韓國數學敎育學會誌 시리즈 A <數學敎育>
2006. 8. 제 45권, 제 3호, 315-318.

J. Korea Soc. Math. Ed. Ser. A: The Mathematical Education
August. 2006, Vol. 45, No. 3, 315-318.

제곱합과 교차곱합의 특성을 이용한 표본분산과 상관계수의 계산

조 태 경 (동국대학교)
신 미 영 (가톨릭대학교)

I. 서 론

주어진 자료를 분석하여 자료에 포함되어 있는 정보를 쉽고 빠르게 파악하기 위해 도수분포표, 히스토그램, 줄기와 잎 그림 등과 같은 도표나 그림으로 자료를 정리, 요약 할 필요가 있다. 그러나 히스토그램의 모양이 계급 구간의 폭 등에 따라 달라지는 것처럼 그림을 통한 분석 방법은 일관성과 객관성이 부족하며 통계적 추론에 대한 이론적 근거를 제시하지 못하는 단점이 있다. 두 연속형 자료의 관계는 산점도를 이용하여 대략적으로 파악할 수 있으나 좌표의 눈금을 변화시킨 산점도로부터 드러나는 두 변수의 관계는 다른 것처럼 보여 질 수 있다. 이러한 단점을 보완하기 위해 자료를 수치로 요약 정리하여 자료의 분포 상태를 파악하게 된다. 분포의 중심 위치를 나타내는 대표값으로서 평균, 중앙값, 최빈값이 주로 사용되고 있으며 분포의 퍼짐 정도를 나타내는 산포도는 분산, 범위 등을 사용되고 있다. 두 연속형 자료의 직선 관계를 측정하는 측도로는 표본상관계수가 있다.

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. Hanson(1975)은 행렬을 이용하여 가중평균과 분산을 계산하는 알고리즘을 소개하였으며, Chan과 Lewis(1979)는 표준편차를 구하는 알고리즘들의 정확성에 대하여 비교 연구하였다.

자료 분석을 할 때 주어진 자료에 새로운 값을 첨가하거나 또는 특정한 값을 제거한 새로운 자료로부터 평균, 분산, 상관계수와 같은 통계량을 구해야하는 경우가 발생 할 수 있다. 예를 들어, 이미 조사한 크기 n 인 자료에 새로운 값을 첨가한 크기 (n+1) 자료의 표본분산은 얼마인가? 또는 크기 n인 자료에서 이상점과 같은 특정한 값을 제거한 크기 (n-1)인 자료에 대한 표본상관계수는 얼마인가? 이런 경우 원자료가 존재한다면 단지 새로운 값을 첨가하거나 특정한 값을 제거한 후에 새로운 통계량값을 구하면 된다. 그러나 원자료의 분실등과 같이 원자료 전체를 사용할 수 없는 경우에는 새로운 정보가 추가된 자료에 대한 통계량값을 구하는 것이 쉽지 않다.

자료의 산포도를 측정하는 표본분산, 두 양적 자료의 직선관계 강도를 측정하는 표본상관계수 등 많은 통계량들은 편차 제곱 합(sum of square product)과 편차 교차곱 합(sum of cross product)의 함수 형태로 이루어져 있다(교육부, 2003; 배현웅, 2001; 심규박 · 조태경 · 신미영, 2002).

본 논문에서는 주어진 자료에 새로운 값이 첨가되거나 또는 특정한 값이 제거된 자료에 대해 표본분산 그리고 표본상관계수와 같이 편차 제곱 합 또는 편차 교차곱 합으로 구성된 통계량을 간단한 식과 단순한 계산기만을 사용해서 구할 수 있는 방법을 제시하였다.

II. 본 론

1. 편차 제곱 합과 교차곱 합의 성질

이 장에서는 크기가 n인 자료에 대한 편차 제곱 합과

* 2006년 5월 투고, 2006년 6월 심사완료.
* ZDM분류 : K15
* MSC2000분류 : 97D99
* 주제어 : 제곱 합, 교차곱 합, 분산, 상관계수

편차 교차곱 합과, 주어진 자료에 새로운 값이 추가되거나 또는 특정한 값이 제거된 새로운 자료에 대한 편차 제곱 합과 편차 교차곱 합의 관계식을 유도한다.

