Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.1 pp. 129 137

Bootstrap Tests for the General Two-Sample Problem[1)]

Kil-Ho Cho[2)] · Seong-Hwa Jeong[3)]

Abstract

Two-sample problem is frequently discussed problem in statistics. In this paper we consider the hypothese methods for the general two-sample problem and suggest the bootstrap methods. And we show that the modified Kolmogorov-Smirnov test is more efficient than the Kolmogorov-Smirnov test.

Keywords : 붓스트랩 방법, 카이제곱 검정, Kolmogorov-Smirnov 검정

1. 머리말

두 모집단으로부터 각각 독립적으로 확률표본을 추출했을 때, 이 추출된 두 집단의 표본을 이용하여 두 모집단이 동일한 분포를 가지는가에 대한 문제는 많은 통계학자들의 관심이 되어왔으며, 특히 비모수적 방법을 이용한 이표본 문제는 많은 연구자들에 의해서 연구되어져 왔다. (Gibbons(1971), Pratt와 Gibbons(1981), Baumgartner와 Schindler(1998), Hollander와 Douglas(1999) 참조)

연속분포이며 서로 독립인 두 모집단으로부터 $X_1, X_2, \cdots, X_m$을 확률분포 F를 가지는 모집단 1에서 추출한 확률표본이라 하고, $Y_1, Y_2, \cdots, Y_n$을 확률분포 G를 가지는 모집단 2에서 추출한 확률표본이라 하자.

이표본 문제에서 관심있는 귀무가설과 대립가설은 다음과 같이 둘 수 있다.

$$H_0 : F(t) = G(t), \text{ 모든 } t\text{에 대해서}$$

---

1. 이 논문은 2000년도 경북대학교의 연구비에 의하여 연구되었음.
2. 대구광역시 북구 산격동 1370번지 경북대학교 통계학과 교수
E-mail : khcho@knu.ac.kr
3. 대구광역시 북구 산격동 1370번지 경북대학교 통계학과

$$H_1 : F(t) \neq G(t), \text{ 적어도 한점의 } t\text{에 대해서} \quad (1.1)$$

붓스트랩 방법은 Efron(1979)에 의해 제안되었으며, Bickel과 Freedman(1981), Davision과 Hinkley(1997)에 의해서 근사적 이론이 규명되었다. 대부분의 통계적 기법은 모집단에 대해서 정규성과 같은 기본 가정을 함에 반하여, 붓스트랩 방법은 별다른 기본 가정이 없이도 사용 가능하다. 또한 현대에 와서 컴퓨터의 지속적인 발전으로 계산 속도가 빨라짐으로 인하여 수리적으로 복잡하거나, 수리적으로 계산이 불가능한 어려운 여러 가지 문제를 컴퓨터의 빠른 계산 능력을 사용함으로써 쉽게 해결할 수 있게 되었다.

본 연구에서는 이표본 문제에 있어서 가설 (1.1)에 대한 통계적 검정방법들을 소개하고, 각 검정방법에 대해서 붓스트랩 방법을 적용하여 그 방법의 유효성을 살펴보고자 한다.

먼저 2장에서는 이표본 문제에 대한 검정방법을 소개하고, 각 방법에 대해서 붓스트랩 검정방법의 알고리즘을 제시하였다. 3장에서는 2장에서 소개한 검정방법들에 대해서 모의실험을 실행하고, 그 결과를 그래프로 나타내어 서로 비교하므로써 붓스트랩 방법의 유효성과 일치성을 보였다.

2. 이표본 검정방법

가설 (1.1)에 대한 이표본 검정방법을 살펴보고자 한다.

2.1 카이제곱 검정

두 모집단에서 추출한 관측값을 r개의 계급으로 분류한 후, 아래와 같은 $2\times r$분할표를 만든다.

계 급	1	2	...	r	합 계
모집단 1	O_{11}	O_{12}	...	O_{1r}	m
모집단 2	O_{21}	O_{22}	...	O_{2r}	n
합 계	N_1	N_2	...	N_r	N

O_{1i}는 계급 $i(i=1, 2, \cdots, r)$에서의 모집단 1의 관측도수이며, O_{2i}는 계급 $i(i=1, 2, \cdots, r)$에서의 모집단 2의 관측도수이고, $N_i = O_{1i}+O_{2i}$, $N=m+n$ 이다.

