양쪽 절단된 정규분포의 평균과 분산의 추정

최윤영[1], 홍종선[2]

요 약

절단된 정규분포의 평균과 분산을 추정하기 위하여 전체 표본에 기초한 최대가능도 추정량을 사용한 방법과 절단된 후에 남아있는 표본만을 고려한 절단된 표본의 표본평균과 표본분산을 시뮬레이션을 통해 비교 연구하였다. 평균을 추정하는 경우에는 놀랍게도 절단된 자료에 기초한 추정량이 전체 표본에 기초한 추정량보다 평균제곱오차가 더 작다는 것을 발견하였다.

1. 서 론

다양한 방법으로 절단된 정규분포(truncated normal distribution)의 평균과 분산이 필요한 경우가 많다. 예를 들어 수험생 전체의 수능점수보다 어느 한 대학에 합격한 수험생들의 점수에 더 관심이 있을 수도 있다. 수험생들은 자신의 수학능력평가시험 점수에 맞추어 진학을 할 것이고 이에 따라 각 대학마다 합격한 수험생들의 상한점수와 하한점수가 존재하며, 복수지원이 가능해지면서 상한점수와 하한점수의 구분은 더욱 명확하게 되었다. 그러므로 수험생들의 점수가 정규분포를 따른다면, 어느 한 대학에 합격한 수험생들의 점수들의 분포는 양쪽이 절단된 정규분포의 형태를 갖게 된다.

원시 모집단의 전체표본을 확보하여 이 자료를 이용할 수 있을 때(어떤 경우에는 원시분포의 모수가 알려져 있는 경우도 있다), 절단 후에 절단된 집단의 평균과 분산을 추정 또는 계산하길 원하는 경우도 있다. 이런 경우에 절단된 분포의 평균과 분산을 추정하는 방법으로 두 가지 방법을 고려하였다. 첫 번째 방법은 최대가능도 추정량의 불변성(invariance)의 성질에 의해 전체 표본자료에 기초한 표본평균과 표본분산 추정량을 절단된 분포의 평균과 분산에 대체하여 구하고, 다른 방법으로는 절단 영역에 속하는 관측값을 제거하여 절단된 모집단에 대한 확률표본으로 간주하여 절단자료로부터 직접 표본평균과 표본분산을 구한다. 더 적은 자료에 기초했기 때문에 절단된 표본에 기초한 추정량이 전체 표본에 기초한 추정량보다 덜 효율적이라고 추측할지 모르지만 다음에서 보여질 것처럼 표본크기가 작은 경우의 분산 추정을 제외하고는 절단된 표본에 기초한 추정량이 더 효율적임이 밝혀졌다.

2. 양쪽 절단의 평균과 분산의 유도

Barr와 Sherrill(1999)은 한쪽 방향(왼쪽 또는 오른쪽)으로 절단된 정규분포의 평균과 분산을 표준정규누적분포함수와 자유도가 3인 카이제곱분포의 적분형태를 이용하여 구하였으며, 비표준정규분포의 평균과 분산에 대하여도 언급하였다. 본 연구에서는 양쪽 방향으로 절단된 정규분포의 경우에 평균과 분산을 다음과 같이 유도하였다.

---

1) 서울. 서초구 서초동 1678-2, 한국정보통신산업협회(KAIT), 조사연구실, 통계분석팀, 연구원. chic602@kait.or.kr

2) 서울. 종로구 명륜동 성균관대학교 통계학과, 교수. cshong@skku.ac.kr

Z가 절단점 t_1과 t_2 ($t_1 < t_2$)의 양쪽이 절단된 표준정규 확률변수라 하면(절단영역은 $(-\infty, t_1] \cup [t_2, \infty)$), Z의 밀도함수는 다음과 같다.

$$f(z) = c(t_1, t_2)\, e^{-\frac{1}{2}z^2} \quad , \quad t_1 < z < t_2 \ ,$$

여기서 $c(t_1, t_2) = 1/[\sqrt{2\pi}\{\Phi(t_2) - \Phi(t_1)\}]$이고 $\Phi(t)$는 표준정규 누적분포함수이다.

Z의 평균을 유도하면 다음과 같다.

$$E(Z) = c(t_1, t_2)\left\{ e^{-\frac{1}{2}t_1^2} - e^{-\frac{1}{2}t_2^2} \right\} . \tag{1}$$

[그림 1]은 (1)식에서 정의되었으며 절단점 t_1과 t_2 ($t_1 < t_2$)의 함수로 표현되는 양쪽방향으로 절단된 표준정규분포의 평균을 이차원 그래프로 나타내었다. [그림 1]을 통하여 왼쪽 절단점 t_1이 감소할수록 평균은 감소하고, 오른쪽 절단점 t_2가 증가할수록 평균은 증가한다는 것을 알 수 있다.

