한국통계학회논문집 제2권 2호
1995년 12월 pp. 221-226

붓스트랩 표준편차 추정량으로 표준화한 U-통계량을 이용한 비모수적 검정법[1)]

이 기 훈[2)]

요 약

본 연구는 붓스트랩에 의한 U-통계량의 분산추정방법을 제안하고, 추정량의 일치성을 증명하였다. 결과적으로 붓스트랩 추정량으로 표준화한 U-통계량의 값이 표준정규분포에 근사함을 보였다. 또한 실제적인 비모수검정에서 이를 응용하여 검정력과 특성을 연구하였다.

1. 서 론

$X_1, X_2, \cdots, X_n$이 분포함수 F를 갖는 모집단에서 얻은 확률표본이라 할 때, 모수의 함수형식 $\theta(F)$를 추정하는 문제를 고려해 보자. $\theta(F)$가 추정가능(estimable)하다면 다음과 같은 식이 성립한다.

$$\theta(F) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \psi(x_1, \cdots, x_n)\, dF(x_1) \cdots dF(x_n) .$$

$\theta(F)$를 추정할 수 있는 n의 최소값이 m일 때, $\psi(x_1, \cdots, x_m)$을 차수(degree)가 m인 $\theta(F)$의 커널(kernel)이라 하며, 이는 각 변수(argument)에 관하여 대칭이라 가정한다. 일표본(one-sample) U-통계량은 다음과 같이 정의된다.

$$U_n = \binom{n}{m}^{-1} \sum_{1 \le a_1 < \cdots} \cdots \sum_{< a_m \le n} \psi(X_{a_1}, \cdots, X_{a_m}) .$$

이 통계량은 $\theta(F)$에 관한 최소분산불편추정량으로 알려져 있다. Hoeffding (1948)은 $X_1, \cdots, X_n$이 독립이라는 조건과 몇 가지 제약조건하에서 U-통계량의 근사분포가 정규분포임을 증명하였다. 즉,

$$(U_n - E(U_n)) / (Var(U_n))^{1/2} \xrightarrow{d} N(0,1) .$$

여기서 $Var(U_n)$의 분산은 다음과 같은 조건부 분산 σ_c^2, $c=0,1,\cdots,m$을 이용하여 정의된다.

---

1) 이 논문은 1993년도 한국학술진흥재단의 공모과제 연구비에 의하여 연구되었음.
2) (560-759) 전북 전주시 완산구 효자동 전주대학교 통계학과.

$$\sigma_0^2 = 0 ,$$

$$\sigma_c^2 = Var[\psi_c(X_1,\cdots,X_c)] , \quad c=1,\cdots,m ,$$

단, $\psi_c(x_1,\cdots,x_c) = E[\psi(x_1,\cdots,x_c,X_{c+1},\cdots,X_m)]$.

이때 U-통계량의 분산은 다음과 같다.

$$Var(U_n) = \sum_{c=1}^{m} \binom{n}{m}^{-1} \binom{m}{c}\binom{n-m}{m-c} \sigma_c^2 . \tag{1}$$

그런데 $Var(U_n)$은 종종 미지의 모집단 분포 F에 의존하게 된다. 비모수적 검정법을 이루는 U-통계량의 경우에도 Lehmann(1951), Potthoff(1974)에서와 같이 $Var(U_n)$의 분포가 F에 종속되어 정규분포근사에 의한 검정을 어렵게 하기도 한다.

이 논문에서는 $Var(U_n)$의 붓스트랩(bootstrap)추정량을 제안하고 그 추정량이 일치추정량(consistent estimator)임을 증명하였다. 이 성질은 표본의 일반적인 경우로 확장될 수 있어서, 이를 이용한 이표본 척도검정법을 제안하고 그 검정법의 모의검정력과 특성을 살펴보았다.

