불균등확률 계통추출에서 분산추정

홍태경[1], 남궁 평[2]

요 약

불균등 확률 계통추출에서는 모집단 총합에 대한 Horvitz-Thompson 추정량의 대안적 분산 추정량들을 사용하게 된다. 이와 같은 모총합에 관한 분산 추정량들의 설계와 관련한 일반적인 방법은 균등 확률 계통추출에 대한 분산 추정량들에서 시작하고 비율 y_i/p_i에 의한 추정량의 정의에서 y_i를 재배치하게 한다. 비선형 조사 통계학에서 추정량들 중의 하나로 테일러 급수 공식을 적용한다. 불균등 확률 계통추출에서의 분산은 8가지 방법으로 추정이 가능하므로 이를 이용한 분산추정량을 구해보고, 비복원 불균등 확률에서의 jackknife방법을 살펴보고자 한다. 또한 이들 분산추정량들에 대한 비교를 몇가지 방법을 이용하여 알아보도록 한다.

주요용어 : 불균등확률, Horvitz-Thompson 추정량, jackknife방법, Taylor series방법

1. 서론

불균등 확률 계통추출은 불균등 비복원 추출 중 가장 널리 사용되는 방법이다. 이것은 수식을 이용하던지 컴퓨터를 사용하던지 추출 설계를 쉽게 해주며 정확하게 적용을 한다면 $\pi_i = np_i$의 예와 같이 πps 추출 설계를 할 수 있다. 또한 불균등 확률 계통추출은 $n=2$ 처럼 어떤 표본크기의 제한 없이 임의의 표본크기 n을 적용할 수 있다는 장점이 있으며 모집단에서 층이 숨겨져 있거나 어떤 함축된 것을 찾는 작은 설계의 분산에서 아주 효율적일 수 있다.

비복원 불균등 확률추출에서 분산추정의 다른 방법으로는 jackknife 방법이 사용될 수 있다. 이 방법의 장점은 jackknife방법이 사용될 수 있는 모든 통계량들의 분산추정에 동일한 절차가 적용된다는 것이다.

따라서 본 논문에서는 불균등 확률 계통추출에서의 분산추정을 살펴보고 대안적으로 비복원 불균등 확률추출에서의 jackknife방법을 통한 분산추정을 살펴보고자 한다. 또한 각각의 분산추정량들을 상대편의, 상대 평균제곱오차, 신뢰구간 포함율 등을 이용하여 비교하고자 한다.

2. 불균등 확률 계통추출의 분산추정

$\hat{\theta}=g(\hat{Y})$ 형태의 비선형 통계량의 분산추정량은 테일러 시리즈 공식을 함께 이용함으로써 구할 수 있다.

분산의 흥미 있는 추정량은 Hartley와 Rao (1962)에 의해 개선된 π_{ii}'의 근사값을 Yates와 Grundy의 공식에 대체함으로써 구할 수 있다. 이와 같은 근사값은

1) 성균관대학교, 통계학과 강사 E-mail : hongstat@skku.edu
2) 성균관대학교, 통계학과 교수 E-mail : namkung@skku.ac.kr

$$\pi_{ii'} = \frac{n-1}{n}\pi_i\pi_{i'} + \frac{n-1}{n^2}(\pi_i^2\pi_{i'} + \pi_i\pi_{i'}^2) - \frac{n-1}{n^3}\pi_i\pi_{i'}\sum_{j=1}^{N}\pi_j^2$$

이며, 이는 다음 조건들 하에서 차수가 $0(N^{-3})$으로 수정된다.

(1) 모집단 목록을 임의의 차수라고 생각한다.

(2) π_{ii}'는 차수가 $0(N^{-1})$ 이다.

π_{ii}'와 연관된 분산추정량은

$$v_1 = \frac{1}{n-1}\sum_{i}^{n}\sum_{i<i'}^{n}(1-\pi_i-\pi_{i'}+\sum_{j=1}^{N}\frac{\pi_i^2}{n})\cdot(\frac{y_i}{\pi_i}-\frac{y_{i'}}{\pi_{i'}})^2$$

이고, 차수 $0(N)$ 항으로 수정된다. Hartley와 Rao는 차수 $0(N^{-4})$로 수정되는 $\pi_{ii'}$의 보다 좋은 근사를 제시했고 그와 연관된 분산추정량은 차수 $0(1)$으로 수정된다.

