사례연구를 통한 표본설계과정

류 제 복*

Ⅰ. 서 론

의사결정이나 정책결정에 있어서 관심 대상이 되는 집단의 특성을 파악하는 것이 우선적으로 필요하다. 이들 특성을 파악하기 위해서 대상 집단에 관한 정보가 필요한데 이들 정보는 흔히 연구 대상 전체에 대한 조사를 통해 얻거나 아니면 전체 집단으로부터 일부를 발췌해서 얻게 된다. 전체에 대한 조사를 전수조사(complete enumeration)라하고 일부에 대한 조사를 표본조사(sample survey)라 한다. 사회가 급변하고 있는 현대 사회에 있어서는 저렴한 비용으로 신속하고 정확한 정보가 요구되므로 자연히 표본조사가 일반적인 조사방법으로 사용되고 있다. 따라서 본 과정에서는 표본조사에 국한해서 다루고자 한다.

전수조사든 표본조사든 간에 조사를 통해 얻은 정보는 실제 대상 집단이 갖고 있는 특성과 차이가 있게 되는데 이를 오차(error)라 한다. 따라서 이들 오차를 어떻게 줄이는가가 통계조사에서 주 관심사사 된다. 이들 오차의 정도가 크게 되면 조사를 통해 얻은 정보는 사용가치를 잃게 된다. 그러므로 가능한 한 오차를 줄여 보다 정확한 정보를 얻기 위한 통계조사과정을 살펴본다.

표본조사를 실시하여 정보를 얻기 위해서는 표본조사 계획, 자료수집, 그리고 자료분석의 3단계로 크게 나눌 수 있다. 연구자들에 따라 다소 달리 나눌 수 있지만 이들 과정을 좀더 구체적으로 나누어 살펴보면 다음과 같다.

* 청주대학교 통계학과 교수

Ⅱ. 조사과정

1. 조사목적

조사목적은 조사에 관계하는 모든 사람들이 명확하게 이해할 수 있도록 가능한 한 단순하게 정해져야 하며, 완전하게 달성될 수 있는 것이어야 한다.

조사목적을 3가지로 나누어 보면,

① 탐색 : 사전에 특별한 지식 없거나 선행연구가 없을 때

② 상황이나 사건을 기술 : 현상의 분포자체에 더 관심을 가지는 경우로. 관찰이나 측정과정에서 정확성과 정밀성이 요구된다. 예를 들면, 사회통계조사, 여론조사, 소비자 동향조사 등

③ 설명 : 단순한 기술에서 더 나아가 상황이나 사건을 여러 측면에서 동시에 검토. 성격상 실험과 유사하지만 실험이 독립변수를 조작할 수 있는 반면에 설명적 조사는 독립변수가 되는 요인을 자연적 현상 속에서 발견해내야 한다.

2. 모집단 정의

표본추출의 대상이 될 조사모집단을 명확하게 규정하여야 하고, 조사모집단을 가장 잘 나타낼 수 있는 추출단위들의 목록인 추출틀(frame)을 결정해야 한다. 여러 개의 추출틀을 동시에 사용하여 표본을 추출하는 것이 더 효율적일 수 있다.

3. 표본설계

조사목적을 위해 모집단으로부터 표본을 추출하는 과정, 추출된 표본 자료로부터 모집단 특성치에 대한 추론과 추정치의 정도 계산, 그리고 추정된 자료의 효율적 활용방법들을 다룬다. 표본설계과정에서 주로 다룰 내용은, 추출방법, 표본크기, 표본배분, 추정오차, 가중치, 표본관리 등이다,

4. 조사방법

적절한 조사방법을 선정해야 하는데 대표적인 조사방법으로는 면접조사, 전화조사, 우편조사 등의 방법들이 있고 최근에는 인터넷조사가 사용되고 있다.

5. 조사도구

조사 목적을 달성하기 위해서 설문지를 작성할 시에는 필요한 정보를 정확하고 충분히 확보할 수 있어야 함은 물론 무응답이나 부적절한 응답으로 인한 편향을 최소로 줄일 수 있도록 하는 설문지를 설계하여야 한다. 이를 위해서 신뢰도와 타당도를 측정한다.

6. 조사원의 선정과 훈련

자료를 수집하는 과정에서 발생하는 오차의 주요 원인은 조사원에 따라 큰 영향을 받게 된다. 그러므로 조사원을 선정하고 훈련하는 일에 세심한 주의를 기울여야 한다.

7. 예비조사

예비조사를 통해 설문지나 조사방법 등이 적절한지를 검토해보고, 조사원의 자질을 점검하며, 실제 조사를 위한 관리계획을 수립한다. 본 조를 시행하기 전에 실시되는 예비조사 결과를 통해 조사계획이나 조사방법 등을 일부 수정할 수 있다.

8. 자료관리 및 통계분석

조사의 모든 과정에 걸쳐서 개개의 자료들을 어떻게 다룰 것인지에 대해 전체적인 윤곽을 그려봐야 한다. 이 계획은 현지에서 조사를 통해 자료를 얻을 때부터 시작하여 최종분석이 완료될 때까지의 전과정을 포함하는 계획이어야 한다. 현지에서 조사된 자료와 처리된 자료가 정확하게 일치하는 지를 점검하기 위한 자료관리계획도 세워야 한다. 또한 전체적으로 어떤 분석을 해야 할 지에 대해서도 계획을 가지고 있어야 한다. 조사가 진행되기 전에 미리 최종보고서에 담아야 할 내용이 어떤 것인지를 충분히 생각한다면 조사항목을 결정할 때 유용할 것이다.

조사 결과로 발생하는 오차는 상기의 8개 과정 모두에서 발생할 수 있다. 따라서 최대한 오차를 줄여 정확성 있는 조사결과를 얻기 위해서는 모든 단계에서 오차가 최소가 되도록 심혈을 기울여야 한다. 본 tutorial에서는 위의 과정 중 표본설계와 모집단 및 추출틀에 국한해서 다룬다.

Ⅲ. 모집단, 표본 그리고 추출틀

1. 모집단

모집단이란 어떤 필요한 정보를 얻기 위해서 연구대상이 되는 집단 전체를 말한다. 그러나 연구 대상은 사람, 가구, 기관, 그리고 단체 등 다양하고, 모집단의 규모도 차이가 크다. 경우에 따라서는 연구모집단을 정의하기가 명확치 않은 경우도 있다. 실제 표본조사에서 사용되는 모집단은 이론적으로 정의한 모집단과 실제 통계조사시에 사용할 모집단이 다른 경우가 많다. 이들을 두 가지 형태로 나눈다.

- 연구모집단(population) 또는 목표모집단(target population) : 연구자가 연구에 적용하기 위해서 가정한 이론적인 모집단.
- 조사모집단(working population or survey population) : 연구모집단에 대한 표본선정 작업

상의 정의이고 목표모집단을 대표한다. 연구자가 합리적으로 신원을 확인할 수 있도록 목표모집단의 가능한 모든 수를 완벽한 명부로 작성함으로써 구체화된다. 대부분의 조사에서는 연구모집단과 조사모집단 간에 차이가 발생하게 되므로 가능한 한 차이가 작게 되도록 조사모집단을 정의하는 것이 중요하다.

2. 표본

모집단에서 추출되어 모집단을 대표하는 추출단위들의 집합으로 모집단의 일부분이 된다. 표본은 모집단의 변화에 의존하므로 모집단이 정확하게 정의되지 못하면 부정확한 모집단에서 추출된 표본이 모집단을 잘 대표하지 못하는 것은 자명하다. 따라서 조사모집단의 정확한 규정이 선행되어야 한다.

