한국통계학회논문집
2012, 19권, 3호, 405–414

DOI: http://dx.doi.org/10.5351/CKSS.2012.19.3.405

층화모집단 평균에 대한 붓스트랩 추론

허태영[a], 이두리[a], 조중재[1,a]

[a]충북대학교 정보통계학과

요 약

층화확률추출은 모집단을 어떤 층화기준에 의해 여러 층으로 분할한 다음 각 층으로부터 독립적으로 표본을 임의추출하는 방법으로 여러 가지 장점을 가지고 있어 실제 조사에서 많이 활용되고 있다. 본 연구에서는 대규모 표본조사에서 많이 사용하고 있는 층화확률추출을 사용하여 추출된 표본을 통해 모평균에 대한 붓스트랩 추정량과 신뢰구간 및 가설검정 등 통계적 추론에 대하여 연구하였다. 층화모집단에서의 모평균의 추정량과 관련된 극한 분포이론들을 기초로 붓스트랩 일치성을 근거로 층화 모평균에 대해 표준 붓스트랩 방법, 백분위수 붓스트랩 방법, 스튜던트화 붓스트랩 방법을 활용한 신뢰구간과 붓스트랩 가설검정 방법을 제안하였으며, 모의실험을 통해 신뢰구간 추정 방법들의 유효성을 확인하였다.

주요용어: 층화확률추출, 붓스트랩 일치성, 붓스트랩 신뢰구간, 붓스트랩 가설검정, ASL_{boot}.

1. 서론

대부분의 사회조사에서 표본조사를 많이 활용하고 있으며 표본조사는 모집단 전체의 자료를 조사하는 것에 비해 여러 가지 장점을 가지고 있다. 이러한 표본조사를 수행하기 위해서는 모집단을 이용하여 먼저 표본추출이 선행되어야 한다.

다양한 표본추출방법 중 층화확률추출은 동일한 수준의 표본오차 내에서 단순임의추출에 비해서 표본수가 작아질 수 있다는 장점을 가지고 있다. 특히 표본수가 적어진다는 것은 조사비용과 함께 시간이 절약되어 조사 관리를 편하게 할 수 있어 비표본오차를 효율적으로 통제할 수 있으며, 각 층별 추정값을 쉽게 얻을 수 있다는 장점을 가지고 있다.

그러나 표본조사에서 층화확률추출의 경우 각 층별로 표본을 배분하다 보면 특정 층에서 표본수가 너무 작아 모집단의 모수 추정의 신뢰성에 문제가 되는 경우가 발생할 수 있다.

따라서 본 연구에서는 통계적 추론에서 모집단의 분포에 관계없이 보다 효율적으로 소표본에서도 비교적 정확하게 관심모수를 추정할 수 있는 장점을 가진 붓스트랩(bootstrap) 기법을 사용하여 층화확률추출에 대한 각 층별 모평균 및 전체 모평균의 추정량에 대한 붓스트랩 추정량 및 그에 대한 극한분포를 제시하고 대표적인 세 가지 붓스트랩 구간 추정 방법을 제안하고자 한다 (Efron, 1979, 1985, 1987; Efron과 Tibshirani, 1986; Hall, 1986, 1988; Martin, 1990; Pons, 2007).

본 논문의 구성은 다음과 같다. 먼저 2장에서는 층화확률추출 모형 및 추정량의 극한분포를 설명하였고 3장에서는 층화임의추출에서의 붓스트랩 알고리즘 소개와 붓스트랩 추정량의 극한분포에 대해서 논하였다. 또한 층화 모평균에 대하여 표준 붓스트랩 신뢰구간과 백분위수 붓스트랩 신뢰구간 그리고 스튜던트화 붓스트랩 신뢰구간 등 붓스트랩 신뢰구간 추정방법을 연구하였다. 나아가 전체 모평

이 논문은 2010년도 충북대학교 학술연구 지원사업의 연구비 지원에 의하여 연구되었음.
[1] 교신저자: (361-763) 충북 청주시 흥덕구 내수동로 52번지, 충북대학교 정보통계학과, 교수.
E-mail: jjcho@chungbuk.ac.kr

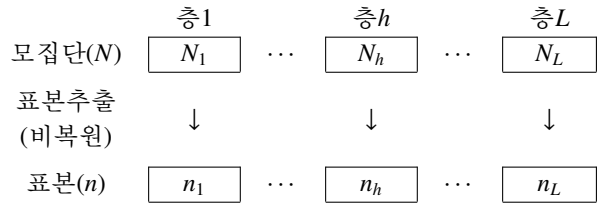

그림 1: *층화 추출 모형*

균에 대한 붓스트랩 가설검정 방법을 제안하였다. 4장에서는 정규분포와 카이제곱분포 가정 하에서 모의실험을 실시하여 3장에서 제안한 추정방법들에 대한 유효성을 연구하였다. 마지막으로 5장은 논문의 연구결과와 향후 연구과제에 대하여 언급하였다.

