한국통계학회논문집 제3권 1호
1996년 4월 pp. 179-188

붓스트랩방법의 실제적활용[1)]

군집표본추출법에 근거한 분할표분석을 중심으로

전 명 식[2)]

요 약

복합조사표본추출법(complex survey sampling)에 근거한 분할표분석에 카이제곱검정법을 사용할 때의 문제점들과 해결방법들을 살펴보았다. 나아가, 군집표본추출의 경우에 붓스트랩방법의 타당성을 보였으며, 실제자료분석을 통하여 실제 활용가능성과 잇점을 제시하였다.

1. 소개

복합조사로부터 구한 자료에 근거하여 만들어진 분할표에 대한 통계적추론은 보건, 의학, 그리고 사회과학등의 분야에서 많이 이용되고 있다. 이들에 대한 통계적방법중 대표적인 것들로는 Pearson의 카이제곱검정법을 비롯하여 이와 근사적으로 동등한 log-likelihood 검정법, Wald 검정법 그리고 log-linear모형에 의한 접근방법등이 있다. 그러나, 조사자료에 근거한 분할표를 분석함에 있어 그 조사가 어떻게 수행되었는지에 각별한 주의가 필요하다. 왜냐하면, 많은 경우에 조사방법은 군집표본추출, 층화표본추출 또는 이들이 복합된 다단계추출방법을 사용하며 가장 통상적인 통계적방법들의 기본이 되는 다항 또는 승적다항(product multinomial) 표본추출법의 가정을 위반하기 때문이다. 따라서 카이제곱검정법등을 포함한 통상적인 추론방법은 부적절하며 그릇된 결과를 제공하게된다. 이에, 복합조사자료에 카이제곱검정법을 사용할 경우의 문제점들이 지적되었고 나아가 수정방법들이 제안되었다. Rao & Scott(1981)은 design-based 접근방법을 사용하여 계획효과(design-effect)를 활용한 수정방법을 제공하였으며, Brier(1976)와 Cohen(1976)은 군집표본추출의 경우, 그리고 Tavare & Altham(1983) 은 계열적으로 종속된 자료에 관하여 각각 적절한 model-based 접근방법을 활용한 근사이론 및 수정방법들을 제시하였다.

Efron(1979)에 의해 제안된 붓스트랩(bootstrap)방법은 주어진 표본에 근거한 재표본(resampling)을 취하여 연구대상이 되는 통계량의 성질을 파악하는 기법으로 통상적인 접근이 어려운 문제들을 포함한 많은 통계문제에 여러가지 이론적 이점들이 밝혀져왔다. 더우기, 컴퓨터의 발전과 더불어 그 효용성이 나날이 높아지고있는 실정이다.

본 논문에서는 복합조사자료에 대하여 전통적인 Pearson의 카이제곱검정법에 의한 적합도검

1) 이 논문은 교육부(학술진흥재단)의 '94대학교수 국비해외파견연구지원으로 이루어졌음
2) (136-701) 서울시 성북구 안암동 고려대학교 통계학과 교수.

정에 수반되는 문제점들을 살펴보고 가능한 붓스트랩(bootstrap)방법들을 제시하고 그들의 타당성과 가능성을 살펴보고자 한다. 이론적인 면보다는 군집표본추출을 중심으로 실제사용에의 응용을 다루기로 한다. (우도비검정법과 카이제곱검정법에 의한 적합도검정법의 비교는 Fay(1985)를 참조할 것.) 물론 여기서 제안되는 적합도검정법은 다차원분할표에 대한 log-linear 모형에도 활용될 수 있을 것으로 기대된다. 2절에서는 복합조사자료 특히 군집표본추출(cluster sampling)에 근거한 분할표에 대한 적합도검정법의 수정에 필요한 접근방법들에 관해 알아보겠으며, 3절에서는 여러가지 가능한 붓스트랩방법들을 제시하고 그 타당성을 살펴본다. 마지막으로 4절에서는 실제자료의 분석을 통해 기존의 방법들과 붓스트랩방법의 결과를 비교하여보기로 한다.