3. 표본추출틀

조사모집단의 명부를 표본추출틀(sampling frame)이라고 하며, 최종적으로 뽑히는 실제 표본은 이 명부에서 얻게 된다.(여러 개의 추출틀을 동시에 사용). 정기적으로 실시되는 조사의 경우 연구 대상이 되는 모집단은 변동하게 되므로 변동된 모집단을 잘 반영할 수 있도록 새로이 작성된 추출틀이 마련되어야 한다. 추출틀을 선정할 때는 틀이 부정확(inaccuracy), 불완전(incomplete), 중복(duplication), 부적정(inadequate), 노후(out of date) 등의 여부를 살펴야 한다.

(예1) 노인실태조사 : 65세 이상의 인구는 시간의 흐름에 따라 변동.

(예2) 소규모사업체근로실태조사 : 종사자수가 5인 미만인 사업체(97년 기준 84만 여개)

(예3) 국민건강·영양조사 : 15세 이상 국민 대상. 전국적으로 매년 30만 가구의 아파트가 신축됨

(기타) 비정형근로자실태조사, 서울시민 건강지표조사, 정보화통계조사 등

Ⅳ. 표본설계과정

표본설계에서는 조사모집단으로부터 일부의 표본을 추출해서 모집단에 대한 정확한 정보를 얻기 위한 일련의 과정을 다루고 있다. 효율적인 표본설계를 위해서는 조사관련 비용, 요구되는 정도, 얻어진 정보에 대한 평가 등 제반 사항들을 고려해야 한다. 일반적으로 표본설계 시 다루는 내용은 다음과 같다.

1. 모집단분석

연구 대상이 되는 모집단에 대한 분석이 요구된다. 이는 지난 조사와의 연계를 고려하고 새로이 변화하는 모집단의 특성을 파악해서 새로운 표본설계를 하기 위한 기초자료를 얻기

위함이다.

2. 기존 표본설계에 대한 분석

과거 설계를 통해 얻어진 자료를 바탕으로 조사연구의 중요 연구변수들에 대해 기본적인 통계적 분석을 실시해서(평균, 분산, CV 등을 계산) 기존 표본설계에서의 문제점을 파악한다.

3. 새로운 표본설계의 특징

변화된 모집단의 특성을 반영하고 과거 표본설계의 문제점을 보완한 새로운 표본설계를 하고 이의 특징들을 살펴본다. 또한 새로운 표본설계에 대한 효율성분석을 한다.

4. 추출방법, 표본크기 및 표본배분

새로운 모집단의 특성을 파악해서 이에 적합한 표본추출방법을 정한다. 이를 위해서는 모집단의 층화변수를 정하여 모집단을 층화하고 요구되는 허용오차수준에 따른 표본크기를 구하여 층에 따라 표본을 배정한다.

5. 추정

표본조사의 궁극적인 목적은 표본자료를 바탕으로 모집단의 특성을 추정하고 추정된 자료의 신뢰정도를 파악하는 것이다. 실제 조사에서 완전한 자료를 얻을 수 없으므로 이를 위해서 가중치를 계산하고, 추정식과 추정량의 분산 및 분산 추정량을 유도한다.

6. 표본관리

조사대상은 시간이 경과함에 따라 변동하므로 이에 따른 표본관리가 필요하며, 변동된 표본의 특성을 감안한 표본의 대체문제를 고려해야 한다. 또한 조사시에 응답자들의 응답거부나 불성실한 응답으로 인한 무응답오차의 발생을 최소화하도록 한다.

위에서 언급한 절차에 따라 표본설계를 수행한 과정은 8절의 사례연구에서 상세히 다루고 있다. 여기서는 표본설계 시 중점적으로 고려해야 할 사항만을 다룬다.

Ⅴ. 표본추출방법

모집단으로부터 표본을 추출하는 방법은 크게 확률추출과 비확률추출로 나눈다. 확률추출은 확률이론에 근거한 추출 방법으로 추출단계에서 각 추출단위의 추출확률이 알려져 있어야 한다. 확률추출은 모집단의 모수를 추정하기 위한 통계적 추론에 적용될 수 있다. 그러나

비확률추출은 모집단의 특정 원소가 표본으로 뽑히게될 확률이 알려져 있지 않아 잠재적인 응답자들이 동일한 확률로 뽑혔다고 확신할 수 없다. 그러므로 비확률추출로 얻은 표본자료는 표본오차를 추정할 수 없기 때문에 일반화하여 사용할 수가 없다. 여기서는 기본적인 확률추출방법만을 다루도록 한다.

1. 단순확률추출

단순확률추출(simple random sampling)은 여러 표본추출방법 중에서 가장 기본이 되며, 그 밖의 다른 추출방법은 단순확률추출법의 응용으로 볼 수 있다. 단순확률추출법이 표본조사에서 그대로 쓰이는 경우는 많지 않지만 다른 모든 추출법의 기초가 되기 때문에 이 방법을 정확히 이해하는 것이 중요하다.

크기가 N인 모집단으로부터 크기 n인 표본을 추출할 때 모든 가능한 표본들이 추출될 가능성이 동일하도록 해주는 추출과정이다. 단순확률표본들은 난수표를 이용하여 뽑을 수 있다. 난수표는 정수들의 집합으로 이루어진 표로서 결과적으로 10개의 정수(0,1, · · · ,9) 모두를 거의 동일한 비율로 포함하고 있어서 숫자들을 생성하는 양식에서 어떠한 경향도 띠지 않는다. 따라서, 난수표 안에서 랜덤하게 한 숫자가 선택되었다면, 그 숫자는 0부터 9까지의 정수들 중에서 균등확률로 선택된 것이다.

단순확률추출을 하기 위하여 먼저 해야 할 첫 단계는 모집단을 구성하는 N개의 조사단위에 1부터 N까지의 숫자를 부여한다. 그런 다음 난수표나 컴퓨터를 사용하여 1부터 N까지의 숫자 중 서로 다른 n개의 숫자를 뽑고 숫자에 해당하는 조사단위를 표본으로 한다. 난수표를 사용하여 단순확률표본을 추출하는 방법을 살펴보자.

(예) 어떤 회사의 직원 200명으로부터 5명을 단순확률추출하여 직원들의 월 평균 사용하는 용돈을 추정하고자 한다. 이 때 크기 200 ($=N$)인 모집단에서 크기 5 ($=n$)인 표본을 단순확률로 추출하는 것을 난수표를 사용하여 실행하여 보자. 먼저 200명 직원 각각에 001로부터 200까지의 일련번호를 부여하여 명단을 만든다. 난수표를 사용하여 5명의 표본을 뽑기 위하여, 난수표로부터 랜덤하게 선정된 다음과 같은 난수표의 일부를 이용한다.

난수표의 예(일부)

10480	15015	01536	02011	81647
22368	46573	25595	85393	30995
24130	48360	22527	97265	76393
42167	93093	06243	61680	07856
37570	39975	81837	16656	06121
77921	06907	11008	42751	27756
99562	72905	56420	69994	98872
96301	91977	05463	07972	18876
89579	14342	63661	10281	17453
85475	36857	53342	53988	53060

직원 200명은 최대 번호가 200인 3자리수가 부여됨으로 위의 난수표에서 처음 3열로부터 차례로 다음 수를 얻어 낼 수 있다.

104 223 241 421 375 779 995 963 895 854

그리고 처음 3열을 제외한 다음 3열(4,5,6열)로부터는 다음 수를 얻을 수 있다.

801 684 304 679 703 210 627 019 791 753

또한, 다음 3열(7, 8, 9열)로부터는

501 657 836 309 997 690 290 197 434 685

을 얻는다. 동일하게 다음 3열로부터

501 325 022 306 581 711 556 705 263 753
536 595 527 243 837 008 420 463 661 342
020 853 972 616 166 427 699 079 102 539

등을 얻을 수 있다. 이상의 나열된 3자리 숫자 중에서 처음부터 차례로 200이내의 숫자를 찾아 기록하면 다음과 같다.

