연구논문

표본의 대표성과 추정의 효율성

Representative of Sample and Efficiency of Estimation

김규성*
Kyu－Seong Kim

본 논문에서는 표본조사에서 흔히 말하여지는 '표본의 대표성'과 추정의 '일치성', '비편향성', '효율성'의 개념을 알아보았다. 표본의 대표성은 표집에 연관된 개념으로 조사모집단의 포함률 및 기초조사의 응답률, 표본섭외 과정의 승락률과 밀접한 관련이 있다. 그리고 추정의 일치성, 비편향성 및 효율성은 표집설계 및 추정량에 동시에 연관된 개념이다. 일치성 및 비편향성은 표본의 대표성을 전제로 한 개념인 반면, 효율성은 표본의 대표성을 전제로 하지 않는다.

표본의 대표성은 포함률, 응답률, 승낙률 등을 제고함으로써 높일 수 있다. 일치성은 관심변수의 일치성과 보조변수의 일치성으로 구분할 수 있으며, 잘 알려진 래킹비 가중법은 모집단 크기를 일치시키는 방법으로 보조변수의 일치성을 높이고자 하는 방법이다. 효율성은 표본의 대표성과는 직접적인 관련이 없으며, 층화표집에서 비례배정과 네이만 배정같은 표본배정, 그리고 사후층화 등은 모두 표본의 대표성이 만족된다는 전제 아래 추정의 효율성을 높이고자 하는 방법들이다.

주제어 : 래킹비 가중법, 비편향성, 사후층화, 일치성, 포함률, 표본배정.

In this paper we investigate some concepts frequently called in sample surveys such as 'representative of sample' as well as 'consistency', 'unbiasedness', and 'efficiency' in estimation. The first is strongly related with sampling procedure including coverage rate of survey population, response rate in establishment survey, and recruit rate of final samples. The others, however, are concerned with both sampling design and corresponding estimators simultaneously. Whereas both consistency and unbiasedness are based on the representative sample, efficiency does not depend on the representative sample.

The representative of sample can be increased by raising the rate of coverage, response and recruit as well. Consistency may be investigated according to variables of interest and auxiliary variables. The well－known

* 교신저자(corresponding author): 서울시립대학교 통계학과 부교수 김규성.
E－mail : kskim@uos.ac.kr

raking−ratio weighting method is a method to increase consistency of auxiliary variables by means of matching population size in each cell. Efficiency is not directly related with the representative of sample, and allocation methods such as proportional and Neyman allocation in stratified sampling and post−stratification are all methods to increase the efficiency of estimation under the condition of satisfying the representative of sample.
key words : consistency, coverage rate, post−stratification, raking−ratio weighting, sample allocation, unbiasedness.

I. 서론

표본조사에서는 모집단의 일부인 표본을 조사하여 모집단 전체에 관한 사항을 추론해야 하므로 표본은 표본 자신뿐 아니라 표본에 포함되지 않은 단위까지 대표해야 한다. 이러한 맥락에서 '대표성 있는 표본(representative sample)'이란 모집단 전체를 잘 대표하는 표본이라는 의미를 함축하고 있으며 통상적으로 표본은 대표성이 있어야 한다고 간주된다. 그러나 실제 많은 조사에서 '표본의 대표성'의 의미는 불분명하게 사용되는 경우가 많으며, 다른 개념, 예를 들면 '일치성(consistency)', '비편향성(unbiasedness)' 혹은 '효율성(efficiency)' 등과 혼용되기도 하고 경우에 따라서는 오용되는 경우도 있다. 따라서 이러한 문제를 해결하기 위해서는 표본의 대표성에 대한 개념을 분명히 할 필요가 있으며 또한 다른 개념들과도 명확하게 구분할 필요가 있다.

문제의 이해를 돕기 위하여 하나의 가구조사에서 두 가지 표집전략(sampling strategy)을 고려하자.

- 표집전략A : 표본가구 선정을 위하여 이중추출법(double sampling)을 사용한다. 전화번호부를 이용하여 1차 표본을 랜덤하게 선정하여 기초조사를 한다. 기초조사에서는 가구의 기본 인구통계적 속성인 지역, 소득, 가구원수, 성별, 연령 등을 질문한다. 1차 표본을 지역으

로 사후층화(post-stratification)를 한 후 지역별로 표본수를 할당한다. 이때 표본수(sample size) 할당에는 1차 표본에서 얻은 지역, 가구원수, 소득 등이 이용된다. 마지막으로 각 지역에서 할당된 수만큼 표본을 섭외하여 표본을 확정한다. 본 조사를 실시한 후, 추정과정에서 가구원의 가중값으로는 포함확률의 역수를 기본 가중값으로 사용한다. 추정량의 형태는 가중값을 이용한 가중평균이다.

- 표집전략B : 표집전략A와 마찬가지로 이중추출법을 사용한다. 전화번호부를 이용한 1차 표본에서는 기초조사로 지역, 소득, 가구원수, 성별, 연령 그리고 기타 보조정보를 수집한다. 2차 표본은 1차 표본을 여러 변수를 이용하여 정렬한 후 계통추출한다. 그리고 가중값으로는 인구속성변수를 이용한 래킹비(raking-ratio) 가중값을 구하여 사용하며 모평균 추정값은 가중평균이다.

표본을 섭외하는 과정에서 선정된 표본의 100%가 표본으로 확정되는 것은 아니다. 예를 들어 표집전략A의 경우 1차 표본 응답률이 60%라 하고 2차 표본 승락률이 70%라고 가정하자. 반면 표집전략B는 1차 표본 응답률이 70%, 2차 표본 승락률이 60%라고 하자. 또한 전략A의 표본수는 1,000이라 하고 전략B의 표본수는 1,100이라고 하자. 두 표집전략에서 추정량의 형태는 동일한 반면, 가중값 부여 방식과 표본수가 서로 다르다. 이제 두 표집전략을 비교하여 다음과 같은 질문을 던질 수 있다. (1) 어느 전략에 의한 표본이 더 대표성이 있는가? (2) 어느 전략으로 구한 추정값의 신뢰도가 더 높은가? 쉽게 생각할 수 있는 부가적인 질문은 다음과 같다. 표집전략B의 표본수가 표집전략A의 표본수보다 많으므로 전략B의 표본이 전략A의 표본보다 더 대표성이 있다고 할 수 있는가? 마찬가지로 전략B에서 구한 추정값을 전략A에서 구한 추정값보다 더 신뢰할 수 있는가?