2. 복합조사자료에 관한 검정법

여기서는 적합도검정에 사용되는 Pearson의 카이제곱통계량과 Wald의 검정통계량을 살펴보자. 독립성에 관한 검정도 아래에 설명될 적합도검정의 경우와 대동소이한 결과를 가지며 3절에서 제안되는 붓스트랩방법을 활용할 수 있다. 이제, 모집단은 흥미의 대상이 되는 변수의 값에 의하여 r개의 범주로 나누어진다고 하고 그에 해당하는 각각의 확률을 $p_1, p_2, \cdots, p_r$ ($\sum_{h=1}^{r} p_h = 1$)로 나타낸다. 또한, 편의상 마지막 범주를 제외하고, $\mathbf{p} = (p_1,\ldots,p_{r-1})'$ 로 표기한다. 이제 다음과 같은 적합도에 관한 가설

$$H : p_1=p_{10}, \cdots, p_r= p_{r0} \qquad \text{vs.} \qquad K : H\text{가 아님}$$

에 대한 검정법을 살펴보자. 복합조사자료에 근거한 모비율의 불편(또는 일치)추정량 $\widehat{p_1}, \widehat{p_2}, \cdots, \widehat{p_r}$을 사용한 통상적인 Pearson의 카이제곱 검정통계량은

$$\mathbf{X}^2 = n \sum_{h=1}^{r} (\widehat{p_h} - p_{h0})^2/p_{h0}$$

$$= n (\widehat{\mathbf{p}} - \mathbf{p_0})' \mathbf{P_0^{-1}} (\widehat{\mathbf{p}} - \mathbf{p_0}) \qquad P_0 = \mathrm{diag}(p_0) - p_0 p_0'$$

으로 표현된다. 단, P/n는 다항표본추출하에서 y/n 의 분산공분산행렬이다. 여기서, $y_1,y_2,\ldots,y_r$은 각 범주의 관찰돗수이며, $n= \sum_{h=1}^{r} y_i$ 으로 표기한다. 그런데, 복합조사자료의 경우 분할표작성에 사용된 개체들이 다항표본추출법에 의하여 구한 것이 아니기 때문에, $\mathbf{X}^2$의 표본분포가 가설 H하에서 근사적으로 $\chi^2(r-1)$ 분포를 따른다고 하는데는 문제가 생긴다. 이에 대하여 Wald의 검정통계량은

$$\mathbf{X}_W^2 = n (\hat{\mathbf{p}} - p_0)' \mathbf{V^{-1}} (\hat{\mathbf{p}} - p_0)$$

(단, V/n는 $\hat{\mathbf{p}}$의 분산공분산행렬 V/n 의 일치추정량)으로 표현되는데, $\mathbf{X}_W^2$의 가설 H하에서의 표본분포를 $\chi^2(r-1)$로 근사할 수 있으나 일반적으로 V/n의 일치추정량을 구하는 것이 문제가 된다. 뒤에 설명될 붓스트랩방법을 이용하여 구한 V의 추정량을 사용하는 것도 고려대상이 될 수 있다.

이제 군집표본추출법에 의해 얻어진 자료를 이용하여 구한 분할표에 대한 Pearson의 카이제곱 통계량의 성질을 살펴보자. 유한모집단(finite population) U는 c 개의 군집들인 U_i, i=1,...,c로 구성되어 있다. 즉, $U = \bigcup_i U_i$, U_i와 U_j는 상호배반이다. 또한, i번째 군집 U_i는 M_i개의 개체로 구성되어있다. 즉, $U_i = (U_{i1},...,U_{iMi})$으로 표기될 수 있다. 따라서 모집단의 개체의 수는 $\sum_{i=1}^{c} M_i = N$ 이 된다. 이 때, i번째 군집의 j번째 개체 U_{ij}는 다변량일 수 있으며, 그 특성에 따라 r개의 범주중의 하나로 분류된다. 나아가, 지시변수를 활용하여,

$$X_{ijh} = \begin{cases} 1 \text{ if } U_{ij}\text{가 h번째 범주로 분류} \\ 0 \text{ 그 외의 경우} \end{cases}$$

$$i=1,...,C,\ j=1,...,M_i,\ h=1,...,r$$

라고 정의하자. 그러면, 모집단의 범주돗수를 $Y_1,...,Y_r$, 그리고 모집단의 범주비율을 $p_1,...,p_r$로 표기할 수 있다.