2. 붓스트랩 분산추정량의 특성

앞에서 $Var(U_n)$의 표현식은 다음과 같이 고칠 수 있다.

$$Var(U_n) = \sum_{c=1}^{m} \binom{n}{m}^{-1} \binom{m}{c}\binom{n-m}{m-c} \sigma_c^2$$

$$= \sum_{c=0}^{m} \binom{n}{m}^{-1} \binom{m}{c}\binom{n-m}{m-c} \eta_c(F) - \eta_0(F) ,$$

여기서, $\eta_c(F) = \int \Psi^{(c)}(x_1,\cdots,x_{2m-c}) \prod_{i=1}^{2m-c} dF(x_i)$; $c=0,1,\cdots,m$,

$$\Psi^{(c)}(x_1,\cdots,x_{2m-c}) = \frac{1}{(2m-c)!} \sum_{(2m-c)!} \psi(x_{i_1},\cdots,x_{i_m})\, \psi(x_{i_{m-c+1}},\cdots,x_{i_{2m-c}}) ,$$

단, $\Sigma_{(2m-c)!}$는 $(1,2,\cdots,2m-c)$의 모든 순열(permutation)에 대하여 계산된다.

그런데 붓스트랩 통계량은 경험적 분포함수(empirical distribution function) F_n에 기초하므로, 위의 식에서 $\eta_c(F)$대신 다음과 같은 $\eta_c(F_n)$을 사용한다면, 원하는 추정량을 얻을 수 있다.

$$\eta_c(F_n) = \frac{1}{n^{2m-c}} \sum_{i_1=1}^{n} \cdots \sum_{i_{2m-c}=1}^{n} \Psi^{(c)}(x_{i_1},\cdots,x_{i_{2m-c}})$$

$$= V_n^{(c)} .$$

위의 식은 von Mises의 V-통계량이라 불린다. 붓스트랩 분산추정량은 다음과 같이 표현된다.

$$\widehat{Var(B)} = \sum_{c=0}^{m} \binom{n}{m}^{-1} \binom{m}{c}\binom{n-m}{m-c} V_n^{(c)} - V_n^{(o)} \quad . \qquad (2)$$

다음은 Lee(1985), Lee(1990)의 결과들을 이용해 이 분산 추정량이 일치추정량임을 증명한 정리이다.

정 리 $E[\psi(X_1,X_2,\cdots,X_c)^2] < \infty$, $c=1,\cdots,m$ 이면

$$\sqrt{n}(\widehat{Var(B)} - Var(U_n)) \xrightarrow{p} 0.$$

증 명 : $\psi(x_1,\cdots,x_m)\psi(x_{m-c+1},\cdots,x_{2m-c})$의 대칭화된 커널인 $\Psi^{(c)}(x_1,\cdots,x_{2m-c})$에 기초한 U-통계량을 $U_n^{(c)}$라 한다면, 이는 다음 식의 불편추정량이다.

$$E[\psi(X_1,\cdots,X_m)\,\psi(X_{m-c+1},\cdots,X_{2m-c})] = \sigma_c^2 + \theta^2 \; ; \; c=0,\cdots,m \; .$$

그런데 Janssen(1981)에 의하면 다음과 같은 부등식이 성립한다.

$$E[U_n^{(c)} - (\sigma_c^2 + \theta^2)]^2 \le kr\, n^{-1},$$

여기서 k는 상수이고, $r = E[\Psi^{(c)}(X_1,\cdots,X_{2m-c}) - \sigma_c^2 - \theta^2]^2 < \infty$.

또한, V-통계량과 U-통계량의 관계와 Minkowski 부등식에 의하여 다음과 같은 사실이 증명된다.

$$E(U_n^{(c)} - V_n^{(c)})^2 = O(n^{-2}) \; .$$

그러므로 $V_n^{(c)}$는 $\sigma_c^2 + \theta^2$에 수렴하고, (1)과 (2)의 관계식에서 이 정리가 증명된다.

즉 위의 정리를 통해서 $(U_n - \theta)/(\widehat{Var(B)})^{1/2}$ 는 표준정규분포에 수렴함을 알 수 있다. 여기서는 일표본의 경우에만 증명하였지만, 이 사실은 일반적인 k-표본의 경우로 바로 확장될 수 있다.

