집단화된 자료의 분위수를 계산하는 수정된 방법[1)]

김 혁 주[2)] 유 지 선[3)]

요 약

본 논문에서는 집단화된 자료의 분위수들을 계산하는 수정된 방법을 제시하였다. 제시된 방법은 각 계급구간 안의 자료들이 그 구간에 걸쳐 균등한 간격으로, 그리고 구간의 중간점에 관하여 대칭으로 분포하고 있다고 가정하고 분위수들을 계산하는 방법이다. 개개의 자료값들이 주어진 자료를 통하여, 제시된 방법과 기존의 방법을 비교하였다.

주요용어 : 집단화된 자료, 분위수, 사분위수, 순서통계량.

1. 서론

주어진 통계자료가 개별적인 자료값들로 구성되어 있지 않고 집단화되어 있는 경우를 종종 접하게 된다. 이러한 경우 이 자료의 특성을 나타내는 통계량의 정확한 값을 구하는 것은 불가능하며, 단지 근사값을 구할 수 있다. 따라서 이것들의 참값에 보다 가까운 값을 얻게 해 주는 계산 방법을 사용하는 것이 중요하다. 이러한 의미에서 김혁주와 김영선(2003)은 집단화된 자료의 평균(mean)과 분산(variance)을 계산하는 새로운 방법을 제시하였다.

자료의 특성을 나타내는 통계량으로는 평균과 분산 외에도 중앙값(median), 사분위수(quartile), 백분위수(percentile) 등이 있다. 이들을 통틀어 분위수(quantile)라 부른다. 여기서 분위수의 정의로는 김우철 등(1998, p.22)에 나와 있는 다음의 정의를 사용한다.

정의 1.1: 사분위수와 백분위수

제1사분위수 Q_1 = 제25백분위수

제2사분위수 Q_2 = 제50백분위수

제3사분위수 Q_3 = 제75백분위수

제P백분위수라 함은 자료를 크기 순서로 늘어놓았을 때 적어도 P%의 관측값이 그 값보다 작거나 같고, 또한 적어도 (100-P)%의 관측값이 그 값보다 크거나 같게 되는 값을 말한다. 이 값이 유일하게 결정되지 않을 때는 그 값들의 평균을 사용한다.

제P백분위수를 (P/100)분위수라고도 하며, 중앙값은 제2사분위수이다. 즉 중앙값은 제50백분위수이며 0.5분위수이다.

---

1) 본 연구는 한국과학재단 목적기초연구(R01-2003-000-10220-0)지원으로 수행되었음.

2) (570-749) 전북 익산시 신용동 344-2 원광대학교 수학·정보통계학부 및 기초자연과학연구소, 교수

E-mail: hjkim@wonkwang.ac.kr

3) (570-749) 전북 익산시 신용동 344-2 원광대학교 교육대학원 수학교육전공, 석사

대부분의 통계학 교재에서는 집단화된 자료의 분위수들을 계산하는 방법으로 다음의 방법을 소개하고 있다. 김우철 등(2000, p.63)에 나와 있는 다음의 자료를 예로 들어 설명하겠다. 표 1.1은 무작위로 추출한 29개의 주식에 대하여 주가와 한 주식당 당해연도 당기순이익의 비율(stockprice-earnings ratio)에 관한 자료를 나타낸 도수분포표이다.

표 1.1: 주가와 당기순이익의 비율에 관한 자료

계급구간	도수	누적도수
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
계	29	

자료값들을 작은 것부터 순서대로 늘어놓았을 때 i번째 위치에 오는 값, 즉 i번째 순서통계량(order statistic)의 값을 $x_{(i)}$ 로 나타내자. 정의 1.1에 의하면 $Q_1 = x_{(8)}$, $Q_2 = x_{(15)}$, $Q_3 = x_{(22)}$이다. 대부분의 교재에서는 다음과 같이 사분위수(엄밀히 말하면 사분위수의 추측값)들을 계산한다.

$$Q_1 = x_{(8)} = 12.5 + (17.5 - 12.5) \times \frac{8-7}{9-7} = 15.0 \qquad (1.1)$$

$$Q_2 = x_{(15)} = 17.5 + (22.5 - 17.5) \times \frac{15-9}{17-9} = 21.25 \qquad (1.2)$$

$$Q_3 = x_{(22)} = 27.5 + (32.5 - 27.5) \times \frac{22-21}{23-21} = 30.0 \qquad (1.3)$$

위의 식들을 살펴보면 이 방법이 각 계급구간 안의 자료값들의 분포 모양을 어떻게 가정하고 있는지 알 수 있다. 식 (1.1)에서는 두 번째 계급구간인 12.5~17.5를 2등분하여 두 값이 이 계급구간의 2등분점(15.0)과 끝점(17.5)에 있다고 간주한 것이며, 식 (1.2)에서는 세 번째 계급구간인 17.5~22.5를 8등분하여 여덟 개의 값이 이 계급구간의 일곱 개의 8등분점과 끝점(22.5)에 있다고 간주한 것이다. 식 (1.3)의 경우도 유사하게 해석된다.

