 + 5 \times \frac{22-21}{23-21} = 30.0$$

위의 식에서 볼 수 있듯이, 이 방법은 각각의 계급구간 안에 있는 자룟값들이 다음과 같은 모양으로 분포하고 있다고 가정하는 셈이다. 두 번째 계급구간 (12.5∼17.5)에서는 두 값이 이 계급구간의 2등분점 (15.0)과 끝점 (17.5)에 있다고 간주한 것이며, 세 번째 계급구간 (17.5∼22.5)에서는 여덟 개의 값이 이 계급구간의 일곱 개의 8등분점과 끝점 (22.5)에 있다고 간주한 것이다. 셋째 식의 해석도 유사하게 할 수 있겠다.

위와 같은 계산법은 개개의 자룟값들이 주어지지 않은 상태에서 자료의 분위수들을 거칠게 구하기 위한 방법으로 생각할 수 있지만, 결과적으로 각 계급구간에서 가장 큰 자룟값이 그 계급구간과 다음 계급구간의 경계에 해당하는 값을 갖는다고 간주하는 셈이므로 개선의 여지가 있다. 이러한 취지에서, 각 계급구간 안의 자룟값들 사이의 간격이 균등하되 자룟값들이 계급구간의 중간점에 대하여 대칭으로 분포하고 있다고 간주하고 분위수들을 계산하는 방법이 Kim과 Yu (2008)에 의해 제시되었다 (앞으로 이 방법을 방법 2라 부르겠다). 모의실험을 통하여 비교한 결과 방법 2는 사분위수와 십분위수의 계산에서 방법 1에 비해 전반적으로 우위를 보였다.

그런데 위의 방법 1과 방법 2는 자료의 대체적인 분포 상태를 고려하지 않고 계급구간의 폭과 도수만을 고려하여 각 계급구간 안의 자룟값들의 위치를 정한 것이다. 자료의 대체적인 분포 상태를 고려하여 자룟값들의 위치를 정한다면 이를 바탕으로 자료의 분위수들도 좀 더 합리적으로 계산될 것이다. 이러한 취지에서 Kim (2013)은 집단화된 자료의 히스토그램으로부터 그려지는 도수다각형을 근거로 자룟값

들의 위치를 정하여 분위수들을 계산하는 방법을 제시하였고, 이 방법은 방법 1과 방법 2보다 우위를 보이는 경우가 많은 것으로 모의실험을 통해 밝혀졌다.

본 논문에서는 역시 집단화된 자료의 도수다각형을 근거로 자룟값들의 위치를 정하여 분위수들을 계산하는 방법이되 Kim (2013)의 방법과는 다른 새로운 방법을 제시하여 기존의 방법들과 비교하고자 한다. 한편 집단화된 자료를 다룬 국내 학자들의 연구로는 Ryu와 Moon (2014), Lee 등 (2014)이 있다.

2. 새로운 분위수 계산법

Table 2.1의 자료는 어떤 집단에서 뽑힌 40명의 몸무게 (단위: kg)의 분포를 나타낸 도수분포표이다. 이 자료는 Kim 등 (2001)에서 인용한 것이다. 또한 이 자료의 히스토그램과 이에 따른 도수다각형을 그린 것이 Figure 2.1이다.

Table 2.1 Frequency distribution of weights

class interval	frequency	cumulative frequency
36.5∼46.5	1	1
46.5∼56.5	14	15
56.5∼66.5	14	29
66.5∼76.5	4	33
76.5∼86.5	6	39
86.5∼96.5	1	40
total	40	