(1.1)의 귀무가설 H_0하에서 계급 i에 대한 모집단 1과 모집단 2의 기대도수를

E_{1i}와 E_{2i}로 나타내면 다음과 같다.

$$E_{1i} = \frac{m}{N} N_i, \quad E_{2i} = \frac{n}{N} N_i, \quad i = 1, 2, \cdots, r.$$

그리고 검정통계량

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

은 귀무가설 H_0하에서 근사적으로 자유도 $(r-1)$을 가지는 카이제곱분포 χ^2_{r-1}를 따른다. 따라서 $\chi^2 \geq \chi^2_{(\alpha, r-1)}$이면 유의수준 α로 H_0를 기각한다. 여기서 $\chi^2_{(\alpha, r-1)}$은 χ^2_{r-1}의 $100(1-\alpha)$ 백분위수를 나타낸다.

2.2 이표본 콜모고로프-스미르노브 검정

확률표본 $X_1, X_2, \cdots, X_m$에 대한 경험적 분포함수 $F_m(x)$은

$$F_m(x) = \begin{cases} 0, & x < X_{(1)} \text{ 일때}, \\ \dfrac{i}{m}, & X_{(i)} \leq x < X_{(i+1)} \text{ 일때}, \\ 1, & x \geq X_{(m)} \text{ 일때}. \end{cases}$$

여기서 $X_{(i)}$는 확률표본 $X_1, X_2, \cdots, X_m$의 i번째 순위통계량이다.

확률표본 $Y_1, Y_2, \cdots, Y_n$에 대한 경험적 분포함수 $G_n(y)$은

$$G_n(y) = \begin{cases} 0, & y < Y_{(1)} \text{ 일때}, \\ \dfrac{i}{n}, & Y_{(i)} \leq y < Y_{(i+1)} \text{ 일때}, \\ 1, & y \geq Y_{(n)} \text{ 일때}. \end{cases}$$

여기서 $Y_{(i)}$는 확률표본 $Y_1, Y_2, \cdots, Y_n$의 i번째 순위통계량이다.

확률표본 $X_1, X_2, \cdots, X_m$와 $Y_1, Y_2, \cdots, Y_n$에 대한 혼합표본의 순위를 다음과 같이 둔다.

$$Z_{(1)} \leq Z_{(2)} \leq \cdots \leq Z_{(N)}, \quad N = m + n$$

가설 (1.1)에 대한 이표본 콜모고로프-스미르노브 검정(Two-sample Kolmogorov-Smirnov test)의 검정통계량은 다음과 같다.

$$D_{m,n} = \max_{i=1,\cdots,N} |F_m(Z_{(i)}) - G_n(Z_{(i)})|$$

따라서 검정절차는 $D_{m,n} \geq d(\frac{\alpha}{2}, m, n)$이면 유의수준 α로 H_0를 기각한다. 여기서

$d(\alpha, m, n)$은 H_0하에서 $D_{m,n}$ 분포의 $100(1-\alpha)$ 백분위수를 나타낸다. 또한

$$J^* = \left(\frac{mn}{N}\right)^{\frac{1}{2}} D_{m,n}$$

으로 두면, J^*의 분포는 $\min(m,n) \to \infty$ 일 때, H_0하에서

$$P_0(J^* < s) = \sum_{k=-\infty}^{\infty} (-1)^k e^{-2k^2 s^2}$$

가 된다.

2.3 붓스트랩 검정

2절에서 소개한 이표본 검정법에 대해서 붓스트랩 방법을 적용하는 절차를 설명하고, 이표본 콜모고로프-스미르노브 검정법을 수정한 새로운 검정통계량을 제시하고자 한다.

2.3.1 붓스트랩 카이제곱검정

2.1절에서와 같은 분할표를 작성하여 검정통계량

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{r} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

을 구한다. m개의 $X = (X_1, X_2, \cdots, X_m)$ 표본으로부터 붓스트랩 표본 $X^* = (X_1^*, X_2^*, \cdots, X_m^*)$를 랜덤추출하고, n개의 $Y = (Y_1, Y_2, \cdots, Y_n)$ 표본으로부터 붓스트랩 표본 $Y^* = (Y_1^*, Y_2^*, \cdots, Y_n^*)$를 랜덤추출한다. 붓스트랩 표본 X^*와 Y^*를 가지고 2.1절에서와 같은 분할표를 작성하여

$$\chi^{2*} = \sum_{i=1}^{2} \sum_{j=1}^{r} \frac{(O_{ij}^* - E_{ij}^*)^2}{E_{ij}^*}$$

를 구한다. 여기서 O_{ij}^*와 E_{ij}^*는 2.1절의 O_{ij}와 E_{ij}와 같은 방법으로 X^*와 Y^*로부터 구한다. 이와같은 과정을 충분히 많은 B회를 시행하여 χ^{2*} 값들을 χ^2에 대한 표본분포의 추정분포로 사용한다. 따라서 χ^2의 값이 χ^{2*} 값들의 $100(1-\alpha)$ 백분위수 보다 크면, 유의수준 α로 H_0를 기각한다.