그런데 이처럼 각 계급구간 안에서 가장 큰 자료값이 항상 그 다음 계급구간과의 경계에 위치하고 있다고 간주하는 것이 합리적이지는 않을 것이다. 본 논문에서는 계급구간 안의 자료값들이 이러한 가정과 약간 다르게 분포하고 있다고 가정하고 자료의 분위수들을 계산하는 방법을 연구하고자 한다.

2. 분위수의 수정된 계산 방법

본 논문에서 제시하는 방법은 각 계급구간 안의 자료값들이 균등한 간격으로 분포하되 계급구간의 중간점에 대하여 대칭으로 분포하고 있다고 간주하고 분위수들을 계산하는 방법이다. 자료가 표 2.1과 같은 도수분포표로 주어졌다고 하자.

표 2.1: 일반적인 형태의 도수분포표

계급구간	계급값	도수
$a_0 \sim a_1$	m_1	f_1
$a_1 \sim a_2$	m_2	f_2
.	.	.
.	.	.
.	.	.
$a_{k-1} \sim a_k$	m_k	f_k
계		n

첫 번째 계급구간의 경우 a_0부터 a_1까지의 구간을 (f_1+1)등분하여 f_1개의 자료값들이 균등한 간격으로 분포하고 있다고 간주하며, 다른 계급구간들의 경우에도 같은 방식으로 생각한다. 즉 $x_{(i)} = a_0 + id_1 \ (i=1,2,\cdots,f_1)$ (단, $d_1=(a_1-a_0)/\ (f_1+1)$) 이며, $x_{(f_1+j)}= a_1+jd_2(\ j=1,2,\cdots,f_2)$(단, $d_2=(a_2-a_1)/\ (f_2+1)$) 이다.

표 1.1의 자료에 이 방법을 적용해 보자. 29개의 자료값이 표 2.2와 같다고 가정한다. 그림 2.1은 표 1.1의 자료를 기존의 방법과 제시된 방법에 따라 점도표로 나타낸 것이다.

표 2.2: 주가와 당기순이익의 비율에 관한 자료값들 (제시된 방법에 따른 것)

8.1250	8.7500	9.3750	10.0000	10.6250	11.2500	11.8750	14.1667
15.8333	18.0556	18.6111	19.1667	19.7222	20.2778	20.8333	21.3889
21.9444	23.5000	24.5000	25.5000	26.5000	29.1667	30.8333	33.5000
34.5000	35.5000	36.5000	39.1667	40.8333			

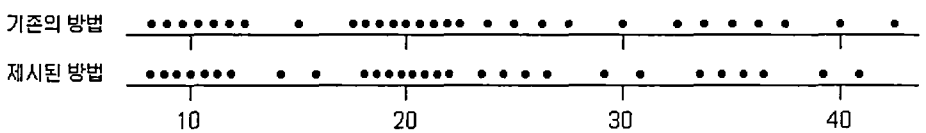

그림 2.1: 주가와 당기순이익의 비율 자료를 나타낸 점도표

이제부터는 기존의 방법을 방법 1이라 부르고 본 논문에서 제시된 방법을 방법 2라 부르겠다. 방법 1과 방법 2의 차이점은 명백하다. 예로서 표 1.1의 두 번째 계급구간의 경우 방법 1과 방법 2에 따라 가정된 자료값들을 그림 2.2에 나타냈다. 이 그림을 보면 두 방법 중 방법 2가 좀 더 합리적이라는 느낌을 갖게 될 것이다.

방법 2에 의한 표 2.2의 자료의 사분위수들을 구해 보면 Q_1=14.1667, Q_2=20.8333, Q_3 =29.1667 로 얻어진다. 이것을 방법 1에 의한 값들과 비교하면 세 가지 모두 조금씩 감소한 값이다. 이러한 대소관계는 일반적으로도 성립한다는 것을 그림 2.1로부터 쉽게 알 수 있다.