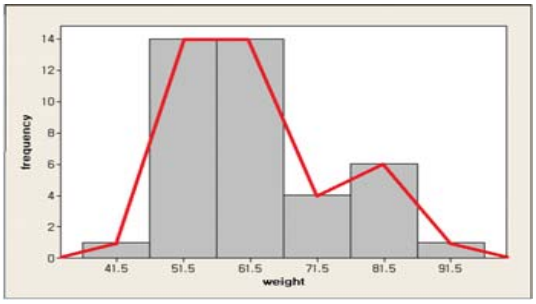

Figure 2.1 Histogram and frequency polygon for the data in Table 2.1

본 논문에서 제시하는 새로운 방법은 다음과 같다. 예를 들어 다섯째 계급구간 (76.5∼86.5)의 경우를 보자. Figure 2.1에서 이 구간만 확대하여 따로 그려서 오각형 모양으로 표시한 것이 Figure 2.2이다. 이 구간에는 6개의 자룟값이 있으므로 6개의 선분을 세로축에 평행으로 그어서 이 오각형의 넓이가 7등분되게 한 뒤, 이 선분들과 가로축의 교점인 6개의 점이 나타내는 값들을 이 계급구간 안의 자룟값들로 간주한다. 구체적으로 계산하면 이 구간에 있는 6개의 자룟값은 77.9237, 79.2746, 80.5627, 81.8014, 83.1283, 84.6521이 된다. 이러한 방법을 나머지 계급구간들의 경우에도 구간의 도수에 따라 적용하여 자룟값들을 계산한다. 자룟값들이 정해지면 1절에서 정의된 바에 의해 분위수들이 계산된다.

참고로 Kim (2013)의 방법을 그림으로 표시하면 Figure 2.3과 같다 (역시 다섯째 계급구간을 나타낸 것이다). 그림에서 보다시피 이 방법은 오각형의 넓이를 12등분하는 11개의 수직 선분을 그은 뒤 1, 3, 5, 7, 9, 11번째의 선분이 가로축과 만나는 점들의 가로좌표를 자룟값으로 간주하는 것이다.

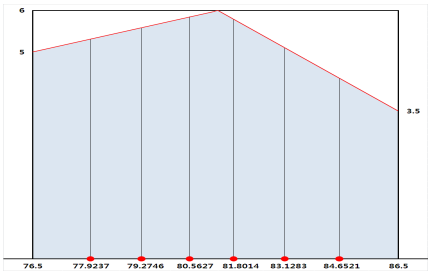

Figure 2.2 The fifth class interval for the data in Table 2.1

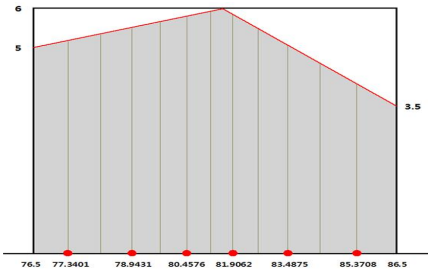

Figure 2.3 The fifth class interval for the data in Table 2.1, by Kim (2013)

3. 모의실험을 통한 비교

방법 1, 2, 3, 4를 모의실험을 통해 비교해보자. 우선 Figure 2.1의 히스토그램과 도수다각형으로 표현된 Table 2.1의 자료를 대상으로 비교한다. 각 계급구간에서 도수다각형이 나타내는 모양의 확률밀도함수를 갖는 분포를 찾아낸 뒤, 이 분포를 따르는 모의자룟값들을 각 계급구간에서 주어진 도수와 같은 개수만큼 생성한다. 생성에 사용되는 방법은 역변환 방법 (inverse transform method)이며, 모의실험에 사용되는 컴퓨터 소프트웨어는 미니탭 14 (Minitab Release 14)이다.

모의실험의 내용을 좀 더 상세히 설명한다. 예를 들어 다섯째 계급구간의 경우를 다시 생각해보자. Figure 2.2에 표시된 오각형의 넓이를 계산하면 51.25이다. 도수다각형을 이루는 윗부분의 두 선분 중 왼쪽 선분은 점 (76.5, 5)와 점 (81.5, 6)을 잇는 선분이므로 방정식이 $y = 0.2x - 10.3$이고 오른쪽 선분은 점 (81.5, 6)과 점 (86.5, 3.5)를 잇는 선분이므로 방정식이 $y = 46.75 - 0.5x$이다. 따라서 확률밀도함수

$$f(x) = \begin{cases} \dfrac{0.2x - 10.3}{51.25}, & 76.5 \le x < 81.5 \\ \dfrac{46.75 - 0.5x}{51.25}, & 81.5 \le x < 86.5 \end{cases}$$

를 갖는 연속형 확률분포로부터 6개의 난수를 생성하면 Figure 2.2에 의해 표시된 다섯째 계급구간에서 6개의 자룟값들을 생성한 결과를 얻는다.