104 019 197 022 008 020 166 079 102

이 중에서 처음 5개의 숫자

104 19 197 22 8

에 해당하는 일련번호의 직원을 뽑아 표본을 구성하는 것이 크기 $N=200$ 인 모집단으로부터 크기 $n=5$ 인 표본을 난수표를 사용해서 단순확률추출하는 것이 된다. 물론 표본추출에서 복원추출과 비복원추출은 다르다.

난수표로부터 얻어진 일련 번호 중 200을 초과하는 수가 있으면 그 수에서 200을 빼고, 400을 초과하는 수가 있다면 400을 빼는 식으로 하여 표본을 선정할 수 있다. 예를 들어 처음 3열에

서 얻어낸 숫자

104 223 241 21 375

로부터 200에 배수를 뺀

104 23 41 21 175

에 해당하는 일련번호의 학생을 표본으로 선정하면 난수표를 절약하면서 표본을 선출할 수 있다.

2. 층화확률추출

층화확률추출(stratified random sampling)은 모집단을 서로 중복되지 않는 여러 개의 층으로 분할한 다음 각 층에서 독립적으로 일정한 수의 표본을 단순확률추출하는 방법이다.

단순확률추출이 종종 적은 비용으로 모집단 값들에 대한 좋은 추정값들을 제공해 준다. 그러나 모집단에 관한 어떤 정보가 있을 때 이 정보를 이용하면 그만큼 모집단을 잘 대표할 수 있는 표본을 추출할 수 있다. 예를 들어, 시민건강지표조사(참고문헌3)의 경우는 지역(구)을, 정보화통계조사(참고문헌11)의 경우는 지역, 산업, 그리고 시장규모를 층화변수로 사용하였다.

모집단에 속하는 모든 단위를 어떤 지표에 의하여 몇 개의 집단으로 분할하되 분할된 집단의 각 단위들이 동질적인 것이 되도록 하는 것을 층화(stratification)라고 하며 그 분할된 집단을 층(stratum)이라 한다. 층화할 때는 층내는 동질적(homogeneous)이 되도록 하고 층과 층간에는 이질적(heterogeneous)이 되도록 하여야 한다.

단순확률표본 대신에 층화확률표본을 선택한 이유는,

첫째, 조사설계의 목적이 고정된 비용 하에서 정보를 최대한 얻는 것이므로, 추정오차의 한계를 최소화하는데 있다. 표본의 크기가 같은 경우 층화는 단순확률표본에 의한 것보다 추정오차의 한계를 작게 할 수 있어 조사결과의 정도를 높일 수 있다. 층화할 때 층내를 동질적으로 하는 것은 층내에 있는 단위들의 값의 변동을 되도록 적게 하기 위한 것이며, 이렇게 되면 층화의 효율이 더욱 커지게 된다.

둘째, 모집단 조사단위들을 서로 편리한 그룹으로 묶음으로서 인터뷰 시간과 여행에 따른 비용, 그리고 전반적인 추출과정을 경영하는데 따르는 비용 등 조사시 관찰비용을 절감할 수 있다.

셋째, 모집단 전체에 대한 정보뿐만 아니라 모집단의 일부분에 대한 정보가 필요한 경우도 있다. 모집단내의 부그룹들에 대한 모수 추정값을 따로 구할 수 있다. 최근에 지방자치가 시행되면서 전국의 결과뿐만 아니라 지역별 자료가 요구된다. 또한 업종별 임금통계 즉 제조업, 광업, 건설업 등의 통계자료가 필요하게 되는데 이를 위해서 모집단을 조사목적에 맞게 업종별로 층을 설정한다.

모집단을 몇 개의 층으로 나누고자 할 때 각 단위가 어느 층에 속하는 지를 구분하기 위해 기준으로 삼는 변수를 층화변수라 한다. 층화확률추출에서 층화변수를 무엇으로 정하는 지에 따라서 추정의 효율이 달라지므로 효과적인 층화변수를 정하여 모집단을 층화하는 일은 층화확률추출에서 매우 중요한 작업이라 할 수 있다. 실제 표본조사를 위한 표본설계시에 모집단에 대해 가능한 한 많은 정보를 얻어 효과적인 층화기준을 마련하는 것이 필요하다. 조사항목이 단순한 조사에서는 조사항목과 상관도가 높은 과거의 자료나 정보를 이용하여 유사한 것끼리 모으면 되나 조사항목이 복잡하고 다양한 경우에는 어떠한 점에 초점을 맞추느냐가 어려운 문제이다. 이러한 경우에는 층화기준으로 다음과 같은 조건을 고려하여 결정하면 효과적이다.

첫째, 조사항목에서 가장 중심이 되는 항목과 관계가 깊은 특성을 기준으로 한다.

둘째, 양적인 특성에 대해서는 모집단의 분포가 편향된 것 또는 표준편차가 큰 것을 기준으로 한다. 표준편차 대신 변동계수를 이용하는 경우도 많이 있다.

셋째, 질적 속성에 대한 계수적인 것은 모집단에 있어서 비율이 적은 것을 기준으로 한다.

넷째, 시간적인 안정성이 없는 특성은 기준으로 삼지 않는다.

이와 같이 층화확률추출의 본질은 얼마만큼의 보조정보를 근거로 모집단을 층으로 분할하고 분할된 각 층으로부터 독립적인 표본을 추출하느냐 하는 문제이다. 그러므로 모집단의 각 층에 대한 정확한 정보가 필요하고 정보가 없을 때에는 오차가 커지게 되며, 모집단에 대한 층화된 목록이 없는 경우에는 그것을 만들기 위한 많은 시간과 노력이 요구된다. 완전한 표본추출틀이 각 층별로 필요하고 각 층내에서 표본이 추출되어야 하므로 단순확률추출보다 시간과 비용이 더 든다.

일반적으로 성별, 연령, 학력, 지역, 종업원수, 매출액, 주택유형, 주택면적, 직업 등이 층화변수로 널리 쓰인다. 층화변수는 조사목적, 조사내용, 분석내용 등에 따라서 적절하게 달리 사용할 수 있다.

3. 계통추출

많은 이름들이 기재된 리스트에서 n명의 이름을 추출하려 할 때 간단한 방법으로는 적절한 크기의 간격을 정한 후 리스트에서 매번 그 간격에 해당되는 이름들을 선택하면 될 것이다. 예를 들면 매 10번째에 해당되는 이름을 선택하는 방법 등을 생각할 수 있다. 처음 출발점을 랜덤하게 선택하였다면 그 결과 나오는 표본이 바로 계통표본이 된다.

모집단으로부터 첫 번째 추출단위를 처음 일정 개수(k)에서 램덤추출하고, 두 번째 추출단위부터는 일정한 간격(k)으로 표본을 추출하는 방법을 ($1/k$)계통추출(systematic sampling)이라고 한다.

계통추출법의 장점은,

첫째, 처음 k개 단위들 중 하나를 랜덤하게 선택한 후 그 다음에는 자동적으로 매 k번째 간격의 단위를 선택하므로 단순확률추출보다 표본추출작업이 용이하여 비전문가도 쉽게 이용할 수 있다

둘째, 단순확률추출법에 비해 일반적으로 단위비용 당 얻는 정보의 양이 더 많다.

셋째, 만일 표본추출틀이 랜덤하게 배열되어 있는 경우에는 계통추출에 의해 표본을 뽑아도 거의 단순확률추출과 같은 효과를 지니게 된다. 따라서 표본의 추출시에는 수월한 계통추출을 사용하고 그 조사로부터 얻어진 자료들을 분석할 때에는 단순확률추출에 의해 분석하면 된다.

계통추출을 사용하는 예;

(예1) 인구주택총조사에서 표본조사는 1/10 계통표본을 추출하여 조사한다.

(예2) 2000년 4월 13일에 실시된 제16대 총선(국회의원 선거)에 대한 출구조사에서는 계통추출법을 사용하였다.

(예3) 회계사들이 회계과정이 적절한지를 점검하기 위해 회계표 중 일부를 표본으로 선택하거나, 재고금액을 확인하기 위해 재고항목 중 일부를 표본으로 선택할 때 편리한 계통추출법을 활용하는 것이 바람직하다.

