

## 표본분산에 대한 고찰

장대홍<sup>1)</sup>

요약

우리는 모분산  $\sigma^2$ 에 대한 추정량으로서 표본분산  $S_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ 을 주로 사용한다. 그러나, 제 7차 교육 과정에 따른 고등학교 수학 교과서(10-가, 수학 I 과 실용수학)에서는 표본분산의 정의를  $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ 으로 사용하고 있다. 이 두 표본분산들의 관계를 알아보고, 시뮬레이션을 통하여 확인하여 본다.

주요용어: 모분산, 표본분산

### 1. 서론

우리는 모분산  $\sigma^2$ 에 대한 추정량으로서 표본분산  $S_{n-1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}$ 을 주로 사용한다. 이러한 표본분산은  $E(S_{n-1}^2) = \sigma^2$  즉, 불편성이라는 성질을 갖는다. 모든 통계학 대학교재(예로, 김우철외 9인(2000), McClave와 Sincich(2000), Ott와 Longnecker(2001), 구자홍외 6인(2003), John과 Kuby(2003), Mann(2004))에서는 모분산에 대한 추정량으로서 이러한 표본분산을 사용한다. 그러나, 제 7차 교육 과정에 따른 고등학교 수학 교과서(10-가, 수학 I 과 실용수학)에서는 표본분산의 정의를  $S_n^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$ 로 사용하고 있다. 이렇게 정의된 표본분산은 모분산에 대한 편의추정량이 된다. 이 두 표본분산들의 관계를 알아보고, 시뮬레이션을 통하여 확인하여 본다.

### 2. 두 종류 표본분산들의 관계

$S_{n-1}^2$ 과  $S_n^2$ 의 정의로부터  $S_n^2 = \frac{n-1}{n} S_{n-1}^2$ 이므로

$$E(S_n^2) = \frac{n-1}{n} E(S_{n-1}^2) = \frac{n-1}{n} \sigma^2, \quad \text{Var}(S_n^2) = \left(\frac{n-1}{n}\right)^2 \text{Var}(S_{n-1}^2)$$

1) (608-737) 부산광역시 남구 대연3동 599-1 부경대학교 수리과학부 통계학전공, 교수  
E-mail: dhjang@pknu.ac.kr

이 됨을 알 수 있다. 여기서,  $Var(S_n^2)$ 과  $Var(S_{n-1}^2)$ 은 각각  $S_n^2$ 과  $S_{n-1}^2$ 의 분산을 가르킨다.  $S_n^2$ 은 모분산  $\sigma^2$ 에 대한 편의추정량이 된다. 반면,  $S_{n-1}^2$ 은 모분산  $\sigma^2$ 에 대한 불편추정량이 된다. 이러한 이유로 모분산  $\sigma^2$ 에 대한 추정량으로서  $S_{n-1}^2$ 이 주로 사용되어진다. 그러나,  $S_n^2$ 이 모분산  $\sigma^2$ 에 대한 편의추정량이라는 하나  $S_n^2$ 의 분산이  $S_{n-1}^2$ 의 분산보다 작게 됨으로 우리는 추정량들의 평균제곱오차(Mean Square Error, MSE)의 입장에서 이러한 표본분산들을 평가하여 볼 필요가 있다. 즉, 불편성과 유효성을 동시에 고려해 볼 필요가 있다.  $n$ 이 커지면  $S_{n-1}^2$ 과  $S_n^2$ 의 값은 비슷하여 진다. 문제가 되는 것은  $n$ 이 작을 때  $S_{n-1}^2$ 과  $S_n^2$  중 어느 추정량을 사용하는 것이 좋으나 하는 것이다. 제 7차 교육 과정에 따른 고등학교 수학 교과서(10-가, 수학 I 과 실용수학)에서는 표본분산을  $S_n^2$ 로 정의하고 수치 예로 모두  $n$ 이 작을 때 ( $n < 30$ )만 예시하고 있다.

표본분산의 평가를 위하여 우리는 다음과 같은 일반화된 표본분산을 정의하여 볼 수 있다.

$$S_{n-c}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-c} \quad (0 \leq c \leq 1) \quad (1)$$

이러한  $S_{n-c}^2$ 은  $S_{n-1}^2$ 과  $S_n^2$ 을 포함한다. 즉,  $c=0$ 이면  $S_n^2$ 이 되고,  $c=1$ 이면  $S_{n-1}^2$ 이 된다.  $S_{n-c}^2$ 과  $S_{n-1}^2$ 의 MSE를 각각  $MSE_{n-c}$ ,  $MSE_{n-1}$ 이라 하자. 그러면, 우리는 다음과 같은 사실을 알 수 있다.

