집단화된 자료의 평균과 분산을 계산하는 방법에 관하여

김혁주[1] 김영선[2]

요 약

본 논문에서는 집단화된 자료의 평균과 분산을 계산하는 새로운 방법을 제시하였다. 제시된 방법은 각 계급구간 안의 자료값들이 그 구간에 걸쳐 균등한 간격으로 분포하고 있다고 가정하고 평균과 분산을 계산하는 것이다. 개개의 자료값들이 주어진 자료와 모의실험에 의해 생성된 자료를 이용하여 제시된 방법과 기존의 방법을 비교하였다.

주요용어: 집단화된 자료, 평균, 분산

1. 서론

주어진 통계자료가 개개의 자료값들로 이루어져 있지 않고 집단화되어 있는 상태인 경우가 종종 있다. 대부분의 통계학 교재에서는 집단화된 자료의 평균과 분산을 계산하는 방법으로 다음의 방법을 소개하고 있다. 김우철 등(2000, p.63)에 나와 있는 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. 대부분의 교재에서는, 각 계급구간의 모든 자료값들은 그 계급구간의 계급값과 같다고(즉 10이 7개, 15가 2개, ……, 40이 2개 있다고) 간주하여 이 자료의 평균 $\bar{x}$와 분산 s^2을 다음과 같이 계산한다.

표 1.1 주가와 당기순이익의 비율에 관한 자료

계 급 구 간	계 급 값	도 수
7.5 ~ 12.5	10	7
12.5 ~ 17.5	15	2
17.5 ~ 22.5	20	8
22.5 ~ 27.5	25	4
27.5 ~ 32.5	30	2
32.5 ~ 37.5	35	4
37.5 ~ 42.5	40	2
계		29

$$\bar{x} = \frac{1}{29}(10\times 7 + 15\times 2 + \cdots + 40\times 2) = \frac{640}{29} = 22.0690$$

$$s^2 = \frac{1}{29-1}\{(10)^2\times 7 + (15)^2\times 2 + \cdots + (40)^2\times 2 - \frac{(640)^2}{29}\} = 93.7808$$

1) (570-749) 전북 익산시 신용동 344-2, 원광대학교 자연과학대학 수학정보통계학부, 교수
E-mail : hjkim@wonkwang.ac.kr
2) (573-500) 전북 군산시 산북동 3581, 군산산북중학교, 교사
E-mail : miljwsun@hanmail.net

그런데 첫 번째 계급구간의 경우 7.5와 12.5 사이에 존재하는 7개의 자료값이 모두 10이라는 동일한 값을 갖는다고 간주하는 것이 유일한 합리적 가정인가 하는 의문을 가질 수 있다. 이것은 다른 계급구간의 경우에도 마찬가지이다. 본 논문에서는 계급구간 안의 자료값들이 기존의 가정과 다르게 분포하고 있다고 가정하고 자료의 평균과 분산을 계산하는 방법을 연구하고자 한다.

2. 새로운 계산 방법

본 논문에서 제시하는 방법은 계급구간 안의 자료값들이 균등한 간격으로 분포하고 있다는 가정하에 평균과 분산을 계산하는 방법이다. 자료가 표 2.1과 같은 도수분포표로 주어졌다고 하자.

표 2.1 일반적인 형태의 도수분포표

계 급 구 간	계 급 값	도 수
$a_0 \sim a_1$	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 첫 번째 계급구간의 f_1개의 자료값을 작은 값부터 큰 값까지 순서대로 $x_1, x_2, \cdots, x_{f_1}$으로 나타내면 $x_i = a_0 + id_1$ $(i=1,2,\cdots,f_1)$ (단 $d_1=(a_1-a_0)/(f_1+1)$)이며, 두 번째 계급구간의 f_2개의 자료값을 $x_{f_1+1}, x_{f_1+2}, \cdots, x_{f_1+f_2}$라 하면 $x_{f_1+j} = a_1 + jd_2$ $(j=1,2,\cdots,f_2)$ (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. 또한 그림 1.1은 표 1.1과 표 2.2의 자료를 점도표로 나타낸 것이다.

표 2.2 주가와 당기순이익의 비율에 관한 자료값들(제시된 방법에 의한 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			

이제부터는 기존의 방법을 방법 1, 본 논문에서 제시된 방법을 방법 2라 부르겠다. 방법 2에 의한 표 2.2의 자료의 평균과 분산을 계산하면 $\bar{x}=22.0690$, $s^2=95.1403$을 얻는다. 이것을 방법 1에 의한 값과 비교하면, 평균은 같고 분산은 약간 증가한 값이다. 이러한 대소관계는 일반적으로도 성립한다. 그 이유는 다음과 같다.