2. 층화확률추출 모형 및 추정량의 극한분포

2.1. 층화확률추출 모형 및 모수 추정

층화확률추출은 모집단을 어떤 적당한 층화기준에 의해 여러 층으로 나눈 다음 각 층으로부터 독립적으로 표본을 임의추출하는 방법으로, L개의 층으로 구성된 모집단에 대한 층화확률추출 모형을 그림으로 표현하면 그림 1과 같다. 여기서 N_h는 h번째 층의 크기이고, $\sum_{h=1}^{L} N_h = N$이다. 그리고 n_h는 h번째 층에서의 임의 추출한 표본의 크기이고, $\sum_{h=1}^{L} n_h = n$이다.

그림 1과 같은 층화확률추출 과정을 통해 모집단의 모수들을 추정하고자 할 경우, h층 내에서의 평균 μ_h와 분산 σ_h^2의 추정량들인 표본평균 $\bar{X}_h$와 표본분산 s_h^2, $(h = 1, 2, 3, \ldots, L)$은 식 (2.1)과 같다.

$$\hat{\mu}_h = \bar{X}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} X_{hj}, \qquad \hat{\sigma}_h^2 = s_h^2 = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(X_{hj} - \bar{X}_h\right)^2. \tag{2.1}$$

또한 각 층으로부터 비복원추출의 경우 전체 모평균 μ의 추정량 $\hat{\mu} = \bar{X}_{st}$와 이 추정량에 대한 분산 $\mathrm{Var}(\bar{X}_{st})$은 식 (2.2)와 같이 나타낸다.

$$\hat{\mu} = \bar{X}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{X}_h, \tag{2.2}$$

$$\mathrm{Var}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \cdot \mathrm{Var}\left(\bar{X}_h\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{S_h^2}{n_h}.$$

단, h번째 층에 대한 모분산 S_h^2은 $S_h^2 = 1/(N_h - 1)\sum_{h=1}^{N_h}(x_{hj} - \bar{x}_{h\cdot})^2$, $(h = 1, 2, \ldots, L)$로 정의되며 자료 $\{x_{hj};\ j = 1, 2, \ldots, N_h\}$는 모집단 내에서 정의된 자료이다.

따라서 분산 $\mathrm{Var}(\bar{X}_{st})$의 추정량은 식 (2.3)과 같이 추정된다.

$$\widehat{\mathrm{Var}}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^2}{n_h}. \tag{2.3}$$

또한 전체 모평균 μ에 대한 $(1 - 2\alpha)100\%$ 근사적인 신뢰구간은 적당한 조건하에서 식 (2.4)와 같이 표현된다.

$$\bar{X}_{st} \pm z_{(1-\alpha)}\sqrt{\widehat{\mathrm{Var}}\left(\bar{X}_{st}\right)} \tag{2.4}$$

여기서, $z_{(1-\alpha)}$는 표준정규분포 $N(0, 1)$에서의 하위 $100(1 - \alpha)$백분위수를 뜻한다.

2.2. 층화확률 추정량의 극한분포

층화확률추출에 의한 전체 모평균 μ에 대한 층화확률 추정량 $\hat{\mu}_{st} = \bar{X}_{st}$와 관련된 극한분포와 관련된 보조정리는 다음과 같다.

보조정리 1. 층화 모집단에서 층 모평균 μ_h와 유한분산 σ_h^2을 가지는 h번째 부모집단에서의 표본 $\chi_{hn_h} = (X_{h1}, X_{h2}, \ldots, X_{hn_h})$에 대해 비복원추출(sampling without replacement)을 사용할 경우 다음의 결과가 성립한다.

$$\sqrt{n_h}\left(\bar{X}_h - \mu_h\right) \xrightarrow{d} N\left(0, \sigma_h^2\right), \quad (h = 1, 2, \ldots, L) \text{ as } n_h \to \infty.$$