본 논문에서는 이와 같은 질문에 답을 하기 위하여 표본조사에서 흔

히 사용되는 표본의 대표성, 추정의 일치성, 비편향성, 효율성 등에 대한 개념들을 심도 있게 고찰하고, 이러한 개념을 바탕으로 위의 질문에 대한 답을 제시하고자 한다. 표본의 대표성은 표본이 모집단을 대표하는 방법에 관한 것으로 대표성을 바라보는 관점에 따라 의미가 다르다. 2절에서는 확률화 관점과 모형화 관점에서 표본의 대표성을 고찰한다. 일치성, 비편향성, 효율성 등은 모수 추정과 관련한 개념으로 표본의 대표성과 밀접한 관련이 있기는 하지만 의미하는 바는 다르다. 3절에서는 일치성의 개념을 관심변수와 보조변수로 나누어 살펴보며 4절에서는 비편향성을 고찰한다. 그리고 5절에서는 효율성의 개념을 표본의 대표성과 연관하여 살펴보고, 표본배정 및 사후층화에 따라 추정의 효율성이 어떻게 바뀔 수 있는지를 예제를 통하여 설명한다. 마지막으로 6절에서는 간단한 요약과 함께 위에서 던져진 두 질문에 대한 답을 찾아본다.

II. 표본의 대표성

유한모집단에서 조사단위 및 조사변수에 부여되는 성질은 유한성과 식별성이다. 유한성에 의하여 불균등 확률추출이 가능해지며 식별성에 의하여 조사단위들의 구분이 가능해진다. 유한성과 식별성으로 인하여 다음과 같은 근본적인 물음이 발생한다 : '조사단위 i는조사단위 j를 대신할 수 있는가?' 유한성과 식별성이 없다면 이러한 문제는 발생하지 않으며 원천적으로 표본조사는 발생하지 않았을지 모른다. 표본에 포함된 단위 i가 표본에 포함되지 않은 단위 j를 대신하려면 두 단위 간에 어떤 연관성이 존재해야 한다. 표본조사 초기에 카이어(Kiaer 1897)는 대표성 있는 방법으로 표본은 모집단의 축소판이어야 한다고 주장하였다(Sarndal et al. 1992, p.526). 표본이 모집단의 축소판이라면 표본

은 자연스럽게 모집단을 대표할 수 있기 때문이다. 그러나 복합조사(complex survey)에서 모집단의 축소판이 되도록 표본을 선정하는 일은 쉽지 않은 일이기 때문에 표본과 비표본(非標本, non-sample) 간의 관계를 규명하는 것이 표집이론 개발에서는 매우 중요한 일이었다.

단위들 간에 연관성을 부여하는 방법을 확률화의 관점과 모형화의 관점으로 나누어 살펴본다.

1. 확률화 관점의 연관성

표본을 확률추출(probability sampling)하면 한 조사단위가 표본에 포함되어 모집단을 대표할 기회는 확률적으로 주어진다. 마찬가지로 다른 조사단위도 표본에 확률적으로 포함되기 때문에 서로가 서로를 대신할 기회는 확률적이다. 이처럼 확률추출 과정에서 조사단위 간에 연관성이 발생한다. 대표성 원리(representative principle)는 만일 표본이 확률추출되면 표본에 포함된 조사단위는 자기 자신뿐 아니라 표본에 선정되지 않는 조사단위까지 대표할 수 있고, 표본에 선정되지 않은 단위의 수는 표본에 선정된 단위들의 포함확률로 추정할 수 있다고 하는 것이다(Brewer 1999, p.36). 이 원리는 전통적인 표집이론의 기초 원리이기도 하다. 대표성의 원리가 구현되려면 모집단의 모든 조사단위는 양수의 포함확률(inclusion probability)을 가져야 한다. 포함확률이 0인 단위들은 표본에 포함될 수 없으므로 다른 단위를 대표할 수 없기 때문이다. 모든 조사단위가 양수의 포함확률을 갖는 표집설계를 확률표집설계(probability sampling design, Sarndal et al. 1992, p.32)라고 하는데, 이를 이용하면 '대표성 있는 표본'이란 확률표집설계에 의해서 선정된 표본을 의미하게 된다. 표집이론 책에 나오는 대부분의 표집법은 확률표집법의 일종이다(예를 들면 Cochran 1977; 박홍래 1999). 확률표본(probability sample)은 정의에서부터 이미 표본의 대표성을 확

보하고 있기 때문에 확률표본에서 표본의 대표성을 따지는 것은 무의미하다. 예를 들어 층화표집에서 비례배정과 최적배정을 고려해 보자. 이 경우 '비례배정 표본과 최적배정 표본 중 어느 표본이 더 대표성이 있는가' 하는 질문은 의미가 없다. 왜냐 하면 두 표본은 모두 대표성이 있으며, 단지 차이는 추정의 효율성에 있기 때문이다.

2. 모형화 관점의 연관성

관심변수에 확률모형을 가정하면 조사단위 간의 연관성은 모형을 통하여 발생한다. 예를 들어 다음과 같은 비모형(比模型, ratio model)을 고려하자.

$$Y_i = \beta x_i + v(x_i)\varepsilon_i, \quad i = 1, \ldots, N \tag{1}$$

여기서 $\varepsilon_i \sim (0, \sigma^2)$이다. 그러면 모든 조사단위에서 관심변수 Y_i와 보조변수 x_i는 선형성을 갖는다는 공통점이 생기고, 이러한 연관성은 표본과 비표본을 연결해 주는 매개체가 된다. 확률화 관점과는 달리 모형화 관점에서는 모든 단위가 양수의 포함확률을 가질 필요는 없다. 왜냐하면 어느 조사단위에서든 관심변수와 보조변수는 연관성이 있고 표본은 모형을 통하여 비표본을 대신하기 때문에 모든 단위가 표본에 포함될 기회가 주어질 필요는 없기 때문이다. 대신 추정의 효율을 높이는 표본이 더 선호될 수 있다. 예를 들어 위의 비모형의 경우 모평균 추정에 적합한 최적의 표본은 보조변수값이 큰 조사단위들로 구성된 표본이다(Royall 1970). 이러한 표본은 확률추출된 표본이 아니라 유의선정(purposive selection)된 표본이다. 유의선정에서 표본에 포함되지 않는 조사단위들은 포함확률이 0이기 때문에 모집단을 대표할 기회가 없지만 유의선정된 단위들이 모형을 통하여 모집단을 대표하므로 대표성

의 문제는 발생하지 않는다.