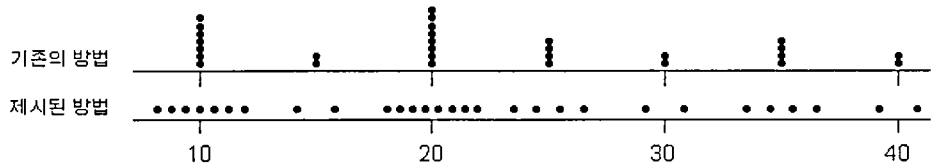

그림 1.1 주가와 당기순이익의 비율 자료를 나타낸 점도표

(1) 평균의 경우

우선 첫 번째 계급구간에 대해서 생각하자. 방법 2에 의한 자료값 $x_1, x_2, \cdots, x_{f_1}$은 등차수열을 이루므로 이 값들의 합계는 $\sum_{i=1}^{f_1} x_i = (f_1/2)(x_1 + x_{f_1})$인데, $x_1 = a_0 + d_1$이고 $x_{f_1} = a_1 - d_1$이므로 $x_1 + x_{f_1} = a_0 + a_1 = 2m_1$이다. 따라서 $\sum_{i=1}^{f_1} x_i = m_1 f_1$이 되어 방법 2에 의한 첫 번째 계급구간의 합계는 방법 1에 의한 것과 동일하다. 나머지 구간에서도 이 관계는 명백히 성립하므로 방법 1과 방법 2에 의한 자료의 총합계는 동일하며, 따라서 자료의 평균도 동일하다.

(2) 분산의 경우

자료 $x_1, x_2, \cdots, x_n$의 분산은 $s^2 = \left\{ \sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i \right)^2 \right\} / (n-1)$임을 상기하자. 본 논문에서 다루는 자료는 모두 표본 자료라고 가정한다.

방법 1과 방법 2에 의한 자료값들의 합계가 같으므로 분산의 대소관계는 자료값들의 제곱합의 대소관계와 일치한다. 첫 번째 계급구간의 경우 자료값들의 제곱합은 $\sum_{i=1}^{f_1} x_i^2$인데, $f_1 \geq 2$인 경우 $x_1^2 + x_{f_1}^2$은 다음과 같이 쓸 수 있다.

$$x_1^2 + x_{f_1}^2 = 2\left(\frac{x_1 + x_{f_1}}{2} \right)^2 + \frac{(x_1 - x_{f_1})^2}{2} > 2m_1^2 \quad \left(\frac{x_1 + x_{f_1}}{2} = m_1 \text{ 이므로} \right)$$

마찬가지로, f_1이 짝수인 경우 $x_{f_1/2}^2 + x_{f_1/2+1}^2 > 2m_1^2$까지 보일 수 있으며, f_1이 홀수인 경우 $x_{(f_1-1)/2}^2 + x_{(f_1+3)/2}^2 > 2m_1^2$까지 보일 수 있고 $x_{(f_1+1)/2} = m_1$이다. 이상의 내용을 종합하면, f_1이 짝수이든 홀수이든 2 이상이면 $\sum_{i=1}^{f_1} x_i^2 > m_1^2 f_1$임을 알 수 있다.

위의 논리는 나머지 계급구간들의 경우에도 그대로 성립한다. 그러므로 우리는 모든 계급구간에 대하여 자료값이 한 개씩만 존재하는 경우만 아니면 방법 2에 의한 자료값들의 제곱합이 방법 1에 의한 제곱합보다 크며 따라서 방법 2에 의한 자료의 분산이 방법 1에 의한 분산보다 크다는 것을 알 수 있다.

3. 개별값들이 주어진 자료와 모의자료를 통한 비교

방법 1과 방법 2에 의한 평균값은 항상 같다는 것이 앞절에서 밝혀졌으므로 이제부터는 두 방법에 의한 분산 중 어느 쪽이 실제 자료의 분산에 더 가까운가 하는 것이 주된 내용이 되겠다.

3.1 개별값들이 주어진 자료를 이용한 비교

<예1> 표 3.1의 자료는 Walpole(1982, p.49)에서 인용한 것으로서 40개의 자동차 배터리의 수명(단위: 년)을 나타낸 것이며, 개개의 값들이 주어져 있는 경우이다. 표 3.2는 표 3.1의 자료를 바탕으로 역시 Walpole(1982, p.50)에서 작성한 도수분포표이다. 또한 표 3.3은 표 3.2를 바탕으로 방법 2에 의해 얻은 자료이며, 그림 3.1은 표 3.1, 표 3.3의 자료를 나타낸 점도표이다.