확실하게 큰 단위들이 표본으로 선택되었다면 이 공식과 공식에 포함된 모든 기호들 $(N, n, \pi_i, \pi_{i'})$은 모집단의 확실치 않은 부분에만 관련된다. 확실한 경우는 실제와 추정된 분산 간의 분포가 0과 같게 된다.

두 번째 분산추정량은 마치 복원 확률비례크기(pps) 표본처럼 표본을 처리함으로써 구할 수 있다. 추정량은 다음과 같다.

$$v_2 = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\frac{y_i}{p_i}-\hat{Y})^2$$

이 추정량은 계통 pps 설계의 상황에서 편의를 가지지만 모집단이 크고 모집단 목록이 근사적으로 무작위 순이며 어떤 모집단 단위도 불균형적으로 크지 않은 경우 그 편의는 매우 미미하다. 또한 v_2 추정량은 계통 pps 표본추출이 복원 pps 표본추출보다 작은 실제분산을 갖는 경우에 보수적(즉 너무 큰) 경향을 띠게 된다.

세 번째 추정량은 마치 표본이 크기가 n/2로 동일한 각각의 층에서 $n_h = 2$ 단위들을 선택하듯이 표본을 처리함으로써 얻어진다. 분산추정량은 다음과 같다.

$$v_3 = \frac{1}{n}\sum_{i=1}^{n/2}(\frac{y_{2i}}{p_{2i}}-\frac{y_{2i-1}}{p_{2i-1}})^2/n$$

또 다른 추정량은 자유도의 수를 증가시키기 위한 것인데 다음과 같다.

$$v_4 = \frac{1}{n}\sum_{i=2}^{n}(\frac{y_i}{p_i}-\frac{y_{i-1}}{p_{i-1}})^2/2(n-1)$$

이 추정량은 v_3가 중복되지 않은 차분에 유용한 것처럼 중복되는 차분에 유용하다.

다섯 번째 추정량은 랜덤그룹원칙에 적용시켜 구할 수 있다. 계통표본이 k 계통 부표본 속으로 나뉘어 지고, 각각은 $m = n/k$의 정수크기를 갖는다고 하자.

$$\hat{Y}_\alpha = \frac{1}{m}\sum_{i=1}^{m_\alpha}\frac{y_i}{p_i}$$

를 α번째 부표본 ($\alpha = 1,\cdots,k$)의 Horvitz-Thompson 총합추정량이라 하자. 그러면 분산추정량은 다음과 같이 정의된다.

$$v_5 = \frac{1}{k(k-1)}\sum_{\alpha=1}^{k}(\hat{Y}_\alpha - \hat{Y})^2$$

대안적으로 계통표본은 계통적 대신 임의의 부표본들 속으로 분할된다. 이와 같은 v_5 식은 동일한 기댓값을 가지지만 복원 pps 추정량 v_2보다 큰 분산을 갖게된다.

각각의 추정량 $v_2,\cdots,v_5$이 유한모집단 수정계수 (이하 fpc) 항이 곱해지면 v_1 추정량은 아마도 표본추출설계 상황에서 복원인 경우로 간주될 것이다. 계산적으로 쉽고 유용한 계통 pps 표본추출의 fpc는 다음과 같다.

$$\widehat{fpc} = (1 - n^{-1}\sum_{i=1}^{n}\pi_i)$$

물론 계통 pps 표본추출의 실제분산에서 정확한 fpc는 없다. 그러므로 위의 식을 사용함은 계통 pps 표본추출이 복원 pps 표본추출보다 효율적이라는 생각에서 추정분산을 줄이는 데는 가장 좋다는 것을 알 수 있을 것이다.

사실 불균등 확률 계통표본추출의 분산추정량을 구성하는 일반적인 방법은 거의 균등 확률의 경우의 분산추정량을 포함하며 추정량의 정의에 y_i를 $z_i = y_i/p_i$로 대치하고 있다.