모형을 이용할 때에는 가정된 모형에 대한 타당성 여부가 항상 검토되어야 한다. 만일 가정된 모형이 모집단에 적합하지 않으면 당연한 결과로서 적합하지 않은 모형을 통한 표본의 대표성은 손상될 수밖에 없기 때문이다. 그런데 현실적으로 모집단을 충분히 설명하는 모형을 설정하기는 쉽지 않으므로 모형 가정이 다소 틀리더라도 추론에 큰 영향을 받지 않는 표본이 선호될 수 있다. 그러한 표본을 모형에 강건(robust)한 표본이라고 하는데, 균형표본(balanced sample, Royall & Herson 1973, p.884)이 그 중의 하나이며 할당표본(quota sample)도 강건성을 염두에 둔 표본이다. 할당표본은 미리 정한 할당표에 의하여 표본수를 고정하고 조건에 맞는 단위를 정해진 수만큼 조사하는 것이므로 확률표본은 아니다. 그러나 주요 보조변수를 이용하여 만든 할당표에 의하여 할당표의 셀별로 표본수를 확보하기 때문에 모형의 오류에 강건하게 반응하는 표본이라고 할 수 있다.

3. 대표성 저해 요인

확률화 관점에서 표본의 대표성을 저해하는 요인은 확률표집방법에 있는 것이 아니라 포함확률이 0이 되도록 하는 요인에 있다. 예를 들어 전화조사에서 전화번호부가 모집단 전체를 포함하지 않으면 전화번호부에 등재되지 않은 가구는 표본에 포함될 확률이 0이다. 우리나라의 경우 2003년 현재 전화번호부의 포함률이 대략 70% 정도라는 보고가 있다(허명회 외 2인 2004, p.5). 이 경우 30%의 가구는 표본에 포함될 수 없기 때문에 표본의 대표성은 저하될 수밖에 없다. 랜덤번호전화걸기(random digit dialing, RDD)는 전화번호부의 포함률(coverage rate)의 약점에 착안하여 표본의 대표성을 개선하기 위한 방법으로 볼 수 있다.

앞 절의 예제에서 표집전략A, B는 전화번호부를 기초조사 표본 선

정에 이용했는데, 전화번호부의 포함률이 70% 정도이므로 나머지 30%에 속하는 가구는 원천적으로 표본에 포함될 가능성이 없다. 즉, 전화번호부에서 선정된 표본이 대표할 수 있는 범위는 모집단의 70%라고 할 수 있다. 또한 기초조사 및 표본섭외 과정에서 통화 안 됨, 무응답, 표본섭외 거절 등으로 인하여 두 전략 모두 초기 선정된 표본 중 60% × 70% = 42%만이 표본으로 확정되었다. 즉, 전화번호부 포함률까지 고려하면 두 전략의 표본의 대표성은 70% × 60% × 70% = 29.4%에 불과하다고 할 수 있다. 그런데 전략 A에서는 할당표본을 선정했으므로 확률화 관점에서 보면 2차 표본의 대표성이 70%라고 하기는 힘들다. 할당표본은 표본 선정 조건을 충족하는 가구가 표본수만큼 순차적으로 선정되기 때문에 모집단 단위들의 포함확률이 모두 양수의 값을 취한다고 단언하기 힘들기 때문이다. 할당표본에 대해서는 확률화 관점이 아닌 다른 관점에서 대표성 부여가 추가되어야 한다.

대표성을 향상시키는 방법은 포함확률이 0인 조사단위의 비율을 가능한 한 줄이는 것이다. 즉, 표집에서는 모집단의 포함률을 높이는 것이며, 표본가구 접촉에서는 접촉률을 높이는 것이고, 표본섭외 과정에서는 승낙률을 높이는 것이다. 표본의 대표성 저하가 가져오는 추론상의 문제는 포함확률이 0인 단위들의 특성이 표본조사에 반영되지 않는다는 것이다. 예컨대, 모평균이 관심사라고 하면 표본조사의 추정값은 포함확률이 양수인 단위들의 모집단 평균을 추정하는 수치일 뿐 전체 모집단 평균을 추정하는 수치는 아닌 것이다. 모평균 차이가 클수록 대표성 저하는 추론의 오류를 가져올 가능성이 크다.

모형화 관점에서 볼 때, 표집방법은 대표성과 관련이 없다. 단지 모집단에 가정된 모형이 얼마만큼 잘 적합하는지가 대표성을 좌우한다. 이와 같은 맥락에서 보면 전화번호부의 70% 포함률은 다음과 같이 설명될 수 있다. 등재가구의 특성값에 관한 평균모형이 다음과 같다고 하고,

$$Y_i = \mu_{1+} \varepsilon_i, \ \ \varepsilon_i \sim (0, \sigma_1^2)$$

(2)

비등재가구의 평균모형은 다음과 같다고 하자.

$$Y_j = \mu_{2+} \varepsilon_j, \ \ \varepsilon_j \sim (0, \sigma_2^2)$$

(3)

따라서 만일 두 모평균 및 모분산이 동일하다면, 즉, $\mu_1 = \mu_2$, $\sigma_1^2 = \sigma_2^2$면 모평균 추론에서 전화번호부의 70% 포함률은 문제가 되지 않는다. 그러나 두 모형이 동일하지 않다면, 전화번호부를 통해서 얻어진 추정값은 단지 μ_1을 추정하는 것이므로 비등재가구의 특성을 설명하지는 못한다. 따라서 두 모평균의 차이가 크면 70% 포함률은 대표성에 문제가 될 것이고 두 모평균의 차이가 작으면 대표성 문제는 크지 않을 것이다.

표본의 대표성이 표본 선정 및 모집단에 부여된 모형에 관련한 기준이라면 앞으로 언급할 일치성, 비편향성, 효율성 등은 표집설계 및 추정량에 동시에 관련된 기준이라고 할 수 있다. 일치성, 비편향성, 효율성 등은 동일한 용어를 사용하긴 하지만 표집설계를 기준으로 할 때와 모집단에 부여된 모형을 기준으로 할 때 그 의미가 서로 다르므로 본 논문에서는 표집설계를 기준으로 하는 설계기반 일치성, 비편향성, 효율성만을 다루기로 한다.

III. 일치성

일반 통계학에서 추정량의 일치성은 추론의 대상이 되는 관심변수를 대상으로 한다. 그런데 표본조사에서는 활용 가능한 보조변수를 확보할 수 있는 경우가 많으므로 관심변수뿐만 아니라 보조변수에도 일치성을

부여할 수 있다. 래킹비 가중법(혹은 반복비례가중법: iterative proportional weighting method)은 보조변수에 일치성을 부여하는 대표적인 방법이다. 아래의 소절에서 추정량의 일치성을 관심변수를 대상으로 하는 경우와 보조변수를 대상으로 하는 경우로 나누어 살펴보기로 한다.