표 3.1 개별값들이 주어진 자동차 배터리 수명 자료

2.2	4.1	3.5	4.5	3.2	3.7	3.0	2.6	3.4	1.6	3.1	3.3	3.8	3.1
4.7	3.7	2.5	4.3	3.4	3.6	2.9	3.3	3.9	3.1	3.3	3.1	3.7	4.4
3.2	4.1	1.9	3.4	4.7	3.8	3.2	2.6	3.9	3.0	4.2	3.5		

표 3.2 자동차 배터리 수명의 도수분포표

계 급 구 간	계 급 값	도 수
1.45 ~ 1.95	1.7	2
1.95 ~ 2.45	2.2	1
2.45 ~ 2.95	2.7	4
2.95 ~ 3.45	3.2	15
3.45 ~ 3.95	3.7	10
3.95 ~ 4.45	4.2	5
4.45 ~ 4.95	4.7	3
계		40

표 3.3 자동차 배터리 수명에 관한 자료값들 (방법 2에 의한 것)

1.61667	1.78333	2.20000	2.55000	2.65000	2.75000	2.85000	2.98125	3.01250	3.04375
3.07500	3.10625	2.13750	3.16875	3.20000	3.23125	3.26250	3.29375	3.32500	3.35625
3.38750	3.41875	3.49545	3.54091	3.58636	3.63182	3.67727	3.72273	3.76818	3.81364
3.85909	3.90455	4.03333	4.11667	4.20000	4.28333	4.36667	4.57500	4.70000	4.82500

표 3.1, 표 3.2, 표 3.3의 자료에 대하여 평균 $\bar{x}$와 분산 s^2을 계산해 보면 다음 결과를 얻는다.

원자료(표 3.1) : $\bar{x}$=3.41250, s^2=0.49394

방법 1(표 3.2) : $\bar{x}$=3.41250, s^2=0.48574

방법 2(표 3.3) : $\bar{x}$=3.41250, s^2=0.50134

이 자료에서는 공교롭게 원자료의 평균도 방법 1과 방법 2에 의한 평균과 정확히 같게 나왔는데, 이것은 항상 그런 것은 아니다. 방법 1에 의한 분산과 방법 2에 의한 분산을 비교해 보면, 방법 1에 의한 분산과 원자료의 분산의 차는 0.48574-0.49394= -0.00820, 방법 2에 의한 분산과 원자료의 분산의 차는 0.50134-0.49394=0.00740으로서 방법 2에 의한 분산이 방법 1에 의한 분산보다 원자료의 분산에 더 가깝다.

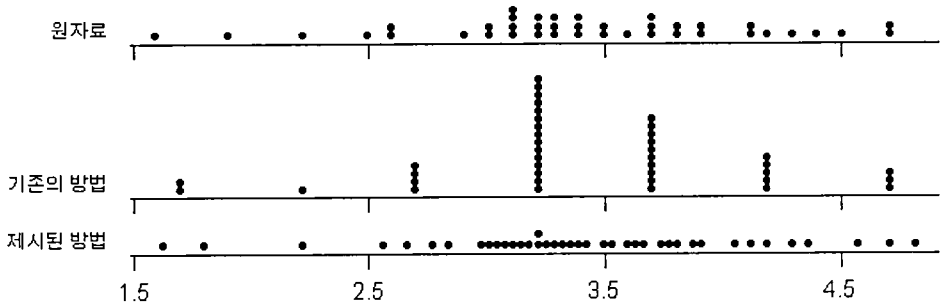

그림 3.1 자동차 배터리 수명 자료의 점도표

예 1에서는 방법 2가 방법 1보다 더 좋은 결과를 주었다. 그러나 항상 이런 결과가 나오는 것은 아니다. 개별적인 자료값들이 주어진 경우 자료가 어떻게 집단화되느냐에 따라 결과가 달라질 수 있다. 즉 계급구간의 개수와 간격 등에 따라 결과가 달라질 수 있다. 따라서 개별값들이 주어진 경우 방법 1과 방법 2의 비교 결과는 단정적으로 말할 수 없으므로 다른 방식으로 두 방법을 비교할 필요가 있다.

3.2 모의자료를 이용한 비교

이 절에서는 각 계급구간에서 난수(random number)를 발생시킴으로써 모의자료를 만들어 방법 1과 방법 2를 비교해 보겠다.

<예 2> 주가와 당기순이익의 비율에 관한 표 1.1의 자료를 다시 생각하자. 개개의 자료값들이 알려져 있지 않으므로 다음과 같은 방법으로 모의자료를 만든다. 먼저 7.5와 12.5 사이에 7개의 값이 있으므로 구간 (7.5, 12.5)에서 균일분포(uniform distribution)를 따르는 난수를 7개 발생시켜 첫 번째 계급구간의 자료값으로 삼는다. 나머지 계급구간들에 대해서도 같은 방식으로 균일분포를 사용하여 난수들을 발생시킴으로써 모의자료를 생성한다. 각 계급구간 안에 존재하는 자료값들의 개수만이 주어져 있고 다른 정보는 없는 상태이므로 이러한 가정에는 충분한 합리성이 있다. 생성된 하나의 모의자료가 표 3.4에 나와 있다. 난수들은 미니탭(Minitab)의 난수발생 기능을 사용하여 발생시켰다.