$$\text{단, } p_h = Y_h/N \quad Y_h = \sum_{i=1}^{c}\sum_{j=1}^{M_i} X_{ijh} \qquad p_h > 0\ , \quad \sum_{h=1}^{r} p_h = 1 \quad \text{이다.}$$

이에 대응되는 표본을 표현해 보자. 여기서 고려하는 표본추출방법은 군집표본추출이다. c개의 군집을 비복원 단순랜덤표본추출하여 $u_i = (u_{i1},...,u_{imi})$ i=1,...,c를 얻었다고 하자.(편의상 당분간 $m_i=M_i=M$(=상수)라고 한다. 따라서 전체 표본의 개체수는 $\sum_{i=1}^{c} m_i = cM = n$ 이 된다.

이제, 표본에 속하는 개체 u_{ij} 에 대하여 지시변수 X_{ijh}를 마찬가지로 정의하면, 표본 범주돗수 $y_1, \ldots, y_r$와 표본 범주비율 $\widehat{p_1}$, $\widehat{p_2}$, ⋯ , $\widehat{p_r}$을 구할 수 있다. 이렇게 구해진 표본에 근거하여, 적합도 검정과 독립성 검정에 관한 문제를 다루어보자. 그런데, 일반적으로 군집내의 개체들은 서로 상관되어있는 경우가 대부분의 많은 표본조사에서의 실제상황이다. 즉, $\rho(U_{ij},U_{ij'})\neq 0$ 이다. 따라서, 통상적으로 분할표분석에 사용되는 카이제곱검정법의 주된 가정중의 하나인 개체들 사이의 독립성이 만족되지 않으며 수정없이 카이제곱검정법을 사용하는 경우 실제의 유의확률보다 작은 유의확률을 제공한다. 이러한 경우에, design-based 접근방법과 model-based 접근방법을 고려

할 수 있을 것이다. Rao & Scott(1981) 은 design-based 접근방법에 대한 일반적인 근사이론을 제공하였다. 그의 내용은 가설 H하에서 통계량 $\mathbf{X}^2$는 자유도가 1인 서로 독립인 카이제곱확률변수들의 가중합인 $\sum_{h=1}^{r-1}\lambda_h Z_h^2$ 로 표현된다. 단, Z_h는 근사적으로 N(0,1)인 확률변수이고 λ_h는 모형효과 행렬(design effects matrix)인 $P_0^{-1}V_0$ 의 고유값을 나타낸다. 따라서, 이러한 결과를 이용하기 위해서는 λ_h에 대한 추정이 필요하게 된다. 만일 λ_h들의 값이나 또는 이들의 일치추정량 λ_h를 구할 수 있으면 $\mathbf{X}^2$의 근사분포에 대한 정확한 근사(exact approximation)를 구할 수 있다. 그러나 이는 V 또는 그의 일치추정량 을 구하는 것과 동일하며(물론 이런 경우에는 Wald의 검정법을 사용할 수 있다.) 이런 상황은 매우 드물다. 따라서 V의 제한된 정보만을 사용하는 방법으로 X^2를 $\lambda_\cdot = \sum_{h=1}^{r-1}\lambda_h/(r-1)$의 추정량 $\hat{\lambda}_\cdot$로 나눈 $X^2/\hat{\lambda}_\cdot$를 검정통계량으로 사용한다. 이 경우 $X^2/\hat{\lambda}_\cdot$는 자유도가 (r-1)인 원래의 카이제곱확률변수와 같은 일차적률을 갖는다는 점에서 타당성을 찾을 수 있다.