1. 관심변수의 일치성

무한 모집단을 대상으로 하는 일반 통계학에서 추정량의 일치성은 표본의 수가 증가하면 추정값은 추정하고자 하는 모수로 접근해 가는 성질을 의미한다. 이러한 개념을 동일하게 유한모집단에 적용하면, 표본의 수를 증가시켜 결국 모집단 전체를 조사할 때 모집단 전체에서 계산된 추정값이 모수와 일치하면 이러한 추정량을 일치추정량이라고 할 수 있을 것이다. 이 개념은 직관적으로 받아들여지기 쉬운 개념으로 보통 사용하는 가중평균, 표본분산, 표본회귀계수 등은 모두 이 성질을 만족하고 있다.

위에서 언급한 일치성은 표본의 수가 증가하는 것을 전제로 하는 데 비하여, 네이만(Neyman 1934)이 사용한 일치성의 개념은 주어진 표본수에서 가능한 모든 표본을 전제로 하고 있다. 즉, 가능한 각각의 표본에서 추정값과 $(1-\varepsilon)$의 오차한계를 갖는 신뢰구간을 구한 후, 모수가 신뢰구간에 $(1-\varepsilon) \times 100\%$ 포함되면 그 추정량은 일치성을 갖는다고 하였다. 이 개념은 관심변수에 관계없이 확률표집만으로 신뢰구간을 구하고 추정량을 구하는 방법으로 설계기반 이론의 새로운 장을 여는 역할을 하였다. 그러나 실제적으로 표본만으로는 이러한 신뢰구간을 구할 수 없기 때문에 네이만이 소개한 일치추정량을 구현하기 위해서는 추가적인 개념의 도입을 필요로 한다.

표집이론 발전 초기부터 유한모집단은 더 큰 초모집단(super-population)의 표본으로 간주할 수 있다는 견해가 있었다(Deming & Stephsn 1941; Hartley & Sielken 1975). 예컨대 센서스를 하나의 큰

표본조사로 볼 수도 있다는 견해이다. 이러한 관점에서 보면 크기가 커지는 유한모집단의 수열을 생각할 수 있고 이러한 수열에서 다음의 성질을 만족하는 추정량의 수열을 생각할 수 있다(Isaki & Fuller 1982, p90) : 주어진 ε에 대하여,

$$\lim_{t\to\infty} \text{Prob}\{|\hat{\Theta}_t - \Theta_t| > \varepsilon\} = 0 \quad (4)$$

여기서 $\hat{\Theta}_t$, Θ_t는 t시점에서의 추정량과 모평균이며, 확률은 표집확률을 의미한다. 이러한 성질을 만족하는 추정량을 설계기반 일치추정량(design-based consistent estimator)라고 부른다. 이러한 설계일치성은 설계기반추론에서는 반드시 갖추어야 할 성질로 받아들여지고 있는데, 그 이유는 설계일치성은 추정량의 편향이 추정량의 표본오차보다 작아질 수 있는 것을 보장해 주기 때문이다(Hansen외 2인 1983, p.779).

어떤 추정량이 설계일치성을 갖기 위해서는 만족해야 하는 성질 중의 하나가 양수의 포함확률을 가져야 하는 것이다(예를 들면 Isaki & Fuller 1982; Robinson & Sarndal 1983). 바꿔 말하면 설계일치성은 표본의 대표성을 전제로 하는 개념이라는 뜻이다.

2. 보조변수의 일치성

관심변수에 부여하는 설계일치성은 점근적 성질이기 때문에 유한 표본을 다루는 실제조사에서는 큰 매력이 아닐 수도 있다. 대신 보조변수를 통해서 추정에 일치성을 부여하는 방법이 있다. 만일 유용한 보조변수가 있다면 보조변수의 일치성은 다음을 의미한다.

$$\sum_{k\in s} w_k x_k = T_x \quad (5)$$

여기서 w_k는 단위 k에 붙는 가중값이며, $T_x = \sum_k x_k$는 보조변수 x의 총계이다. 즉, 위의 성질을 만족하는 가중값 w_k를 사용하면 보조변수의 모총계(population total) 추정량은 모총계와 정확하게 일치한다. 위의 식(5)를 전제로 구한 추정량이 보정추정량(calibration estimator)인데 보조변수가 연속형 변수일 때는 일반화회귀추정량(generalized regression estimator)이 대표적이며 범주형 변수일 때는 사후층화 추정량과 래킹비 추정량이 대표적이다(예를 들면 Deville & Sarndal 1992; Deville외 2인 1993).

앞의 1절에서 소개한 표집전략B는 추정과정에서 래킹비 가중법을 사용하고 있다. 래킹비 가중법은 사회조사에서 종종 사용되며 표본의 대표성을 높여주는 방법으로 설명되곤 한다. 실제로 래킹비 가중법은 표본의 대표성을 높여주는가? 그리고 일치성과는 어떤 연관이 있는가? 이러한 물음에 답을 하기 위하여 래킹비 가중법을 구체적으로 살펴보기로 하자.

우선 래킹비 가중법을 기술하자. 표본조사에서 표본을 조사한 후 사후층화하여 다차원 분할표를 만드는 경우를 고려하자. 예를 들어 성별/연령별/소득별 분할표 등은 사전층화 보다는 사후층화를 통하여 만드는 것이 보통이다. 만일 인구변인이 3개인 3차원 분할표를 고려하면 관심변수의 모총계 추정량으로는 다음과 같은 사후층화 추정량을 고려할 수 있다.

$$t_y = \sum_{ijk} N_{ijk}\, \bar{y}_{ijk}, \quad \bar{y}_{ijk} = \sum_{l \in s_{ijk}} w_{ijkl} y_{ijkl} \tag{6}$$

여기서 N_{ijk}는 (ijk)칸 속하는 모집단 단위의 수이며 s_{ijk}는 표본에 속하는 단위 중 (ijk)칸에 속하는 단위의 모임이다. 위의 식(6)의 사후층화 추정량은 이해하기 쉽고 계산이 간단해서 사용하기 편리한 장점이 있

