본 연구 결과는 통계청 학술연구용역에 의해 수행되었음.

조사통계의 정확성지표 품질관리 매뉴얼

2008. 12. 20

(사)한국통계학회

- 연구 과제명 : 조사통계의 정확성지표 품질관리 매뉴얼 개발

- 연구기간 : 2008년 7월 1일 ~ 2008년 12월 20일

- 연구 수행자 :

연구수행기관		(사)한국통계학회
연구책임자	김영원	숙명여자대학교 통계학과 교수
공동연구원	이용희	서울시립대 통계학과 교수
협력연구원		

본 연구결과는 통계청 학술연구용역에 의해 수행되었음.

조사통계의 정확성지표 품질관리 매뉴얼

2008. 12. 20

연구수행기관 : (사)한국통계학회

책임 연구원 : 김영원 (숙명여자대학교 통계학과 교수)

공동 연구원 : 이용희 (서울시립대학교 통계학과 교수)

▮머 리 말▮

본 매뉴얼은 통계청으로부터 연구용역을 받아 한국통계학회에서 수행한 『조사통계의 정확성지표 품질관리 매뉴얼 개발 연구용역』에 대한 최종결과보고서이다. 본 매뉴얼을 작성하게 된 주된 목적은 정확성 높은 조사통계를 작성하기 위해 유의해야 할 사항들을 알기 쉽게 설명하여, 국내 통계작성기관 실무자들이 보다 신뢰할 수 있는 국가통계를 생산하는 데 도움을 주고자 하는 것이다.

조사통계의 품질은 여러 가지 요소로 구성되지만, 조사통계의 품질을 결정하는데 있어서 정확성(accuracy)이 가장 중요한 요소라는 것에는 이견이 없는 것 같다. 조사통계의 정확성을 확보하기 위해서는 다양한 원인에 의해 발생하는 오차를 줄이고 계속적으로 조사과정을 철저하게 관리해야 한다. 즉, 자료의 품질은 생산과정상의 품질관리를 통해 얻어질 수 있다. 전체 조사과정에서 오차를 줄이고 정확성을 높이기 위해서는 각 과정별로 오차발생 원인이 무엇이고 이런 오차를 줄이기 위해서는 어떤 대책을 강구하는 것이 필요한지 파악하는 것이 중요하다. 아울러 각 조사단계별로 정확성에 영향을 주는 요인들을 정량화해서 설명해 줄 수 있는 관련 지표들을 산출해 검토해 보는 것도 큰 도움이 된다.

본 매뉴얼에서는 다양한 형태의 오차발생 원인을 정리하고, 이들 오차를 정량화할 수 있는 관련 지표를 소개한다. 아울러 조사의 정확성을 제고하기 위한 방편으로 각종 오차를 줄이기 위해서는 조사단계별로 어떤 대책을 강구하는 것이 도움이 되는지 정리하고 있다. 본 매뉴얼을 작성하는 과정에서 많은 부분은 Lessler & Kalsbeek의 “Nonsampling Error in Surveys”와 Biemer & Lyberg의 “Introduction to Survey Quality"의 내용을 참고했다는 점을 밝혀둔다.

본 매뉴얼을 작성할 수 있도록 각종 관련 자료를 제공하고 많은 조언을 해 주신 통계청 품질관리과 관계자 여러분께 진심으로 감사를 드린다.

2008년 12월

연구책임자 : 숙명여자대학교 통계학과 교수 김영원
공동연구자 : 서울시립대학교 통계학과 교수 이용희

▌차 례▐

Ⅰ. 조사통계의 정확성

Ⅱ. 표본추출오차

Ⅲ. 추출틀 오차

Ⅳ. 무응답 오차

V. 측정 오차

VI. 자료처리 및 기타 오차

부록 A. 사용자 가이드라인

부록 B. 주요 용어 설명

참고문헌

▌표 차례 ▐

▌그림 차례 ▐

Ⅰ. 조사통계의 정확성

1. 개 요

통계결과의 품질은 여러 가지 요소로 구성된 개념으로 정확성(accuracy), 시의성(timeliness)/정시성(punctuality), 관련성(relevance), 접근성/편리성(accessibility/convenience), 비교성/일관성(comparability/coherence), 효율성(efficiency) 등의 요소를 포함한다. 전통적으로 조사품질은 조사오차(조사결과의 정확성)에 의해 결정된다고 강조되어 왔다. 최근 조사기관들이 조사품질과 관련해 다양한 요소들을 고려하고 있지만, 조사의 품질을 결정하는데 있어서 정확성이 가장 중요한 요소라는 것에는 이견이 없는 것 같다.

조사통계의 경우 품질을 결정해 주는 요소가 많이 있는데, 왜 우리는 자료의 정확성에만 초점을 맞춰야 하는가? 이에 대한 해답은 정확성이 품질의 기초이고, 정확성 없는 조사 자료는 별로 쓸모가 없다는 것에서 찾을 수 있다. 만일 자료가 잘못되었으면 시의성, 관련성, 접근성, 비교성 등이 충족되어도 별로 도움이 안 된다. 사실 정확성 이외의 다른 품질 요소도 중요하긴 하지만 이것들은 최적화해야 할 속성이기 보다 일종의 제약조건으로 볼 수 있다. 예를 들어 통계를 공표하는 시간을 최대한 단축해야 하는 것이 아니라, 주어진 날짜에 통계를 공표할 수 있어야 한다. 이런 관점에서 보면 비용이 제약조건이듯이 적합성이나 적시성 등 다른 품질 요소도 일종의 제약조건으로 볼 수 있다. 따라서 비용과 함께 다른 품질 요소들에 대한 제약조건을 만족하면서 가능한 정확한 자료를 제공하는 것이 통계조사의 목표이다. 이 매뉴얼에서는 정확성을 최대화하는 조사를 설계하는 문제를 다루고자 한다.

정확성은 모집단의 특성을 나타내는 참값(true value)과 조사를 통해 얻은 추정값(estimated value)과의 차이 즉 오차(error)로 설명될 수 있다. 정확성은 다양한 원인에 의해 발생하는 오차를 종합한 개념인 총조사오차(total survey error)로 표현될 수 있지만, 참값을 모르는 상황에서는 다양한 형태의 오차를 측정하는 것이 어렵기 때문에 정확성을 하나의 수치로 정량화하는 것은 불가능하다. 따라서 조사의 정확성을 설명할 때는 통계의 생산과정이 얼마나 신뢰할 수 있도록 관리되고 수행되었는지를 살펴봄으로써 조사의 정확성을 파악할 수밖에 없다.

정확성을 확보하기 위해서는 다양한 원인에 의해 발생하는 오차를 줄이고 계속적으로 조사과정을 철저하게 관리해야 한다. 즉, 자료의 품질은 생산과정상의 품질관리를 통해 얻어진다. 전체 조사과정에서 오차를 줄이고 정확성을 높이기 위해서는 각 과정별로 오차발생 원인이 무엇이고 이런 오차를 줄이기 위해서는 어떤 대책을 강구하는 것이 필요한지 아는 것이 중요하다. 아울러 각 조사단계별로 정확성에 영향을 주는 요인들을 정량화해서 설명해 줄 수 있는 관련 지표들을 산출해 검토해 보는 것도 큰 도움이 된다.

본 매뉴얼에서는 다양한 형태의 오차발생 원인을 정리하고, 이들 오차를 정량화할 수 있는 관련 지표를 소개한다. 아울러 조사의 정확성을 제고하기 위한 방편으로 각종 오차를 줄이기 위해서는 조사단계별로 어떤 대책을 강구하는 것이 도움이 되는지 정리해 보기로 한다.

2. 정확성과 총조사오차

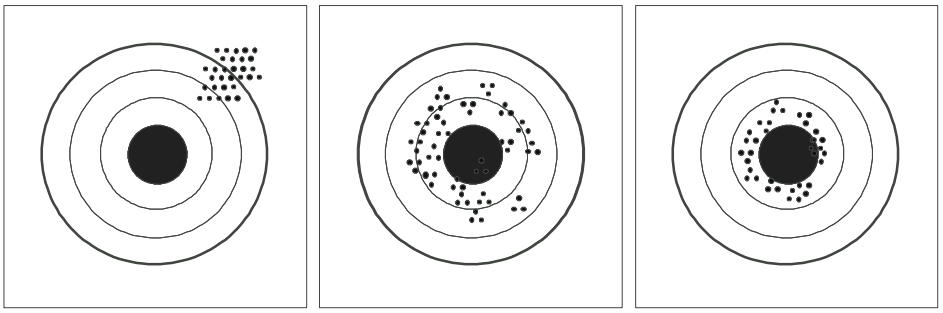
표본조사에서 발생하는 오차는 크게 표본추출오차(sampling error)와 비표본추출오차(non-sampling error)로 구분할 수 있다. 표본추출오차(또는 표집오차)는 표본조사의 경우 모집단 전체 대신 일부 표본을 조사해 얻은 결과를 일반화시키는 과정에서 생기는 오차로 표본조사에서는 필연적으로 표본추출오차가 발생하게 된다. 아울러 이런 표본추출오차는 과학적인 추출법(확률추출법)을 적용한 경우 확률개념을 이용해 통계이론으로 설명이 가능하다. 한편 비표본추출오차는 조사과정 전반에 걸쳐 다양한 원인에 의해 발생할 수 있으며, 조사의 정확성을 저해하는 중요한 요인이 될 수 있다. 비표본추출오차는 발생 원인에 따라 추출틀 오차, 무응답 오차, 측정 오차, 자료처리 오차 등으로 구분할 수 있다.

조사의 정확성은 통계 추정값과 목표로 하고 있는 모집단의 참값 간의 차이를 의미하며, 흔히 이런 차이를 평균제곱오차(mean squared error)로 나타낸다. 평균제곱오차는 정확성 측면에서 조사품질을 설명하는 지표로 사용된다. 평균제곱오차

로 설명되는 총조사오차는 모집단에 대한 전수조사를 실시하는 대신 단지 표본을 통해 결과를 산출함으로써 발생하는 표본추출오차(sampling error)와 자료 수집, 측정, 처리과정 등 조사전반에 걸쳐 발생하는 비표본추출오차(nonsampling error)로 구성된다. 총조사오차를 쉽게 이해하기 위해서는 표본추출부터 자료수집 및 추정까지 전체 통계조사 수행절차를 하나의 과정(process)으로 보는 것이 필요하다. 이런 조사과정을 수행하는 중에 다양한 원인에 의해 오차가 발생할 수 있다. 동일한 과정을 정상적으로 수행하더라도 구체적으로 어떤 표본이 뽑히느냐 또는 자료수집 중에 어떤 상황이 실제로 발생하는지 등에 따라 동일한 방법을 사용하더라도 조사과정을 수행할 때마다 추정결과가 달라진다.

표본조사에서 발생하는 오차의 유형를 좀 더 구체적으로 살펴보면, 총조사오차는 추정결과의 변동(variation)을 나타내는 분산(variance)과 추정결과가 체계적으로 한 쪽으로 치우치는 경향을 가짐으로써 발생하는 편향(bias)으로 구분될 수 있다. 이런 두 가지 유형의 오차를 종합한 개념이 평균제곱오차(mean squared error)이다. 이는 [그림 1-1]과 같은 형식으로 표현될 수 있다.

○ 평균제곱오차(MSE) = (추정결과의 분산) + (추정결과의 편향)^2



(a) (b) (c)

[그림 1-1] 총조사오차: 분산과 편향

[그림 1-1]에서 표본조사과정에 따라 얼마나 정확한 결과를 얻을 수 있는지는 과녁의 중앙으로부터 탄흔이 얼마나 흩어져 맞았는지를 보고 알 수 있다. 여기서 (a)와 같은 결과를 가져오는 표본조사과정은 변동을 의미하는 분산은 작지만 한쪽 방향으로 추정결과가 치우치는 오차, 즉 편향이 큰 경우를 나타내고, (b)는 편향은 발생하지 않지만 분산이 너무 크기 때문에 정확성이 떨어지는 표본조사과정을 보여준다. (c)는 편향도 발생하지 않고 상대적으로 분산도 작기 때문에 (a)나 (b)의 조사과정에 비해서 정확성이 높은 결과를 얻을 수 있는 조사과정이다.

과학적인 방법으로 표본추출을 하고 적절한 가중치를 적용한 추정방법을 사용한

다면 표본추출오차는 변동오차의 형태로 나타나고, 따라서 추정량의 분산으로 표본추출오차를 설명할 수 있다. 하지만 앞으로 다루게 될 추출틀이 갖고 있는 결함, 조사과정상의 무응답, 관측 오차, 자료처리상의 오류 등에 의해서 발생하는 비표본추출오차의 경우 변동오차를 증가시키기도 하지만 무엇보다 심각한 편향을 발생시킬 가능성이 높다는 점에 유의해야 한다. 따라서 품질관리차원에서 보면 과학적인 설명이 가능한 표본추출오차 관리와 더불어 비표본추출오차에 의한 편향을 줄이고 관리하기 위해 많은 시간과 노력을 기울이는 것이 필요하다.

예를 들어 표본추출오차를 제거하기 위해 전수조사를 한다고 해도 우리가 정확한 참값을 조사를 통해서 얻을 수는 없을 것이다. 왜냐하면 이 경우에도 조사과정에서 다양한 원인에 의해 비표본추출오차가 발생하기 때문이다. 아울러 비표본추출오차는 예측할 수 없고 쉽게 통제되지 않는다. 표본크기가 커짐에 따라 표본추출오차는 줄어들지만 오히려 비표본추출오차는 커질 수 있다. 표본크기가 커지면 이에 따라 더 많은 조사원을 신규로 채용해야 하므로 어쩔 수 없이 경험이 적고, 특정 유형의 오류를 많이 발생시키는 면접원을 활용해야 하는 동시에 감독이 어렵기 때문에 오차가 커질 수 있다. 다시 말해 비표본추출오차 측면에서 보면 조사 규모가 커지면 품질관리 체계가 제대로 작동하기 어려워 실사과정이나 자료의 처리과정 등에서 문제가 발생하기 쉽고 그 결과 더 큰 오차가 생길 수 있다.

또한 비표본추출오차는 어떤 자료수집방법을 사용하느냐에 따라서도 많은 영향을 받게 된다. 방문면접조사를 하는 경우와 전화조사 또는 우편조사를 하는 경우 발생하게 되는 비표본추출오차의 원인과 유형이 달라진다. 방문면접조사를 하는 대신 전화조사를 하면 동일한 비용으로 표본크기를 크게 늘릴 수 있기 때문에 오차를 줄일 수 있다고 생각하기 쉽다. 하지만 표본추출오차는 일부 줄일 수 있지만 비표본추출오차는 오히려 증가할 가능성이 높기 때문에 이런 방식으로 총조사오차라는 측면에서 조사의 정확성을 높일 수는 없을 것이다.

최근 조사연구 분야 전문가들은 표본조사에서 비표본추출오차가 표본추출오차보다 추정결과에 훨씬 더 악 영향을 미치는 경우가 많다는 것을 지적하고 있다. 따라서 이 매뉴얼에서도 표본추출오차보다는 비표본추출오차가 발생하는 원인과 이런 오차를 통제하는 것이 왜 필요한지 설명하는데 초점을 맞추고 있다.

3. 통계조사 오차의 원인

표본조사에서 발생하는 총조사오차는 변동오차와 편향으로 구분할 수 있다고 했다. 변동오차는 전수조사 대신 표본조사를 함으로써 불가피하게 발생하는 표본추출

오차에 의해 주로 발생한다. 반면에 체계적인 오차에 해당하는 편향은 조사과정이 제대로 관리되지 않음으로써 발생하는 오류에 가까운 오차이기 때문에 조사설계를 하는 입장에서는 체계적인 오차가 발생하는 원인을 확인하고 관련 오류를 제거하기 위해 노력하게 된다. 아울러 예산이나 인력 등 주어진 자원이 제한적이기 때문에 체계적인 편향을 발생시키는 원인이 다양한 경우 그 중 어떤 오차 발생 요인을 우선적으로 제거하는 것이 효과적인지 판단해야 한다.

예를 들어 추출틀이 갖고 있는 결함을 제거하기 위해 추출틀을 재정비하는데 예산을 사용하는 것이 바람직한지, 아니면 무응답률을 감소시킬 수 있도록 재조사에 예산을 좀 더 배분하는 것이 효과적인지 판단해야 한다. 이런 경우 상황에 따라 변동오차 또는 편향을 가장 많이 발생시키는 오차 원인이 무엇인지를 파악해 필요한 조치가 무엇인지 결정해야 한다. 대부분의 경우 특히 체계적인 오차인 편향을 줄이는 것을 변동오차를 줄이는 것보다 우선적으로 고려하게 된다. 이와 관련해 어떤 오차 발생 요인이 주로 편향을 발생시키는지를 이해하는 것이 조사의 정확성을 제고하는데 도움이 된다.

무응답 오차는 변동오차보다 체계적인 오차에 가깝다. 왜냐하면 무응답은 랜덤하게 발생하는 것이 아니라 일반적으로 특정 그룹에서 더 많이 발생하기 때문에 무응답이 발생하면 추정결과가 편향될 가능성이 높다. 무응답에 의한 편향은 무응답률과 응답자와 무응답자의 특성상의 차이에 따라 달라진다. 한편 추출틀에 모집단 구성원 중 일부 구성원이 누락되는 경우, 예를 들면 농업조사 추출틀에서 규모가 작은 농장은 누락되는 경우를 생각해 보면, 불완전한 추출틀을 사용함으로써 추정결과는 한쪽으로 편향되는 경향이 있게 된다. 이와 관련된 추출틀 오차와 무응답 오차에 대해서는 3장과 4장에서 자세히 살펴보기로 한다.

어떤 오차원인은 주로 변동 오차를 발생시킨다. 예를 들면 자료를 입력하는 사람들이 범하는 오차는 우연이거나 방향성을 갖는다고 볼 수 없기 때문에, 자료 입력 오차는 전형적인 변동오차에 해당한다. 자료를 입력하는 사람들이 어떤 의도를 갖고 자료 값을 늘리거나 또는 줄이려는 경향을 갖는 경우는 흔치 않다. 이런 경우 발생한 오차는 대부분 서로 상쇄되는 경향이 있어 크게 문제가 안 될 수 있다.

어떤 경우에는 변동오차와 편향을 둘 다 발생 시킬 수도 있다. 예를 들어 면접원이 범하는 오차는 변동오차와 편향을 동시에 야기할 수 있다. 가구를 방문해 소득조사를 하는 경우 면접원들의 태도나 복장 등에 따라 어떤 경우는 응답자가 실제 소득을 부풀려 응답하게 유도할 수도 있고, 다른 면접원의 경우 반대의 영향을 줄 수도 있다. 특히 고소득층의 응답자들은 그들의 소득을 실제보다 적게 응답하거나 응답을 거부하는 경향이 강하다. 그러므로 종합적으로 소득은 실제보다 적게 관측되는 경향이 있으며, 동시에 면접원의 태도나 복장 등에 따라 추가적인 변동오차

가 발생할 수 있다.

한편 확률추출법을 이용한 과학적인 조사에서는 표본추출오차에 의해 생기는 편향은 이론적으로 제거하는 것이 가능하지만 변동오차의 발생은 피할 수 없다. 확률추출법을 적용한 표본조사에서 편향이 없는 추정을 위해서는 추출확률을 고려한 올바른 가중치를 작성해 사용해야 한다는 점에 유의해야 한다.

표본조사를 하는 경우 항상 예산과 인력 등 자원이 제한적이기 때문에 정확성을 관리하기 위해서는 다양한 오차 발생원인 중 어떤 것을 우선적으로 관리할 것인지를 판단해야 한다. 이와 관련해 흔히 추정량의 표준오차로 설명되는 표본추출오차를 판단기준으로 사용하는 경우가 많다. 하지만 이는 올바른 접근방법이 아니다. 변동오차에 해당하는 표본추출오차 보다는 앞으로 다루게 될 다양한 원인에 의해 발생하는 비표본추출오차가 오히려 심각한 편향을 발생시켜 조사의 정확성을 크게 저해할 소지가 많다는 점을 항상 염두에 두기 바란다.