한편, model-based 접근방법은 개체 U_{ij}에 의해 정의된 X_{ijh}를 다음과 같은 모형으로부터의 실현값으로 여긴다.

$$E(X_{ijh}) = p_h$$

$$E(X_{ijh}\, X_{i'j'h'}) = p_{hh'} \qquad i=i' \;\; j\neq j' \text{ 인 경우}$$

$$\text{단, } p_{hh'} = \begin{cases} \theta\, p_h + (1-\theta)p_h^2 & h=h' \text{ 인 경우} \\ (1-\theta)\, p_h\, p_{h'} & h\neq h' \text{ 인 경우} \end{cases}$$

이와 같은 모형은 Cohen(1976)에 의해 군집의 크기가 M=2인 경우가 고려되었다. 모형을 정한 댓가로 $\mathrm{var}(p_h)$와 $\mathrm{cov}(p_h,p_{h'})$에 대한 구체적인 결과를 구할 수 있다. 나아가, 이들을 고려한 수정계수가 $(1+\theta)^{-1}$로 구해지며, 앞의 모형하에서, $\mathbf{W}^2 = \mathbf{X}^2/(1+\theta) \sim \chi^2(r-1)$가 근사적으로 성립함이 보여진다. 물론 이와 같은 결과는 $\theta=0$, $\theta=1$인 경우를 고려하면 쉽게 수긍이 간다. 한편 $\rho(U_{ij},U_{ij'})\neq\theta(>0)$이나 그 방향에서는 같은 의미를 가진다. 이러한 결과를 실제적으로 사용하는데는 모형의 적합성과 모수 (p_h,θ)의 추정이 필요하다. 이와 같은 내용은 design-based 접근방법에 의해 design effect를 고려한 것과 근사적으로 일치한다. 이러한 결과는 Altham(1976)에 의해 군집의 크기가 M > 2인 경우로 연장되었다. 또한, Brier(1980)는 Dirichlet-multinomial 분포를 모형으로 사용하여 피어슨의 카이제곱통계량 $\mathbf{X}^2$과 우도비검정통계량이 근사적으로 카이제곱확률변수에 상수를 곱한 것과 같은 분포를 가짐을 밝혔다. 이는 Altham(1976)과 Cohen(1976)의 결과를 연장한 것으로 군집의 크기가 서로 다른 경우에도 활용될 수 있다. Choi & McHugh(1989)도 같은 모형을 활용하여 공분산행렬을 구체적으로 구하여 이를 사용하여 수정된 카이제곱통계량을 제시하였다. 한편, Tavare & Altham(1983)은 계열적으로 종속된 관찰값들에 대한 2x2 분할표에의 경우 통상적인 카이제곱 적합도검정이나 독립성검정이 적절치 못함을 지적하고 Markov chain모형을 사용하여 그에 대한 수정방법을 보였다. 이와 같이 복합조사에 근거한 분할표분석에 통상적인 Pearson의 카이제곱통계

량의 사용에 발생하는 많은 문제점들과 모형을 사용한 해결책들이 제시되어왔다. 그러나, 이러한 모형선택들은 일반적으로 매우 제한되어 있으며 사용된 모형이 틀린 경우에 대한 대비책이 요구된다. 뒤에 다룰 붓스트랩방법은 근본적으로 model-free라는 점에서 분명한 잇점을 가지고 있다고 말할 수 있다.

3. 붓스트랩방법

Efron(1979)에 의해 제안된 붓스트랩(bootstrap)방법은 주어진 표본에 근거하여 재표본(resampling)을 취하여 연구대상이 되는 통계량의 성질을 파악하는데 활용될 수 있다. Fay(1985)는 역시 재표본방법의 일종이라 할 수 있는 jackknife 방법을 사용하여 복합조사자료에 근거한 카이제곱검정법의 보완책을 제시하였다. 여기서는 군집표본추출의 경우에 붓스트랩의 활용을 고려해보자. 기존의 붓스트랩에 대한 방법들은 주로 분산의 추정에 중점을 두었으나 여기서는 검정통계량의 표본분포를 직접 추정하는데 촛점을 맞추기로 한다.