계통추출과정을 단계별로 살펴보면 다음과 같다;

첫째, 단순확률추출법에서와 같이 모집단의 크기 N에 대하여 일련번호를 부여하는 표본추출틀을 작성한다.

둘째, 크기 n인 표본을 추출한다고 가정할 때, 추출율 $\frac{n}{N}$ 의 역수인 $k = \frac{N}{n}$ 을 계산하여 추출간격 k를 정한다. 일반적으로 모집단의 크기 N으로부터 계통표본을 추출할 때 k는 $\frac{N}{n}$ 과 같거나 작아야 한다.

셋째, 난수표 등을 이용하여 랜덤한 방법으로 k보다 같거나 작은 수 r을 첫 번째 단위로 선택한다. 이때, r을 임의출발점이라고 한다.

넷째, 두 번째부터의 표본단위 선택은 일정한 간격으로 k만큼 증가시킨 $r+k$, $r+2k$, $\cdots$, $r+(n-1)k$에 해당하는 단위를 선택한다.

(예) 시민지표조사(2001) : 표본조사구에서 10가구를 표본가구로 추출.

표본조사구인 청운동(행정구역번호:1101051)의 일반가구는 총 43개의 가구로 구성되어 있다. 따라서 $k = \frac{N}{n} = \frac{43}{10} = 4.3$ 이므로 1과 4.3사이에 임의의 출발점을 정한다. 만약 출발점으로 난수 1.2가 뽑혔다면 표본가구는 $[1.2]$, $[5.5]$, $[9.8]$, $\cdots$, $[39.9]$가 되어 2, 6, 10, $\cdots$, 40번째 가구를 표본가구로 선정한다. 이때 $[x]$는 x보다 큰 정수 중 가장 작은 정수이다.

계통추출법은 추출과정이 편리하기 때문에 실용성이 매우 높으나 모집단의 단위들이 어떠한 형태로 구성되어 있는가에 따라 효율이 달라진다. 그러므로 계통추출을 사용하기 전에 모집단의 특성을 살피는 것이 중요하다.

(1) 랜덤모집단

랜덤모집단은 모집단 단위가 상호간 어떤 순서나 경향, 상호작용이 거의 존재하지 않는 랜덤한 순서로 구성된 모집단이다. 여기서 추출된 계통표본의 단위들은 랜덤성을 갖고 있으므로 $\rho \approx 0$이라 할 수 있다. 그리고 N이 크다고 하면 $V(\bar{y}) = V(\bar{y}_{sy})$가 성립된다. 이런 경우에는 표본 추출시에는 계통추출법을 쓰고, 추정을 할 때에는 단순확률추출법의 공식을 사용하는 것이 편리하다.

(2) 순서모집단

모집단의 단위들이 어떤 틀에 따라 크기 순서로 구성된 모집단을 순서모집단이라 한다. 추출된 계통표본의 단위들은 $\rho \leq 0$인 이질적인 표본이 되며, N이 클 때

$$V(\bar{y}_{st}) \leq V(\bar{y}_{sy}) \leq V(\bar{y})$$

이 성립한다. 여기서, $V(\bar{y}_{st})$와 $V(\bar{y}_{sy})$는 층화확률추출과 계통추출에 의한 분산을 의미한다.

(3) 주기모집단

모집단의 단위들이 주기적인 변동을 갖으며, 주기적인 모집단에서 추출된 계통표본의 단위들은 $\rho > 0$인 동질적인 표본이 된다. N이 클 때 $V(\bar{y}_{sy}) > V(\bar{y})$가 성립된다.

4. 집락추출

집락이란 조사단위들이 모여있는 그룹이나 집합을 의미하므로 집락추출은 추출단위가 조사단위들의 집합 또는 집락으로 된 확률표본을 추출하는 방법이다.

집락표본을 사용하는 몇 가지 예를 들어보면 다음과 같다.

(예1) 어떤 도시에 있는 가구들이 공공요금으로 평균 얼마만큼을 지출하는 가를 추정하기 위해서 조사를 실시하려고 한다. 그런데 가구들에 대한 목록을 이용할 수 없기 때문에 구나 면 등을 집락으로 하는 집락추출을 사용하는 것이 바람직하다.

(예2) 택시회사가 자기 회사 택시에 대해서 안전치 못한 타이어의 비율을 추정하고자 한다. 이때 타이어에 대한 단순확률표본을 추출하는 것은 비현실적이기 때문에 자동차를 하나의 집락으로 하는 집락추출을 사용하는 것이 좋다.

그리고 이들 예로부터 집락추출을 사용하게 되는 이유를 살펴보면;

첫째, 모집단을 구성하고 있는 모든 조사단위에 대한 틀 또는 목록이 없는 경우

둘째, 모든 조사단위들에 대한 틀을 만들 때 많은 비용이 들지만, 집락에 대한 틀이 있거나 얻는 데 많은 비용이 들지 않을 경우

셋째, 모든 조사단위에 대한 목록은 있지만 조사단위들이 서로 멀리 떨어져 있어 이동비용이나, 관리비용, 등 조사비용이 많이 드는 경우

이상과 같은 이유에서 단순확률추출이나 층화확률추출 등을 사용하기보다는 집락추출을 사용하게 되면 비용도 적게 들고 조사의 관리감독이 수월하여 전체적으로 조사가 용이하게 된다.

집락추출에서의 추출방법은 모집단이 어떻게 구성되어 있는 가에 달려 있다. (예1)의 경우는 1단계로 조사 대상 지역을 여러 개의 구(또는 면)인 집락으로 만들고 이들 중에서 표본집락을 단순확률추출한 후, 추출된 집락내에 있는 모든 가구들을 조사한다. 이러한 집락추출을 1단계 집락추출이라 한다. 통상적으로 많은 조사에서는 다단계 집락추출을 사용하게 된다. 단순확률추출의 경우는 추출단위에 대한 목록(또는 틀)이 있어 이들로부터 난수표 등을 사용하여 표본을 추출하고 층화확률추출인 경우에는 먼저 전체 모집단을 상호 배반인 층으로 나눈 다음 각 층에서 표본을 단순확률추출한다. 집락추출인 경우도 조사단위들의 집합인 집락을 만드는 데, 이들 집락은 층과 마찬가지로 상호 배반적인 부모집단 그룹으로 만든다. 층이나 집락을 구성하는 데 있어서 통상적으로 층들은 상호 이질적으로 만들고 층내의 단위들은 상호 동질적인 것으로 구성하는 것이 바람직한 반면에 집락은 층을 만든 경우와 반대로 집락들은 상호 동질적이고 집락내의 단위들은 상호 이질적으로 만드는 것이 바람직하다. 왜냐하면, 층화확률추출의 경우에는 모든 층을 사용하지만 집락추출의 경우는 모집단을 구성하고 있는 많은 집락 중에서 일부 표본 집락만을 대상으로 조사를 하기 때문에 가급적 집락들이 유사하고 집락내는 이질적일수록 좋게 된다. 이렇게 집락을 구성하면 집락추출은 다른 추출방법에 비해 효율적이 된다. 실제적으로 집락은 행정구역이나 지도상의 지역 등 지리적으로 집락을 구성하여 사용하는 경우가 많다. 물론 집락추출의 효율성은 집락의 크기나 집락의 형태에 영향을 받는다. 집락의 크기가 보조변수로 사용되어 집락의 추출확률을 결정할 때 사용되기도 한다.

Ⅵ. 표본크기 및 배분

표본설계의 어느 단계에서는 반드시 모집단으로부터 뽑을 표본의 크기를 결정해야만 한다. 관측은 비용을 수반하므로 표본이 지나치게 크면 시간과 인력의 낭비를 가져온다. 반대로, 표본의 크기가 너무 작으면, 들인 시간과 노력에 비하여 부적절한 정보를 얻게되어 결과

적으로 낭비가 된다.