기존의 방법 ● ●

집단화된 자료의 분위수를 계산하는 수정된 방법

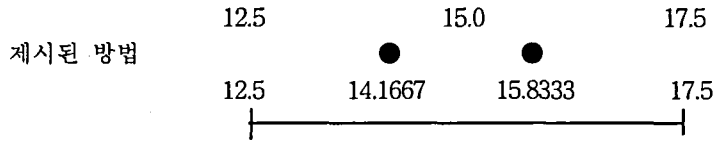

그림 2.2: 표 1.1의 자료의 두 번째 계급구간의 가정된 자료값들

3. 개개의 값들이 주어진 자료를 통한 비교

표 3.1의 자료는 김우철 등(2001, p.57)에서 인용한 것으로서, 어떤 집단에서 뽑힌 40명의 몸무게(단위: kg)를 측정한 것이다. 이 자료는 개개의 값들이 주어져 있는 경우이다. 표 3.2는 표 3.1의 자료를 바탕으로 하여 같은 책에서 작성된 도수분포표이다. 또한 표 3.3과 표 3.4는 표 3.2를 바탕으로 각각 방법 1과 방법 2에 따라 얻은 자료이며, 그림 3.1은 표 3.1, 표 3.3, 표 3.4의 자료를 나타낸 점도표이다.

표 3.1: 개개의 값들이 주어진 몸무게 자료

82	52	52	55	93	57	60	48	50	49
79	73	60	66	74	82	57	54	57	55
50	63	60	49	63	57	57	77	54	69
50	57	54	59	79	54	79	70	37	65

표 3.2: 몸무게의 도수분포표

계급구간	도수	누적도수
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
계	40	

표 3.3: 몸무게의 도수분포표를 바탕으로 한 자료값들 (방법 1에 따른 것)

46.5000	47.2143	47.9286	48.6429	49.3571	50.0714	50.7857	51.5000
52.2143	52.9286	53.6429	54.3571	55.0714	55.7857	56.5000	57.2143
57.9286	58.6429	59.3571	60.0714	60.7857	61.5000	62.2143	62.9286
63.6429	64.3571	65.0714	65.7857	66.5000	69.0000	71.5000	74.0000
76.5000	78.1667	79.8333	81.5000	83.1667	84.8333	86.5000	96.5000

표 3.4: 몸무게의 도수분포표를 바탕으로 한 자료값들 (방법 2에 따른 것)

41.5000	47.1667	47.8333	48.5000	49.1667	49.8333	50.5000	51.1667
51.8333	52.5000	53.1667	53.8333	54.5000	55.1667	55.8333	57.1667
57.8333	58.5000	59.1667	59.8333	60.5000	61.1667	61.8333	62.5000
63.1667	63.8333	64.5000	65.1667	65.8333	68.5000	70.5000	72.5000
74.5000	77.9286	79.3571	80.7857	82.2143	83.6429	85.0714	91.5000

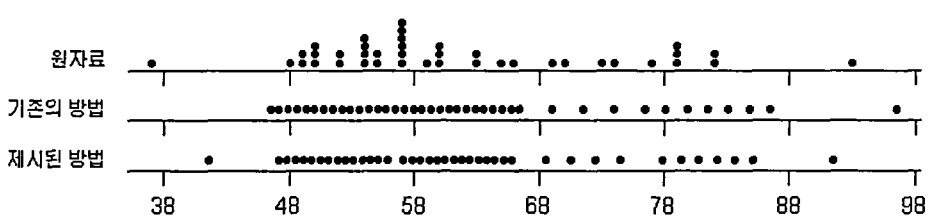

그림 3.1: 몸무게 자료의 점도표

표 3.1, 표 3.3, 표 3.4의 자료에 대하여 사분위수들인 Q_1, Q_2, Q_3를 구해 보면 다음과 같다. 이것은 정의 1.1에 따라 $Q_1=\frac{1}{2}(x_{(10)}+x_{(11)})$, $Q_2=\frac{1}{2}(x_{(20)}+x_{(21)})$, $Q_3=\frac{1}{2}(x_{(30)}+x_{(31)})$ 의 식을 사용한 결과이다.

원자료 (표 3.1) : Q_1= 54.0000, Q_2= 57.0000, Q_3= 69.5000

방법 1 (표 3.3) : Q_1= 53.2857, Q_2= 60.4286, Q_3= 70.2500

방법 2 (표 3.4) : Q_1= 52.8333, Q_2= 60.1667, Q_3= 69.5000

이 결과를 보면 방법 1의 Q_1과 방법 2의 Q_1 중 방법 1의 Q_1이 원자료의 Q_1에 더 가깝다. 그런데 Q_2와 Q_3의 경우를 보면 원자료의 값에 더 가까운 쪽은 방법 2다. 따라서 이 자료의 경우 사분위수의 추정에서 방법 2가 좀 더 좋은 결과를 준다.