앞서 제안된 기존의 방법들은 검정통계량 자체를 적절히 수정함으로써 수정된 검정통계량이 다항표본추출하에서 카이제곱분포를 따르도록 하는 반면, 붓스트랩은 주어진 검정통계량의 분포를 직접 근사하는데 활용될 수 있다. 또한, 앞에서 언급된 바와 같이 붓스트랩방법은 model-free라는 이점을 가지고 있다. 가능한 몇가지의 붓스트랩방법을 다음과 같이 고려해본다.

3.1 Naive Bootstrap

표본으로 주어진 c개의 군집 $u_i=(u_{i1},...,u_{iM})$, i=1,...,c로부터의 복원재표본추출(resampling with replacement)을 생각할 수 있다. c개의 군집들에 대해서는 iid 가정이 큰 무리없이 받아들여지므로 이들을 M-변량변수로 고려하여 붓스트랩방법을 사용할 수 있다. 또한, u_i값들에 의해 만들어진 적절한 분할표로부터 재표본추출하여도 같은 결과를 구할 수 있다. (즉, $V_{h1...hm}$ 을 첫번째 개체가 h_1에 분류되고,..., m번째 개체가 h_m에 분류되는 군집의 갯수라고 하여 $r\times\cdots\times r$ 분할표를 만들고 이로부터 다항재표본추출을 해도 마찬가지 결과를 얻는다(참조: Cohen(1976)). 이러한 과정을 위한 Monte Carlo 근사법을 단계적으로 기술하면 다음과 같다.

단계 1) 군집 $u_i=(u_{i1},...,u_{iM})$, i=1,...,c에 각각 확률질량 1/c를 부여하여 군집의 경험적분포 F_c를 만든다.

단계 2) F_c로부터 크기가 c인 단순랜덤표본 u_i^* i=1,...,c 을 복원추출하여 이에 근거한 표본비율 p_h^*을 추정하고 $X^{*2} = n \sum_{h=1}^{r} (\widehat{p_h}^* - \widehat{p_h})^2/\widehat{p_h}$ 를 계산한다.

단계 3) 단계 2의 과정을 독립적으로 B회 시행하여 계산된 B개의 X^{*2}값들의 붓스트랩 분포를 카이제곱통계량 X^2의 표본분포를 추정하는데 활용한다.

이는 모형을 설정하지 않은 data-based simulation에 근거한 방법이라 볼 수 있으며, model-free, 로버스트성질, 장애모수 추정의 불필요성등의 여러가지 통계적 이점들을 지니고 있다.

이러한 결과는 적합도검정

$$H : p_1=p_{10},\ldots,p_r=p_{r0} \quad \text{vs.} \quad K : \text{not } H$$

에 다음과 같이 사용될 수 있다. 앞에서 구한 붓스트랩분포의 상위 α-백분위수를 b(α)라고 할 때, $X^2 = n \sum_{h=1}^{r} (\widehat{p_h} - p_{h0})^2/p_{h0} \geq b(\alpha)$ 이면 귀무가설 H를 유의수준 α에서 기각한다.

그러나 이는 매우 단순한 것 같다. 재표본과정이 원래의 표본과정을 잘 닮아야 된다는 기본적인 원칙에서 볼 때, 원래 비복원추출을 하였다면 붓스트랩표본도 역시 비복원으로 구하는 것이 타당할 것이다(참조:Chao & Lo,1994). 그러므로, 주어진 표본 군집들로부터 복원추출하는 것은 비합리적인 것으로 여겨진다. 여기서 생각할 수 있는 방법으로 mirror-match와 enlarged bootstrap을 고려해보기로 한다.

3.2 Mirror-match bootstrap

재표본과정이 원래의 표본과정과 가능한 유사하여야 된다는 점에서 붓스트랩표본을 비복원(without replacement)으로 구하는 방법이 Sitter(1992)에 의해서 다음과 같이 제안되었다.