**사실 1.**

$$MSE_{n-c} \leq MSE_{n-1} \text{을 만족하려면 } \frac{Var(S_{n-1}^2)}{\sigma^4} \geq \frac{\left(\frac{c-1}{n-c}\right)^2}{1 - \left(\frac{n-1}{n-c}\right)^2} \text{ 이 되면 된다.}$$

다.

(증명) 
$$MSE_{n-c} = \left(\frac{n-1}{n-c}\right)^2 Var(S_{n-1}^2) + \left(\frac{n-1}{n-c} - 1\right)^2 \sigma^4 \quad \text{이고,}$$

$MSE_{n-1} = Var(S_{n-1}^2)$  이다. 이 식들을  $MSE_{n-c} \leq MSE_{n-1}$ 에 대입하여 정리하면

$$\frac{\text{Var}(S_{n-1}^2)}{\sigma^4} \geq \frac{\left(\frac{c-1}{n-c}\right)^2}{1 - \left(\frac{n-1}{n-c}\right)^2} \text{ 이 된다.}$$

모집단이 정규분포인 경우 사실 1을 적용하여 보자. 모집단이 정규분포이면  $\text{Var}(S_{n-1}^2) = \frac{2}{n-1}\sigma^4$ 이므로 사실 1로부터  $MSE_{n-c} \leq MSE_{n-1}$ 을 만족하려면  $n \geq \frac{3c+1}{c+3}$ 이 되면 된다는 것을 알 수 있다.  $0 \leq c \leq 1$ 일 때  $\frac{1}{3} \leq \frac{3c+1}{c+3} \leq 1$ 이 된다. 즉, 모집단이 정규분포이면  $S_{n-c}^2$ 은  $c \neq 1$ 일 때 모분산  $\sigma^2$ 에 대한 편추정량이라는 하나 MSE의 입장에서는  $S_{n-1}^2$ 보다 나은 추정량임을 알 수 있다. 얼마나 나은 지를 보기 위하여  $\frac{MSE_{n-c}}{MSE_{n-1}} = \frac{(n-1)(2(n-1) + (c-1)^2)}{2(n-c)^2}$ 를  $n$ 과  $c$ 의 함수로 보고 그림을 그리면 다음 그림 2.1과 같다.  $c \neq 1$ 일 때 모든  $n$ 에 대하여  $S_{n-c}^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE보다 작음을 알 수 있고,  $n$ 이 작을 때는  $c$ 가 작을 수록  $S_{n-c}^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE보다 크게 작아짐을 알 수 있다.  $n$ 이 커지면  $S_{n-c}^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE와 거의 같아짐을 알 수 있다.

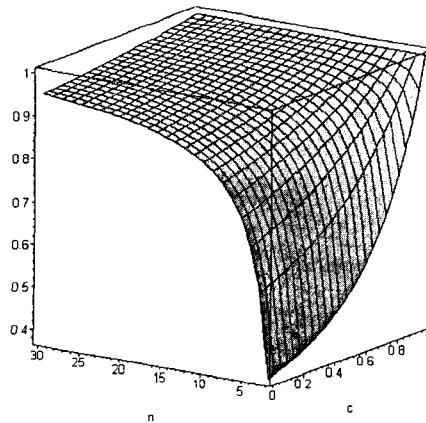


그림 2.1.  $\frac{MSE_{n-c}}{MSE_{n-1}}$ 의 그래프

다음 표 2.1은  $n$ 이 2, 5, 15, 30이고,  $c$ 가 0, 0.5, 1인 경우의  $\frac{MSE_{n-c}}{MSE_{n-1}}$  값들이다.

표 2.1.  $\frac{MSE_{n-c}}{MSE_{n-1}}$  값들

| $n$ | $c$   |       |   |
|-----|-------|-------|---|
|     | 0     | 0.5   | 1 |
| 2   | 0.375 | 0.500 | 1 |
| 5   | 0.720 | 0.815 | 1 |
| 15  | 0.902 | 0.941 | 1 |
| 30  | 0.951 | 0.971 | 1 |

사실 1로부터 임의의 모집단 하에서  $c = 0$ 일 때  $MSE_n \leq MSE_{n-1}$ 을 만족하려면

$$\frac{Var(S_{n-1}^2)}{\sigma^4} \geq \frac{1}{2n-1} \quad (2)$$

이 되면 된다는 것을 알 수 있다.