표본크기는 어떤 추출 방법을 사용하는 가에 따라 달라진다. 여기서는 단순확률추출의 경우를 예로 든다.

1. 오차의 한계가 주어진 경우

모평균 μ를 추정하기 위하여 오차의 한계가 B일 때 표본의 크기 n을 계산.오차의 한계 B와 신뢰계수 $z_{\alpha/2}$, 그리고 표준오차 $\sqrt{V(\bar{y})} = \sqrt{\frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)}$ 사이에는 다음 기본공식이 성립한다.

$$B = z_{\alpha/2}\sqrt{V(\bar{y})}.$$

위 식으로부터 n을 구하면,

$$n = \frac{N\sigma^2}{(N-1)D+\sigma^2} \text{ (비복원)}, \quad n = \frac{\sigma^2}{D} \text{ (복원)}$$

여기서, $D = \left(\frac{B}{z_{\alpha/2}}\right)^2$.

모비율 p를 추정하기 위하여 필요한 표본의 크기를 구하는 공식은,

$$n = \frac{Npq}{(N-1)D+pq}.$$

실제로 모집단 비율 p를 모르기 때문에 p를 추정값으로 대체시켜 근사적인 표본의 크기를 구한다. 모집단 단위수 $N \to \infty$ 이면 상기 식은 다음과 같이 된다.

$$n = \frac{pq}{D} = \left(\frac{z_{\alpha/2}}{B}\right)^2 p(1-p).$$

사전에 p의 값을 알 수 없는 경우에는 $p = 0.5$을 대입한다.

2. 변동계수를 사용한 경우

조사에 사용되는 문항의 형태는 범주형, 순서형, 연속형 등 여러 형태의 문항들로 이루어져 있고 측정단위도 다양하므로 변수들의 허용오차를 가지고 표본 크기를 결정하기가 곤란한 경우 변동계수를 사용한다. 변동계수는 고유의 단위에 의존하지 않고, 두 조의 단위가 다르거나 단위는 같지만 평균의 차이가 클 때 두 조의 자료의 산포를 비교하는데 유용하다. 또한 모집단이 시간적, 공간적으로 변하여도 변동계수의 값이 크게 변하지 않으므로 표본크기를 결정하는데 일반적으로 널리 사용된다.

$\bar{y}$가 정규분포를 한다는 가정 하에서 상대표준오차를 허용오차로 할 경우 최대허용오차 E는 다음과 같이 표시된다.

$$\left| \frac{\bar{y}-\mu}{\mu} \right| = E = z_{\alpha/2} C_{\bar{y}}$$

여기서 $C_{\bar{y}} = \frac{\sqrt{V(\bar{y})}}{E(\bar{y})} = \frac{\sqrt{V(\bar{y})}}{\mu}$ 는 $\bar{y}$에 대한 변동계수이다. 한편 모집단 평균에 대한 변동계수는 $C = \frac{\sigma}{\mu}$ 가 된다. 따라서 비복원일 경우와 복원일 경우 표본 크기 n에 대해 정리하면 다음과 같다.

$$n = N\left(\frac{z_{\alpha/2} C}{E}\right)^2 \Big/ \left[(N-1) + \left(\frac{z_{\alpha/2} C}{E}\right)^2\right] \quad \text{(비복원)}$$

$$n = \left(\frac{z_{\alpha/2} C}{E}\right)^2 \quad \text{(복원)}$$

위 식에서 모집단 변동계수 C는 사전에 알 수 없는 경우가 많으므로 과거의 유사한 조사자료나 예비조사자료로 추정한 값을 사용한다. 물론 신뢰계수 $z_{\alpha/2}$ 나 허용오차 E는 사전에 주어져야 한다.

3. 표본배분

표본조사설계의 목적은 가장 적은 비용으로 작은 분산을 갖는 추정량을 제공하는데 있다. 표본 n이 결정된 다음에 각 층에 표본 $n_1, n_2, \cdots, n_L$을 배분하는 방법에는 여러 가지가 있고, 이때마다 분산의 크기는 서로 다르게 나타난다. 그러므로 최소의 비용으로 어떤 특정량의 정보를 얻을 수 있는 배분방법을 찾는 것이 필요하다. 이러한 관점에서 가장 좋은 배분방법을 찾기 위해서 다음 3가지 요인을 고려해야 한다.

1. 각 층내의 조사단위의 수
2. 각 층내의 관측값들 간의 변동
3. 각 층내에서 관측에 드는 비용

- **비례배분** : 비례배분은 각 층내에 있는 조사단위들의 크기에 비례하여 각 층에 표본을 배분하는 방법이다. 예를 들어 층의 크기가 N_i $(i = 1, 2, \cdots, L)$일 때 n개의 표본을 각 층별로 다음과 같이 배분한다.

$$n_i = \frac{N_i}{N} n$$

비례배분법은 단지 층의 크기만을 고려하는 것이기 때문에 쉽고 간편하여 가장 널리 사용되는 방법이다. 따라서 층별 변동의 차가 그다지 심각하지 않고 또 층별 조사비용이 비슷하게 드는 경우에는 이 방법을 사용하는 것이 효과적이다.

- **최적배분** : 표본배분시 각 층별 조사단위의 조사비용이 달라 조사비용까지 고려해야 하는 경우에 주어진 비용 하에서 추정량의 분산을 최소화하거나 주어진 분산의 범위 하에서 조사비용을 최소화시키는 방법을 최적배분이라 한다. 다시 말하면 최적배분은 층의 크기와 층내변동의 크기 그리고 층별 조사비용 등 3가지 요인을 모두 고려하여 각 층별 표본을 배분하는 방법이다.

$$n_i = n\left(\frac{N_i\sigma_i/\sqrt{c_i}}{\sum_{k=1}^{L} N_k\sigma_k/\sqrt{c_k}}\right)$$

여기서 N_i는 i번째 층의 크기, σ_i은 i번째 층의 모집단 표준편차, c_i는 i번째 층으로부터 하나의 관측값을 얻는데 드는 비용을 나타낸다. n_i는 N_i 와 σ_i 에 정비례하고, $\sqrt{c_i}$에 반비례한다.

- **네이만배분** : 어떤 표본조사의 경우에는 각 층마다 조사비용 c_i가 크게 다르지 않고 거의 비슷한 경우가 있다. 다시 말하면 모든 층에 대해서 c_i가 같아 $c_1 = c_2 = \cdots = c_L$이 된다. 이와 같이 층마다 조사비용이 일정할 때 분산을 최소화하도록 층별 표본의 크기 n_i를 배분하는 방법을 네이만배분이라 한다.

$$n_i = n\left(\frac{N_i\sigma_i}{\sum_{i=1}^{L} N_i\sigma_i}\right)$$

Ⅶ. 추 정

표본조사 자료를 이용해서 연구모집단의 특성을 추정한다. 모집단의 특성으로 모집단 평균, 모집단 총계, 또는 모집단 비율 등이 많이 사용된다. 추정량과 추정량의 추정오차는 어떤 추출 방법을 사용했는가에 따라 달라진다. 여기서는 비복원 단순확률추출을 사용한 경우만을 다룬다.

<모집단 평균>

추정량 : $\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (표본평균)

추정량의 분산 : $V(\bar{y}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

추정분산 : $\hat{V}(\bar{y}) = \frac{s^2}{n}\left(\frac{N-n}{N}\right)$, $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

<모집단 총계>

모집단 총계를 τ로 표기한다. 즉, $N\mu = \tau$

추정량 : $\hat{\tau} = N\bar{y}$

추정량의 분산 : $V(\hat{\tau}) = N^2 V(\bar{y})$

추정분산 : $\hat{V}(\hat{\tau}) = N^2 \hat{V}(\bar{y})$

<모집단 비율>

모집단 비율 p의 추정량을 $\hat{p}$로 표시한다. $\hat{p}$의 성질은 측정값 y_i가 0 또는 1의 값을 가질 때 표본평균 $\bar{y}$의 성질과 동일하다.