II. 표본추출오차

1. 개 요

표본추출오차를 쉽게 이해하기 위해서는 표본추출부터 추정까지의 절차를 하나의 표본조사 과정(process)으로 보는 것이 필요하다. 다시 말해 어떤 표본조사 과정이 주어졌다면 이런 과정을 반복적으로 적용함으로써 무수히 많은 표본을 뽑을 수 있고 이에 따라 무수히 많은 추정값(estimate)을 얻을 수 있다. 이런 관점에서 보면 표본조사 과정을 가상적인 반복 수행을 전제로 한 추정(estimation) 행위로 볼 수 있으며, 이런 반복적인 작업을 전제로 표본추출이론을 전개하기 위해서는 특정 표본이 추출될 확률을 파악하는 것이 필수적이다.

실제 표본조사에 있어서는 조사과정을 반복적으로 적용하는 것이 아니라 하나의 표본만이 추출되고 이를 기초로 모집단에 대한 추정을 하게 된다. 주어진 표본조사 과정의 정확성을 우리가 갖고 있는 단지 하나의 표본에서 얻을 수 있는 정보를 갖고 설명해야 하기 때문에 표본추출오차를 정확히 산출할 수는 없다. 단지 우리가 사용하는 조사과정의 정확성을 통계적인 확률이론을 근거로 표본추출오차를 계산해 봄으로써 추측해 볼 수밖에 없다.

실제 어떤 표본이 추출되는지에 따라 얻어지는 추정값, 즉 추정결과는 변하게 되며, 이런 변동은 표본조사에서는 필연적이다. 이런 변동을 표본추출오차라고 하며, 이런 변동을 기초로 특정 표본조사과정의 정확성을 설명하게 된다. 이런 변동을 과학적으로 설명하기 위해서는 특정 표본이 추출될 확률을 계산하는 것이 필요하고 따라서 특정 표본의 추출확률을 계산할 수 있는 확률추출법(probability sampling)으로 표본을 추출하는 경우, 일반적으로 이런 변동은 추정량의 분산(variance)이나

표준오차(standard error), 상대표준오차(relative standard error)를 의미하는 변동계수(coefficient of variation; CV), 또는 주어진 신뢰수준에서의 오차의 한계(margin of error)로 설명하게 된다.

일반적으로 각 단위가 표본으로 추출될 확률을 계산할 수 있다면 이론적으로 편향이 발생하지 않는 추정량(estimator)을 구할 수 있다. 추정량은 자료가 주어졌을 때 추정값을 계산하기 위해 사용하는 수식을 의미한다. 따라서 확률추출법을 사용하는 대부분의 표본조사에서는 편향이 발생하지 않는 추정량(다시 말해, 추정식 작성과정에서 편향을 제거할 수 있음)을 사용하기 때문에 표본추출오차는 순수하게 변동오차를 나타내는 표준오차 또는 변동계수로 설명하게 된다.

표본추출오차에 대해 이해하기 위해서는 특히 다음 물음에 대해 생각해 볼 필요가 있다.

- 표본은 어떻게 추출하는 것이 바람직한가?
- 표본추출에서 확률화(randomization)의 중요성은 무엇인가?
- 표본추출오차는 무엇이고 그것은 어떻게 발생되는가?
- 조사에서 표본추출오차는 어떻게 측정되는가?
- 어떤 설계 요소들이 표본추출오차의 크기에 영향을 주는가?
- 설계효과, 집락, 층화 등과 같은 표본추출 관련 개념의 의미는 무엇인가?

여기서는 표본추출오차 관련 주제를 광범위하게 다루려는 것은 아니다. 또한 표본을 어떻게 설계할 것인지, 표본추출을 어떻게 할 것인지, 표준오차를 어떻게 추정할 것인지 또는 어떤 추정량을 사용하는 것이 효율적인지 등에 대한 문제를 구체적으로 다루지 않는다. 이런 주제들은 김영원 등(2006), 박홍래(2000), Lohr (1999), Cochran(1977) 등과 같은 표본추출 이론과 활용에 관한 전문서적을 활용할 수 있을 것이다. 여기서는 조사의 정확성을 향상시키기 위해 표본추출오차를 줄이고 관리하기 위해서는 표본조사를 수행하는 과정에서 어떤 점에 유의해야 하고, 어떤 개념들을 이해하는 것이 필요한지 간단히 정리하고자 한다.

2. 표본추출오차의 이해

1) 확률표본추출과 비확률표본추출

확률표본추출(probability sampling)은 모집단을 구성하는 모든 추출단위가 표본

으로 추출될 확률을 계산할 수 있는 추출법을 말한다. 따라서 확률표본추출의 경우 특정한 표본이 선정될 확률을 토대로 표본추출오차를 확률개념을 이용하여 과학적으로 설명하는 것이 가능하다. 단순확률추출(simple random sampling), 계통추출(systematic sampling), 집락추출(cluster sampling), 층화확률추출(stratified random sampling) 등이 대표적인 확률추출법에 해당하고 대규모 표본조사의 경우 흔히 이런 추출법들을 단계별로 적용한 좀 더 복잡한 형태의 복합표본설계(complex sample design)를 사용한다.

한편 특정 표본이 선택될 확률을 알 수 없는 경우, 이런 추출법을 비확률표본추출(non-probability sampling)이라고 한다. 이 경우 특정 표본이 선택될 확률을 알 수 없기 때문에 통계적 추론의 근거가 되는 확률이론을 적용할 수 없고, 따라서 추정결과의 정도(precision)를 나타내는 표본추출오차가 어느 정도 되는지 과학적으로 설명할 수 없다. 아울러 이론적으로 비편향(unbiased) 추정량을 구하는 것이 불가능하기 때문에 이런 추출법을 사용하는 경우 변동오차 이외에 예기치 못한 심각한 편향이 발생할 우려가 있다. 따라서 국가승인통계와 같이 정확성을 요구하는 과학적인 대규모 사회조사를 위해서는 확률표본추출법을 적용하는 것이 필수적이다.

하지만 비확률추출법은 간편하고 비용이 적게 든다는 이유로 다양한 소규모 사회조사에서 광범위하게 사용되고 있는 것이 현실이다. 편의(convenience)추출, 유의(purposive)추출, 할당(quota)추출 등이 대표적인 비확률추출법이다. 참고로 이들 추출법의 특성을 정리하면 다음과 같다(한국통계학회 조사통계연구회, 2005).

○ 편의(convenience), 우연(accidental), 무계획(haphazard) 추출법

조사자가 손쉽게 접촉할 수 있는 조사대상을 표본으로 추출하는 경우이다. 예를 들어 어떤 사거리 또는 쇼핑센터 앞에서 자발적 참여자를 대상으로 한 조사, 특정 포털 사이트에서 자발적 참여자들을 대상으로 한 인터넷 조사 등이 이에 해당한다. 한편 눈덩이추출(snowball sampling)도 편의추출의 일종이지만 폭력서클 가입 청소년들, 또는 불법체류 외국인 근로자들과 같이 현실적으로 조사대상자를 쉽게 접촉할 수 없는 사회조사에서 매우 효과적으로 활용될 수 있는 표본추출 방법이다.

○ 유의(purposive), 판단(judgement) 추출법 또는 전문가 선택(expert choice)

주관적인 판단에 의해 모집단을 대표할 수 있는 대상들을 표본으로 추출하는 방법으로 실제 표본추출에서 표본크기가 작은 경우 확률추출법에 비해 유리할 수 있다. 예를 들어, 서울시내 고등학교를 대표하기 위해 예산상의 문제로 4~5개 학교만을 표본으로 추출해야 하는 상황이라면 단순확률추출법에 비해 교육청 전문가의 선택에 의해 고등학교를 선택하는 것이 표본의 대표성 측면에서 유리할 수 있다.

○ 할당(quota) 추출법

관심 변수에 영향을 주는 주요 특성에 대한 모집단 비율이 표본에 그대로 유지되도록 하는 추출법을 말한다. 여기서 구성 비율을 제어하는 것은 조사원에 의한 선택편향을 제어하기 위한 것으로 볼 수 있다. 확률추출법에 해당하는 층화추출과 유사한 형식이지만 할당추출에서는 층(그룹)내 조사대상을 표본으로 추출하는 과정에서 랜덤화(randomization)과정을 거치지 않는다는 것이 핵심적인 차이이다.

층화추출과 달리 표본추출틀(sampling frame)을 확보할 필요가 없고, 무응답에 대한 재조사 과정도 고려할 필요가 없는 등 실제 적용상의 편리함 때문에 국내 조사회사에서 널리 사용되고 있다. 특히 표본의 대표성 측면에서 보면, 할당추출을 적용하는 경우 과연 연구대상이 되는 관심변수에 영향을 주는 모든 특성을 할당추출 과정에 반영하는 것이 현실적으로 가능한 것이지 고민해 볼 필요가 있다. 많은 경우 할당추출을 적용하면 표본의 대표성을 담보할 수 없고, 표본의 대표성이 떨어지게 되면 결국 수용할 수 없는 수준의 예기치 못한 편향이 발생할 수 있다는 점에 유의해야 한다.

[예제 2-1] 우리나라에서는 대부분의 전화여론조사에서 할당추출이 적용되고 있다. 예를 들어 선거예측을 위한 전화여론조사에서는 대부분 각 지역별로 성별 및 연령대별 상주인구를 기준으로 연령대 및 성별 표본크기를 사전에 할당하여 일반가구 전화번호부를 기초로 표본을 추출하는 방식으로 조사가 이루어지고 있다. 예를 들어, 이런 할당추출법으로 오전 10시부터 오후 8시까지 여론조사를 수행한다고 가정하자. 시간대별 할당과 같은 추가적인 조치가 없는 경우 30~40대 여성은 오후 5시 이전에 직장을 다니지 않는 주부들로 주어진 쿼터가 다 채워지기 때문에 직장에 다니는 30~40대 여성이 표본으로 추출될 가능성은 거의 없게 될 것이다. 다시 말해 이런 할당추출법을 적용하게 되면 성별과 연령대이외의 다른 주요 특성(예를 들어, 소득, 교육수준 등)에 있어서 전체 유권자를 대표할 수 있는 표본을 확보할 수 없기 때문에 심각한 편향이 발생할 소지가 있다.

아울러 할당추출을 사용하는 경우 조사과정의 정확성을 나타내는 표본추출오차도 산출하는 것이 불가능해 진다. 따라서 국가승인통계와 같이 조사의 정확성을 확보하는 것이 요구되는 경우 할당추출법 등과 같은 과학적으로 정확성을 확인할 수 없는 비확률추출법을 사용하는 것은 적절치 않다. 과학적인 조사를 강조하는 통계선진국에서는 주요 통계를 전화조사를 통해 생산하는 경우 할당추출법이 사용된 사례는 거의 찾을 수 없으며, 대신 무작위전화걸기(random digit dialing; RDD) 기법이 폭넓게 사용되고 있다. □

2) 표본크기

표본으로 선택된 개체를 관찰하는 데는 시간과 비용이 든다. 표본의 크기가 크면 클수록 추정량의 분산이 줄어들게 되어 표본추출오차가 감소한다는 장점이 있는 반면 조사를 위한 시간과 비용이 많이 든다는 단점이 있다. 따라서 가능한 예산과 시간의 범위 내에서 효율을 극대화할 수 있도록 표본크기(sample size)를 결정해야 한다.

표본조사를 계획할 때에는 먼저 그 조사에서 얻고자 하는 추정량의 정도(精度)의 한계를 미리 정해주게 되는데 이것을 목표정도(target precision)라고 한다. 목표정도는 일반적으로 오차의 한계나 변동계수의 형태로 주어진다. 예를 들어, 소득에 대한 표본조사에서 95% 신뢰수준에서 추정결과에 대한 오차의 한계가 ±30만원을 넘지 않도록 해야 한다든지, 아니면 근로자 임금조사에서 근로자 평균임금에 대한 변동계수(CV)가 3%를 넘지 않도록 해야 한다는 식으로 사용 목정에 따라 표본조사를 기획할 때 목표정도를 정하게 된다.

표본추출방법에 따라 표본추출오차를 나타내는 표준오차나 변동계수를 산출하는 방법이 다르기 때문에 어떤 표본추출방법을 적용할 것인지를 미리 결정한 후 설정된 목표정도를 만족하는 표본크기를 정하게 된다. 단순확률추출법이외의 추출법을 사용하는 경우 추출법의 효율성을 흔히 설계효과(design effect)로 나타낸다. 또한 비슷한 개념을 이용해 해당 표본추출과 동일한 수준의 정도를 확보할 수 있는 단순확률추출법에서의 표본크기를 나타내는 유효표본크기(effective sample size)를 산출하여 표본추출오차가 어는 정도인지를 판단하기도 한다. 표본추출법에 따라 목표정도를 만족하는 표본크기를 구하는 방법이나 설계효과를 산출하는 방법 등에 관한 구체적인 내용은 김영원 등(2006), 박홍래(2000), Lohr(1999) 등과 같은 표본추출이론 관련 서적에 소개되어 있기 때문에 자세한 내용은 여기서 다루지 않기로 한다.

대부분의 통계조사에서는 전체 모집단에 대한 추정뿐만 아니라 모집단의 일부 구성원들로 정의되는 부차모집단(sub-population) 또는 영역(domain)별 추정을 하게 된다. 예를 들어 1,000명의 성인들을 대상으로 조사를 한 후, 20대, 30대 등 연령대별로 추가적인 분석을 하는 경우, 또는 전국에서 5,000가구에 대한 조사를 한 후 16개 시도별로 통계를 작성하는 경우 등이 이런 사례에 해당한다. 이런 영역별 추정을 하는 경우 추정결과의 표본추출오차는 해당 영역에 배정된 표본크기에 의해 결정된다. 따라서 전국단위 통계의 정확성에는 문제가 없더라도 시도별 통계의 표본추출오차가 너무 커져서 신뢰할 수 있는 시도별 통계 생산이 불가능한 경우가

흔히 발생한다. 수용할 수 없는 수준으로 표본추출오차가 커지는 경우 신뢰성이 떨어져 해당 영역별 통계는 실제 사용할 수 없다는 점을 항상 염두에 두어야 한다. 이런 경우 분석대상이 되는 영역을 층(stratum)으로 정의하는 층화추출법을 적용하면 층별 다시 말해 시도 등과 같은 영역별 표본크기를 사전에 조정할 수 있기 때문에 큰 도움이 된다.

한편, 많은 경우에는 실제 사용이 가능한 예산이나 조사 인력에 제약이 있기 때문에 사전에 설정된 목표정도에 따라 표본크기를 정하는 것이 아니라 주어진 예산과 인력으로 수용이 가능한 범위 내에서 적절한 수준으로 표본크기를 정하기도 한다. 이런 경우에도 표본설계 단계에서 예산에 따라 결정된 표본크기로 조사를 하게 되면 결과적으로 어느 정도의 표본추출오차가 발생하게 될 것인지 사전에 충분히 검토할 필요가 있으며, 만약 주어진 예산에 따라 실현 가능한 표본크기로는 원하는 표본추출오차를 충분히 만족시킬 수 없다고 판단이 되는 경우 추가 예산을 확보해 표본크기를 늘리는 것이 필수적이다.

[예제 2-2] 단순확률추출에서 신뢰수준 $100(1-\alpha)\%$일 때 모평균의 추정량에 대한 오차의 한계가 B로 주어진 경우 표본크기(n)를 구하는 공식은 다음과 같다.

$$n = \frac{N(z_{\alpha/2}S)^2}{B^2 + (z_{\alpha/2}S)^2} = \frac{n_o}{1 + \frac{n_o}{N}}, \quad n_o = \frac{(z_{\alpha/2}S)^2}{B^2}$$

여기서 N은 모집단크기, $z_{\alpha/2}$은 표준정규분포에서 백분위수, S^2은 모집단 분산을 나타낸다. 일반적으로 S^2은 알 수 없기 때문에 표본크기를 정하는데 어려움이 생기게 된다. 과거에도 비슷한 조사를 실시하여 그 자료가 유용한 경우에는 과거의 자료로부터 얻은 표본분산 s^2을 사용한다. 과거 유용한 자료가 없고 처음 조사를 계획하는 경우에는 작은 규모의 예비조사를 통해 s^2을 구해 사용하기도 한다. □

[예제 2-3] 국가 주요통계 생산을 위해 정부부처나 연구기관 등에서 다양한 형태의 전국단위 가구조사들이 수행되고 있다. 대부분의 조사에서는 전국단위 통계뿐만 아니라 시도별 통계의 작성도 목표로 하고 있기 때문에 시도를 층으로 하는 층화추출을 사용한다. 이런 경우 주어진 표본크기를 시도별 가구수에 따라 배분하는 경우, 다시 말해 층별로 표본을 비례배분(proportional allocation) 하게 되면 가구수가 적은 시도에 배정되는 표본크기가 너무 적어 해당 시도에 대한 신뢰할 수 있는 통계 생산이 어렵다. 따라서 이런 경우 시도별 통계의 표본추출오차가 일정 수준 이상이 되도록 시도별 표본크기를 조정하는 것이 필수적이다.

참고로 우리나라 대부분의 정부승인통계의 경우 전국단위 통계에 대해서는 목표정도를 변동계수 2~3%, 시도별 통계의 경우 목표정도를 변동계수 4~5% 정도가 되도록 표본설계를 하고 있다. 하지만 관심변수가 무엇인지에 따라 어떻게 층별로 표본을 배분하는 것이 효율적인지, 또한 어떤 표본추출방법을 적용하는 것이 바람직한지 등이 달라질 수 있기 때문에 대규모 표본조사의 경우 이런 작업은 표본설계 전문가로부터 자문을 받는 것이 필수적이다. □

3) 추출확률과 가중치

표본조사에서 추출확률(selection probability)에 따른 가중치(weight)가 제대로 반영되지 않은 추정식을 사용하는 경우 편향이 발생하게 된다. 적절한 가중치를 사용하지 않음으로써 발생하는 편향은 일반적인 표본추출오차와는 다른 형식의 오차이지만 가중치의 산출과정이 표본추출과정과 관련된 것이고 추정량의 선택과 관련된 내용이기 때문에 표본추출오차와 함께 다루기로 한다. 적절한 가중치는 추출단위가 표본으로 추출될 확률을 기초로 산출되기 때문에 확률표본추출의 경우 가중치를 산출할 수 있지만 비확률추출의 경우 이런 작업이 근본적으로 불가능하다.

확률표본추출에서는 모집단을 구성하는 추출단위가 표본으로 추출된 확률을 알 수 있다고 했다. 예를 들어 10,000가구 중 500가구를 랜덤(random)하게 표본으로 추출하는 경우, 특정 가구가 표본에 추출된 확률은 500/10,000 = 1/20이다. 이런 경우 표본으로 추출된 1개 가구는 모집단에서 20가구를 대표한다고 볼 수 있으며 여기서 추출확률의 역수에 해당하는 20이라는 수를 확대승수 또는 가중치라고 부른다.

따라서 각각의 표본 관측값에 해당 단위의 추출확률의 역수에 해당하는 가중치를 곱하여 이를 모두 합치면 모집단 총계에 대한 추정결과를 얻을 수 있다. 좀더 이론적으로 설명하면 일반적으로 표본으로 추출된 특정 단위가 표본에 포함될 확률, 즉 포함확률(inclusion probability)을 π_i라고 하면 해당 단위에 대한 가중치는 포함확률의 역수, 즉 $w_i = 1/\pi_i$ 로 표현되고 총계에 대한 비편향(unbiased) 추정량인 Horvitz-Thomson(HT) 추정량은 다음과 같이 표현된다.

$$\hat{Y} = \sum w_i y_i$$

표본추출방법에 따라 추출단위가 표본으로 뽑힐 확률이 모두 같을 수도 있고, 그렇지 않은 경우도 생길 수 있다. 만약 단위별로 추출확률이 같지 않은 경우에는 추정단계에서 각 단위에 대한 포함확률, 다시 말해 가중치가 반드시 반영되어야 편향이 없는 정확한 추정을 할 수 있다. 모든 추출단위에 대해 추출확률이 동일하게 유

지되도록 추출된 표본을 자체가중표본(self-weighted sample)이라고 하며, 이런 추출방법을 등확률추출법(epsem; equal probability selection method)이라고 한다.