위와 같은 사실을 검토하기 위하여 3절에서는 모집단이 정규모집단인 경우를 포함하여 정규분포와 아주 다른 분포인 연속일양분포와 감마분포인 경우 세 가지 경우로 나누어 시뮬레이션을 행하여 보았다.

### 3. 시뮬레이션

모집단을 세 가지 경우(정규분포, 연속일양분포, 감마분포)로 나누고 표본의 크기  $n = 2, 5, 15, 30$ 인 경우 각각 1000번씩 시행하여  $S_{n-1}^2$ 과  $S_n^2$ 을 비교하였다.

#### 3.1. 정규모집단인 경우

모집단이 표준정규분포를 이룬다고 가정하자. 표 3.1은  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편倚, MSE를 나타낸 표이다.  $S_{n-1}^2$ 에 대해 먼저 살펴보면 이론적으로는  $\frac{(n-1)S_{n-1}^2}{\sigma^2}$ 이  $\chi^2(n-1)$  분포를 이룬다.  $n$ 이 작을 때는  $S_{n-1}^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지면서 완전하지는 않지만 정규분포를 닮아감을 알 수 있다. 표본의 크기에 관계없이  $E(S_{n-1}^2) = \sigma^2$ 이 성립함을 알 수 있다.  $S_n^2$ 에 대해 살펴보면  $n$ 이 작을 때는  $S_n^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지면서 완전하지는 않지만 정규분포를 닮아감을 알

수 있다. 그러나,  $S_{n-1}^2$ 일 때와는 다르게  $E(S_n^2) \neq \sigma^2$ 이 되나  $n$ 이 커지면서 이 편차는 많이 준다. 모든  $n$ 에 대하여  $S_n^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE보다 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는 거의 없어진다.

표 3.1.  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편차, MSE

| n  | variance | $S_{n-1}^2$<br>bias | MSE    | variance | $S_n^2$<br>bias | MSE    |
|----|----------|---------------------|--------|----------|-----------------|--------|
| 2  | 1.7706   | -0.0194             | 1.7710 | 0.4427   | -0.5097         | 0.7025 |
| 5  | 0.4787   | -0.0237             | 0.4793 | 0.3063   | -0.2190         | 0.3543 |
| 15 | 0.1414   | 0.0008              | 0.1414 | 0.1232   | -0.0660         | 0.1276 |
| 30 | 0.0711   | 0.0049              | 0.0711 | 0.0664   | -0.0286         | 0.0673 |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 1000번의 시행으로 얻어진 표본에 기초한  $\sigma^2$ 에 대한 95% 신뢰구간을 구하면 다음 표 3.2와 같다. 모든  $n$ 에 대하여  $S_n^2$ 에 기초한 신뢰구간의 폭이  $S_{n-1}^2$ 에 기초한 신뢰구간의 폭보다 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는 거의 없어진다.

표 3.2. 표본에 기초한 95% 신뢰구간

| n  | $S_{n-1}^2$      | 신뢰구간의 폭 | $S_n^2$          | 신뢰구간의 폭 |
|----|------------------|---------|------------------|---------|
| 2  | (0.0017, 5.0564) | 5.0547  | (0.0008, 2.5282) | 2.5274  |
| 5  | (0.1247, 2.7415) | 2.6168  | (0.0997, 2.1932) | 2.0935  |
| 15 | (0.4140, 1.8240) | 1.4100  | (0.3864, 1.7024) | 1.3160  |
| 30 | (0.5552, 1.5673) | 1.0121  | (0.5367, 1.5151) | 0.9784  |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 (2) 식을 확인하여 보면 (2) 식을 모두 만족함을 알 수 있다.