2.3.2 붓스트랩 이표본 콜모고로프-스미르노브 검정

주어진 표본 $X = (X_1, X_2, \cdots, X_m)$와 $Y = (Y_1, Y_2, \cdots, Y_n)$로부터 2.2절과

같은 과정으로 검정통계량

$D_{m,n} = \max_{i=1,\cdots,N} |F_m(Z_{(i)}) - G_n(Z_{(i)})|$

를 구한다. m**개의** $X = (X_1, X_2, \cdots, X_m)$ **표본으로부터 붓스트랩 표본** $X^* = (X^*_1, X^*_2, \cdots, X^*_m)$**를 랜덤추출하고,** n**개의** $Y = (Y_1, Y_2, \cdots, Y_n)$ **표본으로부터 붓스트랩 표본** $Y^* = (Y^*_1, Y^*_2, \cdots, Y^*_n)$**를 랜덤추출한 후,** X^***와** Y^***에 대한 경험적분포함수를** $F^*_m(x)$ **와** $G^*_n(y)$ **로 나타내면 다음과 같다.**

$$F_m^*(x) = \begin{cases} 0, & x < X^*_{(1)} \text{ 일때}, \\ \dfrac{i}{m}, & X^*_{(i)} \le x < X^*_{(i+1)} \text{ 일때}, \\ 1, & x \ge X^*_{(m)} \text{ 일때}. \end{cases}$$

여기서 $X^*_{(i)}$**는 확률표본** $X^*_1, X^*_2, \cdots, X^*_m$ **의** i**번째 순위통계량이다.**

$$G^*_n(y) = \begin{cases} 0, & y < Y^*_{(1)} \text{ 일때}, \\ \dfrac{i}{n}, & Y^*_{(i)} \le y < Y^*_{(i+1)} \text{ 일때}, \\ 1, & y \ge Y^*_{(n)} \text{ 일때}. \end{cases}$$

여기서 $Y^*_{(i)}$**는 확률표본** $Y^*_1, Y^*_2, \cdots, Y^*_n$ **의** i**번째 순위통계량이다.**

X^***와** Y^***에 대한 혼합표본의 순위를 다음과 같이 둔다.**

$Z^*_{(1)} \le Z^*_{(2)} \le \cdots \le Z^*_{(N)}, \quad N = m + n$

$D^* = \max_{i=1,\cdots,N} | F^*_m(Z^*_{(i)}) - G^*_n(Z^*_{(i)})|$

의 값을 구한다. 이와같은 과정을 충분히 많은 B**회 시행하고,** D^***의 값들을** $D_{m,n}$**에 대한 표본분포의 추정분포로 사용한다. 따라서 검정절차는** $D_{m,n}$**의 값이** D^* **값들의** $100\frac{\alpha}{2}$ **백분위수보다 작거나,** $100(1-\frac{\alpha}{2})$ **백분위수보다 크면 유의수준** α**로** H_0**를 기각한다.**

2.3.3 수정된 붓스트랩 이표본 콜모고로프-스미르노브 검정

이표본 콜모고로프-스미르노브 검정에서 콜모고로프-스미르노브 검정통계량

$D_{m,n} = \max_{i=1,\cdots,N} |F_m(Z_{(i)}) - G_n(Z_{(i)})|$

대신에

$$MD_{m,n} = \frac{1}{N}\sum_{i=1}^{N}(F_m(Z_{(i)}) - G_n(Z_{(i)}))^2$$

을 생각하고, 이 통계량을 수정된 이표본 콜모고로프-스미르노브 검정통계량이라 부른다. $D_{m,n}$은 두 경험적분포함수의 차이들 중에서 가장 큰 값을 검정통계량으로 하며, $MD_{m,n}$은 두 경험적분포함수의 차이를 제곱한 것의 평균을 검정통계량으로 한다. $MD_{m,n}$의 분포를 구하기 어려움으로, 붓스트랩 방법으로 $MD_{m,n}$의 표본분포를 추정하고자 한다. 즉, 붓스트랩 표본 X^*와 Y^*로부터

$$MD^*_{m,n} = \frac{1}{N}\sum_{i=1}^{N}(F^*_m(Z^*_{(i)}) - G^*_n(Z^*_{(i)}))^2$$

를 구한다. 이와같은 과정을 충분히 많은 B회 시행하고, $MD^*_{m,n}$의 값들을 $MD_{m,n}$에 대한 표본분포의 추정분포로 사용한다. 따라서 검정절차는 $MD_{m,n}$의 값이 $MD^*_{m,n}$ 값들의 $100(1-\alpha)$ 백분위수보다 크면 유의수준 α로 H_0를 기각한다.