다. 반면, 사후층화 추정량을 사용하기 위해서는 각 칸의 모집단 크기 N_{ijk}를 알아야 하며 또한 각 칸에 배정된 표본을 모두 조사할 수 있어야 한다. 이제 각 칸의 모집단 크기 N_{ijk}이 알려져 있지 않다고 가정하고 대신 각 보조변수의 주변 모집단 크기는 알려져 있다고 하자. 즉, $N_{i++}=\sum_{jk}N_{ijk}$, $N_{+j+}=\sum_{ik}N_{ijk}$, $N_{++k}=\sum_{ij}N_{ijk}$은 알려져 있다. 그러면 모집단 주변분포는 일치시키면서 각 칸의 모집단 크기를 추정하는 방법을 생각할 수 있다. 그러면 각 칸의 모집단 크기 추정량 $\widehat{N}_{ijk}$은 다음의 조건을 만족한다.

$$N_{i++}=\sum_{jk}\widehat{N}_{ijk},\quad N_{+j+}=\sum_{ik}\widehat{N}_{ijk},\quad N_{++k}=\sum_{ij}\widehat{N}_{ijk}$$

이러한 조건을 만족시키면서 각 칸의 모집단 크기를 추정하는 방법 중 잘 알려진 방법이 래킹비 가중법 혹은 반복비례가중법이다. 각 칸의 모집단 크기는 일종의 보조변수에 해당하므로 래킹비 가중법은 보조변수의 일치성을 갖는다.

이제 래킹비 가중법과 표본의 대표성 그리고 관심변수의 일치성과의 연관성을 알아보자. 앞의 표집전략B에서는 지역/소득/가구원수/성별/연령 등으로 구성한 다차원 분할표에서 각 칸의 모집단 크기를 알 수 없기 때문에 래킹비 가중법으로 각 칸의 크기를 추정하고 있다. 또한 실제 조사된 표본의 모집단 대표성 비율은 포함률(70%), 응답률(70%), 승낙률(60%) 등으로 인하여 29.4%이므로 표집전략B에서는 사후층화 추정량을 직접 사용할 수 없다. 대신 각 칸의 모집단 크기 N_{ijk}를 래킹비 가중법으로 추정하고 완전응답을 받은 경우의 표본평균 $\bar{y}_{ijk}$을 응답 데이터로 추정한 추정량 $\tilde{y}_{ijk}$을 이용하여 관심변수의 모총계를 추정할 수 있다.

$$\tilde{t}_y=\sum_{ijk}\widehat{N}_{ijk}\,\tilde{y}_{ijk},\quad \tilde{y}_{ijk}=\sum_{l\in s_{ijk}}w^*_{ijkl}y_{ijkl}$$

(7)

여기서 $\widehat{N}_{ijk}$는 래킹비 가중법을 이용한 추정량이고 s^*_{ijk}는 (ijk)칸에서 원래 표본 s_{ijk} 중에서 응답을 받은 표본이며, 즉 $s_{ijk}=s^*_{ijk}\bigcup s^{*c}_{ijk}$ 여기서 c는 여집합(complement set)을 의미, w^*_{ijkl}은 포함률, 응답률, 승낙률을 고려하여 사후적으로 조정된 가중값이다. 래킹비 가중법으로 모집단 크기를 추정한 후 만든 식 (7)의 추정량은 래킹비 추정량이 된다. 그리고 래킹비 추정량 $\tilde{t}_y$이 추정하고자 하는 대상은 모총계 $T_y=\sum_{ijk}N_{ijk}\ \overline{Y}_{ijk}$이다. 여기서 $\overline{Y}_{ijk}$는 (ijk)칸의 모평균이다.

먼저 래킹비 가중법과 표본의 대표성의 관계를 살펴보자. 래킹비 추정량에 포함된 표본평균 $\tilde{y}_{ijk}$가 추정하고자 하는 대상은 (ijk)칸의 모평균 $\overline{Y}_{ijk}$이긴 하지만 실제로 $\tilde{y}_{ijk}$가 추정하는 대상은 (ijk)칸에서 표본 섭외를 승낙하고 응답을 한 집단의 모평균이다. 따라서 만일 응답 그룹과 무응답 그룹의 속성이 유사하여 응답 그룹의 평균과 무응답 그룹의 평균이 동일하다면 $\tilde{y}_{ijk}$으로 (ijk)칸의 모평균을 추정할 수 있을 것이다. 그러나 만일 두 그룹의 평균일 일치하지 않는다면 응답 평균으로 무응답 그룹의 모평균을 대신할 수는 없다. 이러한 현상은 래킹비 가중법을 사용한다 하더라도 변하지 않는다. 왜냐하면 래킹비 가중법은 단지 가중값을 조정하는 방법일 뿐이기 때문이다. 따라서 래킹비 가중법을 사용하여 표본의 대표성을 높일 수 있다고 하기는 어렵다.

두 번째로 래킹비 가중법과 관심변수의 일치성과의 연관성을 살펴보자. 각 칸에서 완전응답을 받은 경우를 가정하자. 표본의 수가 증가하면 래킹비 추정량, $\tilde{t}_y=\sum_{ijk}\widehat{N}_{ijk}\ \overline{y}_{ijk}$은 모평균, $T_y=\sum_{ijk}N_{ijk}\ \overline{Y}_{ijk}$에 일치해 간다고 할 수 있는가? 각 칸에서 표본평균 $\overline{y}_{ijk}$는 모평균 $\overline{Y}_{ijk}$에 일치해 가는 성질이 있다. 그러나 래킹비 가중값 $\widehat{N}_{ijk}$은 모집단 크

기 N_{ijk}에 일치해 간다고 하기 어렵다. 그 이유는 다음과 같다. 래킹비 가중법은 분할표에서 보조변수의 주변분포만을 이용하여 각 보조변수의 1차원 효과만 이용할 뿐 보조변수의 결합분포는 이용하지 않는다. 따라서 다음과 같은 표현이 가능하다. $\widehat{N}_{ijk} = (\sum_{l} w_{ijkl})\widehat{\alpha}_i \widehat{\beta}_j \widehat{\gamma}_k$, 여기서 $\widehat{\alpha}_i$, $\widehat{\beta}_j$, $\widehat{\gamma}_k$는 세 변인에 대한 일차원 효과이다. 그런데 각 칸의 모집단 크기 N_{ijk}는 세 변인의 결합효과가 내재되어 있으므로 표본의 수가 증가한다고 하여도 래킹비 가중법으로 구한 추정량 $\widehat{N}_{ijk}$이 모집단 크기 N_{ijk}와 일치하지는 않는다. 추정량 $\widehat{N}_{ijk}$가 N_{ijk}에 일치하는 경우는 보조변수의 결합효과가 없는 경우이다. 이상과 같이 래킹비 가중법을 사용하면 각 칸의 모집단 크기 추정값의 합이 주변분포의 합과 일치하는 보조변수의 일치성은 확보한다고 할 수 있으나 표본의 대표성이나 관심변수의 일치성이 확보된다고 하기는 어렵다. 단, 응답 그룹의 모평균과 무응답 그룹의 모평균이 동일하면 표본의 대표성을 말할 수 있으며, 각 칸의 모집단 크기가 보조변수의 일차원 효과로 표현이 가능하면 관심변수의 일치성을 말할 수 있다.