단계 1) 군집 $u_i=(u_{i1},\ldots,u_{iM})$, i=1,...,c에 각각 확률질량 1/c를 부여하여 군집들의 경험적분포 F_c를 만든다.
단계 2) F_c로부터 크기가 b_1 (< c)인 단순랜덤표본 u_i^* i=1,...,b_1 을 비복원추출한다. 단, b_1은 b_1/c가 원래의 표본추출율이 되도록 선택한다.
단계 3) 단계 2를 독립적으로 b_2회 반복하여 u_i^* ,i=1,...,c를 구하고 이에 근거한 표본비율 $\widehat{p_h}^*$을 추정하고 $X^{*2} = n \sum_{h=1}^{r} (\widehat{p_h}^* - \widehat{p_h})^2/ \widehat{p_h}$ 를 계산한다. (단, b_2는 $b_1b_2=c$를 만족하는 자연수.)

단계 4) 단계 2와 단계 3의 과정을 독립적으로 B회 시행하여 계산된 B개의 X^{*2}값들의 분포를 카이제곱통계량 X^2의 표본분포를 추정하는데 활용한다.

그런데, 위의 재표본과정에서 b_1=1인 경우는 단순붓스트랩과 같으며, b_1이 커질수록 단순붓스트랩에 비하여 붓스트랩분포의 꼬리부분이 짧아지는 그래서 유의확률이 더 작아지는 경향이 있다. 극단적인 경우 b_1=c이면 붓스트랩분포는 퇴화하게 된다. 따라서, 이에 대한 적절한 선택이 요구된다.

3.3 Enlarged bootstrap

주어진 표본과 같은 크기의 붓스트랩표본을 비복원으로 추출하면 결국은 모든 붓스트랩표본이 원래의 표본과 같으므로 아무런 통계적의미가 없다. 이에, 확장된 경험적분포로부터 재표본을 취하는 방법을 고려할 수 있다.(Babu & Singh(1985)) 이를 단계적으로 설명하면 다음과 같다.

단계 1) 표본 군집 $u_i=(u_{i1},\ldots,u_{iM})$, i=1,...,c들의 k개의 복사본을 만든다. 여기서 k=C/c로 놓고 각 군

집에 확률질량 1/C를 부여하여 표본군집의 경험적분포 F_c^E를 만든다.
단계 2) F_c^E로부터 크기가 c인 단순랜덤표본 u_i^* i=1,...,c 을 비복원추출하여 이에 근거한 표본비율 $\widehat{p_h}^*$을 추정하고 $X^{*2} = n \sum_{h=1}^{r} (\widehat{p_h}^* - \widehat{p_h})^2 / \widehat{p_h}$ 를 계산한다.

단계 3) 단계 2의 과정을 독립적으로 B회 시행하여 계산된 B개의 X^{*2}값들의 분포를 카이제곱통계량 $\mathbf{X}^2$의 표본분포를 추정하는데 활용한다.

주) Bickel & Freedman(1984)은 이와 같은 수정된 붓스트랩방법을 층화랜덤표본추출의 경우에 대하여 적용하였다. 한편, model-based 접근방법이 타당한 경우에는 추정된 모형으로부터의 재표본추출하는 것도 설득력이 있다. 가령, 앞의 모형의 경우 θ와 p_h를 통하여 $p_{hh'}$를 추정하면 이로부터 비복원다항재표본추출을 생각할 수 있다.

4. 실제자료분석사례

제안된 붓스트랩방법을 이용한 카이제곱검정통계량의 표본분포추정의 성능을 알아보기 위하여 다음과 같은 실제자료에 적용하고 기존의 model-based 접근방법의 결과와 비교하여 보았다.

사례 1) 다음의 <표 4.1>은 '이웃에 대한 전반적인 만족도'와 '자신의 가정에 대한 만족도'를 조사한 것으로 대상은 5명으로 구성된 20개의 가족이다. 이 자료는 Brier(1980)에 의해 이미 분석된 바가 있으며, 여기서 각 가족은 하나의 군집으로 볼 수 있다. 이 경우에 가족내의 구성원들 사이에 어느 정도의 상관관계가 존재하며 모든 개체들이 독립이라는 가정은 그리 타당해 보이지 않는다. 또한, 2개의 가족은 3명의 구성원에 대해서만 조사결과가 주어져있다. 따라서 통상적인 카이제곱검정법의 사용에는 문제가 따른다.