[예제 2-4] 모집단을 구성하고 있는 단위 중 표본으로 선택된 단위의 비율을 흔히 추출률(sampling fraction)이라고 한다. 단순확률추출의 경우 추출률(n/N)이 바로 포함확률에 해당하지만 집락추출이나 층화추출 또는 크기비례확률(probability proportional to size)추출 등 추출방법이 달라지면 포함확률도 달리 계산되어야 한다.

단순확률추출법과 같은 등확률추출법을 적용한 경우 가중치는 추출률의 역수로 표시되기 때문에 $w_i = N/n$ 이 된다. 따라서 다음과 같이 단순평균에 모집단 크기를 곱한 것이 모집단 총계에 대한 비편향 추정량이 된다.

$$\hat{Y} = \sum w_i y_i = \frac{N}{n}\sum y_i = N\ \bar{y}$$

한편, 층화확률추출의 경우 i층에서는 추출확률의 역수에 해당하는 가중치가 $w_{hi} = N_h / n_h$ 이고, 따라서 모집단 평균에 대한 비편향 추정량인 HT 추정량은 다음과 같이 표현된다. 여기서 $\sum\sum w_{hi} = N$ 이라는 점에 유의하기 바란다.

$$\overline{y_{st}} = \sum_h\sum_i w_{hi} y_{hi} \,/\, \sum_h\sum_i w_{hi}$$

위 추정식을 다시 표현하면 다음과 같다.

$$\overline{y_{st}} = \sum W_h \overline{y_h} = \frac{1}{N}\sum_{i=1}^{L} N_i \overline{y_i}$$

여기서 $W_h = N_h/N$ 은 층별 구성비를 나타냈다. □

3. 표본추출오차 관련 지표

- 표본추출률

전체 모집단을 구성하는 단위 중 표본으로 추출된 단위들의 비율 또는 백분율을 표본추출률(sampling fraction)이라고 한다. 특히 층화추출을 하는 경우 층별 목표오차를 관리하기 위해 흔히 크기가 작은 층에서 상대적으로 추출률이 높아지도록 표본을 추출하는 경우가 흔히 발생한다. 따라서 각 지역 또는 그룹별로 표본추출률이 다른 경우가 흔히 발생하며, 특히 이런 경우 층별 또는 그룹별로 추출률을 제시하는 것이 바람직하다. 표본추출률은 다음과 같이 계산한다.

$$\bullet \text{ 표본추출률} = \frac{\text{표본으로 추출된 단위의 수}}{\text{모집단 단위의 수}}$$

● 표준오차

표본추출오차를 설명하는 대표적인 지표는 추정량의 분산(variance) 또는 표준오차(standard error)이다. 표본추출오차는 추정과정의 반복적인 적용을 전제로 한 것이기 때문에 하나의 표본에 대한 조사결과를 바탕으로 이를 정확히 파악할 수는 없다. 하지만 표본추출오차를 나타내는 대표적 지표인 표준오차를 계산해 제시함으로써 이용자들에게 조사의 정확성에 대한 정보를 제시해 줄 수 있다. 표준오차는 변수에 따라 달라진다. 따라서 모든 조사변수에 대한 표준오차를 계산해 보는 것이 필수적이다. 하지만 모든 변수에 대한 표준오차를 이용자들에게 제공하는 것이 현실적으로 어려운 경우 최소한 주요 변수에 대한 표준오차는 정확히 계산해 제시하는 것이 필요하다.

표준오차는 표본추출방법에 따라 계산하는 방식이 다르고, 복합표본설계의 경우 단순하지 않기 때문에 조사자료 분석 전문 패키지를 이용하거나 표본이론 전문가의 도움을 받는 것이 바람직하다. 참고로 가장 간단한 단순확률추출(SRS)의 경우 분산과 표준오차는 다음과 같이 계산한다.

$$\bullet \text{ 분산(variance)} : \ V_{SRS}(\bar{y}) = (1 - \frac{n}{N})\frac{s^2}{n}$$

$$\bullet \text{ 표준오차(standard error)} : \ SE_{SRS}(\bar{y}) = \sqrt{V(\bar{y})}$$

● 상대표준오차

표본추출오차를 설명하기 위한 지표로 상대표준오차(relative standard error)를 나타내는 개념인 변동계수(coefficient of variation; CV)를 흔히 사용한다. 변동계수는 표준오차를 해당 추정값으로 나누어 백분율(%)로 표시한 것으로, 추정값 대비 상대적인 변동을 설명해 준다. 표본조사 전문가들은 표준오차에 비해 변동계수를 더 많이 사용한다. 평균 및 총계 추정에 대한 변동계수는 다음과 같이 구한다.

$$\bullet \text{ 변동계수}(CV) = \frac{\text{표준오차}}{\text{추정치}} \times 100\,(\%)$$

예를 들어, 단순확률추출(SRS)의 경우 변동계수는 다음과 같이 구한다.

$$\bullet \text{ 변동계수}(CV_{SRS}) = \frac{\sqrt{V(\bar{y})}}{\bar{y}} \times 100\ (\%)$$

여기서 $\bar{y} = \frac{1}{n}\sum y_i$ 이고, $V(\bar{y}) = (1 - \frac{n}{N})\frac{1}{n-1}\sum (y_i - \bar{y})^2$ 이다.

[예제 2-5] 가구소득조사를 한 결과 평균소득이 200만원인 경우 표준오차가 20만원인 경우와 평균소득이 500만원인 경우 표준오차가 20만원인 경우를 비교해 보자. 두 경우 모두 표준오차가 20만원으로 같지만, 평균소득이 200만원 수준일 때 오차가 20만원인 경우와 평균소득이 500만원 수준일 때 오차가 20만원인 경우는 평균소득 대비 상대적인 오차를 기준으로 보면 후자의 경우가 훨씬 더 정확성이 높은 조사가 이루어졌다고 볼 수 있다. 이 경우 변동계수를 구하면 각각 10%와 4%로 상대표준오차 측면에서 보면 평균소득이 200만원인 경우 20만원이란 표준오차는 상당히 오차가 크다고 볼 수 있지만, 평균소득이 500만원일 때 20만원이란 표준오차는 충분히 수용할 수 있는 수준의 표본추출오차라는 것을 알 수 있다. □

● 변동에 대한 표준오차

동일 목적을 갖는 표본조사를 매년 또는 일정 주기를 갖고 실시하거나 패널조사(panel survey)를 하는 경우 특정 변수가 시점에 따라 얼마나 변했는지를 추정하게 된다. 이와 같이 시점 간의 변화 또는 변동을 추정하는 경우 표준오차는 다음과 같이 구한다.

- 절대 변동 : $var(\widehat{Y}_2 - \widehat{Y}_1) = var(\widehat{Y}_2) + var(\widehat{Y}_1) - 2cov(\widehat{Y}_1, \widehat{Y}_2)$
- 상대 변동 : $var(\frac{\widehat{Y}_2}{\widehat{Y}_1}) \approx (\frac{\widehat{Y}_2}{\widehat{Y}_1})^2 [\frac{var(\widehat{Y}_1)}{\widehat{Y}_1^2} + \frac{var(\widehat{Y}_2)}{\widehat{Y}_2^2} - \frac{2cov(\widehat{Y}_1, \widehat{Y}_2)}{\widehat{Y}_1 \widehat{Y}_2}]$

● 설계효과

단순확률추출(SRS)과 달리 집락을 추출단위로 사용하는 집락추출(cluster sampling)을 하거나 층화(stratification)를 하는 경우 단순확률추출과 동일한 표본크기로 조사를 하더라도 표본추출오차는 달라진다. 실제 조사에 적용한 표본추출방법의 효율성을 설명하기 위해 흔히 설계효과(design effect)라는 개념을 사용한다. 주어진 표본추출법의 설계효과는 해당 추출법을 적용할 때 얻어지는 추정량의 분산을 동일한 표본크기를 갖는 단순확률추출법을 적용했을 때의 분산과 상대적으로 비교한 것이다. 설계효과는 다음과 같이 정의된다.

- 설계효과$(Deff) = \frac{\text{주어진 표본추출법에서 } Var(\widehat{Y})}{\text{동일한 표본크기의 } SRS\text{에서 } Var(\widehat{Y})}$

일반적으로 학급이나 통/반 또는 조사구 등을 추출단위로 하는 집락추출의 경우 설계효과가 1보다 커지게 되는데, 그 이유는 같은 집락내의 단위들은 보통 동질성을 갖고 있기 때문이다. 따라서 집락추출을 하는 경우 일반적으로 동일 표본크기를

갖는 단순확률추출법에 비해 효율성이 떨어진다는 점에 유의해야 한다.

● 유효표본크기

설계효과를 기반으로 주어진 표본설계의 효율성을 나타내기 위해 유효표본크기(effective sample size; n_{eff})라는 개념을 사용하기도 한다. 주어진 표본추출법으로 얻은 결과와 동일한 수준의 표본추출오차를 갖는 조사를 단순확률추출로 하는 경우 필요한 표본크기를 유효표본크기라고 한다. 유효표본크기는 주어진 표본추출법의 효율성을 나타내는 설계효과를 이용해 구할 수 있다. 실제 적용된 표본추출법에서 표본크기가 n일 때 설계효과를 $Deff$라고 하면 유효표본크기는 다음과 같이 정의된다.

• 유효표본크기 : $n_{eff} = \frac{n}{Deff}$

[예제 2-6] 통계청에서 실시하는 사교육비 조사의 경우 학교급별로 학급을 추출해 표본으로 추출된 학급내의 모든 학생들을 대상으로 사교육비 지출액을 조사한다. 이와 같이 사교육비 조사를 위해 학생을 직접 추출하는 대신 학급을 집락추출해 6,000명을 조사한 결과 설계효과가 3.0이 나왔다고 가정하자. 이 경우 학생을 직접 추출하는 방식으로 6,000명의 학생을 단순확률추출하는 경우에 비해 학급을 추출단위로 한 집락추출을 하기 때문에 분산이 3배로 증가해 정도가 많이 떨어진다는 것을 알 수 있다. 이 경우 유효표본크기를 구하면 6,000/3 = 2,000이다. 이는 학급을 집락으로 사용해 6,000명을 표본으로 추출해 조사하는 것과 학생들을 직접 추출하는 단순확률추출법으로 2,000명을 조사하는 것이 표본추출오차 측면에서 동일한 수준의 조사라는 것을 의미이다. 학급당 학생수가 30명이라고 가정할 때 이 경우 200개 학급을 추출해 6,000명을 조사하는 것과 학생을 직접 뽑아 여기저기 산재되어 있는 2,000명의 학생을 조사하는 경우 비용과 시간이란 관점에서 어떤 조사방식이 더 유리한지를 비교해 보아야 할 것이다. □

4. 표본추출오차의 관리

1) 효율적인 표본추출방법의 선택

표본추출오차는 단순히 표본크기에 의해 결정되는 것이 아니다. 표본크기가 같더

라도 어떤 표본추출방법으로 표본을 추출하는지에 따라 표본추출오차가 달라진다. 예를 들어 동질적인 단위들을 묶어 층을 구성한 후 표본을 추출하는 층화추출은 표본조사의 효율성을 높이는 데 크게 기여할 수 있다. 아울러 계통추출(systematic sampling)하는 경우에도 사전에 추출단위를 특정 속성에 따라 정렬을 한 후 계통추출함으로써 표본추출오차를 줄일 수 있다. 보조정보가 있는 경우 크기비례확률추출을 적용함으로써 동일한 표본크기를 유지하면서 효율성을 제고할 수도 있다. 따라서 주어진 비용으로 조사의 표본추출오차를 최소로 하는 표본추출방법이 무엇인지를 사전에 충분히 검토해 표본추출을 하는 것이 표본추출오차를 줄이는 방법이 될 수 있다. 주어진 조건하에서 가장 효율적인 표본추출방법을 구현하기 위해서는 표본추출이론을 충분히 이해하고 활용하는 것이 요구된다. 따라서 표본조사를 기획하는 단계에서 표본설계 전문가의 도움을 받아 어떤 표본추출법을 적용하는 것이 효율적인지 사전에 충분히 검토하는 것이 바람직하다.

한편 표본조사를 통해 구한 추정결과의 정확성을 이용자들이 파악하고 활용하게 하기 위해서는 구체적인 표본추출방법이 이용자들에게 제시되어야 한다. 조사자료를 분석하는 과정에서 표본추출오차를 산출하기 위해서는 특히 어떤 방식으로 층화추출이 이루어졌으며, 추출단위로 어떤 것들이 단계적으로 사용되었는지, 아울러 표본추출과정에 따른 가중치 산출과정이 제시되어야 올바른 표본추출오차를 파악하는 것이 가능하다. 특히 연구자들이 추가적인 분석을 할 수 있도록 마이크로 자료를 공개하는 경우, 표본추출오차의 산출 및 관련 통계분석이 이론적으로 타당한 방식으로 이루어질 수 있도록 표본추출방법과 관련된 내용이 구체적으로 제공되는 것이 필요하다.

다음 두 가지 사례를 통해 국가통계를 생산하기 위한 조사에서 표본추출오차를 비롯해 전반적인 조사오차를 줄이기 위해 어떤 방식의 표본추출법을 실제 적용하고 있는지 참고하기 바란다.

[예제 2-7] 통계청의 경제활동인구조사조사에서는 추출틀로 인구주택총조사 결과 자료를 사용하고 있으며, 우선 조사구를 추출하고 조사구 내에서 가구를 추출하는 다단계추출법을 적용하고 있다. 아울러 인구주택총조사 실시 이후 모집단 변화를 반영하기 위해 신축 아파트 명부를 보조 자료로 사용하고 있다. 통계의 공표단위를 고려해 16개 시도로 층화한 후 지역별 고용형태 및 소득·소비 구조에 따른 차이를 감안하기 위해 동부와 읍면부로 층화하여 추정의 정확성을 높이고 있다.

실업자 수에 대한 상대표준오차가 전국의 경우 1%이하, 서울과 경기도는 2%이하, 나머지 광역시와 도는 3~5%이내가 되도록 층별 표본크기를 결정한다. 층내에서 조사구 추출은 각 조사구의 주택특성, 산업특성, 실업자수, 행정구역 및 조사구

번호를 기준으로 조사구를 정렬한 후 계통추출함으로써 표본추출오차를 줄이고 있다. 표본으로 추출된 조사구에 대해서는 현지 확인을 통해 조사구 요도와 가구명부를 재작성한 후 표본가구를 선정하고 있다(통계청, 2003). □

[예제 2-8] 노동부의 고용형태별근로실태조사의 경우 매년 실시되는 통계청의 사업체기초통계조사 결과 자료를 추출틀로 사용하고 있다. 추출과정에서 표본추출오차를 줄이기 위해 동질성을 갖는 사업체들을 하나의 층으로 묶기 위해 산업대분류와 사업체규모에 따라 층을 구성하고, 층내에서 다시 비정규근로자 수를 기준으로 2개의 층을 추가로 구성하고 있다. 층별로 근로자 월평균 임금총액에 대한 변동계수가 3% 수준이 되도록 층별 사업체 표본크기를 결정한다. 아울러 층내에서 사업체를 추출하는 과정에서 사업체의 산업중분류와 사업체 소재지 코드를 기준으로 정렬한 후 계통추출함으로써 표본추출오차를 줄이는 동시에 필요에 따라 산업중분류에 따른 노동 관련 통계 생산이 가능하도록 표본을 추출했다. 한편 표본 사업체 내의 조사대상 근로자 추출은 임금대장 순서에 따라 근로자를 계통추출했으며 표본 사업체의 근로자 수에 따라 조사대상 근로자 추출률을 달리하여 표본추출오차는 거의 차이가 없도록 하면서 조사 비용과 업무량을 줄이는 표본추출법을 적용하고 있다(노동부, 2008). □

2) 표준오차 추정방법 및 결과 제시

표본조사의 정확성 수준을 참조해 활용할 수 있도록 변수별 표준오차나 변동계수 계산 결과를 이용자에게 제공하는 것이 필수적이다. 특히, 복합표본설계를 적용하는 경우 표준오차 또는 변동계수를 산출하는 과정은 간단하지 않기 때문에 선형근사식을 이용해 근사적으로 산출하기도 하고, 잭나이프 등의 재추출법(resampling method)을 이용하여 계산하기도 한다. 표준오차 산출과정의 정확성을 파악할 수 있도록 표준오차 또는 변동계수를 산출할 때 사용한 방법도 표준오차 추정결과와 함께 제시하는 것이 바람직하다.

아울러 일반적인 표본조사에서 작성되는 통계를 보면 전체 모집단에 대한 통계뿐만 아니라 영역별 통계를 산출하게 된다. 예를 들어 전체 가구를 대상으로 소득조사를 한 후 전체 가구의 평균 소득뿐만 아니라 가구주의 연령대별 평균 소득 통계를 작성하기도 한다. 이런 경우 전체 가구 평균 소득에 대한 표본추출오차와 가구주 연령대별로 구분된 영역별 평균 소득에 대한 표본추출오차는 큰 차이가 있기 때문에 영역별 주요 통계에 대해서는 해당 통계의 표준오차나 변동계수를 산출해 제공하는 것이 필요하다.

만약 표준오차 계산방법이나 각 변수별 표준오차와 같은 기술적인 내용을 일반 보고서에 수록하는 것이 현실적으로 어렵다면 관련 내용을 정리한 기술보고서(technical report)를 작성해 인터넷 등을 통해 제공함으로써 필요에 따라 이용자가 참조할 수 있도록 조치하는 것이 바람직하다.

3) 가중치 작성 및 보정

표본조사에서 편향이 없는 추정결과를 얻기 위해서는 적절한 가중치를 반영한 추정식을 사용하는 것이 필요하다. 일반적으로 조사자료의 분석을 위해 적용되는 가중치 작성과정은 다음과 같이 3단계로 나누어 정리할 수 있다. 특히 복합표본설계의 경우 아래와 같은 과정을 통해 추출확률에 의한 설계가중치를 보정한 최종 가중치를 산출해 조사자료의 분석에 사용하는 것이 일반적이다.

- 추출확률에 따른 가중치

모든 단위가 추출될 확률이 같은 등확률추출이 아닌 경우 단위별로 추출확률의 상이함에 따른 가중치 조정이 필요하다. 이런 목적으로 설정된 가중치를 흔히 설계가중치(design weight), 기초가중치(base weight) 또는 표본추출 가중치라고 한다. 설계 가중치를 w_1으로 표시하면 모집단 총계에 대한 비편향성을 만족하는 HT 추정량은 다음과 같다.

$$\hat{Y} = \sum w_{1i} y_i$$

[예제 2-9] 모집단이 도시 지역 20만 가구와 농촌 지역 80만가구로 구성되어 있는 경우, 도시에서는 4천 가구, 농촌에서는 8천 가구를 단순확률추출하면 추출확률은 도시가구의 경우 4/200=1/50, 농촌가구의 경우 8/800=1/100이다. 따라서 표본으로 선택된 한 가구가 모집단에서 도시가계의 경우 50가구를, 농가의 경우 100가구를 대표하고 있다고 볼 수 있으며, 이를 확대승수 또는 가중치라고 이해하면 된다. 이 경우 등확률추출(epsem)에 해당하지 않기 때문에 단순평균을 이용하면 편향이 발생하므로 반드시 가중치를 적용한 추정식을 사용해야 한다. □

- 무응답 가중치 조정

실제 조사를 하다 보면 그룹별로 응답률에 차이가 발생하게 된다. 예를 들어 선거예측을 위한 출구조사를 하는 경우 고연령층 투표자의 경우 다른 연령층에 비해 무응답률이 높으며, 이런 경우 무응답을 무시하고 응답 자료만을 이용하여 추정을

하게 되면 편향이 발생한다. 무응답에 따른 오차에 대해서는 다음 장에서 자세히 살펴보기로 한다.