표 3.5: 몇 개의 자료에 대한 방법 1과 방법 2의 사분위수 비교

자료	n	k	a_0	계급간격	원자료	방법 1	방법 2
김우철 등 (2001, p.57)	40	6	36.5	10	Q_1=54.0000 Q_2=57.0000 Q_3=69.5000	Q_1=53.2857(○) Q_2=60.4286 Q_3=70.2500	Q_1=52.8333 Q_2=60.1667(○) Q_3=69.5000(○)
김우철 등 (2001, p.79)	64	12	2.05	1.1	Q_1=6.0500 Q_2=6.9500 Q_3=8.4500	Q_1=5.9000(○) Q_2=7.1375 Q_3=8.3933(○)	Q_1=5.8694 Q_2=7.0846(○) Q_3=8.3406
김병휘 등 (2002, p.55 #4)	60	9	9.5	10	Q_1=54.5000 Q_2=71.5000 Q_3=80.5000	Q_1=56.5000 Q_2=71.2857(○) Q_3=82.0000	Q_1=55.3333(○) Q_2=71.1667 Q_3=81.8333(○)
김병휘 등 (2002, p.55 #5)	50	8	0.15	0.8	Q_1=1.5000 Q_2=2.2000 Q_3=3.1000	Q_1=1.5100(○) Q_2=2.2567 Q_3=3.2500	Q_1=1.4591 Q_2=2.2250(○) Q_3=3.1722(○)
김병휘 등 (2002, p.56)	30	6	36.065	5.21	Q_1=40.1500 Q_2=45.9900 Q_3=60.2900	Q_1=40.2330(○) Q_2=46.0509(○) Q_3=60.8125	Q_1=39.8541 Q_2=45.3686 Q_3=60.0310(○)
Walpole (1982, p.49)	40	7	1.45	0.5	Q_1=3.1000 Q_2=3.4000 Q_3=3.8500	Q_1=3.0667(○) Q_2=3.4000(○) Q_3=3.8750	Q_1=3.0594 Q_2=3.3719 Q_3=3.8364(○)

개개의 값들이 주어진 몇 개의 자료에 대하여 위와 같은 방식으로 비교한 결과가 표 3.5에 나와 있다. 이 표에서 n은 자료값의 수, k는 계급구간의 수이며, a_0는 첫 번째 계급구간의 하한을 나타낸다. 그리고 Q_1, Q_2, Q_3의 값 뒤의 괄호 안에 있는 기호(○)는 이 값이 원자료의 사분위수 값에 더 가깝다는 것을 의미한다. 이 표에 의하면, Q_1의 경우는 여섯 개의 자료 중 다섯 개에서 방법 1이 우세했고, Q_2의 경우는 세 개씩으로 동수였으며, Q_3의 경우는 다섯 개에

서 방법 2가 우세했다. 그런데 이러한 차이, 특히 Q_1의 경우 방법 1쪽에서 원자료의 Q_1에 가까운 것이 더 많이 나온 것이 크게 의미있는 것으로 보이지는 않는다. 다양한 자료들을 통하여 계속 연구할 문제라고 생각된다.

표 3.5에서 다룬 자료들을 대상으로 비교한 결과 방법 1과 방법 2가 대등하게 나왔다. 우리는 여기서 사분위수 Q_1, Q_2, Q_3만을 기준으로 하였으나, 분위수에는 사분위수 외에도 십분위수, 백분위수 등이 있고 이것들은 모두 순서통계량을 구함으로써 얻어지는 것이므로 이번에는 자료의 순서통계량들을 기준으로 하여 방법 1과 방법 2를 비교해 보자.