3. 이표본 척도검정에서의 응용

$X_1,X_2,\cdots,X_m$과 $Y_1,Y_2,\cdots,Y_n$이 분포함수가 각각 $F(\cdot)$와 $G(\cdot)$인 모집단에서 뽑은 확률표본이라 가정할 때 $G(x)=F(x/\delta)$의 관계식을 갖는다 하자. 이때 두 모집단의 척도모수가 같으냐는 검정의 가설은 다음과 같다.

$$H_0 : \delta = 1 \quad \text{대} \quad H_1 : \delta \neq 1$$

이를 검정하는 모수적 방법은 다음과 같은 검정통계량을 이용한 F-검정법이다.

$$F = \frac{\sum(Y_i - \overline{Y})^2/(n-1)}{\sum(X_i - \overline{X})^2/(m-1)} \; .$$

Lehmann(1951)이 제안한 U-통계량은 다음과 같다.

$$U = \binom{m}{2}^{-1} \binom{n}{2}^{-1} \sum_{i<j} \sum_{k<l} \phi(|X_i - X_j| < |Y_k - Y_l|) ,$$

여기서, $\phi(t) = \begin{cases} 1, & t > 0 \\ 0, & t \leq 0 \end{cases}$.

그러나 위의 U 통계량의 분산은 F에 종속되므로 Sukhatme(1959)가 다음과 같은 통계량을 제안하였다.

$$T = \frac{1}{mn} \sum_i \sum_j K(X_i, Y_j) ,$$

여기서, $K(X, Y) = \begin{cases} 1, & \text{만약 } 0<X<Y \text{ 또는 } Y<X<0 \\ 0, & \text{그외의 경우} \end{cases}$.

위의 U 통계량과 T 통계량은 정규분포에 근사한다는 사실을 이용해 표준화하여 표준정규분포에 의한 검정을 시행한다. U의 분산은 붓스트랩 분산추정량을 사용하였고, T의 평균과 분산은 다음과 같다.

$$E(T) = \frac{1}{4} ,$$

$$Var(T) = \frac{m+n+7}{48mn} .$$

U-통계량의 붓스트랩 분산추정량은 (2)식을 이용해 직접 계산이 가능하지만 방대한 계산량을 필요로 하기 때문에, 다음과 같은 Efron(1982)의 방법을 사용하였다.

1. 미지의 분포 F와 G에서 얻은 확률표본 $x_1, x_2, \cdots, x_m$과 $y_1, y_2, \cdots, y_n$에 의하여 경험적 분포함수 F_m과 G_n을 구하고, 통계량 U를 계산한다.

2. F_m과 G_n에서 각각 붓스트랩 확률표본, $x_1^*, x_2^*, \cdots, x_m^*$과 $y_1^*, y_2^*, \cdots, y_n^*$을 얻는다.

3. $x_1^*, x_2^*, \cdots, x_m^*$과 $y_1^*, y_2^*, \cdots, y_n^*$에 따른 U-통계량 U^{*i}를 계산한다.

4. 2-3의 절차를 NB번 반복한 뒤, 분산을 다음과 같이 구한다.

$$\widehat{Var(B)} = \frac{1}{NB} \sum_{i=1}^{NB} (U^{*i} - U)^2 .$$

모의실험에서 표본의 수는 $m=n=10,20,30$으로 하였고, 분포는 균일분포, 정규분포, 이중지수분포에서 살펴보았다. 척도모수의 값은 $\delta = 1, \sqrt{2}, \sqrt{3}, 2$ 등의 값으로 변화시켰으며 유의수준은 $\alpha = 0.05(0.10)$으로 하였다. 붓스트랩 반복횟수는 $NB = 100$으로 하였고, 500번 반복실험하여 기각된 횟수를 세어 모의검정력을 구하였다. 표에서 나타난 숫자는 경험적 검정력에 1000을 곱한 숫자이다.