증명: 자세한 증명은 Bickel과 Freedman (1984)을 참조하기 바랍니다. □

식 (2.2)로부터 층화 추정량 $\hat{\mu}_{st} = \bar{X}_{st}$에 대한 분산 σ_{st}^2는 다음과 같으므로

$$\sigma_{st}^2 = \mathrm{Var}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{S_h^2}{n_h}$$

분산 σ_{st}^2에 대한 바람직한 플러그-인 추정량 $\hat{\sigma}_{st}^2$는 식 (2.5)와 같다.

$$\hat{\sigma}_{st}^2 = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2 \cdot \widehat{\mathrm{Var}}\left(\bar{X}_h\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s_h^2}{n_h}, \tag{2.5}$$

여기에서, h번째 층에 대한 표본분산 s_h^2은 $s_h^2 = 1/(n_h - 1)\sum_{j=1}^{n_h}(X_{hj} - \bar{X}_h)^2$, $(h = 1, 2, \ldots, L)$로 정의한다.

한편, 극한분포이론에서 필요한 정칙조건(regularity condition)을 명시하기 위하여 h번째 층에 대한 분산 가중치 v_h^2과 효과적인 표본크기 ρ_h를 각각 식 (2.6)에서와 같이 정의하도록 하자.

$$v_h^2 = \left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{S_h^2}{n_h}\frac{1}{\sigma_{st}^2} = \mathrm{Var}\left(\frac{N_h}{N}\frac{1}{\sigma_{st}}\bar{X}_h\right); \qquad \rho_h = \frac{n_h N_h}{(N_h - n_h)} \tag{2.6}$$

그리고 h번째 층에 대한 모분산 σ_h^2을 $\sigma_h^2 = (1/N_h)\sum_{j=1}^{N_h}(x_{hj} - \mu_h)^2$로 놓고, 새로운 확률변수 y_{hj}를 식 (2.7)과 같이 정의한다.

$$y_{hj} = \frac{(x_{hj} - \mu_h)}{\sigma_h}, \quad h = 1, 2, \ldots, L;\ j = 1, 2, \ldots, N_h. \tag{2.7}$$

극한분포이론에 필요한 정칙조건을 소개하면 다음과 같다.

[정칙조건] 임의의 양수 ϵ에 대하여 모든 층에서의 표본크기 $n_h\,(h = 1, 2, \ldots, L)$가 무한히 크게 될 때$(\min(n_h) \to \infty)$ 다음의 결과가 성립한다는 조건을 의미한다.

$$\sum_{h=1}^{L}\frac{1}{N_h}\sum_{j=1}^{N_h}\phi^2\left(v_h y_{hj}, \epsilon\sqrt{\rho_h}\right) \to 0,$$

여기서, 함수 $\phi(x, \epsilon)$는 식 (2.8)과 같다.

$$\phi(x, \epsilon) = \begin{cases} x, & |x| \geq \epsilon, \\ 0, & \text{그외의 영역.} \end{cases} \tag{2.8}$$

정리 1. 앞에서의 정칙조건이 충족된다면, 모든 층에서의 표본크기 $n_h\,(h = 1, 2, \ldots, L)$가 무한히 크게 될 때$(\min(n_h) \to \infty)$ 다음의 극한분포 결과가 성립한다.

(1) 확률변수 $\dfrac{(\bar{X}_{st} - \mu)}{\sigma_{st}} \xrightarrow{d} N(0, 1)$,

(2) 분산비 $\dfrac{\hat{\sigma}^2_{st}}{\sigma^2_{st}} \xrightarrow{p} 1$.

증명: 자세한 증명은 Bickel와 Freedman (1984)를 참조하면 됨. □

따라서 슬럿츠키 정리(Slutsky's Theorem)에 의해 식 (2.9)의 극한분포 결과도 성립함을 알 수 있다.

$$\frac{\left(\bar{X}_{st} - \mu\right)}{\hat{\sigma}_{st}} \xrightarrow{d} N(0, 1). \tag{2.9}$$

3. 붓스트랩을 활용한 통계적 추론

Efron (1979)을 시작으로 수많은 학자들에 의해 다양한 통계학 분야에 전반적으로 널리 연구된 붓스트랩 방법은 컴퓨터 계산능력 발전과 더불어 모집단에 대한 최소한의 가정 하에서 통계량의 표본분포와 통계적 추정이나 가설검정 문제에 효과적으로 활용될 수 있다. 우선 본 논문에서의 층화모집단 평균에 대한 붓스트랩 추론과 관련한 붓스트랩 알고리즘을 소개하면 다음과 같다.