크기가 n인 자료 $(x_1, y_1), \cdots, (x_n, y_n)$ 에서 편차 제곱 합과 편차 교차곱 합은 각각 다음과 같이 정의된다.

$$S_{xx}^n = \sum_{i=1}^{n}(x_i - \bar{x}_n)^2,\ S_{yy}^n = \sum_{i=1}^{n}(y_i - \bar{y}_n)^2 ,$$

$$S_{xy}^n = \sum_{i=1}^{n}(x_i - \bar{x}_n)(y_i - \bar{y}_n) .$$

크기가 n인 자료에 새로운 값 (x, y) 가 추가된 크기가 (n+1)인 자료 $(x_1, y_1), \cdots, (x_n, y_n), (x, y)$ 에 대한 편차 제곱 합 S_{xx}^{n+1} 은 다음과 같이 나타낼 수 있다.

$$S_{xx}^{n+1} = \sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^2 + (x - \bar{x}_{n+1})^2$$

$$= \sum_{i=1}^{n}\{(x_i - \bar{x}_n) + (\bar{x}_n - \bar{x}_{n+1})\}^2 + (x - \bar{x}_{n+1})\}^2,$$

여기서 $\bar{x}_{n+1}$ 은 새로운 값 x 가 추가된 크기 (n+1)인 자료에 대한 산술평균으로 다음과 같다.

$$\bar{x}_{n+1} = (n\bar{x}_n + x)/(n+1) .$$

$\bar{x}_{n+1}$ 을 위의 S_{xx}^{n+1} 에 대입하여 정리하면 S_{xx}^{n+1} 과 S_{xx}^n 의 관계식은 (1)과 같아진다.

$$S_{xx}^{n+1} = S_{xx}^n + \frac{n}{n+1}(\bar{x}_n - x)^2 . \qquad (1)$$

새로운 값 (x, y) 가 추가된 크기 (n+1)인 자료에 대한 편차 교차곱 합 S_{xy}^{n+1} 과 S_{xy}^n 의 관계식은 편차 제곱 합과 같은 방법으로 구할 수 있으며 그 결과는 식(2)와 같다.

$$S_{xy}^{n+1} = S_{xy}^n + \frac{n}{n+1}(\bar{x}_n - x)(\bar{y}_n - y) . \qquad (2)$$

크기 n인 이변량 자료 $(x_1, y_1), \cdots, (x_n, y_n)$ 에서 이상점과 같은 특정한 값 $(x_j, y_j), (j = 1, \cdots, n)$, 가 제거되어 크기가 (n-1)인 자료에 대한 편차 제곱의 합과 편차 교차곱 합도 식(1)과 (2)를 유도하는 과정과 같은 방법으로 구하면 식(3)과 식(4)와 같이 구할 수 있다.

$$S_{xx}^{n-1} = S_{xx}^n - \frac{n}{n-1}(\bar{x}_{n-1} - x_j)^2 . \qquad (3)$$

$$S_{xy}^{n-1} = S_{xy}^n - \frac{n}{n-1}(\bar{x}_{n-1} - x_j)(\bar{y}_{n-1} - y_j) . \qquad (4)$$

여기서 $\bar{x}_{n-1} = (n\bar{x}_n - x_j)/(n-1)$ 과 $\bar{y}_{n-1}$ 은 원자료에서 특정한 값 $(x_j, y_j), (j = 1, \cdots, n)$, 가 제거된 크기 (n-1)인 자료에 대한 산술평균이다.

2. 표본분산

자료의 분포 상태를 수치로 나타낼 때에는 분포의 중심 위치뿐만 아니라 분포의 퍼짐 정도도 함께 나타내어야 한다. 분포의 퍼짐 정도를 나타내 산포도로서 직관적으로 생각할 수 있는 것은 자료의 최대값과 최소값의 차이인 범위이다. 범위는 계산의 간편성은 있으나 극단적인 값에 영향을 받는 단점이 있다. 일반적으로 산포도를 측정하는 측도로서 표본분산을 사용하고 있다. 크기가 n인 자료 $x_1, x_2, \cdots, x_n$ 에 대한 표본분산 s_n^2 은 다음과 같이 정의 된다.