[예제 2-10] 소득조사에서 가구주의 학력이 대졸이상인 경우 응답률이 60%이고, 고졸이하의 경우 80%라고 하면, 그룹별 응답률의 역수에 해당하는 무응답 보정 가중치(w_2)를 적용하는 것이 필요하다. 따라서 이 경우 무응답 보정 가중치는 대졸 가구의 경우 10/6, 고졸 가구의 경우 10/8이다. 평균 소득을 추정하는데 있어서 대졸 가구와 고졸 가구의 소득에 차이가 없다면 이런 무응답 보정 가중치를 적용하는 것이 큰 의미가 없다. 하지만 대졸 가구와 고졸 가구의 속성상 큰 차이가 있다면 소득뿐만 아니라 다양한 변수에 대한 추정에 있어서 무응답 가중치 보정을 통해 무응답에 따른 편향을 줄일 수 있다. □

● 사후층화 가중치 조정

조사가 완료된 후 가중치를 적용한 표본의 분포가 성별이나 연령대별 등 특정 속성에 대한 이미 알려진 모집단 분포와 일치하지 않는 경우가 흔히 발생한다. 이런 경우 특정 속성에 대한 가중 표본 분포(구성 비율)가 알려진 모집단에서의 분포(구성 비율)와 같아지도록 하는 작업을 흔히 벤치마킹(benchmarking) 가중치 보정이라고 한다. 사후적인 가중치 보정을 위해 흔히 사후층화(post-stratification) 또는 레이킹 비(raking ratio) 방법을 사용한다. 이와 관련된 구체제인 내용은 김영원 등(2006), Lohr(1999) 등을 참고하기 바란다.

[예제 2-11] 전체 모집단 10만 가구 중 아파트 거주 가구수가 4만 가구이고, 일반주택 거주 가구수가 6만 가구라고 가정하자. 실제 이런 정보는 통계청의 인구추계 자료 등을 통해 확인할 수 있는 경우가 많이 있다. 하지만 조사자료의 실제 가중치를 적용하여 표본에서 아파트 거주가구 비율을 구해보니 50%인 것으로 나타났다면, 아파트 가구 가중치와 단독주택 가구 가중치를 일부 보정해 주거형태별 구성비율이 벤치마킹 대상인 인구추계 자료와 일치하게 할 수 있다. 이 경우 아파트 가구와 단독주택 가구의 구성 비율을 알려진 대로 40:60으로 유지하기 위해서는 아파트 가구에 대해서는 사후층화 보정승수로 4/10, 표본 단독주택의 경우 6/10을 설계 가중치에 곱하는 방식으로 가중치를 보정해 주면 된다. □

Ⅲ. 추출틀 오차

1. 개 요

모집단은 표본조사를 통해 특성을 파악하고자 하는 연구대상 전체 집단을 말하며, 구체적으로 보면 관심대상이 되는 모든 기본단위들의 집합을 모집단(population)이라고 한다. 표본조사 결과는 오직 표본을 추출할 때 설정한 모집단에 대한 정보만을 제공할 수 있기 때문에, 표본조사 기획단계에서 연구목적에 따른 모집단을 명확히 정의하는 것이 필요하다. 모집단 설정에 있어서 특히 공간적 또는 시간적 개념 등을 포함한 모든 사항들이 정확하고 세밀하게 정의되어, 특정 단위가 모집단에 포함되는지 여부가 명확해야 한다.

연구목적에 따라 개념적으로 정의된 모집단을 목표모집단(target population)이라고 한다. 표본조사를 위해서는 현실적인 여건하에서 모집단을 구성하는 모든 개체가 접촉이 가능할 뿐만 아니라 실제 조사가 이루어질 수 있어야 한다. 따라서 표본추출과정에서 개념상 정의된 모집단은 현실적으로 일부 수정이 불가피하게 되는 경우가 대부분이다. 따라서 실제 표본추출을 하는 과정에서 현실적인 제약을 반영하여 구체적으로 표본추출 대상이 될 수 있는 기본단위들로 구성한 모집단을 조사모집단(survey population) 또는 조사가능모집단(accessible population)이라고 한다.

대부분의 표본조사에 있어서 개념상 규정된 목표모집단과 실제 조사대상이 되는 조사모집단은 일치하지 않으며, 조사모집단은 어떤 표본추출틀(sampling frame)을 사용하는지에 의해 결정된다. 따라서 연구목적에 의해 설정된 목표모집단과 추출틀에 의해 규정되는 조사모집단이 일치하지 않기 때문에 다양한 오차가 발생할 수

있으며 이런 오차를 추출틀 오차(frame error)라고 부른다. 추출틀 오차는 표본추출에 사용하는 추출틀이 실제 연구대상인 목표모집단을 구성하는 단위들을 얼마나 정확하게 포함하고 있는지에 의해 크게 좌우된다. 따라서 추출틀이 갖고 있는 결함에 의해 발생하는 오차는 주로 포함오차(coverage error)로 구분된다. 추출틀 오차는 추출틀이 갖고 있는 다양한 형태의 결함에 의해 발생하게 된다. 예를 들어 미포함, 과다포함, 중복등재, 접촉정보 부정확 등이 추출틀 오차의 주요한 원인에 해당한다.

추출틀 오차를 정확히 이해하기 위해서는 표본추출단위(sampling unit)와 조사단위(observational unit)를 구분해 이해할 필요가 있다. 조사단위는 표본조사에서 실제 조사대상이 되는 개체를 나타내고 추출단위는 표본을 추출하는 개체를 의미한다. 따라서 조사단위와 추출단위가 다를 수도 있다. 예를 들어 가구를 추출하고 표본 가구내의 모든 가구원을 대상으로 조사를 수행하는 경우 조사단위는 가구원이지만 추출단위는 가구이며, 사업체를 추출해 해당 사업체 내의 모든 근로자를 대상으로 임금을 조사하는 경우 조사단위는 근로자이지만 사업체가 추출단위이다.

모집단에서 실제 표본을 추출하기 위해서는 표본추출단위를 결정해야 한다. 표본추출단위는 추출단위, 표집단위라고도 부르며 각 단계별 표본추출과정에서 실제적인 표본추출 대상이 되는 개체를 말한다. 추출단위가 결정되면 표본을 추출하기 위해서는 모든 추출단위가 나열된 명부 또는 목록이 있어야 한다. 이런 명부를 표본추출틀(sampling frame) 또는 간단히 추출틀 또는 표집틀이라고 한다. 참고로 특정 표본이 추출될 확률을 계산할 수 있는 과학적인 확률추출법(probability sampling)을 구현하기 위해서는 정확한 표본추출틀을 확보하는 것이 필요하다.

많은 표본설계에서는 단계별 추출과정을 통해 최종적인 표본을 구성하게 된다. 예를 들어 고등학생 1학년을 대상으로 표본을 추출해 학력평가를 실시한다고 하면, 추출단계는 '고등학교(PSU: Primary Sampling Unit) → 학급 → 학생'으로 이루어지며, 이 경우 '고등학교', '학급', '학생'이 각 단계별로 추출단위가 된다. 이와 같은 추출방법을 다단계추출(multi-stage sampling)이라고 한다. 다단계 추출에서는 추출단계별로 추출단위가 다르기 때문에 별개의 추출틀을 사용하게 된다. 이 경우 1단계에서의 추출틀은 전체 고등학교 명부, 2단계에서는 표본으로 추출된 고등학교의 1학년 학급 명부, 마지막 단계에서는 표본으로 추출된 학급의 학생 명부가 단계별 추출틀이 된다.

조사시점을 기준으로 모든 추출단위들에 대한 정보가 완벽하게 반영된 추출틀을 확보하는 것은 현실적으로 거의 불가능하다. 실제 표본조사에서 연구주제에 따라 적절한 표본추출틀을 확보하는 작업은 쉽지 않으며, 많은 표본설계에 있어서 구체적으로 확보할 수 있는 추출틀이 무엇인가에 따라 구체적인 추출단위와 표본추출

방법을 결정한다.

국가통계의 정확성을 제고하기 위해서는 국가차원에서 모집단을 관리할 필요가 있다. 이런 측면을 고려해 통계청에서는 조사관리국에 '모집단관리팀'을 최근 신설하여 국가통계 전반에 활용할 수 있는 가구 모집단을 비롯해 사업체 및 기업체 모집단, 농어업 가구 모집단, 경지 모집단 등 국가통계 생산을 위해 필요한 모집단 관리 업무를 총괄하고 있다. 이와 같이 통계청에서도 추출틀 오차 관리의 중요성을 인식하고 추출틀 결함에 의해 발생하는 오차를 줄이기 위해 노력하고 있다.

2. 추출틀 오차의 이해

표본조사를 위한 추출틀을 마련하는 과정에서 어떤 결함이나 비효율적인 요소가 발생하면 이는 조사의 정확성에 큰 영향을 미치게 된다. 조사설계의 다른 과정과 마찬가지로 추출틀을 결정할 때에도 추출틀을 작성하는 데 드는 비용과 특정 추출틀을 사용하는 경우 활용할 수 있는 표본추출방법과 이에 따른 추정방법의 효율성 등을 모두 종합적으로 고려해야 한다.

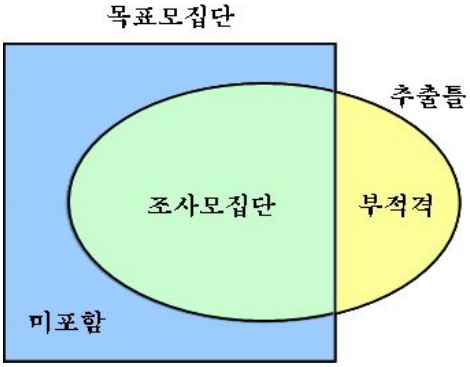
어떤 추출틀을 사용하는 것이 다른 추출틀을 사용하는 것에 비해 더 효율적인가 하는 문제는 추출틀이 갖고 있는 결함뿐만 아니라 각 추출틀을 사용하는 경우 구현 가능한 표본추출방법의 효율성과도 매우 밀접한 관련을 갖는다. 일반적으로 어떤 조사를 위해 활용 가능한 추출틀의 종류가 매우 제한되어 있기는 하지만, 동일한 형식의 추출틀을 갖고 있더라도 어떤 추가적인 보조정보가 추출틀에 담겨 있는지에 따라 추출틀의 효율성은 달라질 수 있다.

예를 들어 사업체 명부를 추출틀로 사용한다면, 명부상에 단지 사업체명과 주소와 연락처 등만 갖고 있는 경우와 여기에 추가해 각 사업체의 산업분류와 종업원 규모 등에 대한 보조정보를 추가적으로 갖는 경우 실제 적용할 수 있는 표본추출방법이 크게 달라질 수 있다. 만약 산업분류와 사업체 규모에 대한 정보를 갖고 있는 경우 효과적인 층화추출을 함으로써 단순확률추출법으로 사업체를 추출하는 것에 비해 표본추출오차를 줄여 추정의 정확성을 대폭 향상시킬 수 있다. 경우에 따라서 사업체 규모를 이용한 크기비례확률추출법(probability proportional to size sampling) 등의 보다 효율적인 추출방법을 적용할 수도 있기 때문에 추출틀에 어떤 정보가 담겨 있는지 여부가 조사의 정확성에 많은 영향을 줄 수 있다. 물론 이런 추가적인 정보가 담긴 추출틀을 확보하는 데 드는 비용도 추출틀을 선택하는 과정에서 고려되어야 할 것이며 결국 주어진 자원을 효율적으로 배분하는 문제가 추출틀의 선택에 있어서도 적용된다.

무엇보다 추출틀을 선택하고 작성하는 과정에서 중요하게 고려해야 할 사항은 실제로 표본설계에 사용하는 추출틀에 결함이 있는 경우 결과적으로 편향된 추정결과를 얻을 가능성이 매우 높다는 점에 유의해야 한다. 따라서 조사를 설계함에 있어서 어떤 추출틀을 사용할 것인지 결정하는 문제는 표본추출뿐만 아니라 조사수행과정 및 통계 작성과정 전반에 걸쳐 큰 영향을 끼치는 매우 중요한 의사결정 과정이라는 점을 간과해서는 안 된다.

예를 들어 2008년도 현재 우리나라 가구의 소비행태를 파악하기 위해서 전국의 가구를 대상으로 표본을 추출한다고 생각해 보자. 이 경우 전국의 가구에 대한 추출틀이 필요하고 이를 위해 2005년도 인구주택총조사에서 조사된 가구 명부를 추출틀로 사용한다고 하면, 실제 총조사가 실시된 후 3년이 경과했기 때문에 일부 가구의 경우 이미 재개발 등을 통해 가구가 사라졌거나, 이사 등을 통해 가구주가 변경되었을 가능성이 있으며, 특히 3년 동안 신축된 아파트는 추출틀상에 누락되는 문제가 있다. 이와 같은 결함을 갖는 추출틀을 사용하여 표본을 추출하는 경우 목표모집단과 추출틀이 일치하지 않기 때문에 결과적으로 편향이 발생할 여지가 있다. 따라서 조사를 설계하는 과정에서는 추출틀이 어떤 결함을 갖고 있고, 이런 결함이 추정결과에 어떤 영향을 미치게 되는지 항상 사전에 충분히 검토하는 과정을 갖는 것이 필요하다.

일반적으로 조사설계에서 추출틀의 결함으로 인해 생기는 편향이나 조사의 효율성 저하 문제를 살펴보기 위해 추출틀이 갖게 되는 문제점을 몇 가지 유형으로 구분하여 살펴봄으로써 추출틀에 의해서 발생하는 오차를 이해할 수 있을 것이다. 일반적인 표본조사에서 목표모집단과 추출틀의 관계는 [그림 3-1]과 같이 나타낼 수 있다.



[그림 3-1] 목표모집단과 추출틀의 관계

1) 모집단 구성원의 누락

목표 모집단의 구성원이 추출틀에 누락되는 문제가 추출틀 결함 중 가장 빈번하게 발생하면서 가장 문제가 심각한 오차 발생원인이라고 볼 수 있다. 특히 추출틀이나 표본을 가지고서는 이런 결함을 감지할 수 없다는 점에 유의해야 한다. 모집단 구성원이 누락되었는데도 이를 파악하지 못한다면 필연적으로 조사 결과는 편향될 수밖에 없다. 총계를 추정하는 경우라면 과소추정(under estimation)될 것이며, 다른 모수를 추정하는 경우에도 그에 따른 편향이 생긴다.

이러한 종류의 오차를 설명하기 위해 미포함(under-coverage), 불포함(non-coverage), 불완전포함(incomplete coverage) 등 다양한 용어들이 사용되고 있다. 이런 용어들은 원래 센서스(census)에서 집계의 품질이나 정확성을 나타내는 목적으로 사용되었다. 센서스에서 누락된 조사대상이 없이 자료가 완벽하게 수집되었는지를 점검하는 과정이나, 추출틀에 누락된 모집단 구성원을 검토하는 과정은 서로 비슷한 것으로 볼 수 있다. 추출틀은 실제 조사대상이 되는 추출단위들(가구, 사업체, 조사구 등)을 명부로 작성하는 것이 보통이다. 조사단위와 추출단위가 다르지만 사용하는 추출틀에 명시된 추출단위들을 이용하면 모집단을 구성하는 모든 조사단위들이 특정 추출단위와 연계되어 최종 표본으로 추출될 수 있어야 한다. 이와 같이 목표모집단을 구성하는 모든 조사단위가 추출틀을 통해 표본으로 추출이 가능한 경우 모집단 구성원의 누락에 의한 오차가 없다고 볼 수 있다.

실제 표본조사를 하는 경우 연구목적에 따라 설정된 모집단 구성원 중 일부가 추출틀에 누락되는 경우가 흔히 발생하기 때문에 모집단을 일컫는 용어도 앞에서 언급한 것과 같이 이론적으로 설정된 목표모집단과 실제 표본추출 과정에서 사용하는 추출틀에 의해 정의되는 조사모집단으로 구분해 사용한다는 점에 유의하기 바란다.

[예제 3-1] 우리나라 15세 이상 성인들이 새로운 정부정책에 대한 찬성비율을 알아보기 위해 전화조사를 실시한다고 하자. 현재 우리나라 대부분의 조사회사에서는 전화조사를 위한 추출틀로 한국통신 등에서 제공하는 일반가구 전화번호 CD를 사용하고 있다. 이 경우 일반전화를 보유하고 있지 성인은 물론이고 일반전화를 보유하고 있더라고 현재 제공되는 전화번호 CD에 전화번호가 수록되지 않은 가구에 거주하는 15세 이상 성인들은 추출틀에 누락되게 된다. 특히 휴대폰 사용이 최근 급증함에 따라 청년층 가구의 경우 일반전화를 보유하지 않는 경향이 높기 때문에 이들 일반전화를 보유하지 않는 청년층은 추출틀에 포함되지 않기 때문에 오차가

발생하게 된다. 이런 유형의 오차가 모집단 구성원이 추출틀에 누락됨으로써 발생하는 미포함오차에 해당한다.

참고로 허명회 등(2008)에 의하면 우리나라의 경우 일반 가구용 전화번호 CD에 전화번호가 수록된 가구의 비율은 전체 가구 중 60% 수준인 것으로 알려져 있다. 따라서 전화번호부(CD)에 의존하는 전화조사를 하는 경우 전체 가구 중 40% 정도가 추출틀에서 누락되는 문제가 발생한다. 이에 따라 미국이나 영국 등 좀 더 과학적인 전화조사가 이루어지는 국가에서는 전화번호부를 추출틀로 사용하는 대신 무작위전화걸기(RDD; random digit dialing) 기법을 사용한 전화조사가 일반적으로 사용되고 있다. □

[예제 3-2] 어업생산통계조사의 경우 1차 추출단위(primary sampling unit; PSU)로 통계청에서 5년 간격으로 실시하는 어업총조사 결과를 기초로 구성한 어업조사구를 사용하고 있다. 이 경우 실제 어업생산 관련 조사대상이 되는 조사단위는 어가들이지만 1차 추출단위가 어업조사구이기 때문에 추출틀로 전체 어업조사구 명부를 사용하게 된다. 2005년 어업총조사의 경우 전국 어업조사구수는 3,340개이고 이들 조사구 내에 전체 어가 79,942가구를 포함하게 된다. 하지만 어가에 대한 이들 어업조사구 중에는 일부 내륙지역에 위치한 어업조사구가 존재하는 경우도 있고, 실제 농업을 위주로 하는 지역이지만 극히 일부 어업을 겸하는 어가들이 있는 지역도 있기 때문에 실제 표본추출단계에서는 조사의 편의를 위해 이들 조사구 중 어가수가 10개 이하인 조사구 148개는 추출대상에서 제외했다. 이 경우 목표모집단은 3,340개 조사구로 구성되지만 실제 표본추출 대상이 되는 조사모집단은 148개 조사구를 제외한 3,192개 조사구로 구성했다. 이 경우 어가수 기준으로 보면 조사모집단에서 제외되는 어가는 1,162가구로 전체 어가의 1.45%에 불과하기 때문에 최종 추정결과에 미치는 영향은 무시할 수준이다. 참고로 148개 조사구가 제외됨으로써 어업생산통계조사 결과에 미치는 영향을 사전에 충분히 분석할 필요가 있다(해양수산부, 2007). □

[예제 3-3] 통계청뿐만 아니라 정부부처나 연구기관에서 수행하는 가구 면접조사의 경우 대부분 인구주택총조사 조사구를 1차 추출단위(PSU)로 사용하고 있다. 인구주택총조사는 매 5년을 주기로 실시되고 있으며 가장 최근에는 2005년에 실시되었다. 2005년 인구주택총조사의 경우 조사대상을 “2005년 11월 1일 0시 현재 우리나라에 거주하는 모든 사람과 이들이 살고 있는 주택이 조사대상입니다”라고 정의하고 있다(통계청, 2005). 따라서 예를 들어 2008년도에 인구주택총조사 조사구 명부를 추출틀로 사용한 가구표본을 추출하는 경우 2005년 11월 이후에 신축

된 아파트 지역은 추출틀에 누락되고 이에 따라서 미포함오차가 발생하게 된다. 아울러 조사의 편의를 위해 도서지역의 조사구를 흔히 추출대상에서 제외하는 경우(제주도를 조사대상에서 제외하기도 함)가 많기 때문에 이에 따라 추가적으로 미포함오차가 발생하게 된다. 이런 조사의 경우 도서지역 또는 신축아파트 단지가 조사대상에서 누락됨에 따라 발생하는 오차를 면밀히 분석할 필요가 있으며 이런 미포함오차 때문에 발생하는 편향이 무시할 수 없는 수준이라면 이에 대한 대책을 강구할 필요가 있다. □

2) 모집단 구성원이 아닌 것을 포함

목표모집단의 구성원이 아닌 조사단위들이 추출틀에 포함되는 것을 과다포함(over-coverage)이라고 한다. 이런 과다포함 오차를 제거하지 않는다면 총계는 과대추정이 될 가능성이 높으며, 그 밖의 다른 통계량의 값에도 편향이 생기게 될 것이다. 하지만 목표모집단에 해당하지 않는 단위들을 조사과정에서 구별하여 제거할 수 있어 조사과정이 적절하게 관리되면 미포함인 경우에 비해 편향이 심각하게 문제되지는 않을 수 있다. 과다포함의 경우 목표 모집단은 추출틀로 정의된 조사단위들의 부분집합의 형태가 되므로 조사과정에서 필요한 정보를 얻는 경우 비편향 추정도 가능해진다. 목표모집단 구성원이 아닌 단위를 구분할 수만 있으면 과다포함 추출틀로 인해서 생기는 오차는 단지 추정의 효율이 좀 떨어지는 것으로 볼 수도 있다.