위의 확률밀도함수를 갖는 분포의 누적분포함수는

$$F(x) = \begin{cases} \dfrac{(0.2x - 10.3)^2 - 25}{20.50}, & 76.5 \le x < 81.5 \\ \dfrac{63.5 - (46.75 - 0.5x)^2}{51.25}, & 81.5 \le x < 86.5 \end{cases}$$

로 얻어지며, 따라서 역누적분포함수는 다음과 같이 구해진다.

$$x = F^{-1}(y) = \begin{cases} 5\sqrt{20.5y + 25} + 51.5, & 0 \le y < 0.536585 \\ 93.5 - \sqrt{254 - 205y}, & 0.536585 \le y < 1 \end{cases}$$

균일분포 $U(0,1)$로부터 확률변수 U의 값 6개를 발생시킨 뒤 $X = F^{-1}(U)$에 의해 X의 값 6개를 구하면 다섯째 계급구간에서 6개의 자룟값들을 생성한 결과가 된다. 나머지 계급구간들의 경우에도 동일한 방법으로 자룟값들을 생성한다.

Table 3.1은 1회의 모의실험 결과 생성된 모의자료의 예이다. 이 모의자료의 Q_1, Q_2, Q_3와 방법 1, 2, 3, 4에 의한 Q_1, Q_2, Q_3의 값들을 구하여 Table 3.2에 정리하였다 (굵은 글씨는 가장 좋은 결과를 보인 것을 나타낸다). 따라서 이 모의자료의 경우 Q_1과 Q_2의 계산에서는 네 가지 방법 중 방법 4가 가장 좋은 결과를 줬으며, Q_3의 계산에서는 방법 3이 가장 좋은 결과를 줬다.

Table 3.1 An example of simulated data obtained as a result of a single simulation

class interval	weight	class interval	weight	class interval	weight	class interval	weight
1	43.2244	2	54.2046	3	60.6342	4	68.8032
2	47.2349	2	54.8239	3	63.0995	4	69.5980
2	47.4119	2	55.9639	3	63.1185	4	71.0730
2	48.7579	2	56.2352	3	63.6173	5	77.9295
2	49.4004	2	56.4924	3	63.9928	5	78.1678
2	49.9219	3	56.8272	3	64.6770	5	80.8852
2	51.1836	3	57.2732	3	64.9573	5	80.9480
2	51.4787	3	58.1474	3	65.3281	5	83.4023
2	51.8730	3	58.3745	3	65.7282	5	84.4875
2	52.0170	3	59.3680	4	67.2990	6	87.9500

Table 3.2 Quartiles of the data in Table 3.1 and quartiles by methods 1∼4

quartile	Table 3.1	method 1	method 2	method 3	method 4
Q_1	53.1108	53.2857	52.8333	53.3431	**53.2590**
Q_2	60.0011	60.4286	60.1667	59.7526	**59.8393**
Q_3	68.0511	70.2500	69.5000	**68.2473**	68.6181

그런데 위의 결과는 1회의 모의실험의 결과일 뿐이므로 신뢰성을 갖지 못한다. 모의실험의 결과가 신뢰성을 갖기 위해서는 많은 횟수의 실험을 실시해야 하므로 위와 같은 방식으로 10,000회의 모의실험을 실시하였다. 사분위수의 경우의 비교 결과가 Table 3.3에 정리되어 있다. 비교 기준으로 두 가지를 사용하였는데, 첫째 기준은 모의자료의 사분위수와 네 가지 방법에 의한 사분위수의 차를 제곱하여 모두 합한 값이고, 둘째 기준은 10,000번의 모의실험 중 더 우세한 횟수로 하였다. 첫째 기준으로 비교한 결과, Q_1, Q_2, Q_3의 경우 모두 방법 4가 네 가지 방법 중 가장 좋은 결과를 보였다 (표에서 $\sum$은 10,000회의 모의실험에 걸쳐 모두 합한 것을 나타낸다. $Q_{i(1)}$, $Q_{i(2)}$, $Q_{i(3)}$, $Q_{i(4)}$는 각각 방법 1, 2, 3, 4에 의한 Q_i를 나타내며, D_i의 경우도 같은 방식이다. 또한 Table 3.4, Table 3.5, Table 3.6에서도 마찬가지다). 둘째 기준으로 비교해보니, 방법 1과 방법 4의 비교에서는, Q_1의 경우만 방법 1이 방법 4보다 다소 우세했고 Q_2와 Q_3의 경우는 방법 4가 압도적으로 우세했다. 다음으로 방법 2와 방법 4의 비교에서는 Q_1, Q_2, Q_3의 경우 모두 방법 4가 훨씬 우세했다. 방법 3과 방법 4의 비교에서는, Q_1은 방법 3이 우세했고 Q_3는 방법 4가 우세했으며, Q_2는 5002 대 4998로 거의 차이가 없었다. Table 3.4는 십분위수의 경우의 비교 결과이다. 모의자료의 십분위수와의 차의 제곱합에서 D_6의 경우만 방법 3이 우세를 보