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2 .$$

여기서 $\bar{x}_n$ 은 크기가 n인 자료에 대한 산술평균이다.

자료에 대한 평균과 분산만 알고 있으며 원자료값 자체는 분실된 상태에서 새로운 값이 추가되거나 삭제된 새로운 자료의 분산은 어떻게 구할 것인가? 표본분산은 편차의 제곱합의 함수이므로 1절에서 구한 제곱합의 관계식 (1)을 자료의 산포도를 측정하는 표본분산 식에 적용하여 원자료의 분산과 특정 자료가 첨가되거나 혹은 제거된 새로운 자료의 표본분산의 관계를 구할 수 있다.

n개의 자료에 새로운 값 x가 추가 되어 크기(n+1)인 자료 $x_1, x_2, \cdots, x_n, x$ 에 대한 표본분산 s_{n+1}^2 은

$$s_{n+1}^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(x_i - \bar{x}_{n+1})^2 + (x - \bar{x}_{n+1})^2\right)$$

이며 $s_{n+1}^2 = S_{xx}^{n+1}/n$ 이므로, 식 (1)을 이용하여

$$s_{n+1}^2 = \frac{1}{n}\left(S_{xx}^{n+1} + \frac{n}{n+1}(x - \bar{x}_n)^2\right)$$

을 얻을 수 있다. $s_n^2 = S_{xx}^n/(n-1)$ 임을 이용하여 위 식을 정리하면 식(5)와 같이 s_{n+1}^2 와 s_n^2 의 관계식을 구할 수 있다.

$$s_{n+1}^2 = \frac{n-1}{n} s_n^2 + \frac{1}{n+1}(\bar{x}_n - x)^2 . \quad (5)$$

식(4)로부터 새로운 값 x의 크기에 따라 s_{n+1}^2 과 s_n^2 의 크기는 $|x-\bar{x}_n| > \sqrt{(n+1)s_n^2/n}$ 이면 $s_{n+1}^2 > s_n^2$ 임을 알 수 있다.

n개의 자료에서 특정한 값 x_j, $(j=1,\cdots,n)$, 을 제거한 크기 (n-1)인 자료의 표본분산 s_{n-1}^2 과 원자료 표본분산 s_n^2 의 관계를 위와 같은 방법으로 구하면 식(6)과 같다.

$$s_{n-1}^2 = \frac{n-1}{n-2} s_n^2 - \frac{n}{(n-1)(n-2)}(\bar{x}_n - x_j)^2 . \quad (6)$$

3. 표본상관계수

산점도를 통하여 두 변수 사이의 관계를 시각적으로 파악할 때 많은 경우 자료들이 띠의 형태를 갖고 있으며, 이 띠의 형태는 직선, 곡선 등 여러 모습을 가질 수 있다. 산점도에서 자료들이 얼마나 직선에 가까운가의 정도를 나타내는데 쓰이는 측도가 표본상관계수이다. 표본상관계수는 피어슨(Karl Pearson)에 의해 제안되었기에 피어슨의 표본상관계수라고도 한다. 크기가 n인 자료 $(x_1, y_1), \cdots, (x_n, y_n)$에 대해 표본상관계수는 다음과 같이 정의된다.

$$r_{xy}^n = \frac{S_{xy}^n}{\sqrt{S_{xx}^n \; S_{yy}^n}} .$$

크기가 n인 자료에 새로운 값 (x, y) 가 추가된 크기가 (n+1)인 자료 $(x_1, y_1), \cdots, (x_n, y_n), (x, y)$ 에 대한 표본 상관계수 r_{xy}^{n+1} 는

$$r_{xy}^{n+1} = \frac{S_{xy}^{n+1}}{\sqrt{S_{xx}^{n+1} \; S_{yy}^{n+1}}}$$

이 되므로 식 (1)과 식 (2)를 이용하면 식 (7)과 같이 구할 수 있다.

$$r_{xy}^{n+1} = \frac{S_{xy}^n + D_{xy}^{n+1}}{\sqrt{(S_{xx}^n + D_{xx}^{n+1})(S_{yy}^n + D_{yy}^{n+1})}} . \quad (7)$$

여기서 $D_{xy}^{n+1} = n(\bar{x}_n - x)(\bar{y}_n - y)/(n+1)$ 이며 $D_{xx}^{n+1} = n(\bar{x}_n - x)^2/(n+1)$이다. D_{yy}^{n+1}은 D_{xx}^{n+1} 과 같은 방법으로 정의한다.