추정량 : $\hat{p} = \frac{\sum_{i=1}^{n} y_i}{n}$ $(= \bar{y})$

추정량의 분산 : $V(\hat{p}) = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$

추정분산 : $\hat{V}(\hat{p}) = \frac{\hat{p}\hat{q}}{n-1}\left(\frac{N-n}{N}\right)$

Ⅷ. 가중치조정

가중(weighting)이란 조사의 일부 또는 모든 응답자들에 적용되는 승수인자(multiplying factor)이다. 표본자료에 가중치를 주는 목적은 표본의 대표성을 증진시키기 위함이다. 예를 들어, 표본단위들이 추출될 확률이 다를 때, 모집단의 실제 상황을 반영하기 위해서 결과치들에 추출확률에 역비례로 가중치를 주는 것이 일반적이다.

가중치를 주려는 이유들은 다음과 같다.

① 사업체조사에서 규모가 큰 사업체를 높은 확률로 추출하는 것이 종종 효율적이다.

② 가구조사에서 조차도 영역(domain)별로 추출확률을 달리 사용하는 것이 때로는 필요하다.

③ 각 표본지역내에서 표본크기를 고정시키려는 제약들이 추출확률을 변동시키는 결과를 가져온다.

④ 조사단위들이 균등확률로 추출되는 설계에서조차도 추출틀의 결함, 추출오차, 무응답

등에 의해서 표본들이 비자체가중(non-self-weighting)이 될 수 있다.

⑤ 많은 조사들이 가중 없이 모집단 총계에 대한 좋은 추정치를 제공하기가 적절치 않다.

표본자료에 가중치를 주고자 할 때는 비용이 드는 것과 가중치를 주었을 때의 이점간에 균형이 이루어지도록 하여야 한다. 전자의 경우는 복잡성, 불편성, 프로그램하고 분석하는 일, 비용 등의 증가, 오류의 증가, 마구잡이 식 가중으로 인한 분산의 증가, 그리고 심지어 적절치 못한 기준을 사용함에 따른 편향의 증가를 포함하고 있다. 반면에, 후자인 이점에는 편향의 감소와 적절한 추정방법을 사용함으로써 얻게 되는 분산의 감소 등이 포함된다. 서로 다른 추출확률, 추출틀의 결함, 무응답 등과 같은 것들의 복합 효과로 인해서 자체가중으로부터의 편차가 유의할 때(±20% 범위밖에 있을 때)는 가중을 고려해야 한다.

실제 자료의 추정에서 가중치를 주는 경우는 ,

① 설계가중치로 불균등확률추출에 대한 보상

② 무응답이 발생한 경우 조정

③ 사후층화에 대한 조정

표본으로부터 비율, 평균, 그리고 다른 여러 비들을 추정하는 데 있어서 가장 중요한 것은 필요한 가중을 무시했을 때 편향이 증가하거나 적절치 못한 가중을 사용함으로써 분산이 증가하는 것이다.

제대로 가중치를 주지 못하면 다음과 같은 요소에 의해서 분산이 증가한다.

$$D_w^2 = \sum_h (W_h w_h) \sum_h (W_h / w_h) = \frac{n \sum_h n_h w_h^2}{(\sum_h n_h w_h)^2} \qquad (8.1)$$

여기서, $n = \sum_h n_h$; $\sum_h W_h = 1$.

위에서, w_h 는 가중치이고, n_h 는 층 h 에 있는 표본 수이다. 그리고 W_h 는 모집단에서 층의 상대크기를 나타낸다. 한 편 위 식은 다음과 같이 가중치의 변동계수(CV)로 나타낼 수 있다.

$$D_w^2 = \frac{n \sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j) \qquad (8.2)$$

여기서, $CV^2(w_j) = \frac{1}{n \bar{w}^2} \sum_j (w_j - \bar{w})^2$; $\bar{w} = \frac{\sum_j w_j}{n}$.

가중치를 무시한 경우 편향효과를 알아보기 위해서 모집단을 구성하고 있는 두 층 (a 와 b 이고, 가중치는 W_a 와 $1 - W_a$)의 평균에 적절한 가중치를 주어 계산한 전체평균과 가중치를 주지 않은 평균에 있어서 편향차이를 구하면,

$$y_w = W_a\, y_a + (1 - W_a)\, y_b$$

$$\bar{y}_u = (\bar{y}_a + \bar{y}_b)/2$$

$$bias = \bar{y}_w - \bar{y}_u = \frac{1}{2}(\bar{y}_b - \bar{y}_a)(W_b - W_a). \qquad (8.3)$$

편향은 평균값들과 다른 가중치를 가진 그룹의 크기의 차이에 의존한다. 가중을 무시함으로써 생기는 편향은 다른 형태의 통계량에 대해서 일정하지 않다. 분산에 대한 가중효과의 상대적 크기는 고려할 통계량의 형태에 따라 크게 변할 수 있다.

(예) 어떤 지역에 거주하는 주민들을 대상으로 지난 1년간 건강검진을 받은 비율을 조사하기 위해서 두 지역을 A, B 두 개의 층으로 나누어 두 지역에서 각각 100명과 400명을 추출하였다. A, B 지역의 추출률은 1/100 과 1/200 이다.

(1) 추출률과 무응답을 고려한 가중치를 계산하고 이를 이용해서 그 지역에서 건강검진을 받은 비율을 추정하라.

(2) 지역 A와 B지역에 거주하는 주민 수가 각각 11,500과 75,800명이라 할 때, 사후층화에 의한 가중치를 구하고 이를 이용해서 건강검진을 받은 비율을 추정하라.

(풀이)

(1)

지역	n_h (표본수)	n_h' (응답수)	y_h (검진 받은 사람수)	w_{dh} (추출률 역수)	$w_{nh} = n_h / n_h'$	$w_h = w_{dh} w_{nh}$	$w_h n_h'$	$w_h y_h$
A	100	70	60	100	1.43	143	10,010	8,580
B	400	380	210	200	1.05	210	79,800	44,100
	500	450					89,810	52,680

$$\Rightarrow \hat{p} = \frac{52,680}{89,810} = 0.5866.$$

(2) 위의 표에서부터 가중표본 주민 수는 10,010과 79,800명이다.

지역	N_h	$w_{ph} = N_h / w_h n_h'$	$w_h^* = w_{ph} w_h$	$w_h^* n_h'$	$w_h^* y_h'$
A	11,500	1.149	164.31	11,501	9,859
B	75,800	0.950	199.5	75,810	41,895
	87,300			87,311	51,754

$$\Rightarrow \hat{p} = \frac{51,754}{87,311} = 0.5928$$

Ⅸ. 사례연구

[사례1] 임업 업종별 경영실태조사를 위한 표본설계(산림청, 2000)

① **조사목적** : 1999년 11월에 실시된 임업 총 조사 자료를 근거로 업종별 경영실태를 정확히 파악하여 임업 진흥정책을 적기에 수립하고 시행하는 것을 목적.

② **모집단 정의와 추출틀** : 1999년 11월에 실시한 임업 총 조사에서 임업가구로 분류되어 조사된 31,274가구를 모집단으로 정의하였다. 임업을 벌목업, 송이 채취업, 유실수 재배업, 야생화 재배업 그리고 제재업인 5개 업종으로 분류하고 이중 유실수 재배업은 4개로 세분하여 총 8개 업종별로 모집단을 구성하였다. 그리고 표본 추출틀은 각 업종별로 작성한다.