추출틀이 최근 업데이트가 되지 않은 경우 빈 집과 같이 명백하게 목표 구성원이 아닌 단위들이 추출틀에 기재되어 있을 수 있다. 또 어떤 경우에는 목표 구성원은 아니지만 그와 유사한 단위가 추출단위로 기록될 수도 있다. 예를 들어 우리나라 성인들을 대상으로 하는 가구조사에서 추출틀에 올라 있는 가구들 중에는 조사대상이 아닌 외국인이 거주하는 가구가 존재할 수 있다. 아울러 가구조사를 위한 추출단위로 가장 흔히 사용되는 인구주택총조사 조사구를 사용하는 경우 총조사는 5년을 주기로 이루어지기 때문에 총조사 후 일정 기간이 경과한 후에는 조사 시점에 존재했던 조사구 전체가 재개발 등으로 없어질 수가 있지만 조사구 명부상에는 이런 조사구가 그대로 남아 있을 수 있다. 이런 경우에 역시 과대포함 문제가 발생하게 된다.

또한 인구주택조사구는 지형지물을 이용해 우리나라 전체 국토를 지도상에서 일정 구획으로 구분하는 형식으로 만들어진다. 따라서 조사구를 추출단위로 사용하는 것은 지도상에서 지역을 구분하여 추출단위로 사용하는 구역추출법(area sampling)을 사용하는 것이고, 이 경우 일반적으로 표본으로 추출된 조사구(구획)

내의 가구 중 일부를 2차 추출단계에서 표본으로 추출하게 된다. 이런 구역추출법을 적용할 때 조사구별 경계가 명확하지 않으면, 실제 모집단 구성원이 표본 조사구의 경계 바깥에 위치하는 경우에도 경계 내에 있는 것으로 간주하게 되어 과대포함 또는 중복포함 문제가 발생할 수 있다. 실제 센서스를 실시할 때에도 이런 유형의 오차가 생기기 않도록 조사구 경계를 정교하게 설정하는 것이 필요하며, 이런 오류는 센서스 조사구를 이용한 다른 표본조사에도 영향을 주게 된다. 어떤 조사에서는 미포함과 과다포함이 함께 일어나게 되어 그 효과가 서로 상쇄될 수도 있다. 예를 들어 센서스에서 어떤 사람들은 실수로 누락되는 반면 어떤 사람들은 중복집계되면 총 인구수에 있어서는 큰 차이가 발생하지 않을 수도 있다.

일반적인 추출틀의 과대포함과는 차이가 있지만 추출틀을 통해 접촉이 가능한 조사단위 중 특정 조건을 만족하는 단위들만을 조사대상으로 하는 경우가 있다. 이런 경우 실제 연구목적에 따른 조사대상 단위가 아닌 부적격 단위들이 추출틀에 상당 부분 포함되어 있기 때문에 조사대상이 되는 조사적격 단위를 조사과정에서 구분해 내고 실제 추정에 필요한 정보를 얻은 것이 필요하다. 이 경우 일반적인 과대포함 문제와는 약간 성격이 다르지만 실사과정에서 엄격한 규칙을 적용하여 조사적격 여부를 판정하고, 특히 추정과정에서 추출틀의 과다포함으로 인해 발생할 가능성이 높은 편향을 방지하기 위해 실사과정에서 모집단 중 조사적격 단위의 비율이 얼마인지 구체적으로 파악하여 이를 가중치 작성과정에 반영하는 것이 필수적이다. 특히 실사과정에서 표본으로 추출된 단위를 대상으로 조사적격 여부를 관측해야 하는 경우 현장에서 해당 조사단위와 접촉을 하지 못할 때는 조사적격 여부를 확인할 수 없기 때문에 이에 따른 조사상의 어려움도 고려해야 한다.

[예제 3-4] 2008년도에 농가경제조사를 위한 농가 표본추출을 위해 2005년 농업총조사에서 구성된 농업조사구를 1차추출단위로 사용하는 경우, 일부 조사구의 경우 재개발 등으로 인해 농가가 더 이상 존재하지 않는 다시 말해 모집단 구성원이 아닌 조사구가 추출틀에 포함되어 이에 따른 오차가 발생할 수 있다. 또한 2차 표본추출 단계에서 표본으로 추출된 농업조사구내에 거주하는 농가 중 일부 농가를 표본으로 추출하는 경우에도 2005년도에 사용했던 조사구내 농가 명부를 추출틀로 사용하면 명부상의 농가 중 일부 농가는 이사 등으로 인해 실제 존재하지 않을 수도 있고, 일부 농가는 더 이상 농업에 종사하지 않기 때문에 농가에 해당하지 않는 경우가 발생하게 된다. 따라서 이런 경우 표본으로 추출된 조사구에 대해서는 조사시점 현재 해당 조사구에 실제 거주하는 농가에 대한 명부 작성 작업을 다시 수행함으로써 추출틀의 과대포함 또는 미포함에 따른 편향을 줄일 수 있다. □

[예제 3-5] 노동연구원의 고령화연구패널(KLoSA)의 경우 우리나라의 고령화 과정을 파악하기 위해 45세 이상 중고령자와 이들이 거주하는 가구를 조사대상으로 하고 있다. 하지만 우리나라 전체 가구 중 45세 이상 중고령자가 거주하는 가구명부를 추출틀로 확보하는 것은 현실적으로 불가능하기 때문에 인구주택총조사 조사구를 1차추출단위로 하고 표본으로 추출된 조사구에서 다시 2차적으로 일부 중고령자 가구를 표본으로 추출했다(한국노동연구원, 2007).

이런 경우 45세 이상 중고령자가 거주하는 가구가 조사적격 가구에 해당하고 이에 해당하지 않는 가구는 모집단 구성원에 해당하지 않지만 추출틀에 포함된 것으로 볼 수 있다. 하지만 전체 명부에서 조사 부적격 가구를 사전에 확인해서 제거하는 것은 현실적으로 불가능하기 때문에 조사 부적격 가구가 포함된 추출틀을 사용하고 있다. 실사 과정에서는 표본 가구의 조사적격 여부를 확인해 최종 표본을 구성하게 된다. 이런 과정을 거쳐 최종 표본을 구성하는 경우, 추출틀에 포함된 가구 중 조사적격 가구의 비율을 파악할 수 있는 정보를 확보해야 추정단계에서 가중치를 조정함으로써 과대포함된 추출틀을 사용함에 따른 편향을 제거할 수 있다. 이를 위해서는 최소한 표본으로 추출된 조사구 내의 가구 중 조사적격 가구의 비율을 파악할 수 있도록 실사가 진행되는 것이 필수적이다. □

3) 추출단위의 중복

추출틀상에 있는 두 개 이상의 추출단위가 목표모집단의 동일한 단위와 대응이 되는 경우 추출틀의 결함에 의해 오차가 발생할 수 있다. 다시 말해 추출틀에 중복된 단위가 존재하는 경우로 목표 모집단을 구성하는 개체 중 일부가 추출틀상의 두 개 이상의 추출단위에 연결되는 경우를 말한다. 이와 같은 경우 과대포함과 유사한 형태의 오차가 발생할 수 있다. 물론 이런 문제를 명확하게 해결하는 방법은 추출틀에서 중복을 확인해서 표본추출을 하기 전에 추출틀을 수정하는 것이다. 중복문제가 있을 때 각 모집단 구성 개체들이 표본에 추출될 확률을 정확하게 알려면 각 개체들이 몇 개의 추출단위와 연결되어 있는지 파악해야 한다. 중복 문제를 다루는 문헌들에서는 이런 경우를 중복 기재(duplicate listings, erroneous inclusions)라는 용어로 설명하고 있다. 추출틀에서 특정 단위가 중복된 횟수를 중복수(multiplicity)라고 하며, 만약 표본에 있는 단위가 추출틀에 등재된 횟수(중복수)를 조사과정에서 명확하게 파악할 수 있다면 중복수를 감안한 추정식을 사용함으로써 추출틀상에서 발생하는 중복에 따른 편향을 제거할 수 있다.

특히 적절한 추출틀을 확보하기 어려울 때 미포함 오차를 줄이기 위해 2개 이상의 추출틀을 결합해 사용하는 경우, 다시 말해 다중추출틀(multiple frame)을 이용

해 표본을 추출하는 경우 중복 문제가 흔히 발생한다. 두 개의 추출틀을 사용하는 경우 만약 조사과정에서 어떤 단위들이 두 개의 추출틀에 포함되었는지를 확인할 수 있으면 이를 기초로 표본으로 추출된 단위에 대한 추출확률을 계산하여 가중치에 반영함으로써 중복에 따른 편향을 제거할 수 있다.

[예제 3-6] 가구조사를 위해 가구용 전화번호부를 이용해 가구표본을 추출하는 경우 어떤 가구는 부부 명의로 각자 전화번호를 보유할 수 있기 때문에 두 개 이상의 일반 전화회선을 사용하는 가구가 존재할 수 있다. 이런 경우 추출틀에 동일 가구가 여러 번 중복 기재된 상황이기 때문에 해당 가구에 대한 추출확률은 다른 가구에 비해 높아지게 된다. 만약 전화조사 과정에서 해당 가구에서 사용하는 일반 전화번호 회선수(이를 중복수라고 함)를 파악하면 추정과정에서 해당 가구에 대한 추출확률을 기초로 한 가중치를 사용함으로써 편향을 제거할 수 있다. □

[예제 3-7] 지난 1년간 전국 종합병원에 내원한 환자들의 입원일수, 진료과목, 내원일수 등을 조사하기 위해 의료기관 이용실태조사를 하는 경우를 고려해 보자. 표본추출을 위해 환자를 추출단위로 하는 경우 지난 1년간 전국 종합병원을 이용한 모든 환자의 명부를 작성해야 하기 때문에 엄청난 시간과 비용이 든다. 따라서 환자 명부를 작성해 추출틀로 사용하는 것은 현실적으로 불가능하다. 이런 경우 먼저 종합병원들을 표본으로 추출하고 표본으로 선택된 병원 이용 환자들의 명부에서 일부 환자를 표본으로 추출하는 2단계 추출법을 적용하는 것이 효과적일 것이다. 따라서 첫 번째 단계의 추출을 위해서는 전국 종합병원 명부가 추출틀로 사용될 것이고, 두 번째 단계 추출을 위해서 표본으로 추출된 종합병원의 환자 명부가 추출틀로 사용된다.

이런 경우 동일 병원이 중복되는 경우가 없고, 특정 병원의 환자 명부에서도 중복이 없도록 추출틀을 완벽하게 정비하더라도 결과적으로 표본추출 과정에서 중복문제가 발생하게 된다. 왜냐하면 어떤 환자들은 지난 1년간 하나의 종합병원에서만 치료를 받은 것이 아니라 다수의 병원에서 치료를 받을 수 있기 때문이다. 결국 동일 환자가 다수의 종합병원에서 표본으로 추출될 수 있기 때문에 이는 추출틀상에서 중복 문제가 발생한 것으로 볼 수 있다. 이런 표본조사에서 편향이 없는 추정결과를 얻기 위해서는 표본으로 추출된 환자가 지난 1년 동안 몇 개의 병원에서 진료를 받았는지를 나타내는 중복수를 파악하고, 중복수를 반영한 추정방법을 사용해야 편향을 제거할 수 있다. □

[예제 3-8] 서울에 개업 중인 음식점들을 대상으로 표본을 추출해 전화조사를

실시한다고 가정하자. 상당 부분의 음식점 전화번호는 업종별 전화번호부에 기재되어 있기 때문에 전화번호부에서 음식점 전화번호 명부를 작성해 추출틀로 사용할 수 있을 것이다. 하지만 최근 개업한 음식점 등 실제 개업 중인 음식점 중 업종별 전화번호부에 등재되어 있지 않은 음식점들이 많이 있을 수 있기 때문에 미포함 오차를 줄이기 위해 업종별 전화번호부와 함께 음식점 협회에서 보유하고 있은 회원 명부를 추출틀로 함께 사용할 수 있을 것이다.

이런 경우 많은 음식점은 전화번호부와 협회 명부에 중복 수록되어 있을 것이고, 두 개의 명부를 결합해 음식점 전화번호를 중복이 없도록 사전에 충분히 정비하지 않는다면 추출틀 중복에 따른 문제가 발생하게 된다. 현실적으로 많은 시간과 노력이 필요하기 때문에 사전에 추출틀을 정비하는 작업이 어렵다면, 두 개의 명부를 결합해 추출틀로 사용하는 대신 조사과정에서 표본으로 추출된 음식점에 대해 두 가지 명부에 중복 등재되었는지 여부를 파악하고 이를 추정과정에 반영함으로써 추출틀 중복에 따른 편향을 제거해야 한다. □

4) 부정확한 정보

추출틀은 단순히 모집단을 구성하는 조사단위들 또는 추출단위들의 목록에 그치는 것이 아니라 보다 더 많은 정보를 포함해야 한다. 무엇보다 추출틀에는 해당 단위를 실제로 접촉할 수 있는 정보가 포함되어 있어야 한다. 예를 들어 추출틀로 사업체 명부를 사용하는 경우 추출틀에서 사업체를 추출한 후에 해당 사업체를 접촉할 수 있는 정확한 정보가 있어야 조사가 가능해 질 수 있다. 이런 경우 흔히 사업체의 주소나 전화번호 등이 이런 정보에 해당한다. 마찬가지로 가구 명부를 추출틀로 사용하는 경우 표본으로 추출된 가구를 명확하게 접촉할 수 있는 정보를 추출틀에서 얻을 수 있어야 한다. 예를 들어 추출틀에 가구의 주소만 기록되어 있는 경우 실제 해당 주소(거처)에 여러 가구가 거주하고 있을 수 있기 때문에 표본으로 추출된 가구인지 여부를 확인할 수 있도록 가구주 이름 등이 추출틀상에 명시되는 것이 필요하다. 또한 미포함이나 과다포함 없이 모집단 구성원을 완벽하게 포함하고 있는 추출틀을 갖고 있더라도 이 추출틀이 낡은 것이라 표본으로 추출된 구성원을 접촉할 수 있는 정보가 정확하지 않다면 이런 명부는 추출틀로 사용하기에 부적합할 수 있다.

또한 추출틀에 어떤 정보가 담겨 있는지에 따라 적용할 수 있는 표본추출법이 달라질 수 있다. 추출틀에 얼마나 많은 보조정보(auxiliary information)가 포함되어 있는지에 따라 얼마나 정확성이 높은 표본조사가 가능한지가 결정된다. 예를 들어 추출단위들을 동질적인 것들로 묶을 수 있는 보조정보가 주어진다면 층화추출

을 통해 보다 효율적인 표본설계가 가능할 수 있고, 관심변수와 상관관계가 높은 보조정보가 추출틀에 포함되어 있다면 크기비례확률추출법을 적용하거나 추정과정에서 보조정보를 활용한 비추정(ratio estimation)이나 회귀추정(regression estimation) 등의 기법을 사용함으로써 조사의 정확성을 대폭 향상시킬 수도 있다. 표본조사에서 어떤 정보가 담긴 추출틀을 사용하는지에 따라 표본조사 과정 전체가 영향을 받게 되고 조사의 정확성이나 조사비용 등이 달라질 수 있을 정도로 추출틀은 조사과정 전반에 걸쳐 많은 영향을 주게 된다.

하지만 이런 보조정보가 정확하지 않다면 오히려 추정의 정확성을 떨어뜨릴 수도 있다는 점에 유의해야 한다. 예를 들어 사업체 표본을 추출하기 위해 산업분류를 층화변수로 사용하는데 일부 사업체의 경우 산업분류에 오류가 있다면 이로 인해 오차가 발생할 수 있다. 이와 같은 추출틀에서 발생하는 분류오류(classification error) 때문에 결과적으로 특정 범주에 대해서는 과대 또는 과소 추정하는 편향이 발생할 수 있다.

[예제 3-9] 한국고용정보원에서는 대학졸업자의 경력개발 및 직업(직장)이동 경로를 추적 조사하여 교육-노동시장간 신뢰성 있는 인력수급정보를 제공하고, 인력수급불일치를 완화할 수 있는 정책수립을 위한 기초자료를 수집하기 위해 '대졸자 직업이동 경로조사(GOMS; Graduates Occupational Mobility Survey)'를 실시하고 있다. 2006년 10월에서 12월에 실시된 조사의 경우 2004년 9월 및 2005년 2월 전문대, 교육대학, 4년제 대학 졸업자를 모집단으로 하고 있으며, 이 조사에서는 표본을 추출하기 위해서 각 대학에서 제공하는 해당 년도 졸업생 명부를 기초로 작성한 학교급, 지역, 전공 및 성별 등의 정보가 담긴 대학 졸업생 명부를 추출틀로 사용하고 있다(한국고용정보원, 2008).

이 경우 많은 대학에서 제공하는 졸업생 명부에는 재학 당시의 주소와 전화번호가 거의 누락 없이 기록되어 있다(일부 대학의 경우 졸업 후 연락처가 업데이트 된 경우도 있음). 하지만 대학 재학 당시의 거처나 전화번호가 졸업 후 변경되는 경우가 빈번하게 발생하기 때문에 실제 이 추출틀에서 표본을 추출하더라도 해당 졸업생을 접촉하는데 많은 어려움을 갖게 된다. 따라서 이런 추출틀의 경우 미포함 또는 과대포함 등의 문제는 없지만 표본으로 추출된 조사대상자를 접촉하는데 필요한 정확한 정보를 제공하지 못한다는 측면에서 일부 결함을 갖고 있다. □

[예제 3-10] 노동부에서 사업체를 대상으로 실시하는 사업체임금근로시간조사, 인력수요동향조사, 고용형태별근로실태조사 등은 매년 통계청에서 전국의 모든 사업체를 전수조사하는 사업체기초통계조사에서 얻은 사업체 명부를 추출틀로 사용

하고 있다. 이 추출틀에는 사업체의 주소 및 연락처를 비롯해 산업분류, 종사자 수 등의 보조정보가 포함되어 있다. 따라서 노동부의 사업체 대상 표본설계에서는 표본의 효율성을 높이기 위해 산업대분류 또는 산업중분류와 상용종사자 수를 기준으로 한 사업체 규모 등을 이용한 층화추출을 하고 있다. 아울러 필요에 따라서 사업체를 종사자 수에 비례하는 확률로 추출하는 크기비례확률추출법의 적용도 가능하다. 이와 같이 추출틀에 유용한 보조정보가 포함되어 있는 경우 보다 효율적인 표본설계를 통해 조사의 정확성을 제고할 수 있다. □

3. 추출틀 오차 관련 지표

추출틀이 갖고 있는 오류 수준을 정량화하기 위해서 다음과 같은 지표들을 고려할 수 있다. 제시된 지표를 산출하기 위해서는 목표모집단 현황을 구체적으로 파악하기 위한 추가적인 시간과 비용을 필요로 하기 때문에 실제로 이런 지표를 산출하는 것은 현실적으로 매우 어렵다는 점에 유의하기 바란다.