<표 4.1>

us-us	us-s	us-vs	s-us	s-s	s-vs	vs-us	vs-s	vs-us
1	0	0	2	2	0	0	0	0
1	0	0	2	2	0	0	0	0
0	2	0	0	2	0	0	1	0
0	1	0	2	1	0	1	0	0
0	0	0	0	4	0	0	1	0
1	0	0	3	1	0	0	0	0
3	0	0	0	1	0	0	1	0
1	0	0	1	3	0	0	0	0
3	0	0	0	0	0	1	0	1
0	1	0	0	3	1	0	0	0
1	1	0	0	2	0	1	0	0

0	1	0	4	0	0	0	0	0
0	0	0	4	1	0	0	0	0
0	0	0	1	2	0	0	0	2
2	0	0	2	1	0	0	0	0
1	0	0	1	1	0	0	0	0
0	0	0	1	1	1	0	2	0
0	0	0	1	0	1	0	0	1
2	0	0	2	1	0	0	0	0
2	0	0	2	0	0	1	0	0

(us:불만족, s:만족, vs:매우 만족)로 표기한다.

위의 표로부터 두 변수 '이웃에 대한 전반적인 만족도'와 '자신의 가정에 대한 만족도'에 대한 다음과 같은 분할표를 구할 수 있으며 관심대상인 두변수의 독립성여부에 관한 가설이다.

<표 4.2> 표 4.1에 근거한 분할표

		자신의 가정에 대한 만족도		
		불만족	만족	매우 만족
이웃에 대한 만족도	불만족	18	6	0
	만족	28	28	3
	매우만족	4	5	4

이 때, 독립성여부의 가설에 관한 카이제곱검정통계량의 값은 $\mathbf{X}^2 = 17.87$ 이며 앞에서 설명된 수정 카이제곱통계량의 값은 $\mathbf{W}^2 = 15.38$으로 예견했던대로 수정된 값이 더 작다. 이제, 자유도가 4인 카이제곱분포에 근거하여 구한 유의확률 p는, $0.001 < p = P[\chi^2(4) \geq \mathbf{W}^2 = 15.38] < 0.005$ 이다. 이제, 제안된 붓스트랩방법에 의하여 카이제곱검정통계량 $\mathbf{X}^2$의 표본분포를 직접 근사하고 그에 따르는 유의확률을 구해보자. 본 사례에서는, 몬테칼로 근사에 필요한 반복수를 B=10000으로 하여, 단순붓스트랩과 mirror-match 붓스트랩을 활용해 보았다. mirror-match 붓스트랩에서는 원래의 표본비율을 모르므로 임의로 $b_1=2$, $b_2=10$, $b_1b_2=20$ 사용하였다. 그 결과 단순붓스트랩분포과 mirror-match 붓스트랩분포에 근거한 유의확률은 각각 p=0.0454와 p=0.0312로 구해졌다. 이는 수정 카이제곱검정통계량을 사용한 경우와 큰 차이를 보이지 않으나, 붓스트랩방법은 모형을 설정하지 않았으므로 모형을 잘못 지정할 때 따르는 문제점으로부터 벗어나는 이점이 있다.

사례 2) 다음은 '성'과 '정신분열증'의 관계를 조사한 것으로 대상은 71쌍의 형제들이다. 여기서, 형제사이에는 일종의 상관관계가 존재한다고 볼 수 있으며 모든 개체들이 서로 독립이라는 가정은 타당치 않다. 따라서 통상적인 카이제곱검정법은 수정하는 것이 옳다고 생각된다. 이 자료는 Cohen(1976)에 의해 이미 분석된 바가 있으며, 여기서 각 형제는 하나의 군집으로 볼 수 있다.

<표 4.3>

		동생			
		SM	SF	NM	NF
형	SM	13	5	1	3
	SF	4	6	1	1
	NM	1	1	2	4
	NF	3	8	3	15

<표 4.4> 표 4.3에 근거한 분할표

		성별	
		M	F
정신분열증	S	43	32
	N	15	52

S:정신분열증이 있음, N:정신분열증이 없음, M:남자, F:여자

위의 <표 4.3>으로부터 두 변수 '성'과 '정신분열증'에 대한 <표 4.4>와 같은 분할표를 구할 수 있으며 관심대상인 두변수의 독립성여부에 관한 가설이다.