3.1. 층화확률추출에서의 붓스트랩 알고리즘

층화모집단 평균에 대한 붓스트랩 추론과 관련한 붓스트랩 알고리즘을 단계별로 나누어 표현하면 다음과 같이 나타낼 수 있다.

(단계1) h번째 층에서의 확률표본 $\chi_{hn_h} = (X_{h1}, X_{h2}, \ldots, X_{hn_h})$으로부터 복원추출로 같은 크기 n_h의 독립적인 붓스트랩 표본 $\chi^*_{hn_h} = (X^*_{h1}, X^*_{h2}, \ldots, X^*_{hn_h})$을 얻는다 $(h = 1, 2, \ldots, L)$.

(단계2) (단계1)의 붓스트랩 표본으로부터 평균 $\bar{X}^*_h$와 분산 s^{*2}_h, $h = 1, 2, \ldots, L$을 다음과 같이 계산한다.

$$\bar{X}^*_h = \frac{1}{n_h}\sum_{j=1}^{n_h} X^*_{hj}, \qquad s^{*2}_h = \frac{1}{n_h - 1}\sum_{j=1}^{n_h}\left(X^*_{hj} - \bar{X}^*_h\right)^2.$$

(단계3) (단계 2)의 평균 $\bar{X}^*_h$와 분산 s^{*2}_h, $(h = 1, 2, \ldots, L)$로부터 전체 모평균 μ에 대한 붓스트랩 평균 $\bar{X}^*_{st}$와 분산 $\hat{\sigma}^{*2}_{st}$를 다음과 같이 계산한다.

$$\hat{\mu}^*_{st} = \bar{X}^*_{st} = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)\bar{X}^*_h, \qquad \hat{\sigma}^{*2}_{st} = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{s^{*2}_h}{n_h}.$$

이와 같은 과정을 적당한 크기 B회 반복하여 층별 붓스트랩 추정량들인 $\bar{X}^*_h(i)$, $(h = 1, 2, \ldots, L)$와 모평균 μ의 붓스트랩 추정량 $\bar{X}^*_{st}(i)$, $(i = 1, 2, \ldots, B)$를 구하여 붓스트랩 추론에 사용한다.

3.2. 붓스트랩 추정량의 극한분포

붓스트랩 추론 문제를 해결하기 위하여 붓스트랩 일치성이 충족되어야 하는 바, 앞에서 소개한 붓스트랩 알고리즘 하에서 다음과 같은 붓스트랩 극한분포결과를 소개한다 (Bickel과 Freedman, 1984).

정리 2. 붓스트랩 추정량 $\bar{X}^*_{st}$의 원래의 표본 $\chi_n = (\chi_{1n_1}, \chi_{2n_2}, \ldots, \chi_{Ln_L})$에 대한 조건부적인 분산을 $\tilde{\sigma}^2_{st}$라고 정의하자. 앞에서의 정칙조건을 포함한 적당한 조건하에서 다음의 극한분포 결과가 성립한다. 모든 층에서의 표본크기 n_h, $(h = 1, 2, \ldots, L)$가 무한히 크게 될 때, 즉, $\min(n_h) \to \infty$ 조건하에서

(i) 확률변수 $\dfrac{\left(\bar{X}^*_{st} - \bar{X}_{st}\right)}{\tilde{\sigma}_{st}} | \chi_n \xrightarrow{d} N(0,1)$

(ii) $\dfrac{\hat{\sigma}^{*2}_{st}}{\tilde{\sigma}^2_{st}} | \chi_n \xrightarrow{p} 1$

여기서, $\hat{\sigma}^*_{st} = \sum_{h=1}^{L} (N_h/N)^2 \{(N_h - n_h)/N_h\}(s^{*2}_h/n_h)$, $s^{*2}_h = 1/(n_h - 1)\sum_{j=1}^{n_h}(x^*_{hj} - \bar{x}^*_h)^2$, $(h = 1, 2, \ldots, L)$이다.

따라서 식 (2.9)와 마찬가지로 슬러츠키 정리에 의해 식 (3.1)의 붓스트랩 추정량의 극한분포 결과도 성립한다.

$$\frac{\left(\bar{X}^*_{st} - \bar{X}_{st}\right)}{\hat{\sigma}^*_{st}} | \chi_n \xrightarrow{d} N(0,1). \tag{3.1}$$

3.3. 붓스트랩 신뢰구간 추정방법

우선 2.2절과 3.2절에서 소개한 극한분포의 결과들을 기초로 한 붓스트랩의 일치성으로 여러 가지 붓스트랩 신뢰구간을 추정할 수 있다. 본 논문에서는 붓스트랩 점 추정량 뿐만 아니라 구간추정을 위해 표준 붓스트랩 방법, 백분위수 붓스트랩 방법, 스튜던트화 붓스트랩 방법 등 세 가지 방법을 제안하여 층화확률추출의 각 층 및 전체 모평균에 대해 추정에 대해 연구하였다.