Z의 분산을 얻기 위해 우선 $E(Z^2)$을 유도하면 다음과 같다.

(1) $t_1, t_2 \geq 0$ 인 경우

$$\begin{aligned} E(Z^2) &= c(t_1, t_2) \int_{t_1}^{t_2} z^2 e^{-\frac{1}{2}z^2}\, dz \\ &= \frac{c(t_1, t_2)}{2}\sqrt{2\pi} \int_{t_1^2}^{t_2^2} \frac{1}{2^{3/2}\Gamma(3/2)}\, u^{\frac{3}{2}-1} e^{-\frac{u}{2}}\, du \\ &= c(t_1, t_2)\sqrt{\frac{\pi}{2}}\{ C_3(t_2^2) - C_3(t_1^2) \} , \end{aligned}$$

여기서 $C_3(\cdot)$는 자유도가 3인 카이제곱 분포함수를 뜻한다.

(2) $t_1 < 0 \leq t_2$인 경우

$$\begin{aligned} E(Z^2) &= c(t_1, t_2)\left\{ \int_{t_1}^{0} z^2 e^{-\frac{1}{2}z^2} dz + \int_{0}^{t_2} z^2 e^{-\frac{1}{2}z^2} dz \right\} \\ &= c(t_1, t_2)\sqrt{\frac{\pi}{2}}\{ C_3(t_1^2) + C_3(t_2^2) \} . \end{aligned}$$

(3) $t_1, t_2 < 0$ 인 경우

$$\begin{aligned} E(Z^2) &= c(t_1, t_2) \int_{t_1}^{t_2} z^2 e^{-\frac{1}{2}z^2} dz \\ &= c(t_1, t_2)\sqrt{\frac{\pi}{2}}\{ C_3(t_1^2) - C_3(t_2^2) \} . \end{aligned}$$

이전의 식을 이용하면 Z의 분산은 다음과 같다.

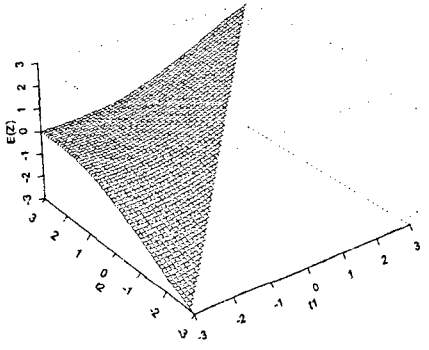
$$V(Z) = c(t_1, t_2)\left[\sqrt{\frac{\pi}{2}}\, C_3(t_1, t_2) - c(t_1, t_2)\left\{ e^{-\frac{1}{2}t_1^2} - e^{-\frac{1}{2}t_2^2} \right\}^2 \right] , \tag{2}$$

여기서 $C_3(t_1, t_2)$는 다음과 같다.

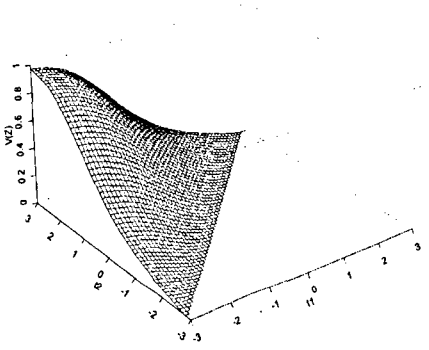
$$\begin{aligned} C_3(t_1, t_2) &= C_3(t_2^2) - C_3(t_1^2), \quad t_1, t_2 \geq 0 \quad \text{인 경우,} \\ C_3(t_1, t_2) &= C_3(t_1^2) + C_3(t_2^2), \quad t_1 < 0 \leq t_2 \quad \text{인 경우,} \\ C_3(t_1, t_2) &= C_3(t_1^2) - C_3(t_2^2), \quad t_1, t_2 < 0 \quad \text{인 경우.} \end{aligned}$$

(2)식에서 정의된 양쪽방향으로 절단된 표준정규분포의 분산을 절단점 t_1과 t_2 ($t_1 < t_2$)에 대하여 그래프로 표현한 그림이 [그림 2]이다. [그림 2]를 살펴보면, 왼쪽 절단점 t_1이 감소할수록

분산은 증가하고, 그리고 오른쪽 절단점 t_2가 증가할수록 분산이 증가한다는 것을 파악할 수 있다. 따라서 t_1과 t_2의 거리가 멀수록 분산이 증가한다는 것을 발견하였다.