③ **표본추출방법, 표본크기 및 표본배분** : 전국을 제주도를 제외한 8개 도와 광역시들을 한데 묶어 총 9개 층으로 나누었고, 제재업의 경우에는 '광역시' 층을 '서울특별시' 및 6개 광역시로 구분하여 총 15개 층으로 세분화하였다. 추출방법은 층화계통추출방법과 절사추출(cut-off sampling)을 함께 적용하였다. 표본크기는 모든 업종에 대해 동일한 CV를 적용할 수 없으므로 표본크기가 모집단 크기의 약 20%정도가 되도록 업종에 따라 CV를 조정해서 1차적으로 표본크기를 계산하였다. 다음에 모든 업종에 대해서 '모평균 + 표준편차'보다 더 큰 핵심변수 값을 갖는 추출단위를 전수조사단위로 선정하고, 모집단 특성상 층내에 개인과 법인이 혼합되어 큰 이상값이 적지 않은 송이 채취업, 밤나무 재배업과 대추나무 재배업의 경우 절사층의 크기를 표본 조사층의 크기와 비슷한 수준으로 조정하였다. 이 결과 얻어진 모평균 추정량의 표준편차는 감소하였는데 이를 이용해서 최종적으로 표본크기를 결정하였다(<표1>참조). 그리고 표본은 Neyman방법을 적용하여 배분하였다.

※ 절사추출법은 층 합계의 추정에 있어서, 변수 값이 큰 몇 개의 추출단위들이 전체의 80~90%의 비중을 차지할 경우, 층 합계에 대한 추정의 효율을 높이기 위하여 변수값이 큰 추출단위들은 모두 조사하고 나머지는 표본을 추출하여 조사하는 방법이다.

<표1> 최종 허용상대오차와 표본크기

업 종	모집단 크기	모평균	표준편차	CV_0	n_s	n_c	n
벌 목 업	1,091	11.35	7.97	0.043	179	37	216
송이채취업	3,277	49.91	29.82	0.024	494	164	658
밤나무 재배업	17,271	871.58	493.87	0.015	1,314	782	2,096
호두나무 재배업	796	183.05	86.93	0.049	94	63	157
대추나무 재배업	4,979	345.72	224.19	0.020	857	136	993
잣나무 재배업	1,935	7,931.63	4,861.27	0.033	280	95	375
야생화 재배업	258	5,865.97	2,811.05	0.050	56	14	70
제 재 업	1,667	1,543.68	1,179.01	0.040	281	33	314
합 계	31,274				3,555	1,324	4,879

④ **추정** : 모집단 총계의 추정량과 추정오차에 대한 공식.

층화절사계통추출법을 사용해서 모집단 총계의 추정식을 구하면 다음과 같다.

$$\hat{Y} = \sum_{h=1}^{L} \hat{Y}_{hc} + \sum_{h=1}^{L} \frac{N_{hs}}{n_{hs}} \sum_{k=1}^{n_{hs}} y_{hsk} \qquad (8.4)$$

여기서 n_{hs}는 h층의 표본, y_{hsk}는 h층의 k번째 표본의 관찰값, $\hat{Y}_{hc}$는 전수조사부분에서 구해지고, $\hat{Y}_{hs}$는 표본조사부분에서 추정한 값이다.

식 (8.4)에 주어진 모집단 총계의 추정량에 대한 추정오차는 표본조사부분에서만 발생하므로 아래 식과 같이 주어질 수 있다.

$$Var(\hat{Y}) = Var(\sum_{h=1}^{L} \frac{N_{hs}}{n_{hs}} \sum_{k=1}^{n_{hs}} y_{hsk})$$

$$= \sum_{h=1}^{L} N_{hs}(N_{hs}-1)\ S_{hs}^{2}\{1-(n_{hs}-\rho_h)\}/n_{hs} \qquad (8.5)$$

여기서 S_{hs}^2은 h층의 표본조사부분의 모분산이고 ρ_h는 h층의 급내상관계수로서 동일한 계통표본내에 있는 조사단위들간의 상관계수이다. 실제 표본조사에서는 S_{hs}^2과 ρ_h를 알 수 없기 때문에 조사된 자료들로부터 추정해야 하지만 일반적인 선형계통추출법에서는 이들의 불편추정량을 계산할 수 없으므로 아래 식으로 분산을 추정할 수 있다.

$$\widehat{Var}(\hat{Y}) = \sum_{h=1}^{L} \widehat{Var}(\hat{Y}_{hs})$$

$$= \sum_{h=1}^{L} \frac{N_h^2}{n_{hs}(n_{hs}-1)} \sum_{k=1}^{n_{hs}} (y_{hsk} - \bar{y}_{hs})^2 \qquad (8.6)$$

⑤ **표본교체** : 표본으로 선정된 조사단위가 조사과정에서 업종변경, 폐업, 타지역 전출, 주거지 불분명, 기타(사망, 수몰지역)사유 등으로 조사가 불가능한 경우에는 조사단위를 교체하였다. 교체 방법은, 특정 업종의 h층에 있는 총 20개의 표본조사단위에서 추출간격 4.0으로 얻은 5개의 계통표본 X_{h3} , X_{h7} , X_{h11} , X_{h15} , X_{h19}에서 X_{h3}가 조사불가능한 경우 특성이 유사한 X_{h2}나 X_{h4}를 교체하여 조사한다. h층에서 전수조사단위 중 일부가 조사 불가능할 경우에는 표본조사단위 중 가장 단위가 큰 유사조사단위로 교체한다. 본 조사에서의 교체율은 전체 표본의 약 4.9%이었다.

[사례2] 시민보건지표조사 표본설계(한국보건사회연구원, 2001)

① **조사목적** : 서울시 전체 및 각 구별로 건강수준 및 의료이용, 보건의식 및 행태 등에 관한 조사를 실시하고, 수집된 자료를 기초로 보건관련 지표를 생산하여 25개 각 구별로 제공함으로써 각 구별 지역 특성에 적합한 지역보건의료계획 수립에 필요한 자료를 제공하고자 한다.

② **모집단 정의와 추출틀** : 서울시민 전체를 모집단으로 하였고 금번 표본설계에서 1차 추출단위는 조사구이고 2차 추출단위는 가구이다. 따라서 조사구모집단으로 2000년 인구주택총조사 후 수정한 조사구를 사용하였다.

③ **표본추출방법, 표본크기 및 표본배분** : 1차 추출단위인 조사구는 먼저 각 구별로 조사구를 동(洞)별로 정렬하고 동(洞)내에서는 아파트조사구와 보통조사구 순으로 정돈한 후에 추출간격을 정하여 확률계통추출하였으며, 2차 추출단위인 가구도 각 조사구로부터 계통추출하였다. 조사비용과 조사원 동원 능력 등을 고려하여 표본가구가 25,000가구로 사전에 정해졌다. 따라서 표본크기는, 한 조사구내 가구수들이 60가구 정도로 대동소이하고 조사여건과 효율성을 감안하여 한 조사구 당 10가구를 표본가구로 정하였다. 한편 표본배분은 각 구별로 최소 표본조사구를 확보하기 위하여 먼저 1500개의 표본조사구를 구별로 60개씩 균일하게 배분한 후 나머지 1000개의 표본조사구는 최적배분법을 적용하였다.

④ **추정** : 서울시 전체 모집단 또는 각 구별 모집단 평균(비율)은 본 조사에서 사용한 층화이단집락추출법의 추정공식을 사용하거나 가중평균을 사용해서 추정할 수 있다. 한편 추정치에 대한 표준오차 추정량도 유도한다. 물론 구별 추정치을 얻은 다음에 이들로부터 서울시 전체에 대한 추정치를 얻는다.

⑤ **표본관리** : 모집단은 시간이 경과함에 따라 변동하게 된다. 즉, 기존의 가구가 없어지거나 새로운 가구가 생기는 등의 조사구내 변동이 생길 수 있고, 기존의 아파트나 가구들

이 철거되고 새로운 아파트나 주택들이 신축되는 경우에도 조사구의 수정 및 보완은 필수적이다. 따라서 모집단에 변동이 생기면 이를 즉시 표본에 반영하여야 한다. 실제 조사를 수행하는 과정에서 조사 단위가 결측되거나 조사 항목에 대한 결측이 생길 수 있다. 조사 단위에 결측이 생기면 1차적으로 표본조사가구를 예비표본조사가구로 교체하고 그래도 결측이 생길 경우는 가중치 조정방법을 사용한다.