몇 가지의 값들을 다음과 같이 정의한다.

$$d_{1i} = (\text{방법 1의 } x_{(i)}) - (\text{원자료의 } x_{(i)})$$

$$d_{2i} = (\text{방법 2의 } x_{(i)}) - (\text{원자료의 } x_{(i)})$$

$$S_1 = \sum_{i=1}^{n} d_{1i}^2$$

$$S_2 = \sum_{i=1}^{n} d_{2i}^2$$

$$M_1 = \frac{S_1}{n}$$

$$M_2 = \frac{S_2}{n}$$

표 3.6: 순서통계량을 기준으로 한 방법 1과 방법 2의 비교

자료	n	S_1	S_2	M_1	M_2	n_1	n_2
김우철 등 (2001, p.57)	40	247.52	113.37	6.1879	2.8341	9	31
김우철 등 (2001, p.79)	64	4.0600	2.5179	0.0634	0.0393	35	29
김병휘 등 (2002, p.55, #4)	60	193.74	81.61	3.2290	1.3602	13	47
김병휘 등 (2002, p.55, #5)	50	1.1597	0.3794	0.0232	0.0076	11	39
김병휘 등 (2002, p.56)	30	12.099	17.867	0.4033	0.5956	18	12
Walpole (1982, p.49)	40	0.2671	0.0939	0.0067	0.0023	14	25

즉 S_j(j=1,2) 는 각 순서통계량의 값에 대해 방법 j와 원자료의 차이를 제곱하여 모두 합한 것이며, M_j는 S_j를 자료값의 수로 나누어 평균한 값이다. 따라서 방법 1과 방법 2 중 S_j와 M_j가 더 작은 쪽이 순서통계량 추정, 나아가서 분위수 추정의 관점에서 더 바람직한 방법이라고 할 수 있다. 또 다른 기준으로, 원자료의 n개의 순서통계량 값들 중 방법 j(j=1,2)에 의한

것에 더 가까운 것의 개수 n_j를 세어 $n_1 > n_2$이면 방법 1이 더 바람직한 방법이고, $n_1 < n_2$이면 방법 2가 더 바람직한 방법이라고 할 수 있다.

표 3.5에서 고려한 자료들에 대하여 위에서 정의한 값들을 구한 결과가 표 3.6에 정리되어 있다. 위의 표 중 Walpole(1982)의 자료에서 $d_{1i}^2 = d_{2i}^2$ 인 경우가 1건 발생하여 $n_1 + n_2 \neq n$ 이 되었다. 위의 표에 따르면, S_j와 M_j를 기준으로 하는 경우에는 고려된 여섯 개의 자료 중 다섯 개에서 방법 2가 우세했으며, n_j를 기준으로 하는 경우에는 여섯 자료 중 네 개에서 방법 2가 우세했다. 더욱이 n_j를 기준으로 할 때, 방법 2가 우세한 경우에는 n_2와 n_1의 격차가 상당히 큰 반면(n_2가 n_1의 두 배에서 네 배 정도), 방법 1이 우세한 경우에는 n_1과 n_2의 차이가 상대적으로 작았다. 따라서 순서통계량을 계산하는 관점에서 생각할 때 방법 2가 방법 1에 비해 더 바람직한 방법인 것으로 판단된다.

4. 결론

자료가 집단화되어 있는 경우 이 자료의 분위수들을 참값에 가깝게 계산하는 것이 중요하다. 대부분의 통계학 교재에서는 각 계급구간에서 가장 큰 자료값이 그 계급구간과 다음 계급구간의 경계에 해당하는 값을 갖는다고 간주하고 분위수들을 계산하고 있다.

본 논문에서는 좀 더 합리적인 가정으로, 각 계급구간 안의 자료값들이 균등한 간격으로 분포하되 계급구간의 중간점에 대하여 대칭으로 분포하고 있다고 간주하고 순서통계량들과 분위수들을 계산하는 방법을 제시하였다. 개개의 값들이 모두 주어진 몇 가지의 자료들을 사용하여 기존의 방법과 제시된 방법을 비교하니, 사분위수의 경우에는 두 방법이 거의 대등하였으나, 모든 분위수 계산의 기초가 되는 순서통계량의 계산을 기준으로 비교한 결과 제시된 방법이 기존의 방법보다 우수한 경우가 훨씬 많았다.

직관적으로 생각해도 그림 2.2에서 볼 수 있듯이 기존의 방법보다 제시된 방법이 더 합리적으로 보이므로, 제시된 방법을 사용하는 것이 더 바람직할 것으로 생각된다.

참고문헌

김병휘 외 6인 편저(2002), <통계학의 이해>, 자유아카데미.

김우철 외 7인 편저(1998), <현대통계학>, 영지문화사.

김우철 외 9인 편저(2000), <통계학개론>(제4개정판), 영지문화사.

김우철 외 8인(2001), <일반통계학>(개정판), 영지문화사.

김혁주, 김영선(2003), 집단화된 자료의 평균과 분산을 계산하는 방법에 관하여, 한국통계학회 2003년 추계 학술발표회 논문집, 227-232.

Walpole, R. E.(1982),. *Introduction to Statistics* (3rd ed.),. Macmillan.