크기가 n인 자료 $(x_1, y_1), \cdots, (x_n, y_n)$에서 특정한 값 (x_j, y_j), $(j=1,\cdots,n)$,이 제거된 크기가 (n-1)인 자료에 대한 표본 상관계수 r_{xy}^{n-1} 는 식 (3)과 (4)를 이용하여 식 (8)과 같이 구할 수 있다.

$$r_{xy}^{n-1} = \frac{S_{xy}^n - D_{xy}^{n-1}}{\sqrt{(S_{xx}^n - D_{xx}^{n-1})(S_{yy}^n - D_{yy}^{n-1})}} . \quad (8)$$

여기서 $D_{xy}^{n-1} = n(\bar{x}_n - x_j)(\bar{y}_n - y_j)/(n-1)$ 이며 $D_{xx}^{n-1} = n(\bar{x}_n - x_j)^2/(n-1)$ 이다. D_{yy}^{n-1} 도 같은 방법으로 정의한다.

Ⅲ. 결 론

주어진 자료에 새로운 값이 첨가되거나 특정한 값이 제거된 새로운 자료로부터 표본분산이나 표본상관계수와 같이 편차 제곱 합이나 편차 교차곱 합으로 구성된 통계량값을 구하는 경우 원자료를 사용하지 않고 원자료의 통계량값만을 이용하여 구하는 방법을 제시 하였다. 따라서 원자료가 분실 되었거나 원자료값을 구할 수 없는 경우 새로운 자료가 첨가 되거나 삭제되었을 때 통계량값의 변화를 본 논문의 방법으로 파악할 수 있겠다.

논문에서 제시한 방법은 표본분산과 표본상관계수 뿐만 아니라 단순회귀모형에서 추정된 회귀계수 또는 결정계수, 분산분석표 등 편차 제곱의 합이나 편차 교차곱 합으로 구성된 모든 통계량에 응용될 수 있다.

참 고 문 헌

교육부 (2003). 확률과 통계, 서울: 천재교육.

심규박 · 조태경 · 신미영 (2002). 통계학-개념과 논쟁거리, 서울: 홍릉과학출판사.

배현웅 (2001). 엑셀을 이용한 통계학의 기초와 활용기법, 서울: 교우사.

Abdous, B. & Theodorescu, R. (1998). Mean, Median, Mode IV, *Statistica Neerlandica*, 52, pp.356-359.

Chan, T. F. & Lewis J. G. (1979). Computing Standard Deviations: Accuracy, *Communications of the ACM* 22(9), pp.526-531.

Groeneveld, R. A. & Meeden, G. (1977). The Mode, Median, and Mean Inequality, *The American Statistician*, 31, pp.120-121.

Hanson, R. A. (1975), Stably Updating Mean and Standard Deviation of Data, *Communications of the ACM* 18(1), pp.57-58.

Updating Sample Variance and Correlation Using Sum of Squares and Sum of Cross product

Cho, Tae-Kyoung

Department of Statistics and Information Science, Dongguk University, Kyoungju, 780-814, Korea

E-mail: tkcho@dongguk.ac.kr

Shin, Mi-Young

Department of Mathematics, The Catholic University of Korea, Bucheon-si, 420-743, Korea

E-mail: sati@catholic.ac.kr

In this paper we present the simple updating formulas for a sum of product and a sum of cross product when a new value is added on or a specific value is eliminated from the original data. The sample variance and correlation for the new data set are derived by new computing formulas. Any statistic which is a function of the sum of product and a sum of cross product also can be updated by proposed method even though the original data is not available.

* ZDM Classification : K15
* 2000 Mathematics Subject Classification : 97D99
* Key Word : sum of square product, sum of cross product, sample variance, sample correlation