### 3.2. 연속일양모집단인 경우

모집단이 확률밀도함수가  $f(x) = 1 (0 \leq x \leq 1)$ 인 연속일양분포를 이룬다고 가정하자. 이 분포는 대칭분포이나 꼬리도 없고, 확률밀도가 중앙에 몰리지 않고 일정한 분포이다. 표 3.3은  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편차, MSE를 나타낸 표이다.  $S_{n-1}^2$ 에 대해 먼저 살펴보면  $n$ 이 작을 때는  $S_{n-1}^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지면서 정규분포를 닮아감을 알 수 있다. 표본의 크기에 관계없이  $E(S_{n-1}^2) = \sigma^2$ 이 성립함을 알 수 있다.  $S_n^2$ 에 대해 살펴

표본분산에 대한 고찰

보면  $n$ 이 작을 때는  $S_n^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지면서 정규분포를 닮아감을 알 수 있다. 그러나,  $S_{n-1}^2$ 일 때와는 다르게  $E(S_n^2) \neq \sigma^2$ 이 되나  $n$ 이 커지면서 이 편차는 많이 준다. 모든  $n$ 에 대하여  $S_n^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE보다 조금 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는 거의 없어진다.

표 3.3.  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편차, MSE

| n  | $S_{n-1}^2$ |          |         | $S_n^2$  |          |         |
|----|-------------|----------|---------|----------|----------|---------|
|    | variance    | bias     | MSE     | variance | bias     | MSE     |
| 2  | 0.00998     | 0.00090  | 0.00998 | 0.00250  | -0.04120 | 0.00420 |
| 5  | 0.00172     | -0.00077 | 0.00172 | 0.00110  | -0.01728 | 0.00140 |
| 15 | 0.00043     | 0.00111  | 0.00043 | 0.00037  | -0.00452 | 0.00039 |
| 30 | 0.00020     | 0.00027  | 0.00020 | 0.00019  | -0.00252 | 0.00020 |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 1000번의 시행으로 얻어진 표본에 기초한  $\sigma^2$ 에 대한 95% 신뢰구간을 구하면 다음 표 3.4와 같다. 모든  $n$ 에 대하여  $S_n^2$ 에 기초한 신뢰구간의 폭이  $S_{n-1}^2$ 에 기초한 신뢰구간의 폭보다 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는 거의 없어진다.

표 3.4. 표본에 기초한 95% 신뢰구간

| n  | $S_{n-1}^2$        | 신뢰구간의 폭 | $S_n^2$            | 신뢰구간의 폭 |
|----|--------------------|---------|--------------------|---------|
| 2  | (0.00017, 0.35100) | 0.35083 | (0.00009, 0.17550) | 0.17541 |
| 5  | (0.01444, 0.17259) | 0.15815 | (0.01155, 0.13808) | 0.12653 |
| 15 | (0.04515, 0.12874) | 0.08359 | (0.04214, 0.12016) | 0.07802 |
| 30 | (0.05439, 0.11047) | 0.05608 | (0.05258, 0.10679) | 0.05421 |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 (2) 식을 확인하여 보면 (2) 식을 모두 만족함을 알 수 있다.

### 3.3. 감마모집단인 경우

모집단이 확률밀도함수가  $f(x) = \frac{1}{\Gamma(\alpha)\theta^\alpha} x^{\alpha-1} e^{-\frac{x}{\theta}}$  ( $0 < x < \infty$ )인 감마분포를 이루고  $\alpha = 2, \theta = 2$ 라고 가정하자. 이 분포는 오른쪽으로 치우친 전형적인 비대칭분포이다. 표 3.5는  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편차, MSE를 나타낸 표이다.  $S_{n-1}^2$ 에 대해 먼저 살펴보면  $n$ 이 작을

때는  $S_{n-1}^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지면서 비대칭성이 상당히 완  
화됨을 알 수 있다. 표본의 크기에 관계없이 거의  $E(S_{n-1}^2) = \sigma^2$ 이 성립함을 알 수 있다.  $S_n^2$   
에 대해 살펴보면  $n$ 이 작을 때는  $S_n^2$ 의 분포가 오른쪽으로 치우친 분포를 이루다가  $n$ 이 커지  
면서 비대칭성이 상당히 완화됨을 알 수 있다. 그러나,  $S_{n-1}^2$ 일 때와는 다르게  $E(S_n^2) \neq \sigma^2$ 이  
되나  $n$ 이 커지면서 이 편차는 많이 준다. 모든  $n$ 에 대하여  $S_n^2$ 의 MSE가  $S_{n-1}^2$ 의 MSE보다  
매우 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는 거의 없어진다.