- 미포함에 의한 누락 비율

$$\frac{\text{추출틀에 누락된 단위의 수}}{\text{목표모집단 전체 단위의 수}}$$

- 중복비율

$$\frac{\text{추출틀에 중복 기재된 단위의 수}}{\text{추출틀에 등재된 전체 단위의 수}}$$

- 부적격 비율

$$\frac{\text{추출틀에 등재된 부적격 단위의 수}}{\text{추출틀에 등재된 전체 단위의 수}}$$

- 오분류(misclassification) 비율

$$\frac{\text{추출틀에 등재된 적격 단위 중 오분류된 단위의 수}}{\text{추출틀에 등재된 적격 단위의 수}}$$

- 미포함에 의한 편향

추출틀에 미포함 오차가 있을 때, 표본에서 구한 모집단 평균에 대한 비편향 추

정량 $\overline{y_a}$ 을 사용하는 경우 발생하는 편향은 다음과 같다.

$$bias(\overline{y_a}) = W_0(\overline{Y_a} - \overline{Y_0})$$

여기서 W_0는 미포함에 의한 누락 비율, $\overline{Y_a}$ 는 추출틀에 등재된 단위들의 모집단 평균, $\overline{Y_0}$ 추출틀에 누락된 단위들의 모집단 평균이다. 따라서 누락비율이 증가할수록 미포함에 의한 편향이 커지게 되고, 추출틀에 등재된 단위들의 평균과 누락된 단위들의 평균의 차이가 크면 미포함에 의한 편향도 커지게 된다. 한편 목표모집단 평균에 대한 상대편향(relative bias)는 다음과 같다.

$$RB(\overline{y_a}) = \frac{W_0(\overline{Y_a} - \overline{Y_0})}{\overline{Y}}$$

여기서 $\overline{Y}$ 는 목표모집단 전체 평균을 나타낸다.

4. 추출틀 오차에 대한 대책

표본추출틀로 인해 발생하는 오차는 대부분 변동오차를 증가시킬 뿐만 아니라 심각한 편향을 발생시킬 가능성이 있기 때문에 조사설계 단계부터 추출틀이 갖고 있는 결함을 줄일 수 있는 대책을 강구하는 것이 필요하다. 추출틀에 의한 오차 중 가장 먼저 고려해야 할 것은 포함오차이다. 포함오차를 줄이거나 제거하는 가장 좋은 방법은 아마도 과다 포함된 단위들을 제거하고 미포함된 단위들을 추가하여 추출틀을 완벽하게 보완하는 것이다. 추출틀을 보완하는 것은 현실적으로 시간과 비용이 많이 들어가는 작업이지만 예산과 시간이 허락한다면 추출틀을 사전에 보완하는 것이 포함오차를 줄이는 최선의 방법일 것이다.

표본추출을 하기 전에 추출틀을 보완해 미포함 문제를 해결할 수 없는 경우 흔히 실사과정에서 미포함 문제를 일부 해결하는 방법을 사용할 수 있다. 예를 들어 1단계로 조사구를 추출하고 2단계에서 조사구내의 가구를 추출하는 경우 표본으로 추출된 조사구에 대해서는 실사과정에서 기존의 가구명부를 사용하는 대신 현장확인을 통해 가구명부를 완벽하게 보완한 후에 가구를 추출함으로써 최소한 2차 표본추출단계에서는 포함오차가 발생하지 않도록 하는 것이다. 먼저 병원을 추출하고 병원내에서 환자를 추출하거나, 학교를 추출하고 학교내에서 학생을 추출하는 일반적인 다단계추출의 경우 2차 추출단계에서부터는 추출틀을 재정비하는 작업이 크게 어렵지 않다.

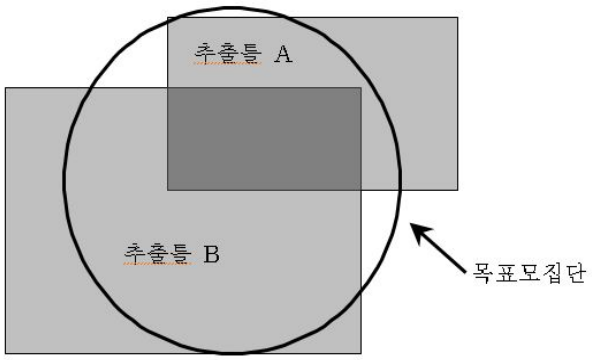
한편 실사과정에서 추출틀에 없는 단위가 있는 경우 어떻게 처리할 것이지를 사전에 규칙을 정해서 처리함으로써 포함오차를 일부 해결할 수도 있다. 예를 들어

사업체 명부를 사용하여 사업체 표본을 추출하는 경우 실제 사용된 추출틀을 작성한 이후에 신설된 사업체를 어떻게 처리할 것인지를 사전에 시나리오를 정해 신설 사업체가 표본으로 추출될 확률을 계산할 수 있도록 표본설계를 하는 것이다. 해당 읍면동 명부의 마지막에 있는 사업체가 추출되는 경우 실사과정에서 추출틀 작성 시점 이후에 신설된 사업체를 찾아서 조사대상에 포함하는 방법을 적용하면 이 경우 추출틀에 누락된 신설 사업체들을 표본에 포함하는 것이 가능하고 신설 사업체가 표본에 포함될 확률도 체계적으로 계산할 수 있다.

유사한 사례로 교육청에서 작년도에 작성된 해당 지역 초등학생 명부에서 학생들을 표본으로 추출하는 경우 만약 학급의 마지막 번호에 해당하는 학생이 표본으로 추출되는 경우 해당 학급의 학생명부 작성 이후에 새로 전학 온 학생도 표본으로 추출하는 방식이다. 이런 경우 전학 온 학생이 표본으로 추출될 확률은 학급의 마지막 번호를 갖고 있는 학생이 추출될 확률과 동일하게 되고 이런 추출확률을 이용해 가중치를 조정함으로써 비편향 추정이 가능하다.

이런 규칙을 통해 추출틀에 누락된 모든 구성원들을 추출틀에 있는 구성원들과 연결한다면 추출틀은 실제로 모든 구성원들을 포함할 수 있을 것이다. 추출틀에 포함된 단위가 표본으로 추출된다면 추출틀에 누락되었던 그와 연결된 모든 단위들도 표본에 포함하는 것이다. 누락된 단위들에 대한 추출확률은 그들과 연결된 추출틀에 포함된 추출단위의 추출확률과 같다.

특정 추출틀을 사용하는 경우 미포함 비율이 너무 높다고 판단되는 경우는 두 개 이상의 추출틀을 결합한 다중추출틀을 사용하는 것도 미포함오차를 줄이는 좋은 대책이 될 수 있다. [그림 3-2]는 두 개의 추출틀 A와 B를 사용하는 경우를 보여주고 있다. 이 경우 두 개의 추출틀 중 어느 하나도 단독으로는 목표모집단을 만족스럽게 포함하지 못한다. 하지만 [그림 3-2]에서 보여주듯이 두 추출틀을 결합하여 사용하면 하나의 추출틀을 사용하는 것보다 포함오차를 상당히 줄일 수 있다. 중간의 작은 직사각형은 두 추출틀이 중복된 부분을 의미한다. 만약 이 부분에 있는 단위들이 중복되는 것을 제거하지 않고 추출틀을 결합시키면 중복된 단위들은 추출틀상의 다른 단위들 보다 추출확률이 높을 것이다. 만약 중복 문제를 사정에 해결하지 못하더라도 만일 조사과정에서 단위들이 어떤 추출틀(A 또는 B) 또는 A와 B에 중복 포함되었는지 여부를 알 수 있다면 이것은 문제가 되지 않는다. 왜냐하면 이런 중복여부가 파악되면 표본 추출단위에 대한 정확한 추출확률을 산출하여 추정과정에서 이것을 가중치로 적절히 반영할 수 있기 때문이다.



[그림 3-2] 다중추출틀에서 목표모집단 포함 상황

물론 몇 개의 추출틀이 결합되는 경우 단위들이 중복되지 않도록 조정하는 것이 최선이다. 이런 작업은 추출틀에 공통적으로 이름, 주소 또는 서로 확인 가능한 정보가 기재되어 있다면 컴퓨터 프로그램을 통해 상당히 쉽게 해결할 수 있다. 특정 추출틀을 보완하여 완벽한 추출틀을 작성하는 것보다 이와 같이 다중추출틀을 이용하는 방식이 비용측면에서 포함오차를 줄이는 효율적인 방법이 될 수 있다.

마지막으로 표본추출틀이 결함을 갖고 있는 경우 간단한 방법은 모집단 정의를 변경하는 것이다. 다시 말해 현재 추출틀로 포함오차가 크게 문제가 되지 않도록 실제 사용하는 추출틀로 접근 가능한 단위들로 목표모집단을 다시 설정하는 것이다. 어떻게 보면 연구목적을 훼손하는 것처럼 보일 수도 있지만 불완전한 추출틀을 갖고 있는 상황에서 가장 손쉽고 경제적인 방법일 수 있다. 하지만 이런 방식으로 목표모집단을 재설정하는 것이 본래의 연구목적에 크게 위배되지 않는 것인지 면밀히 검토해 볼 필요가 있다. 이런 접근방식은 조사편의를 위해 가구조사에서 도서지역의 가구들을 조사대상에서 제외하는 것과 같이 일부 단위들을 인위적으로 목표모집단에서 제외하는 경우와 유사한 것으로 볼 수 있다. 아울러 추출틀 오차가 우려되는 경우 목표모집단을 재설정하지 않더라고, 최소한 추출틀이 갖고 있는 한계점을 자료 이용자들이 알 수 있도록 조사보고서에 관련 내용을 명시하는 것이 추출틀 오차에 따른 문제를 줄이는 올바른 자세일 것이다.

Ⅳ. 무응답 오차

1. 개요

통계조사는 조사를 실시하기 전에 연구목표를 정하고 이를 달성하기 위하여 적절한 자료수집 계획을 세우고 필요한 표본의 크기와 추출 방법을 정하는 동시에 얻고자 하는 자료의 종류와 범위를 미리 정한다. 하지만 추출된 표본들로부터 계획한 자료를 빠짐없이 모두 얻는다는 것은 현실적으로 불가능하다. 실제로 추출된 표본의 모든 단위로부터 의도했던 자료를 하나도 빠짐없이 얻는다는 것은 매우 힘든 일이다.

전화를 이용한 여론조사에서 추출된 전화번호를 통하여 접촉을 여러 번 시도하여도 통화가 이루어지지 않을 수 있으며, 가구가 추출단위인 조사에서는 추출된 가구를 여러 차례 방문해도 상담원에게 문을 열어 주지 않아 자료를 얻지 못하는 경우가 흔히 생긴다. 또한 방문이나 접촉에 성공했다고 하더라도 응답자가 설문지의 일부분에 대하여 응답하지 않는 경우가 많이 발생한다. 이렇게 통계조사에서 추출된 단위(개인, 가구, 기업체 등)와 접촉이 불가능하거나 응답 자체를 거부한 경우, 또는 조사항목 중 일부 항목에 대해서는 응답하지 않은 경우를 모두 무응답(non-response)이라고 부른다.

이러한 무응답은 자료 수집의 결함이 되며 통계조사의 정확성에 영향을 미치는 요인이다. 일반적으로 무응답이 많아지면 조사에서 얻어진 추정결과에 예상치 못한 편향(bias)이 나타날 수 있으며 더 나아가 미리 의도했던 추정결과의 정도(precision)가 낮아질 수 있다. 이렇게 무응답은 통계조사의 전반적 품질을 떨어뜨리는 중요한 요인 중에 하나이므로 무응답의 여러 가지 유형과 원인들을 살펴보고

무응답이 통계의 품질에 어떻게 영향을 미치는 지에 대해 이해하는 것은 통계조사의 관리자로서 매우 중요한 일이다.

2. 무응답의 종류

무응답에는 여러 가지 유형과 종류가 있으며, 이에 대한 이해를 돕기 위하여 다음과 같은 예를 고려해보자.

[예제 4-1] 어느 지역에 7개의 편의점(A, B, C, D, E, F, G)이 있는데 편의점들의 12월 평균 종업원수와 순이익을 알고 싶다고 하자. 전체에서 5개의 편의점을 단순확률추출하여 종업원수와 순이익을 조사하고 그 평균들로 모집단에 해당하는 7개 편의점의 실제 평균 종업원수와 평균 순이익을 추정하고자 한다.

전체 7개의 편의점 중에서 5개의 편의점을 단순확률추출하여 A, C, D, F, G 편의점이 선택되었다. 각각의 편의점을 면접원이 방문하여 업주 또는 그에 준하는 점포 책임자와 면담을 실시하여 조사를 완료했는데 방문과 조사 결과가 다음과 같다.

〈표 4-1〉 추출된 5개 편의점의 조사 결과

추출된 편의점	업주 또는 책임자 접촉 여부	조사 항목	
		종업원 수 (명)	12월 순이익 (백만원)
A	접촉 가능	응답 거부	응답 거부
C	접촉 가능	3	30
D	접촉 불가능 (접촉 5회 시도)	무응답	
F	접촉 불가능 (편의점 폐업)	무응답(부적격 단위)	
G	접촉 가능	7	응답거부

조사 결과 3개의 편의점 A, C, G는 면접 대상인 업주 또는 점포 책임자와 접촉(contact)이 되었다. 편의점 D는 5회를 방문했는데도 책임자와의 접촉에 실패하였다(non-contact). 또한 편의점 F는 면접원이 방문한 결과 폐업한 것으로 판명되었고, 표본에 포함 될 자격이 없는 편의점이었다(non-eligible unit). 접촉이 가능한 3개의 편의점 중에 편의점 A의 책임자는 조사에 응답하는 것을 거부(refusal)하였고 편의점 C의 책임자는 조사항목에 모두 응답하였으며(완전 응답, complete interview), 편의점 G의 책임자는 종업원의 수는 응답을 하였지만 12월 순이익은 응답을 거부하였다(부분응답, partial interview). □

이렇게 실제 조사에서는 표본에 선택된 추출단위의 접촉 여부, 응답에 대한 태도 등에 의하여 여러 가지 종류의 무응답이 발생한다. 이러한 무응답의 여러 가지 형태와 종류를 구분하여 체계적으로 관리하는 것이 통계조사의 품질관리에 있어서 매우 중요하다.

1) 단위 무응답과 항목 무응답

무응답의 유형 중 표본으로 추출된 단위에 대한 조사를 통해 응답거절 등의 이유로 어떤 응답도 얻지 못한 경우를 단위무응답(unit non-response)이라고 부른다. 어떤 경우 완성된 설문지를 얻었더라도 설문지에 있는 모든 항목이 조사되지 않았을 수 있는 데, 이와 같이 일부 항목에 대한 응답이 누락된 경우를 항목무응답(item non-response)이라고 부른다.

단위무응답은 주로 조사대상 단위와 접촉할 수 없거나 접촉에 성공했지만 그들이 조사에 참여하기를 거부하기 때문에 발생한다. 이러한 경우들은 각각 비접촉(noncontact)과 거부(refusal)의 경우로 부르고, 두 그룹은 무응답 추적 과정과 추정 단계에서 모두 별도로 구분하여 처리될 수 있기 때문에 이들을 구별하는 것이 필요하다. 전화조사에서는 전화를 받지 않을 때, 가계조사에서는 집에 아무도 없어서 구성원 중 아무도 접촉할 수 없는 경우 등 조사대상 단위를 접촉하는 것이 불가능할 때가 있으며 이러한 경우를 비접촉이라고 한다. 또한 접촉에 성공했더라도 응답자가 설문에 대한 응답을 처음부터 거부하는 경우를 응답거부라고 부른다. 하지만 비접촉과 거부의 차이가 항상 뚜렷한 것은 아니다. 예를 들어 우편조사에서 응답하지 않는 사람들과 접촉되지 않은 사람들은 실제로는 모두 응답거부에 해당한다. 그 밖에 단위무응답의 원인으로서 언어가 소통되지 않아 인터뷰를 하지 못한 경우, 육체적 또는 정신적으로 응답자가 인터뷰에 응할 수 없는 경우 등이 있다.

2) 응답의 유형

항목무응답은 설문지의 일부 항목에 대한 응답이 누락되어진 경우를 말하며 누락응답의 정도에 따라 응답유형을 세분하여 분류하는 것이 통계조사의 품질관리에 도움을 준다. 예를 들어 인터뷰를 통한 조사일 때 응답자가 응답을 시작하였으나 인터뷰를 시작한지 얼마 되지 않아 응답을 거부하여 중요한 조사항목 대부분에 대한 응답을 거의 얻어내지 못한 경우가 있는가 하면(breakoff; 응답회피), 중요한 문

항들에 대하여 대부분 대답하고 일부 조사항목에 대해 응답을 하지 않는 경우가 있다(partial interview; 부분응답). 또한 대부분의 조사항목에 응답을 하여 모든 항목에 대해 응답을 한 경우와 별 차이가 없는 경우도 있다(complete interview; 완전응답). 이러한 세 가지 경우는 추출단위의 응답에 대한 품질이 다르므로 구별하는 것이 바람직하다. 따라서 통계조사를 실시하기 전에 응답회피, 부분응답, 완전응답에 대한 정의를 정해놓는 것이 바람직하다. 다음은 가계조사인 경우 응답 유형을 구분하기 위해 적용 가능한 예이다.

- 설문지의 모든 항목 중 50%미만 응답한 경우는 회피, 50%~80%를 응답한 경우는 부분응답, 80% 이상을 응답한 경우는 완전응답
- 주요 응답 항목들 중 50%미만 응답한 경우는 회피, 50%~99%를 응답한 경우는 부분응답, 100%를 응답한 경우는 완전응답

항목무응답의 종류에 대한 정의는 조사의 목적과 성격에 따라 관리자가 합리적으로 정하는 것이 바람직하다.

3) 부적격단위

조사에서 추출단위를 접촉하지 못하였거나 또는 거부가 발생하였을 경우 종종 추출단위 자체가 조사범위밖에 있는 경우가 있다. [예제 4-1]에서 폐업한 것으로 드러난 편의점은 조사범위에 포함되지 않는 단위이다. 또한 전화조사에서 선택된 전화번호가 조사지역의 범위를 벗어나 있으나 전화를 받지 않아 접촉이 불가능한 경우가 있는데, 이런 경우 추출단위와 접촉이 안 되었기 때문에 조사 범위 내에 있는 단위인지 여부를 확인하지 못하게 된다. 또한 가구조사인 경우 [예제 3-5]과 같은 고령화연구패널과 같이 모집단이 45세 이상 성인이 거주하는 가구들로 정의되어 있을 때, 실제 접촉 불가능한 가구가 30대 부부가 거주하는 가구라면 이 가구는 조사 부적격에 해당하지만 실사과정에서 파악이 되지 않는다. 더 나아가 접촉이 되어 설문에 응하였다 하더라도 나중에 조사범위에 포함되지 않는다는 사실이 드러날 수 있다. 예를 들어 가임 연령의 여성을 조사하는데 60세의 여성이 응답을 하였다면 이는 조사 적격자가 아니라고 볼 수 있다. 이렇게 조사 범위를 벗어난 추출단위를 부적격단위(ineligible unit)라고 한다. 부적격단위를 접촉하는 것은 근본적으로는 표본추출틀의 부정확성 때문에 발생하게 되며 면접원의 실수 등 다양한 원인에 의해서도 발생할 수 있다.