3.3.1. 표준 붓스트랩 방법(SB; standard bootstrap method)

층화모집단에 대한 관심 모수인 모평균을 μ라고 할 때, 붓스트랩 알고리즘에 의한 B개의 붓스트랩 추정값 $\bar{X}^*_{st}(i)$, $(i = 1, 2, \ldots, B)$들로부터 다음 형태의 붓스트랩 표본평균과 표준분산을 구한다.

$$\bar{X}^*_{st}(\cdot) = \frac{1}{B}\sum_{i=1}^{B} \bar{X}^*_{st}(i), \qquad \widehat{\mathrm{SE}}^2 = \frac{1}{B-1}\sum_{i=1}^{B}\left[\bar{X}^*_{st}(i) - \bar{X}^*_{st}(\cdot)\right]^2$$

따라서 모평균 μ에 대한 $(1 - 2\alpha)100\%$ 표준 붓스트랩 신뢰구간을 식 (3.2)와 같이 설정한다.

$$\left[\bar{X}_{st} - z_{(1-\alpha)} \cdot \widehat{\mathrm{SE}},\ \bar{X}_{st} + z_{(1-\alpha)} \cdot \widehat{\mathrm{SE}}\right]. \tag{3.2}$$

3.3.2. 백분위수 붓스트랩 방법(PB; percentile bootstrap method)

붓스트랩 알고리즘에 의해 크기 B개의 붓스트랩 추정량 $\bar{X}^*_{st}(i)$들을 크기순 $\{\bar{X}^*_{st}(1) \le \bar{X}^*_{st}(2) \le \cdots \le \bar{X}^*_{st}(B)\}$로 가정하였을 경우, 모평균 μ에 대한 $(1 - 2\alpha)100\%$ 백분위수 붓스트랩 신뢰구간을 식 (3.3)과 같이 설정한다.

$$\left[\bar{X}^*_{st}(\alpha B),\ \bar{X}^*_{st}((1-\alpha)B)\right]. \tag{3.3}$$

3.3.3. 스튜던트화 붓스트랩 방법(STUD; studentized bootstrap method)

$(\bar{X}^*_{st} - \bar{X}_{st})/\hat{\sigma}^*_{st}$의 경험적 분포로부터 백분위 지점을 찾아 신뢰구간을 설정하는 것으로 이론적인 근거는 앞에서 논의한 정리 1과 정리 2의 붓스트랩 방법의 일치성이다. 여기에서 모평균 μ에 대한 $(1-2\alpha)100\%$ 스튜던트화 붓스트랩 신뢰구간은 식 (3.4)와 같다.

$$\left[\bar{X}_{st} - \hat{y}_{(1-\alpha)}\hat{\sigma}_{st},\ \bar{X}_{st} - \hat{y}_{\alpha}\hat{\sigma}_{st}\right] \tag{3.4}$$

여기서 $\hat{\sigma}^{*2}_{st} = \sum_{h=1}^{L}(N_h/N)^2\{(N_h - n_h)/N_h\}(s^{*2}_h/n_h)$이며, $\hat{y}_\alpha$는 $\Pr\{(\bar{X}^*_{st} - \bar{X}_{st})/\hat{\sigma}^*_{st} \le \hat{y}_a\} = \alpha$를 만족하는 값이다. 여기서, 분산 $\hat{\sigma}^2_{st}$와 $\hat{\sigma}^{*2}_{st}$는 추정량 $\bar{X}_{st}$의 분산 σ^2_{st}에 대응하는 추정량과 붓스트랩 추정량으로 식 (2.5)와 붓스트랩 알고리즘에 정의되어 있다.