[그림 1] 양쪽 절단된 표준정규분포의 평균



[그림 2] 양쪽 절단된 표준정규분포의 분산

실제로 절단된 분포의 다양한 적용에 있어서 표준정규분포보다는 오히려 비표준 정규분포의 경우가 많다. 그러므로 이런 경우에는 비표준 정규분포를 표준화시켜주어야 한다. 만약 X가 a의 왼쪽과 b의 오른쪽이 절단된, 분포가 $N(\mu, \sigma^2)$인 확률변수라면 Z는 $t_1 = (a-\mu)/\sigma$의 왼쪽과 $t_2 = (b-\mu)/\sigma$의 오른쪽이 절단된 표준정규확률변수이다. 그러므로 비표준 정규분포의 평균과 분산을 표준정규분포의 평균과 분산으로 표현하면 다음과 같다.

$$E(X) = \sigma E(Z) + \mu, \quad V(X) = \sigma^2 V(Z) \tag{3}$$

여기서 $E(Z)$와 $V(Z)$는 각각 (1)식과 (2)식에서 주어졌다.

3. 절단분포의 평균과 분산의 추정

Barr와 Sherrill(1999)은 μ와 σ가 미지의 모수인 모집단 $N(\mu, \sigma^2)$로부터 확률표본이 주어졌을 때, 한쪽 방향으로 절단된 분포의 평균과 분산을 추정하는 방법으로 두 가지를 제시했는데 한가지는 최대가능도 추정량의 불변성에 의해 절단된 평균과 분산의 추정량이 식(3)에서 μ와 σ^2에 대응하는 각각의 최대가능도 추정량 $\overline{X}$와 $(n-1)S^2/n$으로 대체되어 얻어질 수 있다.(이 방법으로 얻어진 추정량을 '전체표본 추정량'이라 하고 이 방법을 '전체표본 추정방법'이라 부른다.) 다른 방법은 절단하고자 하는 영역에 속하는 표본 관측값을 제거하여 절단된 모집단으로부터 추출된 확률표본으로 간주한다. 그리고 이 절단표본자료로부터 직접 표본평균과 표본분산을 계산하여 추정하는 것이다.(이 방법으로 얻어진 추정량을 '절단표본 추정량'이라 하고 이 방법을 '절단표본 추정방법'이라고 부른다.)

본 논문에서는 양쪽방향으로 절단된 분포의 평균과 분산을 추정하기 위하여 시뮬레이션을 이용하여 두 가지 추정방법들을 비교하였고, 그 결과들을 [표 1]에 보여주었다. 우선 표를 살펴보면 표본크기 n의 확률표본을 표준정규분포에서 10,000개 생성하고, 절단 영역 $(-\infty, t_1] \cup [t_2, \infty))$에 대해 절단된 정규분포의 평균과 분산을 추정하였다. 절단의 정도가

크고 표본크기가 작은 경우 절단 후에 자료가 하나만 남거나 하나도 남지 않게 되는 경우가 발생하는데 이런 경우는 분산을 계산할 수 없게 된다. 따라서 절단된 후에 한 개 이하의 값들이 남아있을 경우, 전체표본 추정량으로 그 값을 대체하는 방법과 자료의 손실을 가져오기는 하지만 순수한 추정량으로 두 추정방법을 비교하기 위하여 그에 해당되는 자료를 제거하는 방법을 사용하였다. 절단된 후에 표본크기가 하나 이하인 경우에 전체표본 추정값으로 대체하는데 이런 경우가 발생할 비율을 '대체율'이라 정의하고 표의 맨 오른쪽 열에 나타나 있다.

평균추정량과 분산추정량에 대한 편의를 [표 1]에서 살펴보면 대체율이 큰 경우를 제외하고는 절단표본 평균추정량의 편의는 거의 0에 가까운데 비해 전체표본 평균추정량의 편의는 절단 후의 분포가 비대칭이 되고, 표본크기 n이 작아짐에 따라 점점 커지는 것을 알 수 있다. 이는 편의가

$$b(\hat{\mu}) = E\left[c(t_1, t_2)(e^{-\frac{1}{2}t_1^2} - e^{-\frac{1}{2}t_2^2})(\tilde{\sigma} - 1) + \overline{X} \right], \quad b(\hat{\sigma^2}) = E[(\tilde{\sigma^2} - 1)V(Z)]$$

