응용통계연구 제 **21**권 *1*호, *2008*년, *pp. 177–182*

두 자료들의 평균과 분산을 이용한 혼합자료의 분산 계산*

신미영[1] 조태경[2]

요 약

공통분산을 갖는 두 모집단에서 얻은 두 독립표본 자료로부터 공통분산을 추정하거나, 한 모집단에서 얻는 두 자료의 혼합자료로부터 모분산을 추정할 때 각 표본분산의 가중평균값인 합동추정량(pooled estimator)을 주로 사용한다. 본 논문에서는 동일한 모집단에서 얻은 혼합자료의 표본분산 식을 각 자료의 평균과 분산만 이용하여 구한 후 합동추정량과 비교한다.

주요용어: 혼합자료, 합동분산, 비합동분산.

1. 서론

우리는 일상생활에서 다양한 형태의 자료 혹은 자료를 요약한 정보를 접하게 되며, 수집된 방대한 자료로부터 관심대상에 대한 규칙성과 객관적인 결론을 이끌어내고 이를 토대로 예측을 하기 위해서는 주어진 자료를 이해하기 쉽게 정리, 요약을 하여야 한다.

단순한 숫자의 집합으로만 느껴지는 원자료를 정리하는 효율적인 방법 중의 하나는 그래프를 통한 시각적 방법일 것이다. Chamberes 등 (1983) 은 다변량 자료를 위한 시각적 방법을 제시하였으며 그 외에도 주어진 자료 혹은 정보를 시각적으로 표현하는 연구는 꾸준히 있어왔다 (Robert, 2001; Farebrother, 2002).

또한 숫자로서 분포의 중심위치를 파악하는 값으로서는 평균, 중앙값, 최빈값 등이 있다. Groeneveld와 Meeden (1977)은 연속형이고 단봉인 분포의 경우 왜도가 양수이면 중심측도의 크기가 최빈값 < 중앙값 < 평균 순으로 부등식이 성립되며 왜도가 음수인 경우 반대의 부등식이 성립됨을 보였으며, Abdous와 Theodorescu (1998)는 중앙값 m을 갖는 단봉인 이산형 분포에서 임의의 양수 x에 대하여 $P((X-m)>x) \geq P((X-m)-x)$이 만족되면 최빈값 $\leq$ 중앙값 $\leq$ 평균 부등식이 성립됨을 증명하였다.

조태경과 신미영 (2006)은 주어진 자료에 새로운 값이 첨가되거나 특정한 값이 제거된 새로운 자료로부터 표본분산이나 표본상관계수와 같이 편차 제곱 합이나 편차 교차곱 합으로 구성된 통계량 값을 구하는 경우 원자료(raw data)를 사용하지 않고 원자료의 통계량 값만을 이용하여 구하는 방법을 제시하였다.

* 본 연구는 2007년도 가톨릭대학교 교비연구비의 지원으로 이루어졌음.

1) (420-743) 교신저자. 경기도 부천시 원미구 역곡2동, 가톨릭대학교 수학과, 부교수.
E-mail: sati@catholic.ac.kr
2) (780-814) 경상북도 경주시 석장동, 동국대학교 정보통계학과, 교수.
E-mail: tkcho@dongguk.ac.kr

크기 n인 자료 $x_1, \ldots, x_n$에서 구한 평균과 분산을 각각 아래와 같이 정의하며 크기가 m인 자료 $y_1, \ldots, y_m$에서 구한 평균, $\bar{y}_m$과 분산, s_y^2도 같은 방법으로 정의한다.

$$\bar{x}_n = \sum_i^n x_i/n, \quad s_x^2 = \sum_i^n (x_i - \bar{x}_n)^2/(n-1).$$

기존 수업 방식과 새로운 수업 방식에 따른 학습효과의 차이를 알아본다든가 또는 학부모의 교육수준에 따른 학생들이 언어능력 차이를 알아보는 것처럼 두 모집단간의 차이가 우리의 관심의 대상이 되는 경우 두 모집단에서 얻은 두 독립표본의 평균과 표준오차를 이용하여 통계적 추론을 한다.

두 자료 $x_1, \ldots, x_n$과 $y_1, \ldots, y_m$이 공통분산을 갖는 두 모집단에서 얻은 독립표본이라면 각 자료의 편차제곱합 $\sum_i^n (x_i - \bar{x}_n)^2$과 $\sum_i^m (y_i - \bar{y}_m)^2$을 이용하여 아래의 합동추정량(pooled estimator) s_p^2로 공통분산의 불편추정량을 구하였다 (Hogg와 Tanis, 2006).

$$s_p^2 = \frac{\sum_i^n (x_i - \bar{x}_n)^2 + \sum_i^m (y_i - \bar{y}_m)^2}{n+m-2}.$$

이 추정값은 두 자료의 표본크기를 반영한 두 표본분산 s_x^2과 s_y^2의 가중평균으로, $n = m$이면 s_p^2은 s_x^2과 s_y^2의 평균값이 된다.