3.4. 층화 모평균 μ에 대한 붓스트랩 가설검정

우선 층화 모평균 μ에 대한 가설은 다음과 같은 형태에 관심이 있을 것이다.

$$\begin{aligned} H_0 &: \mu \le \mu_0 \quad vs. \quad H_1 : \mu > \mu_0 \\ H_0 &: \mu \ge \mu_0 \quad vs. \quad H_1 : \mu < \mu_0 \\ H_0 &: \mu = \mu_0 \quad vs. \quad H_1 : \mu \ne \mu_0 \end{aligned}$$

따라서 다음 형태의 가설 $H_0 : \mu \le \mu_0$ *vs.* $H_1 : \mu > \mu_0$을 중심으로 가설검정 문제를 전개하도록 한다. 우선 본 논문에서 고려하고 있는 층화 모평균 μ에 대한 바람직한 층화확률 추정량과 이 추정량에 대한 분산은 다음과 같다.

$$\begin{aligned} \hat{\mu}_{st} &= \bar{X}_{st} = \sum_{h=1}^{L}\frac{N_h}{N}\bar{X}_h \\ \sigma^2_{st} &= \mathrm{Var}\left(\bar{X}_{st}\right) = \sum_{h=1}^{L}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - n_h}{N_h}\right)\frac{S^2_h}{n_h}. \end{aligned}$$

그런데 이러한 추정량뿐만 아니라 관련 검정통계량에 대한 확률분포는 층별 부모집단 분포가 정규성을 갖지 않는 경우에는 매우 복잡할 것이며 가설검정을 위한 표본 자료를 수집한 경우에 유의확률(p-value)을 계산하는 문제도 간단치가 않을 것이다. 물론 부모집단 분포가 정규성을 갖는 경우라 하더라도 믿을 만한 유의확률을 계산하는 문제는 매우 어려운 문제로 중요한 과제라 생각된다.

한편 본 논문에서는 층화 모평균 μ에 대한 가설검정 문제를 붓스트랩 방법을 적용하여 유의확률을 보다 효율적으로 계산하여 유용하게 해결하고자 한다. 붓스트랩 검정방법은 모집단분포의 정규성 문제와 관계없이 이러한 형태의 가설검정문제에 활용할 수 있는 포괄적인 로버스트한 방법으로 보다 효과적이고 편리하게 응용할 수 있을 것이다.

구체적으로 설명하면 층화 모집단에 대한 붓스트랩 알고리즘과 붓스트랩 일치성으로 표현되는 다음의 두 가지 극한 분포 결과들을 활용하게 된다.

$$\begin{aligned} \frac{\left(\bar{X}_{st} - \mu\right)}{\hat{\sigma}_{st}} &\xrightarrow{d} N(0,1) \\ \frac{\left(\bar{X}^*_{st} - \bar{X}_{st}\right)}{\hat{\sigma}^*_{st}} \,|\, \chi_n &\xrightarrow{d} N(0,1) \end{aligned}$$

표 1: 정규분포 붓스트랩 추정값 및 신뢰구간

표본 크기	층	붓스트랩 추정값	표준 오차	표준	백분위수	스튜던트화
20	층1	10.254	0.173	(9.93, 10.60)	(9.86, 10.58)	(9.85, 10.69)
	층2	15.174	0.208	(14.81, 15.63)	(14.76, 15.56)	(14.85, 15.82)
	평균	12.614	0.136	(12.44, 12.79)	(12.44, 12.79)	(12.59, 12.64)
50	층1	10.004	0.154	(9.72, 10.32)	(9.69, 10.29)	(9.71, 10.34)
	층2	14.923	0.125	(14.68, 15.17)	(14.63, 15.14)	(14.68, 15.22)
	평균	12.512	0.092	(12.40, 12.63)	(12.42, 12.63)	(12.50, 12.53)
100	층1	9.952	0.084	(9.79, 10.12)	(9.80, 10.13)	(9.77, 10.12)
	층2	14.988	0.084	(14.82, 15.15)	(14.80, 15.17)	(14.81, 15.17)
	평균	12.530	0.072	(12.44, 12.62)	(12.41, 12.60)	(12.52, 12.54)
200	층1	10.022	0.069	(9.88, 10.16)	(9.90, 10.17)	(9.87, 10.14)
	층2	15.090	0.079	(14.93, 15.24)	(14.94, 15.26)	(14.92, 15.25)
	평균	12.485	0.054	(12.42, 12.55)	(12.42, 12.54)	(12.48, 12.49)

단, $\chi_n = (\chi_{1n_1}, \chi_{2n_2}, \ldots, \chi_{Ln_L})$이고 $\chi_{hn_h} = (X_{h1}, X_{h2}, \ldots, X_{hn_h})$, $(h = 1, 2, \ldots, L)$들은 붓스트랩 알고리즘에 정의되어 있다.