응답률 등 무응답에 대한 여러 가지 관련 지표를 계산하는 경우 접촉을 시도한 전체 표본에서 부적격단위는 제외시켜야 한다. 즉 접촉 여부를 떠나 추출단위가 부적격 단위라면 응답률을 계산할 때 분모로 사용되는 접촉을 시도한 표본의 수에 포함되지 않도록 해야 한다.

그러나 실제 조사에서 접촉 불가능 등으로 단위 무응답이 된 경우 앞에서 설명한 것과 같이 표본추출 단위가 부적격단위인지 아닌지 판단할 수 없는 경우도 많이 있다. 더 나아가 접촉이 되어 응답을 하였더라도 부적격단위인지를 판단하는 것이 불명확할 수도 있다. 이러한 경우 전체 표본 중 접촉이 가능한 단위 중 조사 적격인 단위의 비율을 추정하여 전체 적격단위(eligible unit)의 수를 추정하는 방법이 있다. 필요에 따라 다음과 같은 방식으로 접촉 불능 등으로 적격 여부가 불확실한 단위들 중 적격단위 수를 추정하기도 한다.

- 적격 여부가 불확실한 단위들 중 적격단위의 수

= (적격단위 비율)×(적격여부가 불확실한 단위의 수)

= (적격단위 비율)×(응답한 추출단위 중 적격여부가 불확실한 단위의 수)

+(적격단위 비율)×(접촉 불가능한 추출단위 중 적격여부가 불확실한 단위의 수)

3. 응답률의 계산

표본조사에서 응답률은 조사의 품질을 나타내는 가장 중요한 지표 중 하나이다. 응답률이 낮은 조사는 추정결과에 심각한 편향이 존재할 가능성이 높으며 편향이 작다고 하더라도 실제 조사된 유효 표본의 수가 계획한 수보다 작기 때문에 추정량의 표준오차가 커져서 조사의 정확성에 부정적인 영향을 미친다. 따라서 응답률은 통계조사의 품질을 검증하는 중요한 척도이다.

1) 단위 무응답에 대한 응답률 계산 - 가구조사

이 절에서는 가구조사(household survey)의 경우 단위 무응답에 대한 응답률을 계산하는 방법에 대하여 논의하며 전화, 우편 조사 등도 유사한 개념과 방법을 적용하면 그 응답률을 계산할 수 있다.

단위 무응답에 대한 응답률에는 여러 종류가 있으며 그 계산을 위해서는 추출단위의 자격을 고려하여 다음과 같은 4가지 집단으로 분류할 필요가 있다.

- 응답자(interviews)
- 적격단위인 무응답자(eligible cases that are not interviewed; non response)
- 적격단위인지 판단이 불가능한 경우(cases of unknown eligibility)
- 부적격 단위(cases that are not eligible)

응답률 계산에 있어 응답자와 적격단위 중 무응답자는 반드시 분모에 반영되어야 하고 부적격 단위들은 제외되어야 한다. 어떤 단위가 적격단위인지 부적격단위인지 판단이 불가능한 경우에는 이들 중 조사적격 단위의 비율을 추정하고, 이를 근거로 적격자의 수를 추정하여 분모에 포함시킨다.

응답률 계산을 위하여 추출단위를 적격여부와 응답유형에 따라 분류하는 것이 필요하다. 다음은 응답률 계산을 위해 응답자를 여러 범주로 구분한 것이다. 여기서 I, P, R, NC, O는 모두 적격단위만을 대상으로 한 것이다.

I : 완전응답(complete interview)
P : 부분응답(partial interview)
R : 응답거부(refusal)
NC : 접촉불능 (noncontact)
O : 그 밖의 모든 무응답 (other nonresponse)
UC : 적격단위인지 판단이 불가능 - 접촉 가능한 경우
UN : 적격단위인지 판단이 불가능 - 접촉 불가능한 경우
e_c : 적격단위인지 판단이 불가능(접촉 가능한 경우) 중 적격단위 비율
e_n : 적격단위인지 판단이 불가능(접촉 불가능한 경우) 중 적격단위 비율

응답률은 조사에서 응답을 한 비율을 나타내는 측도이다. 응답률은 기본적으로 표본에서 적격단위 중 응답에 응한 단위의 수를 전체 적격단위 수로 나눈 비율이다. 응답률이 낮을수록 무응답에 의해 편향이 발생할 가능성도 높아진다. 따라서 응답률은 통계조사의 품질을 파악할 수 있는 측정 가능한 지표 중에 가장 중요한 지표이다. 위에서 정의된 응답자들의 유형을 사용한 응답률에는 다음과 같은 두 가지 종류가 있다.

● 완전응답률(full response rate)

완전응답률은 아래와 같이 정의되며 응답자 중 완전응답만을 고려한 비율이므로

여러 가지 응답률 중 가장 작다.

$$\frac{I}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

● 전체응답률(overall response rate)

전체응답률은 완전응답과 부분응답 모두를 응답자로 고려한 비율이며 완전응답률보다 항상 크다.

$$\frac{I+P}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

제시된 완전응답률과 전체응답률을 계산할 경우 실제로 적격여부를 확인할 수 없는 단위들 중 적격단위의 비율 e_c와 e_n을 추정하기 어려운 경우가 많다. 부적격단위일 가능성이 매우 낮은 조사에서는 적격단위의 비율 e_c와 e_n을 1로 보고 응답률을 계산해도 무리가 없을 것이다. 또한 접촉 여부와 상관없이 부적격단위인지 여부를 쉽게 판단할 수 있는 경우에는 비율 e_c와 e_n을 0으로 처리해 분모에서 적격단위인지 판단이 불가능한 경우를 제외할 수도 있다.

응답률 외에 무응답에 관한 품질 지표로 협조율, 접촉률, 거부율 등을 이용할 수 있다.

● 협조율(co-operation rate)

협조율은 접촉 가능한 적격단위들 중에서 조사에 응답한 단위들의 비율을 나타낸다. 즉 응답률에서 접촉 불가능한 적격단위들의 수를 분모에서 제외한 비율이다.

$$\frac{I+P}{(I+P)+(R+O)+e_c UC}$$

● 접촉률(contact rate)

접촉률은 표본에서의 적격단위들 중에서 접촉 가능한 단위들의 비율을 나타낸다. 따라서 이 비율은 실제로 조사를 시행할 때 추출단위에 얼마나 쉽게 접근하여 접촉할 수 있는지에 대한 정도를 나타낸다.

$$\frac{(I+P)+R+O+e_c UC}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

● 거부율(refusal rate)

거부율은 표본에서의 적격단위들 중에서 응답을 거부한 단위들의 비율을 나타낸

다. 따라서 이 비율은 실제로 조사를 시행할 때 추출단위가 응답을 거부하는 비율이 얼마인지 그 정도를 나타낸다.

$$\frac{R}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

2) 단위 무응답에 대한 응답률 계산 - 경기조사

경기조사(business survey)는 기업 또는 기업가를 상대로 일정 시간에 대한 경기예측이나 경영상황에 관한 의견을 응답 받아 그것을 집계해 결과를 도출한다. 통상 단기간(1~3개월) 관측이며 조사항목은 국내외 경기, 국내 경제 전반, 소속 산업과 자기 산업의 현황과 예측이며 산출량, 매출액, 수입액, 설비투자액, 재고자산, 매출채권, 매입채무, 금융기관 차입금, 현금·예금의 기말잔액, 당기와 차기 수입 예상액, 판매가격, 받을 어음의 기간 등을 내용으로 한다. 경기조사는 일정 시간에 대한 조사이므로(예를 들어, 1개월, 3개월, 6개월) 응답자가 모든 기간에 대한 응답을 할 수도 있고 부분 기간에 대한 응답만을 할 수도 있다. 예로 3개월간 매월 매출액조사에서 마지막 달에 대한 매출액만 제출하고 처음 2개월의 매출액을 제출하지 않을 수 있다.

이러한 경기조사의 응답률은 응답 유형을 응답 시간 단위를 고려한 다음과 같은 분류를 기초로 계산하게 된다.

FC : 모든 기간에 대한 완전응답 (Full period return with complete data)
FP : 모든 기간에 대한 부분응답 (Full period return with partial data)
PC : 부분 기간에 대한 완전응답 (part period return with complete data)
PP : 부분 기간에 대한 부분응답 (part period return with partial data)
RNU : 응답하였으나 쓰지 못하는 경우(returned but not used)
NR : 무응답 (non-response)
U : 부적격단위인지 판단이 불가능 (cases of unknown eligibility)
e : 부적격단위인지 판단이 불가능한 단위 중 적격인 비율

위의 분류를 이용하여 경기 조사의 응답률은 다음과 같은 두 가지 종류로 계산할 수 있다.

- 완전응답률(full response rate)

완전응답률은 모든 기간에 대한 완전하게 응답한 경우만을 응답으로 고려하여

계산한 응답률이다.

$$\frac{FC}{(FC+FP+PC+PP)+RNU+NR+e(U)}$$

● 전체응답률(overall response rate)

전체응답률은 부분 기간 또는 부분 응답도 모두 응답한 것으로 간주하여 계산한 응답률이다. 완전응답률보다 항상 크다.

$$\frac{FC+FP+PC+PP}{(FC+FP+PC+PP)+RNU+NR+e(U)}$$

3) 항목 무응답에 대한 응답률 계산

지금까지는 단위 무응답에 대한 여러 가지 응답률의 계산에 대하여 논의했다. 항목 무응답인 경우에도 각각의 항목에 대한 응답률을 계산하는 것도 품질 관리에 큰 도움이 된다. 특히 조사 항목 중 주요 항목에 대해서는 그 응답률을 계산하여 추정의 정확성을 가늠할 수 있는 지표로 고려해야 한다.

항목 응답률은 다음과 같이 그 항목에 응답을 해야 하는 자격을 가진 단위의 수와 그 중에 응답한 단위의 수의 비율로 나타낸다.

● 항목 응답률

$$\frac{\text{항목에 응답한 조사단위의 수}}{\text{항목에 응답해야 하는 조사단위의 수}}$$

4. 무응답에 의한 편향

앞에서 무응답의 종류와 여러 가지 응답률에 대하여 논의했다. 이 절에서는 무응답이 추정량의 정확성에 어떤 영향을 미치는지 간략하게 살펴본다.

실제 조사에서 추출단위가 조사에 응답을 할지 응답거부를 할지는 조사 전에는 알 수가 없다. 그러나 무응답의 영향을 알아보기 위해 미리 응답유형을 알고 있는 모집단을 생각할 수 있다고 가정해 보자. 예로 앞의 [예제 4-1]에서 7개의 편의점 사례를 생각해 보자.

[예제 4-2] 전체 7개 편의점의 실제 평균 순이익(40백만원)을 추정하고자 한다.

모두 접촉 가능하고 12월 순이익에 대한 두 가지 응답유형(I과 II)을 가정하고 비교해 보자.

표본으로 4개의 편의점을 단순확률추출한다고 가정하고 응답한 단위들만의 평균 순이익을 실제 순이익의 추정값으로 사용한다고 가정하자.

〈표 4-2〉 어느 지역의 7개 편의점의 12월 순이익과 응답 유형

편의점	A	B	C	D	E	F	G	평균
12월 순이익(백만원)	10	20	30	40	50	60	70	40
응답유형 I	응답	거부	응답	거부	응답	거부	응답	
응답유형 II	응답	응답	응답	응답	거부	거부	거부	

[응답유형 I]과 [응답유형 II]의 경우 표본에서 산출되는 추정결과에 미치는 영향을 생각해 보자. [응답유형 I]은 편의점 순이익의 크기와 응답여부가 관계가 없음을 알 수 있다. 반대로 [응답유형 II]의 경우는 12월 순이익이 큰 편의점들이 응답을 하지 않는 경우이다. 즉 순이익이 큰 편의점이 응답을 하지 않으므로 추정값이 실제 평균(4천만원)보다 작게 나올 것이다. 예를 들어 4개의 표본으로 A, B, E, F의 편의점이 추출되었다면

- [응답유형 I] 인 경우 A와 E가 응답하고 B와 F가 응답을 거부하므로 추정값은 3천만원이 된다.

$$\frac{A+E}{2}=\frac{10+50}{2}=30 \text{ (백만원)}$$

- [응답유형 II] 인 경우 A와 B가 응답하고 E와 F가 응답을 거부하므로 추정값은 1천5백만원이 된다.

$$\frac{A+B}{2}=\frac{10+20}{2}=15 \text{ (백만원)} \quad \square$$

이와 같이 조사하려고 하는 항목에 대한 응답값(실제값)과 무응답 유형이 밀접한 관계를 가지고 있는 경우에는 일반적으로 응답된 자료만을 이용해 표본에서 산출된 추정값은 무응답에 의한 편향 때문에 정확성이 떨어지게 된다.

이러한 무응답에 의한 편향은 일반적으로 응답률이 낮을수록 증가하는 경향이 있다. 응답률이 낮으면 응답자가 특정 요인에 의해 응답하려고 하는 의지나 경향이

감소된다는 의미이며, 무응답을 일으키는 원인이 응답값(실제값)의 크기와 관련이 있다면 무응답에 의한 편향이 증가할 가능성이 높아진다는 것이다. 한편 응답자와 무응답간의 차이가 큰 경우에는 응답률이 높더라도 무응답에 의한 편향을 무시할 수 없는 경우가 있다.

무응답률이 높아지면 무응답에 의한 편향이 생길 가능성도 높아지고 목표한 추정량의 정확성에도 큰 영향을 미친다. 일반적으로 조사를 실시하기 전에 목표한 추정량의 정도에 의거하여 조사에 필요한 표본크기를 계산한다. 무응답이 많은 경우 실제 목표한 표본의 개수보다 작은 수의 응답자들로부터 얻은 자료를 이용하여 추정값을 구하게 되므로 목표한 추정량의 정도를 확보할 수 없다. 다시 말해 무응답이 발생하는 경우 응답자의 자료만으로 추정을 하면 통계적으로 추정량의 표준오차가 목표로 한 것보다 커지기 때문에 조사의 정확성이 떨어지게 된다.

무응답에 의한 편향이 어느 정도 발생하는지 또는 무응답 발생에 의해 추정량의 정확성이 얼마나 감소하는지 등은 무응답의 특성, 즉 실제값을 조사할 수 없다는 무응답 특성 때문에 수량적으로 측정하는 것은 현실적으로 매우 어렵다. 하지만 응답자와 무응답자의 차이를 여러 가지 다양한 방법으로 분석해 보면 그 차이점이 파악되고 그 차이의 성격과 크기에 따라 무응답에 의한 편향과 추정량의 정확성에 미치는 영향을 어느 정도를 추정할 수 있는 경우도 있다.

[예제 4-3] 앞에 제시된 [예제 4-2]의 경우 종업원의 수와 순이익은 어느 정도 비례관계가 있다는 사실이 알려져 있고, 종업원 수는 추출된 모든 편의점에서 응답결과를 얻었다고 가정하자. 이 경우 순이익 항목에 대해 응답을 거부한 편의점들의 평균 종업원 수와 응답을 한 편의점들의 평균 종업원 수를 비교하면 순이익 항목에 있어서 무응답에 의한 편향이 얼마나 심각한지 파악할 수 있다. 예를 들어 4개의 표본으로 A, B, E, F의 편의점이 추출되었다고 하자.

〈표 4-3〉 어느 지역의 7개 편의점의 12월 순이익과 종업원 수 및 응답 유형

편의점	A	B	C	D	E	F	G	평균
12월 순이익 (백만원)	10	20	30	40	50	60	70	40
응답유형 I	응답	거부	응답	거부	응답	거부	응답	
응답유형 II	응답	응답	응답	응답	거부	거부	거부	
종업원의 수(명)	1	2	3	4	5	6	7	4

- [응답유형 I]인 경우 A, E 가 응답하였으므로 응답한 편의점의 평균 종업원 수는 (1+5)/2=3명이고 B, F가 응답을 거부하므로 응답을 거부한 편의점의 평균 종업원 수는 (2+6)/2=4명이다. 따라서 응답한 편의점과 응답을 거부한 편의점의 평균종업원 수의 차이는 1명이다.
- [응답유형 II]인 경우 A, B가 응답하였으므로 응답한 편의점의 평균종업원 수는 (1+2)/2=1.5명이고 E, F가 응답을 거부하므로 응답을 거부한 편의점의 평균종업원 수는 (5+6)/2=5.5명이다. 따라서 응답한 편의점과 응답을 거부한 편의점의 평균종업원 수의 차이는 4명이다. □

위의 [예제 4-3]과 같이 모집단을 알고 있는 가상적인 경우 무응답에 의한 편향과 무응답이 추정량의 정확성에 미치는 영향이 어느 정도인지를 알 수 있다. 하지만 일반적으로 무응답자에 대한 정보를 얻을 수 없기 때문에 이런 분석이 가능한 경우는 흔하지 않다. 하지만 다양한 자료를 무응답에 의한 영향을 분석하기 위해 사용할 수 있다. 관심 변수와 관련이 높은 항목 중 무응답이 적은 것들, 재조사(follow-up survey)를 통해 얻은 자료, 외부에서 얻은 관련 자료 등을 이용하여 무응답에 의한 편향과 무응답이 추정량의 정확성에 미치는 영향을 다음과 같은 일반적인 방법으로 어느 정도 파악할 수 있다.

- 무응답자와 응답자의 특성 비교
- 무응답에 의한 추정량의 분산 증가분 추정

5. 무응답에 대한 대책

무응답을 처리하는 방법으로는 단위 무응답의 경우 재조사(call-back 또는 follow-up), 무응답가중치조정(non-response weighting adjustment)이나 결측치 대체(imputation) 등이 있다. 항목 무응답인 경우에는 일반적으로 결측치 대체를 통하여 무응답을 처리하는 것이 일반적이고, 단위 무응답인 경우에는 무응답 가중치 조정을 흔히 사용한다.

보통의 경우 재조사는 무응답자를 일정한 횟수만큼 전체적으로 또는 부분적으로 다시 조사하는 것을 말한다. 예를 들어, 전화조사의 경우 응답 거부를 하였거나 접촉 불가능한 번호들을 대상으로 여러 차례에 걸쳐 다시 전화하여 조사하는 방법이다. 재조사를 실시할 때 마다 그 시점에서의 모든 무응답자에게 전화를 걸 수도 있고 시점에 따라 전화를 거는 무응답자들의 비율을 다르게 하여 재조사를 실행할

수도 있다. 재조사를 하는 경우 조사 시간대나 요일을 재조사 시점마다 달리하는 것이 효과적일 수 있다.

무응답가중치조정은 자료 수집이 완료된 후, 응답자의 가중치를 조정하여 무응답에 의한 추정량의 편향을 줄이고자 하는 방법이다. 이 방법은 우선 유사한 속성을 갖는 조사 대상자(무응답 포함)들을 묶어 무응답 조정군(adjustment cell)을 구성하고, 해당 조정군 내에서의 응답확률을 기반으로 응답자의 가중치를 조정해 줌으로써 해당 조정군 내에서 무응답자를 응답자가 대신 설명하도록 해 주는 방법이다. 무응답 가중치 조정을 하려면 조사단위들의 응답 확률을 추정해야 하며, 이를 위해 무응답 조정군을 이용하거나 응답성향점수(response propensity score)를 이용할 수도 있다. 하지만 응답확률을 설명해 주는 확률적 모형을 만드는 것은 통계적인 지식과 관련 조사에 대한 지식을 모두 요구하는 매우 어려운 작업이다. 무응답 가중치 조정을 위해 사후층화(post-stratification)를 이용하는 방법도 있다.