표. 경험적 유의수준 (X1000)

(α = 0.05(0.10))

분포	n	10			20			30		
	δ^2	F	U	T	F	U	T	F	U	T
균	1	10 (28)	32 (78)	52 (94)	8 (30)	42 (64)	46 (80)	4 (30)	50 (110)	56 (110)
일	2	168 (350)	350 (502)	238 (382)	426 (644)	678 (778)	464 (618)	606 (796)	824 (900)	582 (734)
분	3	450 (658)	620 (746)	408 (470)	840 (950)	922 (960)	660 (796)	980 (1000)	990 (1000)	846 (912)
포	4	702 (854)	806 (884)	520 (660)	974 (994)	984 (992)	798 (904)	1000 (1000)	998 (1000)	948 (980)
정	1	46 (104)	88 (138)	48 (108)	30 (90)	46 (100)	44 (94)	74 (112)	92 (130)	52 (70)
규	2	244 (378)	302 (424)	154 (262)	400 (564)	438 (578)	260 (402)	604 (720)	612 (716)	410 (550)
분	3	496 (640)	544 (676)	296 (460)	726 (850)	762 (840)	516 (666)	892 (958)	886 (934)	706 (820)
포	4	608 (750)	666 (770)	394 (540)	896 (954)	910 (944)	700 (800)	986 (996)	978 (994)	882 (942)
이	1	108 (168)	76 (134)	40 (87)	164 (218)	96 (160)	46 (114)	124 (178)	46 (94)	46 (84)
중	2	230 (414)	262 (348)	144 (258)	458 (558)	342 (468)	228 (334)	530 (656)	438 (574)	326 (436)
지	3	512 (614)	470 (576)	250 (374)	678 (768)	586 (710)	394 (548)	772 (852)	682 (802)	492 (656)
수	4	602 (702)	558 (668)	290 (452)	828 (866)	722 (816)	496 (636)	896 (932)	828 (902)	664 (782)

표에서 볼 때 표본의 크기가 작을 때에는 U는 분산의 추정값으로 표준화시키기 때문에 표준정규분포보다 두터운 분포를 갖으리라 추측된다. 그래서 정규분포의 기각값을 사용한 위의 결과에서 경험적 유의수준이 과대하게 추정되는 약점을 보여준다. 표본의 크기가 커지면 이런 현상은 줄어들고 검정력이 우수하게 나타난다. 소표본의 경우에 붓스트랩 분산추정량으로 표준화한 통계량에 의한 검정법에는 새로운 기각역을 찾거나 통계량의 조정으로 정확한 유의수준을 추정하는 개선이 필요하다 할 수 있다.

참 고 문 헌

[1] Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*, SIAM, Philadelphia.

[2] Efron, B. and Stein, C. (1981). The Jackknife Estimate of Variance, *The Annals of statistics*, Vol. 3, 586-596.

[3] Hoeffding, W. (1948). A Class of Statistics with Asymptotically Normal Distribution, *The Annals of Mathematical Statistics*, Vol. 19, 293-325.

[4] Janssen, P. (1981). Rate of Convergence in the Central Limit Theorem and in the Strong Law of Large Numbers for von Mises Statistics, *Metrika*, Vol. 28, 35-46.

[5] Lee, A. J. (1985). On Estimating the Variance of a U-Statistic, *Communications in Statistics, Part A-Theory and Methods*, Vol. 14, 289-301.

[6] Lee, A. J. (1990). *U-Statistics: Theory and Practice*, Marcel Dekker, New York.

[7] Lehmann, E. L. (1951). Consistency and Unbiasedness of Certain Nonparametric Tests, *The Annals of Mathematical Statistics*, Vol. 22, 165-179.

[8] Potthoff, R. F. (1974). A Nonparametric Test for Whether Two Simple Regression Lines are Parallel, *The Annals of Statistics*, Vol. 2, 295-305.

[9] Sukhatme, B. V. (1959). On Certain Two-Sample Nonparametric Tests for Variance, *The Annals of Mathematical Statistics*, Vol. 30, 188-194.