결론적으로 적당한 붓스트랩 알고리즘 하에서 위에서 설명한 이론적 결과를 기초로 층화 모평균 μ에 대한 가설검정에 필요한 식 (3.5)와 같은 형태의 유의확률을 제안하고자 한다.

$$\mathrm{ASL}_{boot} = \frac{\sum_{b=1}^{B} I(t(X^{*b}) \geq t_{obs})}{B} \tag{3.5}$$

단, $t(X^{*b}) = (\bar{X}_{st}^{*b} - \bar{X}_{st})/\hat{\sigma}_{st}^{*b}$, $(b = 1, 2, \ldots, B)$를 나타낸다. 전체 모평균 μ에 대한 점추정치 $\bar{x}_{st}$에 대해 t_{obs}는 $t_{obs} = (\bar{x}_{st} - \mu_0)/\hat{\sigma}_{st}$와 같다. 그리고 ASL_{boot}(Achieved Significance Level)은 붓스트랩 알고리즘에 의해 컴퓨터 실험으로 계산하여 얻게 될 근사적인 유의확률을 의미한다 (Efron과 Tibshirani, 1993).

유의확률 $\Pr(\bar{X}_{st} \geq \bar{x}_{st} \,|\, \mu = \mu_0)$은 식 (3.6)과 같이 표현되기 때문이다.

$$\begin{aligned}
\Pr\left(\bar{X}_{st} \geq \bar{x}_{st} \,|\, \mu = \mu_0\right) &= \Pr\left(\frac{\left(\bar{X}_{st} - \mu_0\right)}{\hat{\sigma}_{st}} \geq \frac{(\bar{x}_{st} - \mu_0)}{\hat{\sigma}_{st}} \,|\, \mu = \mu_0\right) \\
&\approx \Pr\left(\frac{\left(\bar{X}_{st}^{*} - \bar{X}_{st}\right)}{\hat{\sigma}_{st}^{*}} \geq \frac{(\bar{x}_{st} - \mu_0)}{\hat{\sigma}_{st}} \,|\, \mu = \mu_0\right) \\
&= \Pr\left(t(X^{*}) \geq t_{obs}\right). \qquad (3.6)
\end{aligned}$$

4. 모의실험 및 결과

본 절에서는 모의실험을 통해 층화확률추출에 대해 각 층별 및 전체 모평균에 대한 붓스트랩 추정량 및 표준오차를 계산하였다. 다른 통계모형에 대하여 효율적으로 알려진 표준 붓스트랩 방법과 스튜던트화 붓스트랩 방법 그리고 효율성은 떨어지지만 널리 사용되는 백분위수 붓스트랩 방법을 통해 신뢰구간을 계산하였다. 본 모의실험에서는 층화확률추출에서 두 개의 층을 가지고 있는 모집단을 가정하였으며, 왜도를 가지고 있지 않은 정규분포와 왜도의 영향력을 보기 위해 카이제곱분포를 기저분포로 이용하였다. 정규분포는 분산은 같지만 각 층의 평균이 각각 10과 15인 분포와 자유도가 5와 20인 카이제곱분포를 각각 사용하였으며, 해당 분포에서 1000개의 확률표본을 복원추출로 선택하여 그 중 표본의 크기가 각각 20, 50, 100, 200으로 증가시키며 모의실험을 실시하였다.

표 2: 카이제곱분포 표본의 붓스트랩 추정값 및 신뢰구간

표본 크기	층	붓스트랩 추정값	표준 오차	표준	백분위수	스튜던트화
20	층1	5.362	0.511	(4.288, 6.292)	(4.469, 6.558)	(4.189, 6.521)
	층2	23.088	2.451	(18.34, 27.95)	(19.28, 28.73)	(18.73, 29.84)
	평균	12.132	0.679	(11.26, 13.00)	(11.02, 13.15)	(11.48, 13.13)
50	층1	5.401	0.543	(4.333, 6.464)	(4.157, 6.454)	(4.473, 6.965)
	층2	20.183	1.192	(17.93, 22.61)	(17.60, 22.52)	(18.17, 24.23)
	평균	12.29	0.426	(11.75, 12.84)	(11.76, 12.91)	(12.07, 12.57)
100	층1	4.638	0.305	(4.059, 5.255)	(4.004, 5.294)	(4.078, 5.324)
	층2	19.454	0.633	(18.15, 20.64)	(18.30, 20.73)	(18.16, 20.71)
	평균	12.921	0.320	(12.51, 13.33)	(12.52, 13.31)	(12.80, 13.07)
200	층1	5.154	0.213	(4.778, 5.614)	(4.562, 5.466)	(4.863, 5.799)
	층2	20.024	0.512	(19.00, 21.01)	(18.80, 21.03)	(19.05, 21.03)
	평균	11.928	0.271	(11.58, 12.28)	(11.59, 12.29)	(11.86, 12.01)