끝으로 분산추정량의 분산은 자유도와 역관계를 보이는 경향이 있다. 소 표본의 경우 적절한 자유도의 분산추정량을 선택하는데 매우 신중해야 하며 그리하여 분산추정량의 분산이 너무 커서 쓸모없게 되지 않도록 해야 한다.

$\hat{Y}$에 대한 분산추정량의 다른 방향들 모두는 자료가 초모집단 모형에 의해 형성된다는 가정 하에 만들어진다. 이런 방향 중에서 한 추정량은

$$v_8 = X^2\left\{[\hat{\beta}^2 - Var(\hat{\beta}^2)]\sum_k P(k)[\overline{X_k} - n\sum_k P(k)\overline{X_k}] + (N-1)\hat{\sigma}_e^2/Nn\right\}$$

이며, 여기서

$$X = \sum_{i=1}^{N}X_i,\ \hat{\beta} = \frac{\sum_{i=1}^{n}(r_i - \bar{r})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},\ \widehat{Var}(\hat{\beta}) = \hat{\sigma}_e^2/\sum_{i=1}^{n}(x_i - \bar{x})^2,\ \sigma_e^2 = \frac{1}{n-2}\sum_{i=1}^{n}[(r_i - \hat{r}) - \hat{\beta}(x_i - \bar{x})]^2,$$

$r_i = y_i/x_i$, $\bar{r} = \frac{1}{n}\sum_{i=1}^{n}r_i$, $\overline{X_k}$는 k번째 계통표본의 표본평균이고, $P(k)$는 k번째 계통표본이 선택될 확률이다. 이 추정량은 원래 Hartley(1966)에 의한 것이며 다음과 같은 $r_i = y_i/x_i$ 비율과 x_i 크기의 척도와 사이의 선형회귀모형을 가정하여 얻어졌다.

$$r_i = \alpha + \beta x_i + \epsilon_i$$

만일 모집단 N이 초모집단 ($\epsilon(e_i)=0$을 수용하는)으로 부터의 무작위 표본이라면 v_8은 $Var(\hat{Y})$의 설계분산의 불편추정량이다.
Hartley 방법의 확장은 r_i와 x_i의 비율과 관련된 초모집단 모형을 가정하여 얻어질 수 있다.
한편 v_8은 제곱합 간의 계산을 필요로 하며

$$\sum_k P(k)\left(\overline{X_k} - \sum_{k'} P(k')\overline{X_{k'}}\right)^2$$

또한 추정량 $v_1, \cdots v_5$에 비해 엄청난 계산양이 요구된다. 나머지 추정량들은 아마도 비슷한 양의 계산이 필요할 것이다.

근사적 fpc를 포함시켜 수정된 추정량을 생성할 수 있다. 이와 같은 추정량들은 다음과 같이 정의된다.

$$v_6 = \left(1 - \sum_{i=1}^{n} \pi_i / n\right) v_2$$

$$v_7 = \left(1 - \sum_{i=1}^{n} \pi_i / n\right) v_4$$

v_6는 pps wr 추정량을 수정한 것이고, v_7는 중복차분을 기초로 한 추정량이다.

3. 불균등 확률 추출에서의 jackknife방법

불균등확률의 비복원 추출설계에 대한 jackknife방법의 성질에 관해서는 잘 알려져 있지 않다. n을 어떤 비복원 불균등 확률 표본 설계를 이용한 N으로부터 추출된 표본의 크기라고 가정하고, π_i를 모집단에서 i번째 단위와 관련된 포함확률이라고 가정한다.

$$\pi_i = p\{i \in s\}$$

여기서 s는 표본이다. 모집단 총합에 대한 Horvitz-Thompson 추정량은 다음과 같다.