표 3.4 주가와 당기순이익의 비율에 관한 모의자료

10.9115	9.1193	10.1337	8.0296	7.6717	8.6522	10.1596
15.6555	14.1944	18.7716	21.5314	20.9271	18.8791	20.5428
19.0044	18.3387	22.0980	26.2533	26.4294	26.5645	26.1276
27.5074	30.5866	32.5638	34.3741	35.0120	36.8968	40.7916
37.8690						

표 3.4의 자료에 대하여 평균 $\bar{x}$와 분산 s^2을 계산해 보니 $\bar{x}$=21.9171, s^2=97.7166으로 얻어졌다. 방법 1에 의한 분산이 93.7808, 방법 2에 의한 분산이 95.1403으로 1절과 2절에서 얻어졌으므로, 방법 2에 의한 분산이 방법 1에 의한 분산보다 모의자료의 분산에 더 가깝다.

그런데 이것은 1회의 모의실험의 결과일 뿐이므로 신뢰성을 갖지 못한다. 모의실험의 결과가

신뢰성을 갖기 위해서는 많은 횟수의 실험을 실시해야 한다. 그래서 위와 같은 방식으로 10,000회의 모의실험을 실시하여 분산 s^2의 10,000개의 값을 구하였다.

얻어진 10,000개의 분산 s^2의 평균값은 95.8769였으며, 방법 1에 의한 분산 93.7808에 더 가까운 값이 3,935개였고 방법 2에 의한 분산 95.1403에 더 가까운 값이 6,065개였다. 이것은 모의자료의 분산에 더 가까운 것이 바람직하다는 기준에서 볼 때 방법 2가 방법 1보다 더 좋다는 것을 말해 준다. 또한 10,000개의 분산 s^2에 대하여 $(s^2-93.7808)^2$과 $(s^2-95.1403)^2$의 값을 각각 구하여 10,000개의 합을 구한 결과 전자는 322,463.0078, 후자는 283,952.5293이 나왔다. 이 기준으로 보아도 모의자료의 분산이 방법 2에 의한 분산에 평균적으로 더 가깝다는 것을 알 수 있다.

<예 3> 이번에는 또 다른 자료의 경우 모의실험을 통하여 방법 1과 방법 2를 비교해 보자. 표 3.5의 도수분포표 자료는 Daniel(1983, p.33)에 나와 있는 자료로서, 어느 특정한 병에 걸린 환자들의 발병 당시의 나이에 관한 것이다.

표 3.5 환자들의 나이에 관한 자료

계 급 구 간	계 급 값	도 수
4.5 ~ 14.5	9.5	5
14.5 ~ 24.5	19.5	10
24.5 ~ 34.5	29.5	20
34.5 ~ 44.5	39.5	22
44.5 ~ 54.5	49.5	13
54.5 ~ 64.5	59.5	5
계		75

위의 자료에 대하여 방법1과 방법2에 의해 평균과 분산을 구해보니 다음과 같이 얻어졌다.

$$\text{방법 1 : } \bar{x}=35.2333,\ s^2=168.0360$$
$$\text{방법 2 : } \bar{x}=35.2333,\ s^2=175.2628$$

예 2에서와 같은 방법에 의하여 모의실험을 실시하였다. 크기 75인 모의자료를 1,000개 만들어 1,000개의 분산 s^2을 구해 보니 이 분산들의 평균값은 176.4904였으며, 방법1에 의한 분산 168.0360에 더 가까운 값이 298개, 방법 2에 의한 분산 175.2628에 더 가까운 값이 702개로 방법 2에 의한 분산에 더 가까운 값이 압도적으로 많았다. 또한 1,000개의 분산 s^2에 대하여 $(s^2-168.0360)^2$과 $(s^2-175.2628)^2$의 값을 각각 구하여 1,000개의 합을 구한 결과 전자는 150,357.0914, 후자는 80,386.8924로서 방법 2가 방법 1에 비해 훨씬 더 좋게 나왔다.

참 고 문 헌

[1] 김우철 외 9인 편저(2000), 통계학개론(제4개정판), 영지문화사.

[2] Daniel, W. W.(1983), *Biostatistics: A Foundation for Analysis in the Health Sciences*(3rd ed.), John Wiley & Sons, Inc.

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