로 정의되는데 $E(Z)$의 값이 점점 커지고, $V(Z)$의 값이 점점 작아지는 영향때문이다.(여기서 $\tilde{\sigma^2}$는 σ^2의 최대가능도 추정량으로 $\tilde{\sigma^2} = (n-1)s^2/n$이다.) 반면에 절단표본 추정량들의 편의는 우리가 알고있는 것처럼 절단분포에 대한 $\overline{X_t}$와 S_t^2가 불편추정량이므로 대체율이 큰 경우를 제외하고는 0에 가까운 값을 가지는 것이다. 대체율이 큰 경우 절단표본 추정량들의 편의가 존재하는 것은 기대한 것처럼 절단표본 추정량들이 전체표본 추정량들로 대체 되었기 때문인데, 대체율이 0보다 큰 경우에 대해서 그에 해당하는 표본을 제거하는 방법을 이용하여 얻어진 결과를 보면 전체표본 추정량들의 편의는 별로 차이가 없지만 절단표본 추정량들의 편의는 거의 0에 가까워졌음을 알 수 있다.

평균추정량과 분산추정량에 대한 평균제곱오차(MSE)를 살펴보면, 평균추정량의 경우는 전체표본 추정량의 평균제곱오차가 절단표본 추정량의 평균제곱오차보다 크지만, 분산추정량의 경우는 그 반대임을 알 수 있다. 이는 전체표본 평균추정량의 분산이 절단 후의 분포가 비대칭이 됨에 따라 급속하게 커지는 $E(Z)$와 급속하게 작아지는 $V(Z)$의 값에 영향을 받기 때문이다. 평균제곱오차들의 값이 별로 차이가 나지 않는 경우는 대체율이 크다는 것을 알 수 있는데 이는 기대한 것처럼 절단표본 추정량들이 전체표본 추정량들로 대체 되었기 때문이다. [표 1]의 결과를 살펴보면 전체표본 평균추정량의 평균제곱오차는 별로 차이가 없지만 절단표본 평균추정량의 평균제곱오차는 현저하게 작아졌음을 알 수 있다.

4. 결 론

한쪽 또는 양쪽 절단 후의 분포가 비대칭인 경우 평균 추정의 경우 전체표본 평균추정량의 평균제곱오차가 절단표본 평균추정량보다 더 크다. 또한 평균제곱오차 값에 큰 영향은 미치지 못하지만 절단표본 평균추정량은 불편인데 비해 전체표본 평균추정량은 편의가 존재한다. 특히 양쪽 절단 후의 분포가 대칭인 경우, 평균제곱오차가 비대칭인 경우에 비해 약간 작기는 하지만 절단표본 평균추정량의 평균제곱오차에 비하면 여전히 큰 값이다. 그러므로 양쪽 방향으로 절단된 분포의 평균을 추정하는 경우에는 절단표본 추정방법이 훨씬 더 효율적임을 확인할 수 있다. 그러나 분산의 경우 1보다 작은 $V(Z)$의 영향으로 전체표본 분산추정량의 평균제곱오차가 절단표본 분산추정량의 평균제곱오차보다 작아진다. 전체표본 분산추정량의 편의는 평균의 경우와 마찬가지로 평균제곱오차에 크게 영향을 미치지는 못한다. 따라서 양쪽 방향으로 절단된 분포의 분산을 추정하는 경우 특히 표본크기가 작은 경우에 대하여 전체표본에 기초한 최대가능도 추정량을 사용하여 분산을 추정하는 것이 바람직하다.

