The Korean Communications
in Statistics Vol. 11 No. 3, 2004
pp. 513-518

Relationship Between the Mean and Median in a Skewed Frequency Distribution[1)]

Mi-Young Shin[2)] and Tae Kyoung Cho[3)]

Abstract

The well-known mode-mean-median inequality for the unimodal population distribution does not always hold for the frequency distribution. But many elementary statistics text books just mention that the relative location of the mean and median can be used to determine whether a distribution is positively or negatively skewed. In this paper we introduce the method generating data that is positively skewed but mean<median; negatively skewed but median<mean.

Keywords : skewness, mean, median

1. 서론

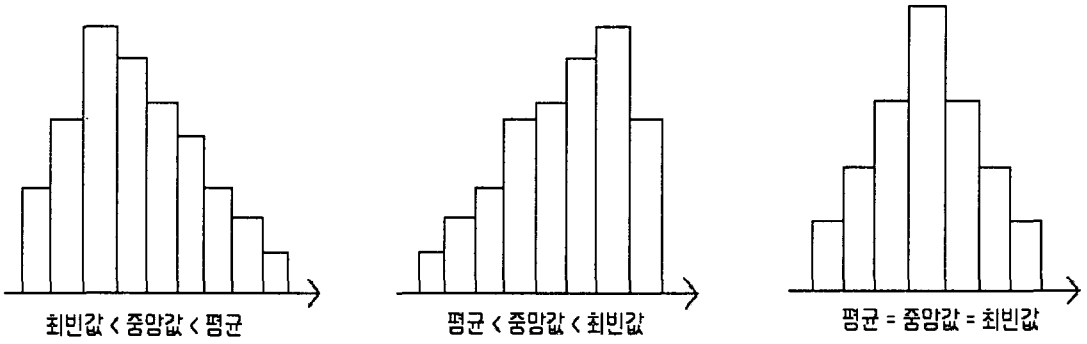
Groeneveld과 Meeden (1977)은 F-분포와 감마 분포처럼 연속형이고 단봉인 분포의 경우 왜도가 양수이면 중심측도의 크기가 최빈값<중앙값<평균 순으로 부등식이 성립되며 왜도가 음수인 경우 반대의 부등식이 성립됨을 증명하였으며 Abdous 와 Theodorescu (1998)는 중앙값 m을 갖는 단봉인 이산형 분포에서 임의의 양수 x에 대하여 $P((X-m)>x) \geq P((X-m) \leq -x)$이 만족되면 최빈값≤중앙값≤평균 부등식이 성립됨을 증명하였다.

자료 분석을 할 때 자료의 구조와 특징을 파악하는 탐색 단계에서 자료의 중심을 측정하는 도구로써 평균, 중앙값, 그리고 최빈값이 주로 사용되고 있다. 자료의 형태를 나타내는 첨도와 왜도에 따라서 각 중심측도의 중요도가 달라 질수 있기에 자료의 중심을 탐색하는 경우 어느 하나의 값만을 고려하는 것보다는 세 개의 값 모두를 고려하는 것이 바람직하다.

왜도의 부호에 따라 세 중심측도의 일반적인 부등관계가 성립하는 모집단 분포 경우와는 달리 표본 자료의 경우 왜도에 따른 중심측도들의 일반적인 부등관계가 항상 성립되지 않음에도 불구

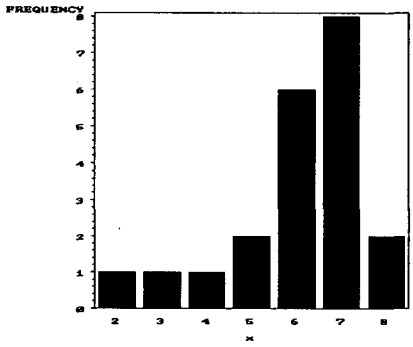
1) This work was supported by the Catholic University of Korea, Research Fund, 2004.
2) Associate Professor, Department of Mathematics, The Catholic University of Korea, Bucheon-si, Gyeonggi-do, 420-743, Korea
Email: sati@catholic.ac.kr
3) Associate Professor, Department of Statistics and Information Science, Dongguk University, Kyoungju, 780-814, Korea