Moser와 Stevens (1992)은 두 집단의 평균차의 t-검정에서 표준오차의 합동(pooled)추정량 대신 비합동(unpolled) 추정량을 사용해야 한다고 주장하였으며 Julious (2005)는 두 집단 이상의 평균차를 비교할 때 왜 합동추정량이 사용되어야 하는가에 대하여 논하였다.

그러나 한 모집단에서 n개의 자료를 구한 후 추정량의 오차를 줄이기 위해 추가로 m개의 자료를 얻은 경우에는 $(n+m)$개의 혼합자료에서 얻은 분산으로 모분산을 추정하여야 할 것이다. 본 논문에서는 이와 같이 동일한 모집단에서 구한 원자료는 분실되고 각 자료의 평균과 분산만 알고 있을 때 두 자료가 결합된 혼합자료의 분산을 구하는 방법을 제시하였다.

2. 혼합자료의 비합동 분산

두 집단 간의 차이를 알아보기 위해 두 표본자료를 이용하는 경우와는 달리 단일 모집단으로부터 얻은 두 자료를 결합한 새로운 혼합자료로부터 정보를 얻어야 할 때도 있다. 예를 들어 전국 중학생의 평균키를 알아보기 위해 n명을 표본 조사하여 평균 키를 구하였다고 하자. 오차를 줄이기 위해 m명의 표본을 더 조사하였을 때, $(n+m)$명의 원자료가 있다면 $(n+m)$명의 자료로부터 통계적 추론에 필요한 표본평균과 표준오차를 계산할 수 있을 것이다. 그러나 원자료는 분실되고 우리가 갖고 있는 정보는 n명과 m명 키의 표본평균과 표본분산 뿐이라면 크기 $(n+m)$인 혼합자료의 분산은 어떻게 구할 것인가? $(n+m)$개의 자료가 한 모집단에서 얻은 자료라면 합동추정량 s_p^2보다 비합동 추정량인 $(n+m)$개의 표본분산이 더 의미가 있을 것이다. 본 논문에서는 동일한 모집단에서 구한 두 자료의 평균과 분산만 이용하여 두 자료를 결합한 혼합자료의 표본분산을 구하는 방법을 알아본다.

두 자료가 합쳐진 크기 $(n+m)$인 혼합자료 $x_1,\ldots,x_n,y_1,\ldots,y_m$의 평균과 분산을 각각 $\bar{w}_t, s_t^2$라고 하자. 자료들의 선형함수 형태인 합동평균은 각 평균의 가중평균으로 아래 식 (2.1)과 같이 구할 수 있다.

$$\bar{w}_t = \frac{\sum_i^n x_i + \sum_j^m y_j}{n+m} = \frac{n\bar{x}_n + m\bar{y}_m}{n+m}. \tag{2.1}$$

혼합자료의 분산 s_t^2은 식 (2.2)와 같이 정의된다.

$$s_t^2 = \frac{1}{n+m-1}\left[\sum_i^n (x_i-\bar{w}_t)^2 + \sum_j^m (y_j-\bar{w}_t)^2\right]. \tag{2.2}$$

각 자료의 평균과 분산만을 알고 있다면 s_t^2은 어떻게 구할 수 있을 것인가?

식 (2.2)의 $\sum_i^n (x_i-\bar{w}_t)^2$에 $\bar{w}_t$을 대입하면 아래 식 (2.3)과 같이 나타낼 수 있으며, $\sum_j^m (y_j-\bar{w}_t)^2$도 같은 방법으로 구한다.

$$\begin{aligned}\sum_i^n (x_i-\bar{w}_t)^2 = &\left(\frac{1}{n+m}\right)^2\left[(n+m)^2\sum_i^n x_i^2 - (n+2m)\left(\sum x_i\right)^2\right.\\ &\left.-2m\left(\sum x_i\right)\left(\sum y_j\right) + n\left(\sum y_j\right)^2\right].\end{aligned} \tag{2.3}$$

식 (2.3)을 (2.2)에 대입하여 분산 s_t^2을 정리하면 식 (2.4)와 같이 각 자료의 평균과 분산만으로 구할 수 있다.

$$\begin{aligned}s_t^2 &= \frac{1}{n+m-1}\left[\sum_i^n (x_i-\bar{w}_t)^2 + \sum_j^m (y_j-\bar{w}_t)^2\right]\\ &= \frac{1}{n+m-1}\left[(n-1)s_x^2 + (m-1)s_y^2 + \frac{nm}{n+m}(\bar{x}_n-\bar{y}_m)^2\right].\end{aligned} \tag{2.4}$$

따라서 원자료를 분실하였다 하더라도 각 자료의 평균과 분산만 알고 있다면 같은 모집단에서 얻은 $(n+m)$개의 혼합자료를 이용하여 모평균에 대한 신뢰구간, 가설검정과 같은 통계적 추론을 할 수 있을 것이다.