[표 1] 양쪽 절단분포의 평균과 분산에 대한 추정량

t_1	t_2	$P[t_1<X<t_2]$	n	평균 전체표본 추정법 bias	평균 전체표본 추정법 MSE	평균 절단표본 추정법 bias	평균 절단표본 추정법 MSE	분산 전체표본 추정법 bias	분산 전체표본 추정법 MSE	분산 절단표본 추정법 bias	분산 절단표본 추정법 MSE	대체율
대체방법												
-1	2	0.82	10	-0.02	0.11	-0.00	0.07	-0.05	0.05	0.00	0.05	0.000
			20	-0.01	0.05	-0.00	0.03	-0.03	0.03	0.00	0.02	0.000
	$\mu_t \simeq 0.230$		36	-0.01	0.03	-0.00	0.02	-0.01	0.01	0.00	0.01	0.000
	$\sigma_t^2 \simeq 0.520$		50	-0.00	0.02	0.00	0.01	-0.01	0.01	-0.00	0.01	0.000
			100	-0.00	0.01	-0.00	0.01	-0.01	0.01	-0.00	0.00	0.000
-1	1	0.68	10	0.00	0.10	-0.00	0.05	-0.03	0.02	-0.00	0.02	0.001
			20	0.00	0.05	0.00	0.02	-0.01	0.01	-0.00	0.01	0.000
	$\mu_t \simeq 0.000$		36	-0.00	0.03	-0.00	0.01	-0.01	0.00	0.00	0.00	0.000
	$\sigma_t^2 \simeq 0.291$		50	-0.00	0.02	-0.00	0.01	-0.01	0.00	-0.00	0.00	0.000
			100	-0.00	0.01	-0.00	0.00	-0.00	0.00	-0.00	0.00	0.000
0	1	0.34	10	-0.06	0.13	-0.01	0.07	-0.03	0.01	0.00	0.03	0.018
			20	-0.02	0.06	-0.00	0.01	-0.00	0.00	-0.00	0.00	0.002
	$\mu_t \simeq 0.466$		36	-0.01	0.03	-0.00	0.01	-0.00	0.00	0.00	0.00	0.000
	$\sigma_t^2 \simeq 0.080$		50	-0.01	0.02	0.00	0.00	-0.00	0.00	0.00	0.00	0.000
			100	-0.00	0.01	0.00	0.00	-0.00	0.00	0.00	0.00	0.000
1	2	0.14	10	-0.11	0.20	-0.17	0.17	-0.01	0.00	-0.01	0.00	0.600
			20	-0.05	0.10	-0.06	0.06	-0.00	0.00	-0.00	0.00	0.221
	$\mu_t \simeq 1.383$		36	-0.02	0.05	-0.01	0.02	-0.00	0.00	0.00	0.00	0.032
	$\sigma_t^2 \simeq 0.073$		50	-0.02	0.04	-0.00	0.01	-0.00	0.00	-0.00	0.00	0.005
			100	-0.01	0.02	0.00	0.01	-0.00	0.00	0.00	0.00	0.000
제거방법												
-1	1	0.68	10	0.00	0.10	0.00	0.05	-0.03	0.02	-0.00	0.02	0.001
0	1	0.34	10	-0.03	0.11	-0.00	0.02	-0.01	0.00	-0.00	0.00	0.018
			20	-0.02	0.06	-0.00	0.01	-0.00	0.00	0.00	0.00	0.002
1	2	0.14	10	-0.10	0.20	-0.00	0.03	-0.01	0.00	-0.00	0.01	0.600
			20	-0.05	0.10	0.00	0.03	-0.00	0.00	0.00	0.00	0.221
			36	-0.03	0.06	0.00	0.02	-0.00	0.00	-0.00	0.00	0.032
			50	-0.02	0.04	0.00	0.01	-0.00	0.00	0.00	0.00	0.005

t = 절단점 　　n = 표본크기

μ_t = 절단된 표준정규분포의 평균 　　bias = 추정량의 표본평균 − 모수값

σ_t^2 = 절단된 표준정규분포의 분산 　　MSE = 추정량의 표본분산 + $(\text{bias})^2$

참고문헌

Barr, D. and Sherrill, E. (1999). "Mean and Variance of truncated Normal Distribution," The American Statistician, 53, 4, 357-361.

Cohen, A. (1949). "On Estimating the Mean and Standard Deviation of Truncated Normal Distributions," Journal of the American Statistical Association, 44, 518-525

Cohen, A. (1950). "Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples," Annals of the Mathematical Statistics, 21, 557-569.

Cohen, A. (1959). "Simplified Estimation for the Normal Distribution when Samples are Singly Censored or Truncated," Technometrics, 1, 217-237.

Cohen, A. (1961). "Tables for Maximum Likelihood Estimation: Singly Truncated and Singly Censored Samples," Technometrics, 3, 433-438.

Cohen, A. (1991). "Truncated and Censored Samples: Theory and Application," New York: Marcel Dekker.

Fisher, R. (1931). "The Truncated Nromal Distribution," British Association for the Advancement of Science, 5.

Gupta, A. (1952). "Estimation of the Mean and Standard Deviation of the Normal Population from a Censored Sample," Biometrika, 39, 260-273.

Hald, A. (1952). Statistical Theory with Engineering Applications.

Halperin, M. (1952). "Estimation in the Truncated Nromal Distribution," Journal of the American Statistical Association, 47, 457-465.

Johnson, N., and Kotz, S. (1970). Continuous Univariate Distribution-1.

Keceioglu, D. (1991). Reliability Engineering Handbook(Vol. 1), Englewood Cliffs.