하고 많은 통계학 교재 (Hines at el(1990); Kirk(1990); Freedman at el(1978); Hildebrand at el(1991); 구자홍 외 (2000); 심규박 외 (2002); 배현웅(2001); 윤상운, 이태섭(2000); 이해용, 이필영(1998);)들은 단봉을 갖는 자료에서 자료의 빈도분포가 오른쪽으로 꼬리가 길면 최빈값<중앙값<평균을 만족하고, 그 반대의 경우는 평균<중앙값<최빈값을 만족하며, 좌우대칭인 경우에는 평균=중앙값=최빈값을 만족한다고 설명하고 있다. 즉, 대부분의 통계학 교재에서는 왜도에 따른 평균, 중앙값 그리고 최빈값의 상대적 위치 관계를 단순히 다음 <그림 1>과 같이 세 가지 경우로만 설명하고 있다.

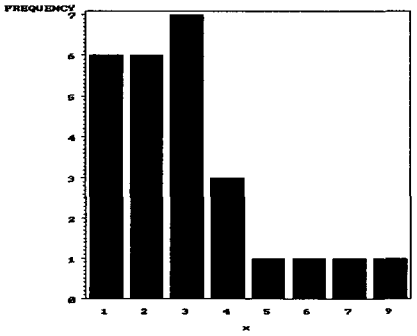


<그림 1> 중심측도의 일반적인 크기 비교

그러나 <그림 2>의 자료와 같이 왜도가 -1.29로 왼쪽으로 꼬리가 긴 분포의 모양을 갖고 있으나 중앙값이 6.00이며 평균은 6.05으로 일반적인 부등관계인 평균<중앙값이 성립되지 않거나, <그림 3>의 자료와 같이 왜도가 1.47로 비대칭을 이루고 있지만 평균과 중앙값은 같은 값 3을 갖는 예를 어렵지 않게 찾을 수 있음에도 불구하고 기초통계 교재에서는 이러한 내용이 간과되고 있는 현실이다.



<그림 2> 평균>중앙값 인 자료



<그림 3> 평균=중앙값 인 자료

본 논문에서는 일반적으로 알려진 왜도에 따른 평균과 중앙값의 크기 관계가 위배되는 모의 자료를 생성하는 방법에 대해 소개한다.

2. 새로운 자료가 추가되었을 때 평균과 왜도의 변화

왜도를 구하는 다양한 공식 중 SAS(1999)에 소개된 공식을 사용하면 크기가 n인 자료, $x_1, x_2, \cdots, x_n$의 평균과 표준편차를 각각 $\bar{x}_n$와 s_n라 할 때, 왜도 sk_n는 다음과 같이 정의된다.

$$sk_n = n\sum_{i=1}^{n}(x_i - \bar{x}_n)^3/(n-1)(n-2)s_n^3 \tag{2.1}$$

식 (2.1)을 사용하는 경우 왜도가 양수이면 오른쪽으로 꼬리가 길며, 음수인 경우에는 왼쪽으로 꼬리가 길고, 그리고 왜도가 0인 경우에는 좌우대칭이 된다. 식 (2.1)에서 왜도의 부호는 $B_n = \sum_{i=1}^{n}(x_i - \bar{x}_n)^3$에 의해 결정됨을 알 수 있다.