3. 비합동 분산과 합동분산의 비교

본 절에서는 2절에서 유도한 비합동 분산 식 (2.4)와 합동 분산 식 (3.1)을 두 표본평균이 같은 경우와 그렇지 않은 경우에 각각 비교한다.

$$s_p^2 = \frac{1}{n+m-2}\left[(n-1)s_x^2 + (m-1)s_y^2\right] \tag{3.1}$$

(1) $\bar{x}_n = \bar{y}_m$인 경우

두 자료의 평균이 같은 경우에는, 혼합자료의 분산 식 (2.4)가 아래 식 (3.2)와 같이 정리되며 식 (3.1)과 (3.2)의 비교를 통해 항상 $s_t^2 < s_p^2$임을 알 수 있다.

$$s_t^2 = \frac{1}{n+m-1}\left[(n-1)s_x^2 + (m-1)s_y^2\right]. \tag{3.2}$$

(2) $\bar{x}_n \neq \bar{y}_m$인 경우

두 자료의 평균이 다른 경우에는 s_t^2/s_p^2이 식 (3.3)과 같이 정리된다.

$$\frac{s_t^2}{s_p^2} = \frac{n+m-2}{n+m-1} + \frac{nm\left(\bar{x}_n - \bar{y}_m\right)^2}{(n+m-1)(n+m)} \frac{n+m-2}{(n-1)s_x^2 + (m-1)s_y^2}. \tag{3.3}$$

이 경우 두 자료의 표본크기가 같은지 다른지 또는 표본 분산이 같은지 다른지에 따라 다양한 값을 갖게 되어 일반적인 크기 비교는 하기 어렵다. 그러나 식 (3.3)을 살펴보면 두 자료의 평균의 차, $(\bar{x}_n - \bar{y}_m)$이 커질수록 합동분산 s_p^2과 비합동 분산 s_t^2의 차이도 커지는 특징이 있다.

4. 결론

추정량의 오차를 줄이기 위해 같은 모집단으로부터 추가로 자료를 얻었다면 두 자료는 같은 정보를 갖고 있게 된다. 이 경우 원자료들이 존재한다면 두 자료를 결합한 혼합자료의 평균과 분산은 쉽게 구할 수 있을 것이다. 본 논문에서는 원자료는 분실되고 각 자료의 평균과 분산만 알고 있을 때 두 자료가 결합된 혼합자료의 분산 s_t^2을 구하는 방법을 제시하였다.

합동분산 s_p^2은 표본크기를 가중치로 갖는 두 표본분산의 가중평균이었다. 본 논문에서 구한 혼합자료의 분산 s_t^2은 표본의 크기뿐만 아니라 두 자료의 평균의 차에 따라 그 값이 달라지는 특징이 있다. 두 자료의 평균이 같은 경우에는 합동분산 s_p^2이 혼합자료의 비합동 분산 s_t^2보다 항상 큰 값을 가지며, 두 자료의 평균의 차가 커질수록 추정량도 더 커짐을 알 수 있었다.

참고문헌

조태경, 신미영 (2006). 제곱합과 교차곱합의 특성을 이용한 표본분산과 상관계수의 계산, 한국수학교육학회지 시리즈 A, <수학교육>, **45**, 317–320.

Abdous, B. and Theodorescu, R. (1998). Mean, median, mode IV, *Statistica Neerlandica,* **52,** 356–359.

Chambers, J. M., Cleverland, W. S. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis*, Duxbury Press, New York.

Farebrother, R. W. (2002). *Visualizing Statistical Models And Concepts,* Marcel Dekker, New York.

Groeneveld, R. A. and Meeden, G. (1977). The mode, median, and mean inequality, *The American Statistician,* **31**, 120–121.

Hogg, R. V. and Tanis, E. A. (2006). *Probability and Statistical Inference*, Prentice Hall.

Julious, S. A. (2005). Why do we use pooled variance analysis of variance?, *Pharmaceutical Statistics,* **4**, 3–5.

Moser, B. K. and Stevens, G. R. (1992). Homogeneity of variance in the two-sample means test, *The American Statistician,* **46,** 19–21.

Robert, S. (2001). *Information Visualization,* Addison-Wesley.

[2007년 9월 접수, 2007년 11월 채택]

Calculating Sample Variance for the Combined Data*

Mi-Young Shin[1] Tae-Kyoung Cho[2]

Abstract

There are times when we need more sample to achieve a more accurate estimator. Since these two sets of sample have the information about the same population, it is necessary to treat both as a single combined data. In this paper we present the unpooled sample variance for the combined data when we just know a sample mean and variance for the each data set without the raw data. It is shown that the pooled variance s_p^2 is always greater than the exact variance s_t^2 when $\bar{x}_n = \bar{y}_m$. And the difference of means for two data, $\bar{x}_n - \bar{y}_m$, is larger, the difference of s_p^2 and s_t^2 is larger.

Keywords: Combined data, pooled variance, unpooled variance.

* This work was supported by the Catholic University of Korea, Research fund, 2007.

1) Corresponding author. Associate Professor, Dept. of Mathematics, The Catholic University of Korea, Bucheon-si 420-743, Korea.
E-mail: sati@catholic.ac.kr

2) Professor, Dept. of Statistics and Information Science, Dongguk University, Kyoungju 780-814, Korea.
E-mail: tkcho@dongguk.ac.kr