표 1과 표 2는 정규분포와 카이제곱분포를 기저분포로 사용하여 모의실험을 통해 두 개의 층으로 구성된 층화확률추출에 대한 붓스트랩 추정량 및 신뢰구간을 나타낸 표이다. 전반적으로 각각의 경우에 신뢰구간이 적절히 계산되었고 표본의 크기가 커질수록 붓스트랩 추정량의 표준오차가 작아짐을 알 수 있으며, 스튜던트화 방법이 대체적으로 신뢰구간의 폭이 야간 작은 것으로 계산되었다. 하지만 효율적으로 알려진 표준 붓스트랩 방법과 스튜던트화 붓스트랩 방법 그리고 효율성은 떨어지지만 널리 사용되는 백분위수 붓스트랩 신뢰구간을 계산하는 방법 모두 좋은 구간 추정 방법이 될 수 있다고 판단된다.

5. 결론

본 연구에서는 층화 모집단에서 모평균에 대한 붓스트랩 추정문제와 가설검정 문제를 연구하였다. 붓스트랩 추론을 위한 붓스트랩의 일치성을 연구, 소개하고 이를 기초로 층화 모평균에 대하여 필요한 붓스트랩 알고리즘과 세가지 붓스트랩 신뢰구간 추정방법 그리고 붓스트랩 가설 검정 방법을 제안, 연구하였다. 또한 두 가지 확률분포 가정하에서 붓스트랩 모의실험을 통해 현실적으로 충분히 활용 가능한 좋은 신뢰구간 추정방법임을 확인하였다.

물론 더욱 정확한 효율적인 붓스트랩 추론과 관련된 의미 있는 결론은 보다 많은 포괄적인 모의실험이 실행되어야 할 것으로 사료된다.

참고 문헌

Bickel, P. J. and Freedman, D. A. (1984). Asymptotic normality and the bootstrap in stratified sampling, *Annals of Statistics*, **12**, 470–482.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife, *Annals of Statistics*, **7**, 1–26.

Efron, B. (1985). Bootstrap confidence intervals for a class of parametric problems, *Biometrika*, **72**, 45–48.

Efron, B. (1987). Better bootstrap confidence intervals, *Journal of the American Statistical Association*, **82**, 171–185.

Efron, B. and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy, *Statistical Science*, **1**, 54–75.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*, Chapman & Hall,

Hall, P. (1986). On the bootstrap and confidence intervals, *Annals of Statistics*, **14**, 1431–1452.

Hall, P. (1988). Theoretical comparison of bootstrap confidence intervals, *Annals of Statistics*, **16**, 927–953.

Martin, M. A. (1990). On bootstrap iteration for coverage correction in confidence intervals, *Journal of the American statistical Association*, **85**, 1105–1118.

Pons, O. (2007). Bootstrap of means under stratified sampling, *Electronic Journal of Statistics*, **1**, 381–391.

2012년 3월 8일 접수; 2012년 3월 28일 수정; 2012년 3월 29일 채택

On Statistical Inference of Stratified Population Mean with Bootstrap

Tae-Young Heo1[a], Doori Lee[a], JoongJae Cho[1,a]

[a]Department of Information Statistics, Chungbuk National University

Abstract

In a stratified sample, the sampling frame is divided into non-overlapping groups or strata (*e.g.* geographical areas, age-groups, and genders). A sample is taken from each stratum, if this sample is a simple random sample it is referred to as stratified random sampling. In this paper, we study the bootstrap inference (including confidence interval) and test for a stratified population mean. We also introduce the bootstrap consistency based on limiting distribution related to the plug-in estimator of the population mean. We suggest three bootstrap confidence intervals such as standard bootstrap method, percentile bootstrap method and studentized bootstrap method. We also suggest a bootstrap test method computing the ASL_{boot}(Achieved Significance Level). The results of estimation are verified using simulation.

Keywords: Stratified random sampling, bootstrap consistency, bootstrap confidence interval, bootstrap test, ASL_{boot}.

This work was supported by the research grant of Chungbuk National University in 2010 .

[1] Corresponding author: Professor, Department of Information and statistics, Chungbuk National University, 52 Naesudongro, Cheongju, Chungbuk 363-761, Korea. E-mail: jjcho@chungbuk.ac.kr