주어진 크기 n인 자료에 새로운 값 x^*이 추가 된 경우에 왜도는 어떠한 변화가 있는지 살펴보자.

n개의 자료에 새로운 값 x^*이 추가 된 크기 $n+1$인 자료, $x_1, x_2, \cdots, x_n, x^*$의 평균은

$$\bar{x}_{n+1} = \bar{x}_n + (x^* - \bar{x}_n)/(n+1) \tag{2.2}$$

으로 나타낼 수 있으며 새로운 $n+1$개 자료의 표준편차를 s_{n+1}이라고 할 때, 왜도 sk_{n+1}는 다음과 같이 표현된다.

$$sk_{n+1} = (n+1)\{\sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^3 + (x^* - \bar{x}_{n+1})^3\}/n(n-1)s_{n+1}^3 \tag{2.3}$$

식 (2.3)에서 왜도의 부호를 결정하는 $B_{n+1} = \sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^3 + (x^* - \bar{x}_{n+1})^3$는 다음과 같이 나타낼 수 있다.

$$B_{n+1} = B_n + \frac{3(n-1)}{n+1}(\bar{x}_n - x^*)s_n^2 - \frac{n(n-1)}{(n+1)^2}(\bar{x}_n - x^*)^3 \tag{2.4}$$

따라서 새로운 자료 값 x^*가 추가되는 경우 왜도의 부호에 어떠한 변화가 생기는지 식 (2.4)을 이용해 알아볼 수 있다.

다음 절에서는 대부분의 기초통계 교재에서 일반화 되어 설명되고 있는 것과 같이 왜도가 음수이면 평균<중앙값 이며, 왜도가 양수이면 중앙값<평균 이라는 관계가 성립되지 않는 자료를 생성하는 방법을 소개하기로 한다.

3. 자료 생성 방법

크기 n인 자료에 새로운 임의의 값이 하나 또는 두개가 추가되어도 중앙값에는 변화가 없이 중심위치 측도가 다양한 위치를 갖는 자료의 생성을 용이하게 하기 위해 처음 생성되는 크기 n인 자료는 다음과 같은 조건이 만족된다고 가정한다.

가정1. 자료는 단봉을 갖는다.
가정2. 중앙값은 자료 값 중 하나이다.
가정3. 중앙값의 앞·뒤에 중앙값과 같은 값이 있다.

왜도의 부호와 중앙값과 평균의 일반적인 크기 관계가 성립되지 않는 자료를 생성하는 방법을 단계별로 소개하면 다음과 같다.

[단계1] 위의 가정들을 만족하는 크기 n인 임의의 자료 $x_1, x_2, \cdots, x_n$을 생성한다.

이 자료의 평균은 $\bar{x}_n$, 중앙값은 Me_n이라 하자.

[단계2] $\bar{x}_{n+1}=Me_{n+1}$이 만족되는 새로운 값 x^*을 추가한다.

위의 가정2와 가정3에 의해 새로운 값 x^*이 추가되어도 $Me_{n+1}=Me_n$이 되므로 식 (2.3)에서 $\bar{x}_{n+1}$ 대신 $Me_n(=Me_{n+1})$을 대입하여 $\bar{x}_{n+1}=Me_{n+1}$이 만족되는 x^*을 다음과 같이 구할 수 있다.

$$x^* = (n+1)Me_n - n\bar{x}_n \tag{3.1}$$

[단계3] 왜도의 부호와 중앙값과 평균의 일반적인 크기 관계가 성립되지 않는 자료를 왜도의 부호에 따라 다음 두 가지 경우로 구분하여 생성한다.

(A) 단계2에서 생성된 자료의 왜도가 양수인 경우;
생성된 자료의 평균값 보다 작은 값 중에서 수치 대입법을 통해 식 (2.4)에 설명된 B_{n+1}을 양수로 하는 값 x^{**}을 추가하여 왜도가 양수 이면서 평균<중앙값을 만족하는 자료를 생성한다.

(B) 단계2에서 생성된 자료의 도가 음수인 경우;
생성된 자료의 평균값 보다 큰 값 중에서 B_{n+1}을 음수로 하는 값 x^{**}을 추가하여 왜도가 음수 이면서 중앙값<평균을 만족하는 자료를 생성 한다.

4. 왜도가 음수이며 중앙값<평균 인 자료 생성

왜도가 음수이지만 중앙값<평균인 자료를 3절에서 소개한 방법으로 생성하도록 한다.

[단계1] 주어진 가정을 만족하는 $n=17$개의 임의의 자료를 생성한다.