IV. 비편향성

일반 통계학에서 비편향성은 비중 있게 다루어지는 개념이긴 하지만 절대적인 개념은 아니다. 편향추정량이라도 평균제곱오차가 작으면 추정량으로 선택될 수 있기 때문이다. 그러나 설계기반 관점에서 설계비편향성은 거의 절대적인 기준이다. 적어도 근사 비편향성이나 설계일치성을 가져야 한다. 표집이론에 등장하는 비추정량(比推定量, ratio estimator)이나 회귀추정량은 편향추정량이긴 하지만 근사 비편향추정량이기 때문에 표본의 수가 크면 비편향성을 만족시킨다고 할 수 있다. 보조변수를 이용한 일반화회귀추정량도 마찬가지다.

확률표집설계에 의하여 표본을 선정하면 모든 조사단위는 양수의 포함확률을 갖는다. 따라서 호르비쯔-톰슨(Horvitz-Thompson) 추정량과 같은 비편향추정량을 만들 수 있다. 즉, 대표성을 확보한 표본(즉, $\pi_i > 0$)으로는 비편향추정량을 만들 수 있다는 뜻이다. 반대로 비편향추정이 가능하려면 모든 조사단위에서 포함확률이 양수여야 하므로 비편향추정량 구성에 사용된 표본은 모집단을 대표하는 대표성이 있다. 따라서 설계비편향추정량을 구할 수 있으면 표본의 대표성은 자연스럽게 만족시킨다고 할 수 있다. 설계비편향성이 강조되는 하나의 이유이다.

V. 효율성

추정의 효율성은 표집설계와 추정량을 동시에 고려한 개념이다. 다음과 같은 유한모집단을 고려하자 : $\{(x_i, y_i): i = 1, \ldots, N\}$. 그리고 모총계, $T_y = \sum_i y_i$, 추정을 위하여 두 표집전략이 있다고 하자 : $(p_1, \hat{T}_1)$, $(p_2, \hat{T}_2)$. 만일

$$E_{p_1}(\hat{T}_1 - T | \mathbf{y})^2 \leq E_{p_2}(\hat{T}_2 - T | \mathbf{y})^2, \tag{8}$$

여기서 $\mathbf{y} = (y_1, \ldots, y_N)'$, 이면 표집전략 $(p_1, \hat{T}_1)$이 표집전략 $(p_2, \hat{T}_2)$보다 더 효율적이라고 한다(Thompson 1997, p.21).

1. 표본의 대표성과 추정의 효율성

대부분의 통계조사에서 표본은 모집단을 잘 대표하기를 바라고, 추정량은 모수를 잘 추정하기를 바란다. 즉, 대표성 있는 표본과 효율적인 추정을 동시에 희망한다. 그러나 불행히도 앞의 정의 (8)은 표본의 대표성과 추정의 효율성은 직접적인 연관이 없음을 암시한다. 예를 들어 위의 유한모집단에서 만일 보조변수 x_i와 관심변수 y_i가 서로 비례

하고 x_i가 커질 때 대응되는 y_i의 산포가 x_i에 비례해서 커지면, 모총계의 가장 효율적인 추정량은 비추정량이며 가장 좋은 표집설계는 x_i가 큰 단위들을 순차적으로 포함하는 표집설계이다. 이 표본은 유의표본이므로 모든 단위에서 포함확률이 양수가 아니다. 따라서 이 표본은 확률화 관점에서 볼 때 대표성이 없다. 그렇지만 정의 (8)에 의하면 가장 효율적인 표집설계와 추정량이다. 따라서 표본의 대표성과 추정의 효율성의 두 가지 기준을 모두 충족하는 표본전략은 찾기 어려우므로 두 기준 중 어느 하나를 먼저 선택하는 것이 현실적이다.

두 기준 중 어느 것이 더 중요한 기준인가는 보는 관점과 구체적인 문제에 따라 다를 수 있다. 그러나 대부분의 통계조사에서 우선적으로 받아들여지는 개념은 표본의 대표성일 것이다. 왜냐하면 표본의 대표성이 전제되지 않는 조사 결과는 일반에게 받아들여지지 않을 가능성이 많기 때문이다. 따라서 표본의 대표성이 전제된 후에, 혹은 현실적으로 전화번호부, 무응답에서와 같이 최대한 표본의 대표성을 높이려는 노력을 한 후에, 추정의 효율성을 높이는 방안을 찾는 것이 현실적인 절차이다.

확률표집을 하면 표본의 대표성이 만족되므로, 이후에는 추정의 효율성이 관심의 대상이 된다. 흔한 물음 중의 하나는 표본수와 관련한 표본의 대표성과 추정의 효율성의 관계이다. 예를 들어 '표본수 1,200인 표본은 표본수 1,000인 표본보다 더 대표성이 있는가?' 만일 두 표본이 단순임의표집 되었다면 위의 질문의 답은 '그렇다'라고 해도 무방할 것이다. 그러나 전자의 표본이 집락표집 되고 후자의 표본이 단순임의표집 되었다면 경우에 따라서는 후자의 표본이 더 효율적일 수 있다. 물론 두 표본은 모두 대표성이 있다. 예에서 보듯이 표본수로 대표성의 많고 적음을 말하는 것은 제한적으로만 타당하다. 일반적인 관점에서 표본의 대표성은 확률표집 여부를 나타내는 기준으로, 그리고 추정의 효율성은 추정값의 신뢰도를 나타내는 기준으로 사용하는 것이 타당해

보인다.

2. 비례배정과 최적배정

층화임의표집에서 비례배정은 표본의 대표성에 중점을 둔 방법이며, 최적배정(Neyman 배정)은 추정의 효율성에 중점을 둔 방법이다. 관심변수가 하나인 경우는 최적배정이 비례배정보다 더 선호될 수 있다. 그러나 관심변수가 여러 개인 경우에 일변량 네이만 배정은 더 이상 최적배정이 아니다.