$$\hat{\theta} = \hat{Y} = \sum_{i=1}^{n} y_i / \pi_i$$

다시 말해, 크기가 m인 k개의 임의 집단들로 나눠진 모표본을 가정하면 $n = mk$가 된다. 이것과 관련하여 Quenouille's 추정량 $\hat{\hat{\theta}}$는

$$\hat{\hat{\theta}} = \sum_{\alpha=1}^{k} \hat{\theta}_\alpha / k$$

이며, 여기서 모조값은

$$\hat{\theta}_{\alpha}=k\hat{Y}-(k-1)\widehat{Y_{(\alpha)}}$$

이고,

$$\widehat{Y_{(\alpha)}}=\sum_{i=1}^{n(k-1)} y_i/[\pi_i m(k-1)/n]$$

은 관찰값들이 α번째 집단으로 이동한 후의 표본에 기초한 Horvitz-Thompson 추정량이다. $\hat{\bar{\theta}}$는 대수적으로 $\hat{Y}$와 동일하다. 따라서 jackknife방법은 총합의 Horvitz-Thompson 추정량이 지닌 불편성을 유지하게 된다.

$\hat{\bar{\theta}}$의 분산을 추정하기 위한 jackknife 추정량은 다음과 같다.

$$v_1(\hat{\bar{\theta}})=\frac{1}{k(k-1)}\sum_{\alpha=1}^{k}(\hat{\theta}_{\alpha}-\hat{\bar{\theta}})^2$$

만약 $i=1,\cdots,n$에 관하여 $\pi_i=np_i$이고 $k=n$이면 다음과 같이 표현할 수 있다.

$$v_1(\hat{\bar{\theta}})=\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i/p_i-\hat{Y})^2$$

보다 일반적으로, 위의 식은 $k<n$일때 다음과 같이 기댓값으로 표현할 수 있다.

$$E\{v(\hat{\bar{\theta}})\}=E\left\{\frac{1}{n(n-1)}\sum_{i=1}^{n}(y_i/p_i-\hat{Y})^2\right\}$$

여기서 기댓값들은 πps 추출 설계에 관한 값이다.

4. 분산 추정량의 비교

불균등 확률 계통추출에서의 분산추정량과 jackknife방법의 분산추정량을 비교하기 위하여 상대편의(relative bias), 상대 평균제곱오차(relative MSE), 그리고 신뢰구간 포함율(confidence interval coverage rates)을 살펴보고자 한다.

4.1. 상대편의

$$\text{Rel Bias}\{v\}=\frac{E\{v\}-Var\{\hat{Y}\}}{Var\{\hat{Y}\}}$$

$$E\{v\}=\sum_{s} v(s)\frac{1}{p}$$

여기서 $p=X/n$로 정의된다.

4.2. 상대평균제곱오차

$$\text{Rel MSE}\{v\} = \frac{E\{(v - Var\{\hat{Y}\})^2\}}{(Var\{\hat{Y}\})^2}$$

여기서

$$E\{(v - Var\{\hat{Y}\})^2\} = \sum_s (v(s) - Var\{\hat{Y}\})^2 \frac{1}{p}$$

4.3. 신뢰구간 포함율

$$c = \frac{100}{p} \sum_s \chi_s$$

여기서

$\chi_s = 1$, 실제 총합 Y가 $Y \in (\hat{Y}(s) \pm z\sqrt{v(s)})$을 만족하는 경우
$= 0$, otherwise

참고문헌

[1] 남궁 평. (1999). 최신 표본조사 설계와 분석, 탐진.

[2] Dippo, C. S. and Wolter, K. M. (1984). A Comparison of Variance Estimators Using the Taylor Series Approximations, in Proceedings of the Survey Research Methods Section, American Statistical Association, 34-42.

[3] Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans, Philadelphia: Society for Industrial and Applied Mathematics.

[4] Kish (1965). Survey Sampling, Wiely.

[5] Risto, Lehtonen. and Erkki, J. P. (1994). Practical Methods for Design and Analysis of Complex Surveys, John Wiely & Sons.

[6] Thompson, S. K. (1992). Sampling, John Wiely & Sons.

[7] Thompson, M. E. (1997). Theory of Sample Surveys, Chapman & Hall.

[8] Wolter, K. M. (1985). Introduction to Variance Estimation, Springer-Verlag, New York.