2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8, 9

이 자료의 평균, 중앙값, 분산, 그리고 왜도는 각각 다음과 같다.

$$\bar{x}_{17}=4.53,\ Me_{17}=4,\ s_{17}^2=4.26,\ sk_{17}=0.82$$

[단계2] 식(3.1)에 의해 구한 새로운 값 $x^*=-5$을 추가하여 $\bar{x}_{18}=Me_{18}=4$인 자료를 생성한다. 이 자료의 왜도는 $sk_{18}=-1.25$, 분산은 $s_{18}^2=9.06$이며 $B_{18}=-516$이다.

[단계3] 새로운 값 $x^*=-5$가 추가된 자료의 왜도가 음수이기 때문에 평균값 보다 큰 값 중에서 (2.4)식, B_{19}의 값을 을 음수로 하는 x^{**}의 범위를 SAS의 DO문을 사용하여 찾은 결과는 다음과 같다.

$$4 < x^{**} \leqq 13.5$$

즉, 부등식을 만족하는 값들 중에서 하나를 새로운 값으로 다시 추가하면 왜도가 음수 이면서 중앙값<평균을 만족하는 자료가 생성된다. 예를 들어 $x^{**}=10$을 선택하면 크기가 $n=19$ 인 자료

2, 2, 3, 3, 3, 3, 3, 4, 4, 4, 5, 5, 6, 6, 7, 8, 9, -5, 10

가 생성되며 이 자료의 왜도는 $sk_{19}=-0.88$, 평균은 $\bar{x}_{19}=4.32$, 그리고 중앙값은$Me_{19}=4$ 으로 왜도는 음수이며 중앙값<평균 이 된다.

5. 결론

많은 통계학 교재에서 일반화하여 설명하고 있는 자료의 왜도에 따른 평균과 중앙값의 상대적 부등호 관계는 항상 성립되는 것은 아니며 또한 평균과 중앙값의 크기 비교에 의한 분포의 왜도 측정도 잘못된 해석을 낳을 수도 있음을 알 수 있었다. 학생들 스스로 왜도의 부호와 평균과 중앙값의 크기 관계가 일치하지 않는 자료를 본 논문에서 제시한 단계를 거쳐 생성해 봄으로써 탐색

적 자료 분석 단계에서 중요한 중심 측도의 역할을 하는 평균과 중앙값의 상대적 위치와 왜도와의 관계에 대해 좀 더 정확한 이해를 얻을 수 있을 것이라 기대된다.

참고문헌

[1] 구자홍, 김진경, 박진호, 박헌진, 이재준, 전홍석, 황진수 (2000). 「통계학-엑셀을 이용한 분석」, 자유아카데미

[2] 배현웅 (2001). 「엑셀을 이용한 통계학의 기초와 활용기법」, 교우사

[3] 심규박, 조태경, 신미영 (2002). 「통계학-개념과 논쟁거리」, 홍릉과학출판사

[4] 윤상운, 이태섭 (2000). 「실용통계학」, 자유아카데미

[5] 이해용, 이필용 (1998). 「통계학입문」, 자유아카데미

[6] Abdous, B. and Theodorescu, R. (1998). Mean, Median, Mode IV, *Statistica Neerlandica,* Vol. 52, 356-359

[7] Freedman, D., Pisani, R. and Purves R. (1978). *Statistics,* New York:Norton

[8] Groeneveld, R. A. and Meeden, G. (1977). The Mode, Median, and Mean Inequality, *The American Statistician,* Vol. 31, 120-121

[9] Hildebrand, D. K. and Ott, L. (1991). *Statistical Thinking for Manager*, Duxbury

[10] Hines, W. W. and Montgomery D. C. (1990). *Probability and Statistics in Engineering and Management Science*, Wiley

[11] Kirk, R. E. (1990). *Statistics: an introduction*, Harcourt Brace College

[12] SAS OnlineDoc Version 8 (1999). SAS Institute Inc., Cary NC

[2004년 7월 접수, 2004년 9월 채택]