표 3.5.  $S_{n-1}^2$ 과  $S_n^2$ 에 대한 분산, 편차, MSE

| n  | $S_{n-1}^2$ |        |         | $S_n^2$  |        |        |
|----|-------------|--------|---------|----------|--------|--------|
|    | variance    | bias   | MSE     | variance | bias   | MSE    |
| 2  | 200.572     | 0.082  | 200.579 | 50.143   | -3.959 | 65.817 |
| 5  | 77.826      | -0.260 | 77.894  | 49.809   | -1.808 | 53.078 |
| 15 | 19.342      | 0.030  | 19.343  | 16.849   | -0.505 | 17.104 |
| 30 | 11.342      | 0.114  | 11.355  | 10.598   | -0.157 | 10.623 |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 1000번의 시행으로 얻어진 표본에 기초한  $\sigma^2$ 에  
대한 95% 신뢰구간을 구하면 다음 표 3.6과 같다. 모든  $n$ 에 대하여  $S_n^2$ 에 기초한 신뢰구간의  
폭이  $S_{n-1}^2$ 에 기초한 신뢰구간의 폭보다 작음을 알 수 있다. 그러나,  $n$ 이 커지면서 이 차이는  
거의 없어진다.

표 3.6. 표본에 기초한 95% 신뢰구간

| n  | $S_{n-1}^2$       |         | $S_n^2$           |         |
|----|-------------------|---------|-------------------|---------|
|    | 신뢰구간의 폭           | 신뢰구간의 폭 | 신뢰구간의 폭           | 신뢰구간의 폭 |
| 2  | (0.0050, 48.3700) | 48.3650 | (0.0027, 24.1850) | 24.1823 |
| 5  | (0.5760, 29.0710) | 28.4950 | (0.4610, 23.2570) | 22.7960 |
| 15 | (2.2507, 18.8599) | 16.6092 | (2.1006, 17.6025) | 15.5019 |
| 30 | (3.3789, 16.1340) | 12.7551 | (3.2663, 15.5963) | 12.3300 |

표본의 크기  $n = 2, 5, 15, 30$  각각에 대하여 (2) 식을 확인하여 보면 (2) 식을 모두 만족함을  
알 수 있다.

#### 4. 결론

우리는 시뮬레이션을 통하여 사실 1을 확인하여 보았다. 확률표본이라는 가정 하에  $S_{n-1}^2$ 은

## 표본분산에 대한 고찰

모분산  $\sigma^2$ 에 대한 불편추정량이 된다. 불편성의 입장에서는  $S_{n-1}^2$ 이  $S_n^2$ 보다 선호된다.  $S_n^2$ 이 모분산  $\sigma^2$ 에 대한 편의추정량이라는 하나 MSE 입장에서는  $S_n^2$ 이  $S_{n-1}^2$ 보다 선호된다. 특히, 표본크기  $n$ 이 작을 때 특히 그렇다. Rohatgi(1984, 654-655p)는 정규모집단인 경우

$$S_{n+1}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n+1}$$
이 가장 MSE가 작음을 보였다. 표본크기  $n$ 이 클 때에는 표본분산으로서

어떤 것을 사용하여도 무방하다. 그러나, 모집단이 유한모집단이 되면 도리어  $S_n^2$ 이 모분산  $\sigma^2$ 에 대한 불편추정량이 된다(Rohatgi(1984, 260-262p)).

## 참고문헌

- [1] 김우철외 9인(2000). <통계학개론>, 제 4개정판, 영지문화사, 서울.
- [2] 고등학교 수학 교과서(10-가) 16종(2004).
- [3] 고등학교 수학 교과서(수학 I) 11종(2004).
- [4] 고등학교 수학 교과서(실용수학) 4종(2004).
- [5] 구자홍외 6인(2003). <통계학>, 자유아카데미, 서울.
- [6] John, R. and Kuby, P.(2003). *Just the Essentials of Elementary Statistics*, 3th ed., Brooks/Cole-Thomson, Pacific Grove.
- [7] Mann, P. S.(2004). *Introductory Statistics*, 5th ed., John Wiley, New York.
- [8] McClave, J. T. and Sincich, T.(2000). *Statistics*, 8th ed., Prentice Hall, Upper Saddle river.
- [9] Ott, R. L. and Longnecker, M.(2001). *Statistical Methods and Data Analysis*, 5th ed., Duxbury, Pacific Grove.
- [10] Rohatgi, V. K.(1984). *Statistical Inference*, John Wiley, New york.