였고, 나머지 십분위수의 경우에는 방법 4가 가장 우세했다. 10,000번 중 우세횟수를 기준으로 하면, 방법 4가 모든 십분위수에서 방법 1보다 좋았고, D_7을 제외한 모든 십분위수에서 방법 2보다 좋았다. 방법 3과 방법 4를 비교해보니, D_1, D_3, D_6, D_7에서는 방법 3이 우세했고 D_4, D_8, D_9에서는 방법 4가 우세하여 우열을 가리기 힘들었다. 이 데이터의 경우 D_2는 방법 3과 방법 4의 결과가 동일한 값으로 계산된다.

Table 3.3 Simulation result regarding the quartiles for Figure 2.1

	Q_1	Q_2	Q_3
$\sum(Q_i - Q_{i(1)})^2$	10530.5	14667.8	43424.4
$\sum(Q_i - Q_{i(2)})^2$	12436.3	12342.6	27030.9
$\sum(Q_i - Q_{i(3)})^2$	10581.3	11465.6	24737.1
$\sum(Q_i - Q_{i(4)})^2$	**10529.3**	**11365.4**	**22145.9**
number of Q_i's closer to $Q_{i(1)}$	**5193**	3780	2868
number of Q_i's closer to $Q_{i(4)}$	4807	**6220**	**7132**
number of Q_i's closer to $Q_{i(2)}$	3923	4246	3569
number of Q_i's closer to $Q_{i(4)}$	**6077**	**5754**	**6431**
number of Q_i's closer to $Q_{i(3)}$	**5072**	**5002**	4906
number of Q_i's closer to $Q_{i(4)}$	4928	4998	**5094**

Table 3.4 Simulation result regarding the deciles for Figure 2.1

	D_1	D_2	D_3	D_4	D_6	D_7	D_8	D_9
$\sum(D_i - D_{i(1)})^2$	13882.8	12232.8	8541.2	4167.2	21659.7	14814.4	87571.7	35406.4
$\sum(D_i - D_{i(2)})^2$	15752.4	14999.4	8734.6	3995.4	15190.1	7160.9	38619.8	23653.1
$\sum(D_i - D_{i(3)})^2$	11824.7	**11798.9**	8103.4	4624.9	**12818.6**	7459.0	38879.1	22829.1
$\sum(D_i - D_{i(4)})^2$	**11641.0**	**11798.9**	**7889.4**	**3924.0**	12825.8	**6425.7**	**35923.2**	**22829.0**
number of D_i's closer to $D_{i(1)}$	4498	4620	4967	3704	3567	3437	3272	3621
number of D_i's closer to $D_{i(4)}$	**5502**	**5380**	**5033**	**6296**	**6433**	**6563**	**6728**	**6379**
number of D_i's closer to $D_{i(2)}$	4205	3989	3918	3903	4368	**5181**	4907	4678
number of D_i's closer to $D_{i(4)}$	**5795**	**6011**	**6082**	**6097**	**5632**	4819	**5093**	**5322**
number of D_i's closer to $D_{i(3)}$	**5075**	0	**5221**	4970	**5123**	**5065**	4880	4939
number of D_i's closer to $D_{i(4)}$	4925	0	4779	**5030**	4877	4935	**5120**	**5061**

이번에는 Kim 등 (2001)의 동일한 자료를 계급구간의 개수를 바꿔 8개의 계급구간을 사용하여 나타내는 경우를 생각해보자. Table 3.5는 도수분포표이며, Figure 3.1은 이 자료의 히스토그램과 이에 따른 도수다각형을 그린 것이다.