비례배정과 최적배정의 차이를 비교해 보기 위하여 간단한 예제를 들어본다. 모집단은 샹달 외 2인(1992)의 부록에 있는 MU284 데이터를 사용하고, 그 중 4개의 변수를 관심변수로 하였다.

- P85 : 1985년도 인구수(단위 1,000)
- RMT85 : 1985년도 세금
- CS82 : 보수당 의석수
- REV84 : 1984년 부동산 가격

지역변수 Reg를 이용하여 3개의 층으로 층화를 하였고(층1: Reg=1,2/ 층2: Reg=3,4/ 층3: Reg=5,6,7), 표본의 수는 30으로 하여 층 크기에 비례한 비례배정과 각 변수에 최적배정을 하였다(예 : 최적배정1은 P85에 최적배정). 아래의 〈표 1〉은 모집단 상황과 표본배정에 따른 층별 표본수이며 〈표 2〉는 5가지 표본배정에 대한 층화평균의 변동계수이다.

〈표 1〉 층별 표본배정 방법

층번호	층크기	층별 표준편차				표본배정				
		P85	RMT85	CS82	REV84	비례	최적 1	최적 2	최적 3	최적 4
1	73	79.25	751.85	6.43	7256	8	13	10	11	13
2	70	32.18	428.53	4.05	2655	7	5	5	7	4
3	141	39.17	575.08	3.70	3763	15	12	15	12	13

〈표 2〉 표본배정 방법에 따른 층화평균의 변동계수

변수명	변동계수				
	비례배정	최적배정1	최적배정2	최적배정3	최적배정4
P85	0.298	0.278	0.285	0.282	0.279
RMT85	0.417	0.421	0.411	0.418	0.421
CS82	0.087	0.086	0.086	0.085	0.088
REV84	0.262	0.244	0.249	0.249	0.244

위의 〈표 2〉에서 보듯이 4가지 최적배정법은 해당 변수에서는 가장 낮은 변동계수를 보이지만 4개 변수 모두에서 최적인 것은 아니다. 보통의 사회·경제조사는 여러 항목을 조사하는 다변량 조사이므로 일변량 최적배정보다는 다변량 최적배정이 더 현실적으로 필요하다. 다변량 최적배정을 구현하기 위해서는 먼저 변수들의 중요도를 나타낼 수 있는 정량적인 기준을 세우는 일이 필요하다. 예를 들어 각 관심변수의 목표분산의 상한같은 수치가 그러하다. 만일 층화임의표집에서 각 변수의 목표분산의 상한이 정해지면 이 조건을 만족하는 최소 표본수 및 최적배정을 구할 수 있다(Bethel 1989). 이 결과는 관심변수를 정량적으로 비교 가능할 때만 유용한 제한점이 있다.

사회·경제조사에서 쉽게 받아들여지는 비례배정법은 대표성을 강조한 방법이다. 카이어의 주장처럼 표본이 모집단의 축소판이 되려면 표본수는 비례배정 되어야 하기 때문이다. 그러나 효율성의 측면에서

비례배정은 최선의 선택이 아니다. 그러나 위에서 언급했듯이 다변량 조사에서 최적의 표집법을 구현하기 어렵기 때문에 차선책으로 비례배정법이 받아들여지고 있다고 보아야 한다. 개별조사에 대한 지식이 축적되고 경험이 쌓이면 최적표집에 근접하는 방법론이 점진적으로 개발될 것이다.

3. 사후층화

사회·경제조사에서 모든 인구변인을 표집에 고려하기 어렵기 때문에 일부 변인은 사후층화를 통하여 추정에 고려하게 된다. 이 논문에서 던지는 일관적인 물음은 다음과 같은 것이다 : 사후층화를 하면 표본의 대표성과 추정의 효율성이 높아지는가?

사후층화 추정량이나 래킹비 보정추정량은 가중값 보정추정량으로서 보조변수를 사후적으로 이용하여 표본수 혹은 보조변수의 일치성을 높일 수 있다. 그렇지만 원천적인 표본의 대표성을 증가시키지는 못한다. 70%의 포함률을 보이는 전화번호부의 경우 사후층화로 인하여 포함률이 높아지는 것은 아니다. 앞의 표집전략B에서 보듯이 래킹비를 통하여 사후층에서 모집단 크기의 일치성을 높이는 역할만 할 뿐이다. 그러나 원래 표본이 대표성 있는 표본이라면 사후층화를 통하여 추정의 효율성을 높일 수 있다. 사후적으로 이용하는 보조변수가 관심변수와 밀접한 연관이 있을수록 추정의 효율을 더 높일 수 있다. 반대로 보조변수가 관심변수와 연관성이 적으면 도리어 사후층화는 효율이 감소할 수 있다.

아래의 〈표 3〉은 4.2절의 예제에서 층을 사후층으로 간주하여 5가지의 사후층화를 고려하였다. 그리고 〈표 4〉는 무조건부(unconditional) 변동계수(coefficient of variation)와 5가지 조건부(conditional) 변동계수를 구한 것이다.

〈표 3〉 층별 표본배정 방법(사후층 분포에서 계산)

층번호	층크기	사후층별 표본수				
		비례표본수	표본수1	표본수2	표본수3	표본수4
1	73	8	10	15	7	7
2	70	7	10	7	15	8
3	141	15	10	8	8	15

〈표 4〉 표본배정 방법에 따른 추정량의 변동계수

변수명	단순임의 표집	사후층화					
		무조건부	비례	표본수1	표본수2	표본수3	표본수4
P85	0.303	0.314	0.298	0.297	0.294	0.343	0.310
RMT85	0.420	0.435	0.417	0.442	0.465	0.500	0.426
CS82	0.093	0.092	0.087	0.088	0.090	0.098	0.090
REV84	0.266	0.276	0.262	0.264	0.262	0.305	0.272

우선 사후층화 후 구한 무조건부 변동계수는 사후층화를 하지 않고 구한 변동계수와 비교하여 4개 변수 중 CS82만 변동계수가 작을 뿐, 나머지 3개는 도리어 변동계수가 크다. 이 결과는 사후층화가 경우에 따라서는 도리어 추정의 효율을 감소시킴을 의미한다. 사후층화가 더 효율적이려면 층 간 변동이 층 내 변동에 비하여 어느 정도 커야 한다 (Cochran 1977) :

$$\left(\frac{1}{n}-\frac{1}{N}\right)\sum_h(\bar{Y}_h-\bar{Y})^2 > \frac{1}{n^2}\sum_h(1-W_h)S_h^2. \quad (9)$$

여기서 $\bar{Y}_h$는 h층의 모평균, S_h^2는 모분산, W_h는 층 비율, $\bar{Y}$는 모평균이다.