[사례3] 비정형근로자실태조사 표본설계(노동부, 2001)

① **조사목적** : 비정형근로자실태조사는 근로자 1인 이상 사업체에 종사하고 있는 비정형근로자에 대해 계약직, 시간제, 파견직, 용역직 등 여러 유형의 비정형근로자 규모 파악과 임금, 근로시간, 고용형태 등 근로조건에 대한 실태를 파악하는 것을 목적으로 한다.

② **모집단 정의와 추출틀** : 비정형근로자근로실태조사의 모집단은 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관, 국제기구 및 기타 외국기관을 제외하고 1999년 12월말을 기준으로 작성된 사업체기초통계조사 결과 중 근로자 1인 이상을 고용하고 있는 전 사업체이다. 일반적으로 해당 근로자가 비정형근로자인지 여부를 판단하기 위해서는 근로형태, 근로시간, 근로장소 등에 대한 변수를 각 근로자로부터 측정해야 한다. 그런데 각 사업체에 대해서 종사상의 지위에 따른 근로자 총수만을 조사하기 때문에 모집단 자료로는 비정형근로자를 정확하게 파악할 수 없다. 따라서 본 표본설계에서는 무급종사자와 임시·일용근로자들을 잠재적인 비정형근로자로 정의해서 이용한다.

③ **표본추출방법, 표본크기 및 표본배분** : 통계 작성의 기본원칙과 표본추출단위가 사업체라는 사실을 고려하여 층화를 산업중분류와 사업체 규모를 층의 기준으로 하였다. 사업체 규모는 사업체 내 상용근로자 수를 기준으로 하였다. 표본 크기는 원칙적으로 현재 모집단 자료에서 비정형근로자를 40,000명 이상 고용하고 있는 산업중분류에 대해서는 월평균 임금총액 추정에 대한 상대표준오차를 5%이하로 10,000명 이상 39,999명 이하인 경우는 7% 이하로, 그 밖의 경우는 10% 이하가 되도록 약 122,230명의 조사 근로자 수와 95,116개의 표본 사업체 수로 결정되었다. 한편 산업중분류 내의 전체 사업체 수가 250개 미만인 산업중분류에 대해서는 전체 사업체를 전수조사하여 추정의 정도를 높였다. 한편 각 산업중분류에 대해서 결정된 조사 비정형근로자 수를 각 사업체 규모에 배분하는 방법은 사업체 규모별 모집단 비정형근로자 수에 비례하도록 배정하는 것을 원칙으로 하였다. 다만 산업중분류 내의 사업체 규모별 비정형근로자 수가 과소한 경우는 해당 산업중분류가 속한 산업대분류의 사업체 규모별 비정형근로자 비율에 따라 배정하였다. 각 산업중분류에서 조사되어야 할 표본 사업체 수는 각 산업중분류별로 조사 근로자수를 먼저 결정하고, 각 산업중분류별 사업체당 평균 비정형 근로자수를 구하여 표본 사업체 수를 결정하였다. 본 조사의 1차 추출단위는 근로자 1인 이상을 고용하고 있는 사업체이고, 2차 추출단위는 각 사업체 내의 비정형근로자이다. 이 조사는 층화이

단추출법에 의해서 먼저 1차 추출단위인 표본 사업체를 추출하고, 사업체 내에서 비정형근로자를 선정하여 조사한다. 표본으로 추출된 사업체 내에 비정형근로자가 많은 경우에는 표본 사업체 내의 전체 비정형근로자 수를 파악하여 그 수에 따라 추출률을 결정해서 전체 비정형근로자 중에서 일부를 추출하여 조사한다. 표본 사업체 내에서 조사근로자를 일부 추출하는 경우에는 계통추출법을 이용한다.

④ **추정** : 비정형근로자 조사는 복합표본조사로 가중치를 이용하면 모집단에 대한 비편향 추정량(unbiased estimator)을 얻을 수 있다.

⑤ **표본관리** : 표본은 조사자료를 얻는데 직접으로 사용되지만 모집단의 변동이 큰 경우에는 표본관리가 어려워진다. 특히 소규모 사업체는 폐업, 휴업, 전업 및 창업 등이 빈번해서 모집단의 증감이 예상되므로 이에 따라 표본크기의 조정이 필요하다. 통상적으로 모집단의 변동이 추출률(사업체 추출률)의 역수만큼 변동이 있는 경우는 변동에 비례해서 표본을 증감시켜야 한다. 표본으로 선정된 사업체(또는 사업체 내의 근로자)가 응답을 거부하여 표본사업체를 그대로 사용할 수 없어 사업체(unit)를 대체하는 경우와 표본 사업체가 일부 항목에 응답을 하지 않아 항목(item)을 대체하는 경우로 구분할 수 있다. 사업체 대체는 다른 사업체로 교체하는 방법과 가중치를 조정해 주는 방법을 사용하고 항목(item) 대체는 어느 조사 대상자가 일부의 항목에 대해서 무응답을 하는 경우에 적용할 수 있는 방법이다. 항목 대체의 가장 중요한 목적은 완비된 조사 데이터를 만드는 것이다. 일반적으로 사용되고 있는 항목 대체방법의 경우는 평균대체, 회귀대체, Hot-deck 대체 등의 방법을 고려한다.

참고문헌

[1] 김영원, 류제복, 박진우, 홍기학 공역(2000), *표본조사의 이해와 활용*, 자유아카데미.

[2] 류제복(2000), 무응답 대체방안, *산업경영연구*, Vol. 23, No. 2, 227-244.

[3] 류제복(2001), 조사설계과정 및 통계분석, *미출간 자료*.

[4] 류제복, 이계오, 김영원(2001), 2001년 국민건강·영양조사 표본설계, *응용통계연구*, 제 14권 제 2호, 289-304.

[5] 보건복지부(1999), *98 국민건강·영양조사 총괄보고서*.

[6] 샘플링아카데미(1999), *표본조사입문*, 자유아카데미.

[7] 서원대학교 통계연구소(2000), *임업 업종별 경영실태조사를 위한 표본설계*.

[8] 조사통계연구회(2000), *무응답오차*, 자유아카데미.

[9] 통계연수부(2001), *표본설계 및 실무과정*.

[10] 한국조사연구학회(2000), *2001년도 국민건강·영양조사 표본설계 및 표본조사구 추출*.

[11] 한국조사연구학회(2001), *2001년 정보화통계조사 표본설계*.

[12] 한국통계학회(1999), *'99 소규모사업체 근로실태조사 표본설계*.

[13] 한국통계학회(2001), *시민보건지표조사 및 건강증진 프로그램개발 표본설계 및 표본조사구추출*.

[14] 한국통계학회(2001), *비정형근로자실태조사 표본설계*.

[15] Kalton, G. and O'Muricheartaigh, C.(2001), Survey Sampling Workshop, Programme of Short Course on Survey Methodology, August 18-22, 2001, Seoul, Korea.

[16] Kish, L.(1965), *Survey Sampling*, John Wiley & Sons, Inc.

[17] Kish, L.(1992), Weighting for unequal P_i, *Journal of Official Statistics*, Vol. 8, No. 2, 183-200.

[18] Lessler, J. T. and Kalsbeek, W. D.(1992), *Nonsampling error in surveys*, John Wiley & Sons, Inc.

[19] Madow, W. G., Nisselson, H., Olkin, I., and Rubin, D. B.(1983), *Incomplete data in sample surveys*, Vol. 1 - Vol. 3, New York : Academic Press.

[20] Verma, V.(1991), *Sampling Methods*, Training Handbook Statistical Institute for Asia and the Pacific, Tokyo.