Table 3.5 Frequency distribution of weights (8 class intervals)

class interval	frequency	cumulative frequency
33.5∼41.5	1	1
41.5∼49.5	3	4
49.5∼57.5	17	21
57.5∼65.5	7	28
65.5∼73.5	4	32
73.5∼81.5	5	37
81.5∼89.5	2	39
89.5∼97.5	1	40
total	40	

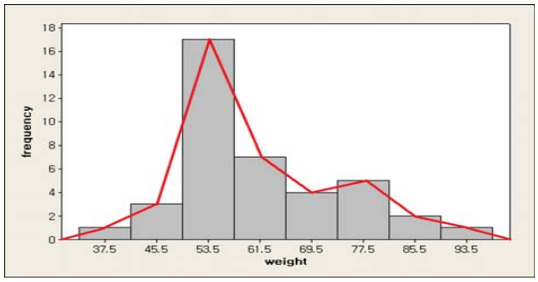

Figure 3.1 Histogram and frequency polygon for the data in Table 3.5

8개 계급구간의 경우 앞에서 기술한 모의실험 방법을 적용하여 역시 10,000회의 모의실험을 실시하였다. 6개 계급구간의 경우와 같은 방식으로 기준을 정하여 비교한 결과를 Table 3.6과 Table 3.7에 나타냈다. 방법 1, 방법 2, 방법 3과 방법 4의 전반적인 비교 결과는 계급구간의 개수를 6개로 할 때와 8개로 할 때가 대동소이하게 얻어졌다.

Table 3.6 Simulation result regarding the quartiles for Figure 3.1

	Q_1	Q_2	Q_3
$\sum(Q_i - Q_{i(1)})^2$	6014.5	4933.6	36717.3
$\sum(Q_i - Q_{i(2)})^2$	6549.3	2423.6	22910.6
$\sum(Q_i - Q_{i(3)})^2$	**5968.8**	2812.8	**22555.1**
$\sum(Q_i - Q_{i(4)})^2$	5973.3	**2367.2**	24271.7
number of Q_i's closer to $Q_{i(1)}$	4933	3592	3092
number of Q_i's closer to $Q_{i(4)}$	**5067**	**6408**	**6908**
number of Q_i's closer to $Q_{i(2)}$	4510	**5548**	**5819**
number of Q_i's closer to $Q_{i(4)}$	**5490**	4452	4181
number of Q_i's closer to $Q_{i(3)}$	**5061**	4971	**5551**
number of Q_i's closer to $Q_{i(4)}$	4939	**5029**	4449

Table 3.7 Simulation result regarding the deciles for Figure 3.1

	D_1	D_2	D_3	D_4	D_6	D_7	D_8	D_9
$\sum(D_i - D_{i(1)})^2$	5545.3	5912.2	5966.1	7206.7	28457.5	15992.1	17435.9	33081.3
$\sum(D_i - D_{i(2)})^2$	5859.9	6435.4	6124.2	5776.7	18840.5	7442.3	7697.5	14104.8
$\sum(D_i - D_{i(3)})^2$	3230.9	5702.5	**5909.1**	5789.3	14052.2	**6765.4**	7726.2	14993.6
$\sum(D_i - D_{i(4)})^2$	**3169.2**	**5662.1**	5909.9	**5749.7**	**13887.3**	6776.1	**7673.2**	**13804.0**
number of D_i's closer to $D_{i(1)}$	3197	4855	4828	4199	3032	2512	2595	3302
number of D_i's closer to $D_{i(4)}$	**6803**	**5145**	**5172**	**5801**	**6968**	**7488**	**7405**	**6698**
number of D_i's closer to $D_{i(2)}$	2667	4555	4583	**5064**	3685	4143	**5183**	**5286**
number of D_i's closer to $D_{i(4)}$	**7333**	**5445**	**5417**	4936	**6315**	**5857**	4817	4714
number of D_i's closer to $D_{i(3)}$	**5131**	**5093**	4967	**5042**	4944	4971	**5123**	**5021**
number of D_i's closer to $D_{i(4)}$	4869	4907	**5033**	4958	**5056**	**5029**	4877	4979