사후층화 추정량의 조건부 변동계수와 단순임의표집의 변동계수를 비교하면 표본수가 사후층에 비례하는 경우는 네 변수의 변동계수가 모두 작게 나타나지만 나머지 표본수에서는 변수별로 효율이 다르게 나타난다. 즉, 사후층화 추정량의 변동계수는 단순평균의 변동계수보다 크게 나타나는 변수도 있고 작게 나타나는 변수도 있다. 이 예가 시사하는 바는 사후층화가 효율적이려면 사후층이 모집단 구조에 적합하게 나누어져야 하고 사후층 내의 표본수도 적절하게 배분되어 있어야 한다는 것이며, 그렇지 못할 경우 도리어 사후층화의 효율성은 저하될 수 있다는 것이다.

VI. 결론

본 논문에서는 표본조사에서 흔히 말하여지는 '표본의 대표성'과 추정의 '일치성', '비편향성', '효율성'의 개념을 알아보았다. 표본의 대표성은 표집에 연관된 개념으로 조사모집단의 포함률 및 기초조사의 응답률, 표본설외 과정의 승낙률과 밀접한 관련이 있다. 그리고 일치성, 비편향성 및 효율성은 표집설계 및 추정량에 동시에 연관된 개념이다. 일치성 및 비편향성은 표본의 대표성을 전제로 한 개념인 반면, 효율성은 표본의 대표성을 전제로 하지 않는다.

표본의 대표성을 높이기 위해서는 포함률, 응답률, 표본 승낙률 등을 높임으로써 대표성 저해 요인을 최대한 제거해야 한다. 일치성은 관심변수의 일치성과 보조변수의 일치성으로 구분할 수 있으며, 잘 알려진 래킹비 가중법은 모집단 크기를 일치시키는 방법으로 보조변수의 일치성을 추구하는 방법에 해당한다. 효율성은 표본의 대표성과는 직접적인 관련이 없으며, 층화표집에서 비례배정과 네이만 배정같은 표본배정, 그리고 사후층화 등은 모두 표본의 대표성이 만족된다는 전제 아래

추정의 효율성을 높이고자 하는 방법들이다. 예제를 통하여 살펴보았듯이 각 방법의 효율성은 경우에 따라 다르게 나타난다.

서론에서 제시한 두 표집전략에 대한 질문의 답을 찾아보자. 두 표집 모두 전화번호부를 기초조사에 사용하고, 응답률 및 승낙률이 100%에 미치지 못하므로 두 전략 모두 표본의 대표성은 개선의 여지가 있다. 특히 전략A는 할당표본을 사용하는데 할당표본에 대표성을 부여하기 위해서는 할당표 및 할당추출 방법을 구체화하여 대표성 부여에 대한 나름대로의 타당성을 확보하여야 한다. 추정의 효율성은 두 전략이 제시하는 추정량에 대한 평균제곱오차를 구하여 비교할 수 있다. 양 전략은 모두 이중추출법을 사용하고 있으며, 1차 표본을 사후층화한 후 전략A는 사후층화추정량을 사용하고, 전략B는 래킹비 추정량을 사용하고 있다. 전체적인 표집방법은 유사하지만 사후층화에 사용된 보조변수, 사후층화 후 가중값 부여 방식(전략A : 칸 가중값, 전략B : 래킹비 가중값) 그리고 표본수 등이 다르므로 표집방법 및 추정량 구성과정만으로는 두 추정법의 효율성을 비교하기 어렵다. 대신 양 전략이 제공하는 추정량에 대한 평균제곱오차 혹은 분산을 구하여 비교할 수 있다. 마지막으로 표본수에 관하여 언급한다. 만일 표집방법 및 추정량의 형태가 동일하다면 표본수가 많을수록 추정의 효율성은 증가한다. 그러나 5절의 예제에서도 보았듯이 표집방법 및 추정량의 형태가 다르면 표본수가 작더라도 효율성이 높은 경우가 발생할 수 있다. 위의 경우도 전략B의 표본수가 전략A보다 크므로 전략B의 추정이 효율성이 높다고 단정하기는 힘들다. 그러다 두 전략이 사용하는 표집방법 및 추정량은 유사하므로 표본수가 크면 추정의 효율성이 높을 가능성은 상당히 많다고 할 수 있다.

참고문헌

박홍래. 1999. 통계조사론. 영지문화사.

허명회 · 강용수 · 손은진. 2004. “사회조사에서 조사방법에 따른 가중 칸 설정에 관한 연구.” ≪조사연구≫ 5권 1호, 1−26.

Bethel, J. 1989. “Sample allocation in multivariate surveys.” *Survey Methodology* 15: 47−57.

Brewer, K.R.W. 1999. “Design−based or prediction−based inference? stratified random vs stratified balanced sampling.” *International Statistical Review* 67: 35−47.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd edition, Wiley.

Deming, W.E. and Stephan, F. 1941. “On the interpretation of censuses as samples.” *Journal of the American Statistical Association* 36: 45−49.

Deville, J.C. and Sarndal, C.E. 1992. “Calibration estimators in survey sampling.” *Journal of the American Statistical Association* 87: 376−382.

Deville, J.C., Sarndal, C.E. and Sautory, O. 1993. “Generalized raking procedures in survey sampling.” *Journal of the American Statistical Association* 88: 1013−1020.

Hartley, H.O. and Sielken, R.L. 1975. “A "super−popualtion viewpoint" for finite population sampling.” *Biometrics* 31: 411−422.

Isaki, C.T. and Fuller, W.A. 1982. “Survey design under the regression superpopulation model.” *Journal of the American Statistical Association* 77: 89−96.

Neyman, J. 1934. “On the two different aspects of the representative method : the method of stratified sampling and the method of purposive selection.” *Journal of the Royal Statistical Society* 97: 558−625.

Robinson, P.M. and Sarndal, C.E. 1983. “Asymptotic properties of the generalized regression estimator in probability sampling.” *Sankhya* B45: 240−248.

Royall, R.M. 1970. “On finite population sampling theory under certain

linear regression models." *Biometrika* 57: 377−387.

Royall, R.M. and Herson, J. 1973. "Robust estimation in finite populations I." *Journal of the American Statistical Association* 68: 880−889.

Sarndal, C.E., Swensson, B. and Wretman, J. 1992. *Model assisted survey sampling*. Springer−Verlag.

Thompson. M. E. 1997. *Theory of sample surveys*. Chapman & Hall.