이 때, 독립성여부의 가설에 관한 카이제곱검정통계량의 값은 $\mathbf{X}^2$ = 17.885 이며 앞에서 설명된 수정 카이제곱통계량의 값은 $\mathbf{W}^2$ = 13.751이다. 이제, 자유도가 1인 카이제곱분포에 근거하여 구한 유의확률 p는, $p = P[\chi^2(4) \geq \mathbf{W}^2 = 13.751] < 0.001$ 이다. 이제, 제안된 붓스트랩방법에 의하여 카이제곱검정통계량 $\mathbf{X}^2$의 표본분포를 직접 근사하고 그에 따르는 유의확률을 구해보자. 본 사례에서는, 몬테칼로 근사에 필요한 반복수 B=10000을 사용하여, 단순붓스트랩과 mirror-match 붓스트랩(b_1=6,b_2=12)을 활용해 보았다. 단순붓스트랩분포과 mirror-match 붓스트랩분포에 근거한 유의확률은 각각 p=0.0026과 p=0.0021로 구해졌다. 앞의 사례에서와 마찬가지로 mirror-match 붓스트랩이 단순 붓스트랩보다 작은 유의확률을 보이고 있으며 이는 앞서 설명한 바와 부합되는 내용이다. 이와 같은 결과는 수정 카이제곱검정통계량을 사용한 경우와 큰 차이를 보이지 않으나, 앞에 언급한 바와 같이 붓스트랩방법은 모형설정이 필요하지 않다는 잇점이 있다. 그런데 앞의 두 사례에서, enlarged bootstrap은 모집단의 군집갯수가 알려지지 않은 관계로 사용하지 않았다.

참고문헌

[1] Altham,P.(1976). Discrete variable analysis for individuals grouped into families, *Biometrics*, Vol. 63, 2, 263-9.

[2] Babu,G.J. and Singh,K.(1985). Edgeworth Expansions for Sampling without Replacement from Finite Populations, *Journal of Multivariate Analysis*, Vol. 17, 261-278.

[3] Brier,S.S.(1980). Analysis of contingency tables under cluster sampling, *Biometrika,* Vol. 67, 3, 591-596.

[4] Chao,M-T and Lo,S-H.(1994). Maximum likelihood summary and the bootstrap method in structured finite population, *Statistica Sinica*, Vol. 4, 389-406.

[5] Choi,J.W. and McHugh,R.B.(1989). A Reduction Factor in Goodness-of-Fit and Independence Tests for Clustered and Weighted Observations, *Biometrics*, Vol. 45, 979-996.

[6] Cochran,W.G.(1977). *Sampling Techniques,* Wiley.

[7] Cohen,J.E.(1976). The distribution of chi-squared statistic under clustered sampling from contingency tables, *Journal of American Statistical Association*, Vol. 71, 665-670.

[8] Fay,R.E.(1985). A Jackknife Chi-Squared Test for Complex Samples. *Journal of American Statistical Association*, Vol. 80, 389, 148-157.

[9] Rao,J.N.K. and Scott,A.J.(1981). The Analysis of Categorical Data From Complex Sample Surveys:Chi-Squared Tests for Goodness of Fit and Independence in Two-Way Tables, *Journal of American Statistical Association*, Vol. 76, 374, 221-230.

[10] Robinson,J.(1978). An Asymptotic Expansions for Samples from a Finite Population, *Annals of Statist.* Vol. 6, 5, 1005-1011.

[11] Sitter,P.R.(1992). A Resampling Procedure for Complex Survey Data, *Journal of American Statistical Association*, Vol. 87, 419, 755-765.

[12] Skinner,C.J. Holt,D. and Smith,T.M.F.(1989). *Analysis of Complex Surveys,* Wiley.

[13] Tavare,S. and Altham,P.(1983). Serial dependence of observations leading to contingency tables, and corrections to chi-squared statistics. *Biometrics*, Vol. 70, 1, 139-44.