이번에는 다른 자료를 대상으로 모의실험을 통하여 네 가지 방법을 비교해보자. Figure 3.2에 그려진 히스토그램과 도수다각형의 근거가 된 자료는Kim 등 (2002)에 나와 있는 것으로서, 한 배관공이 전화 호출 출장 서비스 후 30개 가정에 청구한 금액 (단위: 천 원)에 관한 자료이다. 이 자료를 바탕으로 10,000회의 모의실험을 실시한 결과가 Table 3.8과 Table 3.9에 나와 있다. 이 결과를 보면, 차이의 제곱합을 기준으로 할 때는 방법 2가 D_8에서, 방법 3이 미세한 차이로 D_9에서 우세했고, 나머지 사

분위수와 십분위수에서는 모두 방법 4가 가장 우세했다. 10,000번 중 우세횟수를 기준으로 할 경우 방법 4가 모든 십분위수와 사분위수에서 방법 1보다 좋았고, Q_2를 제외한 모든 사분위수와 십분위수에서 방법 2보다 좋았다.방법 3과 방법 4의 비교에서는, 사분위수의 경우 Q_1과 Q_3는 방법 3이 우세했고 Q_2는 방법 4가 우세했다. 십분위수의 경우에는 D_1, D_2, D_3, D_6, D_9에서는 방법 3이 우세했고 D_4, D_7, D_8에서는 방법 4가 우세하여 어느 한 쪽이 우세하다고 보기 힘들었다.

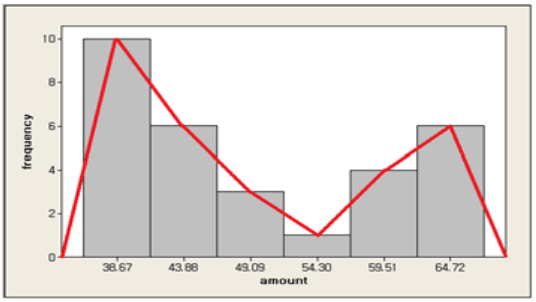

Figure 3.2 Histogram and frequency polygon for the data in Kim *et al.* (2002)

Table 3.8 Simulation result regarding the quartiles for Figure 3.2

	Q_1	Q_2	Q_3
$\sum(Q_i - Q_{i(1)})^2$	4577.9	13781.3	11398.5
$\sum(Q_i - Q_{i(2)})^2$	3747.1	6098.2	10116.0
$\sum(Q_i - Q_{i(3)})^2$	3739.3	6228.4	9422.6
$\sum(Q_i - Q_{i(4)})^2$	**3694.6**	**5627.0**	**9183.5**
number of Q_i's closer to $Q_{i(1)}$	4474	3262	4432
number of Q_i's closer to $Q_{i(4)}$	**5526**	**6738**	**5568**
number of Q_i's closer to $Q_{i(2)}$	4379	**5023**	4093
number of Q_i's closer to $Q_{i(4)}$	**5621**	4977	**5907**
number of Q_i's closer to $Q_{i(3)}$	**5156**	4960	**5037**
number of Q_i's closer to $Q_{i(4)}$	4844	**5040**	4963