3. 모의실험

2장에서 살펴본 이표본 문제에 관한 통계량에 대해서, 정규분포와 지수분포의 경우에 모의실험을 시행하였다. 각 검정에서 유의수준은 α=0.05이고, 모의실험은 1000회 반복 시행하였으며, 붓스트랩 반복 횟수 B 또한 1000회로 하였다. 카이제곱검정과 붓스트랩 카이제곱검정의 경우 계급의 수는 5로 하였다. 각 검정에 대한 검정력은 그림 1 - 그림 6과 같이 나타난다.

μ μ

($m = n = 10$일 때) ($m = n = 50$일 때)

그림 1 $N(0,1)$과 $N(\mu,1)$을 따르는 모집단에 대한 각 검정의 검정력 함수

σ σ

($m = n = 10$**일 때**) ($m = n = 50$**일 때**)

그림 2 $N(0, 1)$**과** $N(\mu, \sigma^2)$**을 따르는 모집단에 대한 각 검정의 검정력 함수**

k k

($m = n = 10$**일 때**) ($m = n = 50$**일 때**)

그림 3 $E(1)$**과** $E(k)$**를 따르는 모집단에 대한 각 검정의 검정력 함수**

μ μ

($m = n = 10$**일 때**) ($m = n = 50$**일 때**)

그림 4 $N(0, 1)$**과** $N(\mu, 1)$**을 따르는 모집단에 대한 각 검정의 검정력 함수**

σ σ

($m = n = 10$일 때) ($m = n = 50$일 때)

그림 5 $N(0,1)$과 $N(\mu, \sigma^2)$을 따르는 모집단에 대한 각 검정의 검정력 함수

k k

($m = n = 10$일 때) ($m = n = 50$일 때)

그림 6 $E(1)$과 $E(k)$를 따르는 모집단에 대한 각 검정의 검정력 함수

그림 1 - 그림 3으로부터 정규분포에 대한 평균과 분산의 변화에 있어서 표본의 개수가 큰 경우 검정력이 큰 차이가 없으며, 지수분포의 경우 또한 표본의 개수가 커지면서 검정력의 차이가 줄어드는 것을 알 수 있다. 그림 4 - 그림 6에서 보듯이 수정된 붓스트랩 이표본 콜모고로프-스미르노브 검정이 붓스트랩 이표본 콜모고로프-스미르노브 검정보다 검정력이 약간 좋은 결과를 나타내고 있다.

결론적으로 붓스트랩 검정법은 소표본일 경우는 분포가 알려진 검정법보다 검정력이 약간 떨어진다고 할 수 있지만, 표본의 수가 증가할수록 기존의 방법과 검정력에 있어서 별 차이를 가지지 않음을 알 수 있다. 그리고 제안된 수정된 이표본 콜모고로

프-스미르노브 검정이 이표본 콜모고로프-스미르노브 검정보다 조금 좋은 검정력을 가진다. 따라서 붓스트랩 검정법은 이표본 문제에 대해서 좋은 검정력을 가짐을 알 수 있으며, 모의실험에서 보듯이 대표본인 경우에 붓스트랩 방법의 일치성을 확인할 수 있다.

참고문헌

1. Baumgarther, W., Weib, P., and Schindler, H. (1998). A Nonparametric Test for the General Two-sample Problem, *Biometrics*, 54, 1129-1135.
2. Bickel, P. J., and Freedman, D. A. (1981). Some Asymptotic Theory for the Boostrap, *The annals of statistic*, 8, 1196-1217.
3. Efron, B. (1979). Bootstrap methods; Another look at the jackknife, *Ann. Statist*, 7, 1-26.
4. Gibbons, J. D., (1971). *Nonparamertic statistical Inference*, McGraw-Hill.
5. Davision, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*, Cambridge Series in Statistical and Probabilistic Mathematics.
6. Hollander, M., and Douglas, A. W. (1999). *Nonparamertic statistical Methods*, Wiley Series in Probability and Statistics.
7. Pratt, J. W., and Gibbons, J. D. (1981). *Concepts of Nonparamertic Theory*. Springer Series in Statistics.