Table 3.9 Simulation result regarding the deciles for Figure 3.2

	D_1	D_2	D_3	D_4	D_6	D_7	D_8	D_9
$\sum(D_i - D_{i(1)})^2$	4166.6	4103.7	3743.5	9208.8	31317.8	8041.3	5309.2	11063.1
$\sum(D_i - D_{i(2)})^2$	4810.2	4051.1	1918.2	6502.3	11487.1	8911.3	**2325.1**	6959.4
$\sum(D_i - D_{i(3)})^2$	4067.6	3845.4	2192.9	5917.1	10355.5	8473.7	2360.6	**6318.3**
$\sum(D_i - D_{i(4)})^2$	**4060.1**	**3842.6**	**1916.5**	**5723.0**	**10229.1**	**8033.9**	2346.3	6318.6
number of D_i's closer to $D_{i(1)}$	4913	4642	3813	2647	2759	4398	2449	3327
number of D_i's closer to $D_{i(4)}$	**5087**	**5358**	**6187**	**7353**	**7241**	**5602**	**7551**	**6673**
number of D_i's closer to $D_{i(2)}$	4404	4402	4051	3958	4626	4784	4789	4267
number of D_i's closer to $D_{i(4)}$	**5596**	**5598**	**5949**	**6042**	**5374**	**5216**	**5211**	**5733**
number of D_i's closer to $D_{i(3)}$	**5085**	**5080**	**5129**	4994	**5027**	4967	4631	**5206**
number of D_i's closer to $D_{i(4)}$	4915	4920	4871	**5006**	4973	**5033**	**5369**	4794

Table 3.3, 3.4, 3.6, 3.7, 3.8, 3.9를 보면, 비교 기준 중의 하나인 제곱합의 크기가 분위수에 따라 상당한 차이를 보이고 있다. 도수다각형의 모양과도 관계가 있지만, 대체적으로 자룟값들이 많이 몰려 있는 구간에 속한 분위수는 제곱합이 작게 계산되고, 도수가 작은 구간에 속한 분위수는 제곱합이 크게 계산되는 것을 볼 수 있다.

4. 결론

자료가 개개의 값들로 주어져 있지 않고 집단화되어 있는 경우 이 자료의 분위수들을 합리적으로 계산하는 것이 바람직하다. 대부분의 통계학개론 교재에서 집단화된 자료의 분위수를 계산할 때 자룟값들의 분포를 처리하는 방식은, Table 1.1에서 예로 든 바와 같이 각 계급구간에서 가장 큰 자룟값이 그 계급구간과 다음 계급구간의 경계에 해당하는 값을 갖는다는 식이다.

본 논문에서는 집단화된 자료의 분위수들을 계산하는 새로운 방법을 제시하였다. 제시된 방법은 히스토그램으로부터 그려지는 도수다각형을 근거로 하여 분위수들을 계산하는 방법이다. 제시된 방법 (방법 4)은 역변환 방법을 이용한 10,000회의 모의실험을 통하여 기존의 방법들과 비교되었다. 비교 기준으로 두 가지를 고려하였는데, 첫째는 모의자료의 분위수와의 차의 제곱합이고, 둘째는 10,000번 중 더 우세한 횟수로 하였다. 첫째 기준에서는 거의 모든 사분위수와 십분위수에서 방법 4가 기존의 방법들보다 우세하게 나타났다. 둘째 기준으로는, 방법 4가 방법 1과 방법 2보다는 압도적으로 우세했고, 방법 3과의 비교에서는 어느 한 쪽이 우세하다고 하기 힘든 것으로 나타났다. 방법 3과 방법 4를 첫째 기준과 둘째 기준으로 비교한 결과를 종합해서 말하자면, 단순한 우세 횟수에서는 큰 차이가 없지만, 모의자료의 분위수와의 차이의 크기에서는 평균적으로 방법 4가 방법 3보다 더 좋은 결과를 보여준 것이라 할 수 있다.

방법 4의 타당성을 다음의 관점에서도 생각할 수 있을 것이다. $X_1, X_2, \cdots, X_n$을 누적분포함수가 $F(x)$인 연속형 분포로부터의 랜덤표본이라 하고 이것으로부터의 순서통계량을 $Y_1 < Y_2 < \cdots < Y_n$이라 하면, $F(Y_1) < F(Y_2) < \cdots < F(Y_n)$을 균일분포 $U(0,1)$로부터의 순서통계량으로 볼 수 있으며,

$$E[F(Y_r)] = \frac{r}{n+1} \; (r = 1, 2, \cdots, n)$$

이 성립한다는 사실이 잘 알려져 있다. 이 결과와 관련해서 생각해도 방법 4가 방법 3보다 좀 더 명분과 타당성이 있다고 본다.

모의실험에서 생성된 자룟값들은 히스토그램과 도수다각형에 의해 표현되는 자료의 분포 경향을 반영한 것이다. 대부분의 경우 우리가 접하는 집단화된 자료는 개별값들이 없이 도수분포표 (즉 히스토그램과 도수다각형)로만 주어진 상태이다. 이 히스토그램과 도수다각형을 가질 수 있는 무한한 개수의 데이터 세트 중 많은 수 (예컨대 10,000개)의 데이터 세트를 뽑아 기존의 방법들과 새로운 방법을 비교해보니 평균적으로 그리고 확률적으로 새로운 방법이 좀 더 우세한 결과를 보인 것이므로, 우리가 접하고 있는 특정한 자료에도 이 새로운 방법의 적용을 고려할 만하다고 본다.

References

Kim, B. H., Choi, K. C., Baek, H. Y., Kim, H. J., Dong, K. H., Park, T. R. and Chang, I. H. (2002). *Understanding statistics*, Freedom Academy, Paju.

Kim, H. J. (2013). A quantile calculation method for grouped data based on the frequency polygon and the related simulation study. *Journal of the Korean Data Analysis Society*, **15**, 3149-3156.

Kim, H. J. and Yu, J. S. (2008). On a method for computing quantiles of grouped data. *Journal of the Korean Data Analysis Society*, **10**, 3453-3464.

Kim, W. C., Kim, J. J., Park, B. U., Park, S. H., Song, M. S., Lee, S. Y., Lee, Y. J., Jeon, J. W. and Cho, S. (2001). *General statistics*, 2nd Ed., Youngji Publishers, Seoul.

Kim, W. C., Kim, J. J., Park, S. H., Park, H. N., Song, M. S., Jeon, J. W., Chung, H. Y. and Cho, S. (2000). *Modern statistics*, 3rd Ed., Youngji Publishers, Seoul.

Kim, Y. D., Kim, W. C., Park, B. U., Park, S. H., Park, T. S., Oh, H. S., Lee, S. Y., Lee, Y. J., Lee, J. Y., Lim, Y. H., Jeon, J. W. and Cho, S. (2008). *Introduction to statistics*, 5th Ed., Youngji Publishers, Seoul.

Lee, W. K., Kim, S. W., Kim, H. I, Chang, H. H., Lee, J. M., Kim, Y. J. and Lee, M. Y. (2014). Development of quality of life with WHOQOL-HIV BREF Korean version among HIV patients in Korea. *Journal of the Korean Data & Information Science Society*, **25**, 337-347.

Ryu, G. Y. and Moon, Y. S. (2014). A case study on verification of internet survey. *Journal of the Korean Data & Information Science Society*, **25**, 11-18.

Journal of the Korean Data & Information Science Society
2017, **28**(2), 383–393

http://dx.doi.org/10.7465/jkdi.2017.28.2.383
한국데이터정보과학회지

A new method for calculating quantiles of grouped data based on the frequency polygon[†]

Hyuk Joo Kim[1]

[1]Division of Mathematics & Informational Statistics and Institute of Basic Natural Sciences, Wonkwang University

Received 28 February 2017, revised 24 March 2017, accepted 27 March 2017

Abstract

When we deal with grouped statistical data, it is desirable to use a calculation method that gives as close value to the true value of a statistic as possible. In this paper, we suggested a new method to calculate the quantiles of grouped data. The main idea of the suggested method is calculating the data values by partitioning the pentagons, that correspond to the class intervals in the frequency polygon drawn according to the histogram, into parts with equal area. We compared this method with existing methods through simulations using some datasets from introductory statistics textbooks. In the simulation study, we simulated as many data values as given in each class interval using the inverse transform method, on the basis of the distribution that has the shape given by the frequency polygon. Using the sum of squares of differences from quantiles of the simulated data as a criterion, the suggested method was found to have better performance than existing methods for almost all quartiles and deciles.

Keywords: Frequency polygon, grouped data, quantile, simulation.

[†] This paper was supported by Wonkwang University in 2016.

[1] Professor, Division of Mathematics & Informational Statistics and Institute of Basic Natural Sciences, Wonkwang University, Jeonbuk 54538, Korea. E-mail: hjkim@wonkwang.ac.kr