단위 무응답의 경우 무응답가중치조정법을 흔히 사용하지만 항목 무응답인 경우에는 결측치 대체법(imputation)을 많이 사용한다. 이는 무응답으로 발생한 결측값을 다양한 방법을 통해 특정한 값으로 대체하는 것을 말한다. 무응답 단위와 유사한 속성을 갖는 응답 단위 중에서 하나를 선택하여 무응답 항목을 대체하는 핫덱대체법(hotdeck imputation) 또는 최근방대체법(nearest neighbor imputation) 등이 있다. 핫덱대체법이나 최근방대체법은 단위 무응답인 경우에도 효과적으로 사용할 수 있다. 또한 결측된 자료를 응답자들의 평균으로 대체하는 평균대체법(mean imputation), 두 개의 관련된 변수의 비율을 사용하는 비대체법(ratio imputation), 관련 변수들을 이용한 회귀모형을 만들어 그 추정값으로 대체하는 회귀대체(regression imputation) 등이 있다. 이와 관련된 구체적인 내용은 Lohr(1999) 등을 참고하기 바란다.

[예제 4-4] 어느 지역에 7개의 편의점(A, B, C, D, E, F, G)이 있는데 편의점들의 12월 평균 종업원수와 순이익을 알고 싶다고 하자. 전체 모집단에서 4개의 편의점을 단순확률추출하여 종업원수와 순이익을 조사하고 그 평균들로 7개 편의점의 실제 평균 종업원수와 평균 순이익을 추정하고자 한다. 예를 들어 4개의 표본으로 A, C, D, G의 편의점이 추출되었고 조사 결과가 <표 4-4>와 같다고 하자.

〈표 4-4〉 추출된 5개 편의점의 조사 결과

추출된 편의점	업주 또는 책임자 접촉 여부	조사 항목	
		종업원 수(명)	12월 순이익(백만원)
A	접촉 가능	1	10
C	접촉 가능	3	30
D	접촉 불가능 (접촉 5회 시도)	무응답	
G	접촉 가능	7	응답거부

- 편의점 중 D는 단위 무응답이므로 핫덱대체법을 이용하는 것이 효과적일 수 있다. 모든 항목에 응답한 편의점 A, C중 랜덤하게 하나를 선택하여 D의 모든 항목을 대체하는 것이다. 만약에 편의점 A가 선택되면 편의점 D의 종업원 수와 순이익은 각각 1명과 10백만원으로 대체된다.
- 편의점 G의 순이익은 항목 무응답이다. 만약에 평균 대체법을 이용한다면 순이익에 응답한 편의점 A와 C의 순이익 평균, 즉 (10+30)/2=20(백만원)으로 편의점 G의 순이익을 대체할 수 있다. □

결측치를 처리하는 무응답가중치조정과 대체법(imputation)과 관련된 자세한 이론은 조사 통계 이론의 핵심적인 부분을 차지하며 그 방법의 개발은 통계적 지식과 관련 통계의 전반적 지식이 결합되어지는 복잡한 절차를 통해 이루어지므로 이 매뉴얼에서는 다루지 않는다. 결측치 처리와 관련된 구체적인 내용은 Lohr(1999), Cochran(1977) 등을 참고하기 바란다.

결측치 처리와 관련된 내용을 이용자들에게 제공하는 것은 통계조사의 품질관리 측면에서 중요하다. 따라서 다음과 같은 결측치 처리 방법이나 측정 가능한 무응답 관련 지표를 산출해 이를 통계 이용자들에게 제공하는 것이 바람직하다.

- 단위 무응답 처리 방법 또는 결측치 대체법
- 결측치 대체로 대체된 단위의 비율

$$\frac{\text{결측치 대체된 단위의 수}}{\text{전체 추출 단위의 수}}$$

- 항목 무응답 처리 방법 또는 결측치 대체법
- 중요 항목에서 결측치가 대체된 비율

$$\frac{\text{결측치 대체된 단위의 수}}{\text{항목에 응답해야 하는 단위의 수}}$$

- 중요 항목의 추정값에서 결측치 대체가 차지하는 비율

$$\frac{\text{결측치 대체된 자료에 대한 가중치 적용 합계}}{\text{모든 단위의 자료에 대한 가중치 적용 합계}}$$

Ⅴ. 측정 오차

1. 개 요

측정 오차는 다양한 원인에 의해 조사단위가 갖고 있는 참값을 측정하지 못하기 때문에 발생하는 오차이다. 사회조사의 경우 측정도구에 해당하는 설문지가 정교하지 못한 경우 측정 오차가 발생할 수 있으며, 응답자가 여러 가지 이유로 정확하지 않은 응답을 하는 경우도 측정 오차로 볼 수 있다.

예를 들어, 의학연구자가 시행하는 보건 관련 조사 중 초기 문진조사에서 개별 환자들이 오래전에 가졌던 자신의 질병을 제대로 기억하지 못하는 경우 면접원에게 잘못된 정보를 제공하게 될 것이다. 응답자는 경우에 따라 의도적으로 또는 무의식적으로 잘못된 정보를 줄 수 있다. 면접 조사원들이 질문을 하고 응답을 기록하는 중에 오류를 범할 수도 있다. 애매한 질문, 혼동되는 지시, 오해하기 쉬운 용어가 사용된 설문지 때문에 측정 오차가 발생할 수 있다. 기업가를 대상으로 한 경기 동향 조사에서 작년 동기 또는 동월 대비 분석을 위해 필요한 기업가의 의견을 제대로 측정할 수 없도록 질문이 작성되었다면 조사 목적에 따라 의도했던 내용을 측정할 수 없을 것이다.

또한 자료수집방법이 측정 오차에 영향을 미칠 수 있다는 것은 잘 알려져 있다. 예를 들어, 어떤 경우에는 전화조사에서 얻은 정보가 면접조사에서 얻은 정보에 비해 덜 정확할 수 있다. 마지막으로 조사를 실시하는 주변여건이나 환경도 측정 오차에 영향을 줄 수 있다. 예를 들어 약물사용, 성적인 행태, 출산력 등 민감한 주제에 대한 자료를 수집하기 위해서는 집안의 다른 식구들이 있는 장소보다 남의

눈을 피할 수 있는 장소에서 면접을 하는 것이 정확한 응답을 얻는데 도움이 된다.

측정 오차는 크게 두 가지 유형으로 구분할 수 있는데, 첫 번째는 참값보다 크거나 작은 방향으로 치우친 관측이 지속적으로 이루어져 편향이 발생하는 경우이고, 두 번째는 랜덤하게 참값보다 크거나 작은 값으로 관측이 되는 경우이다. 첫 번째 경우는 편향이 발생하여 측정 오차가 추정값의 정확성(accuracy)에 심각한 영향을 미칠 수 있는 경우이다. 예를 들어 소득조사에서 대부분의 응답자가 실제보다 소득을 줄여서 응답하는 경우가 이에 해당한다. 두 번째 경우는 참값에 랜덤오차(random error)가 추가되어 변동이 커짐으로써 추정의 정도(precision)에 영향을 주는 경우이다. 이러한 경우 추정량의 표준오차가 증가하게 된다. 측정 오차는 다음과 같은 다양한 원인에 의하여 발생할 수 있다.

- 측정에 필요한 개념화의 문제
- 측정하는 도구 (설문지, 전화, 면접 등)
- 응답자의 응답 능력 (기억력, 거짓말 등)
- 면접원의 기술 (화술, 경험, 친밀도 등)
- 환경 (인터뷰 장소, 시간 등)
- 측정방법의 실행과정에서 발생되는 실수 (코딩 에러 등)

이러한 측정 오차의 원인들은 독립적으로 작용하여 오차를 발생시키기도 하지만 많은 경우에 서로 상호 작용을 하면서 복잡한 경로를 통하여 오차가 발생한다. 측정 오차에 대한 연구는 오랜 기간에 걸쳐서 진행되어 왔으며 오차를 줄이려는 여러 가지 방법들이 개발되어 왔다. 여기서는 측정 오차의 여러 가지 원인 중에 가장 중요한 설문지, 면접원과 측정 방법 등에 대하여 살펴본다.

2. 설문지와 응답과정

질문을 어떻게 표현하는지에 따라 응답결과는 상당한 차이를 보일 수 있다. 특히 개인적인 견해와 관련된 질문의 경우 이런 경향이 많이 나타났다. 이와 관련해 자주 인용하는 유명한 예가 있다(Schuman and Presser, 1981). “미국 정부가 민주주의를 반대하는 연설을 금지해야 한다고 생각합니까?”라고 물었을 때 응답자의 21.4%가 “예”라고 응답했다. 반면에 “미국 정부가 민주주의를 반대하는 연설을 허용해야 한다고 생각합니까?”라는 질문에는 “아니오”라고 응답한 비율이 47.8%인 것으로 나타났다. 결과적으로 단순히 ‘금지해야 한다’라는 강한 표현 대신에 ‘허용하

지 말아야 한다'와 같은 다소 완곡한 표현을 사용하더라도 조사 결과에 있어서 큰 차이기 생길 수 있다는 것을 알 수 있다. 민주주의를 반대하는 연설에 대해 부정적인 생각을 갖고 있기는 하지만 '금지해야 한다'와 같은 강한 표현에 부담감을 느끼는 사람들이 상당 부분 존재하기 때문에 이런 차이가 발생한 것으로 볼 수 있다.

따라서 설문지의 설계는 통계조사의 품질을 결정하는 중요한 요소이며, 설문지 설계단계부터 정확한 목표를 설정하고 지속적으로 관리하는 것이 필요하다. 설문지 설계의 목표는 다음 세 가지로 요약될 수 있다.

- 연구가 의도한 의미를 정확히 전달할 수 있도록 질문을 만드는 것이다.
- 가장 정확한 대답이 나오게 설계된 방법으로 질문을 설문지에 기술해야 한다.
- 자료수집 비용이 예산을 넘지 않도록 설문지를 설계해야 한다.

실제 응답자가 질문에 응답하는 과정은 5개의 뚜렷한 인지단계를 순서대로 거치게 된다. 여기서 5단계는 다음과 같이 정리할 수 있다.

(1) 정보의 생성

정보를 습득해서 기억에 저장하는 과정이다. 사건이나 경험을 측정과정에서 기억하거나 검색해 내려면 그것에 대한 기록이 만들어져 있어야 한다. 예를 들어 조사에서 가구 구성원들의 행동에 관한 질문을 응답자가 정확하게 대답하려면 그 행동을 우선 관찰하고 면접 중에 기억해 낼 수 있도록 암기하고 있어야 한다. 이런 측면에서 보면 실업률 조사에서 본인이 아닌 대리자가 질문에 대신 응답할 경우 조사 대상자의 구직 경험에 대한 정보가 부족하여 측정 오차가 생길 수 있다.

(2) 질문의 이해

응답자는 질문에 대해 깊이 생각해보고, 질문에서 어떤 정보를 요구하는 것인지 이해하려고 노력해야 한다.

(3) 정보의 검색

응답자가 질문을 이해하고 나면, 응답자는 이제 질문에 답하는데 필요한 기억되어 있는 정보를 회상하거나 검색한다. 기억 또는 다른 가족, 직장 동료, 회사 데이터베이스나 파일 등과 같은 출처를 통해 필요한 정보를 찾아내는 것이다.

(4) 응답의 구성

검색한 정보를 평가하고 질문에서 요구한 양식에 따라 응답을 구체화하는 것이

다. 흔히 사회조사 질문들은 선택형 또는 서술형으로 응답이 이루어진다.

(5) 응답의 편집 및 전달

마지막 단계로 응답자는 자신의 응답을 교정하고 전달한다. 그대로 정확한 응답을 하거나, 필요에 따라 응답자가 교정을 한 후에 연구자에게 전달한다. 예로 사회적으로 민감한 부분에 대한 질문은 거짓으로 대답할 수 있다.

3. 면접원과 면접원 변동

모든 통계조사에서 면접원이 필요한 것은 아니다. 예를 들어, 우편, 인터넷, 전자메일 같은 자료수집 방식에서는 설문지를 응답자에게 보내고, 그들은 면접원의 도움 없이 설문지를 완성한다. 하지만 상당수 통계 조사에서 면접원은 조사과정에서 없어서는 안 되는 매우 중요한 역할을 한다. 면접원들의 가장 중요한 역할 중 하나는 조사 대상자와 연락을 해서 그들이 조사에 참여하도록 설득하는 것이다. 면접원이 응답자와 면접하는 방식은 크게 표준화 면접과 대화식 면접으로 나뉜다. 대부분의 조사에서는 아래 두 대비되는 방식을 적절히 타협한 면접기법을 사용한다.

표준화 면접(standardized interviewing)방법은 면접원과 응답자사이의 상호작용을 표준화 하는 것이다. 표준화 면접의 목적은 응답이 면접원에 의해 조금도 영향을 받지 않게 하기 위하여 모든 응답자에게 같은 질문을 정확히 같은 방법으로 하도록 만드는 것이다. 이런 표준화의 중요 이점은 면접원 때문에 생기는 오차의 변동인 면접원 변동(interviewer variability)을 줄이는 것이다. 단점은 응답자의 이해 수준이나 질문의 난이도에 따라 면접원의 설명이나 개입이 필요한 경우가 있을 때 면접원이 개입하지 못 한다는 것이다.

두 번째 방법은 응답자에게 질문들의 의미를 표준화시키기 위해 면접원과 응답자 사이에 훨씬 높은 수준의 상호작용을 요구하는 대화식 면접(conversational interviewing)이 있다. 이 방법에서 면접원은 설문지에 있는 질문의 표현을 바꿀 수 있는데, 예를 들어, 응답자의 상황에 맞추어 질문을 변형시킬 수 있다. 더욱이 질문의 의미를 좀더 명확히 하고 그 질문이 응답자의 특정한 상황에 어떻게 적용되는지 확실히 하기 위하여 필요한 방법을 동원해서 응답자를 자유로이 도울 수 있다. 면접원이 조사의 내용에 대한 폭 넓은 지식을 가지고 질문에 대하여 명확한 이해를 하고 있다면 이 방법은 성공적으로 적용될 수 있지만 그렇지 않으면 부정적인 효과가 나타날 수 있다.

면접원 변동은 같은 질문에 대하여 같은 대상에게 면접하여 응답을 받았을 때

면접원에 따른 차이를 말한다. 예를 들어 같은 답을 가진 응답자가 다른 면접원에게 상이한 응답을 하는 경우에 발생한다. 예를 들어 개방형 질문(open-ended question)에서는 면접원의 태도나 생각에 따라 응답자의 답이 다를 가능성이 크다.

이러한 차이는 면접원의 경험, 교육 정도, 성격, 외모 등에 의하여 나타날 수 있다. 또한 응답자의 특성이나 설문지가 면접원의 특성과 상호 작용하여 나타날 수 있다. 현실적으로 면접원 변동은 측정하기 매우 어려우며 측정을 위한 특별한 연구조사가 이루어지지 않으면 수량적으로 측정하기 어렵다. 하지만 대규모 보건 관련 조사 등 국제적으로 표준화된 도구를 사용하려는 조사는 조사 시작 전에 실시하는 시험조사(pilot study)에 면접원 변동을 측정할 수 있는 방법을 포함하여 면접원 변동을 측정하고 그 원인을 파악하려고 하는 경우가 있다.

4. 자료수집 방법

자료수집 방법(mode of data collection)은 조사 대상자를 접촉하고 질문에 대한 응답을 얻는데 사용되는 수단을 가리킨다. 현재 사용되는 자료수집방법은 응답자와 접촉의 정도, 자료수집자인 면접원의 참여 정도, 자료 수집 도구 등에 의해 다음과 같이 분류될 수 있다.

○ 면접조사

대면 면접조사(face-to-face interviewing)는 비용이 많이 들지만 융통성이 있고, 높은 응답률을 얻을 수 있다. 하지만 사회적 기대부응 편향과 면접원 변동이 발생할 수 있다. 현재는 과거에 이용되었던 종이설문 면접방식(paper and pencil interviewing)에서 측정 오차를 많은 부분에서 줄일 수 있는 컴퓨터보조 개별면접(computer assisted personal interviewing; CAPI)을 활용하는 형식의 조사방법이 점차 많이 사용되고 있다.

○ 전화조사

전화조사의 장점은 면접조사보다 일반적으로 비용이 싸다는 것이며 특정 장소에서 면접원들을 동시에 관리할 수 있기 때문에 면접원 변동이 면접조사보다 더 작다. 그러나 면접조사보다 융통성이 적고 면접 시간이 길 수 없다는 단점이 있다.

○ 우편조사

우편조사는 면접원이 없으므로 면접원 변동을 일으키지 않고 민감한 주제에 대

한 자료를 수집하는데 적당하다. 비용이 저렴하긴 하지만 우편조사는 수행하는데 시간이 많이 걸리고, 응답률이 비교적 낮으며, 항목 무응답률이 높다.

○ 일지조사

일지조사(diary surveys)는 사건들에 대한 정보를 회고하여 수집하는 목적으로 사용된다. 가구 구매, 음식 섭취, 일간 이동, TV 시청, 그리고 시간의 활용 같은 자주 일어나는 일들에 대한 정보를 얻기 위해 구조적인 설문지 대신 응답자가 입력한 일지를 사용하는 것이다. 일지의 성공적인 작성을 위해서는 응답자가 정보를 기록하는데 매우 적극적인 자세를 보여야 하며 응답률을 높이기 위하여 면접원이 응답자들에게 지속적으로 접촉을 해야 한다.

○ 직접관측

직접관측(direct observation)은 수량이나 사건을 관측자의 감각(시각, 청각, 미각)이나 물리적인 측정 장치를 사용해서 기록하는 것이다. 이 방법에서 관측자는 응답자의 역할을 한다. 예를 들어 작물 수확량을 추정하는 것, 좌석 안전벨트를 매지 않는 운전자의 수를 세는 것, 전자 장치로 TV 시청률을 측정하는 것, 기계적인 기구를 통해 공해를 측정하는 것, 항공사진을 해석해서 토지 사용을 평가하는 것 등이 이에 해당한다.

위와 같은 여러 가지 자료수집 방법 중 하나를 선택하려고 할 때 다음과 같은 사항들을 고려해서 적절한 방법을 택한다.

- 원하는 수준의 자료품질
- 조사 예산
- 설문지의 내용 (질문 종류, 응답선택항목 수, 질문 수)
- 자료수집기간
- 모집단 종류
- 표본추출틀에 포함되어 있는 접촉 관련 정보

5. 측정 오차의 관리

통계조사에서 측정 오차의 요인을 파악하고 오차를 줄일 수 있는 개선 방안을 알아내기 위해서는 측정 오차에 대한 지속적인 관리가 필요하다. 측정 오차를 수량

화하는 방법은 매우 어려우며 특별히 고안된 방법과 자료 수집 과정을 통해 이루어지는 경우가 많다. 따라서 조사 관리자는 측정 오차의 크기가 커서 추정값의 정확성에 중대한 영향을 미칠 가능성이 크다고 판단되면 전문가와 함께 측정 오차의 요인을 파악하고 그 크기를 줄일 수 있는 방안을 마련하여 실행해야 한다. 아래의 몇 가지 예는 측정 오차의 요인과 그 영향을 알아볼 수 있는 여러 가지 방법들 중 몇 가지 예이다.

- 면접자 변동이 크다고 판단될 때는 변동의 측정을 위한 소규모 연구를 진행 할 수 있다. 서로 다른 응답자들이 같은 면접자에 대해 또는 동일한 응답자가 서로 다른 면접자들에 대해 얼마나 다르게 반응하는지에 대한 연구를 진행할 수 있다.
- 조사가 끝난 후에 일정 시간이 지난 후 응답자들의 일정 부분을 선택하여 중요 항목에 대하여 재조사를 실시하면 실제 조사에서의 응답과 재조사에서의 응답의 차이를 비교하여 측정 오차의 요인과 그 크기를 가늠할 수 있다.
- 시험조사(pliot study)에서 중요 항목에 대하여 응답자에게 응답을 받고 질문 항목이 이해하기 쉬웠는지에 대한 여부를 물어서 각 질문에 대한 이해 정도를 파악할 수 있다.
- 시험조사나 재조사에서 같은 질문에 대하여 자료의 수집 방법을 달리 하여 서로 비교하고 측정 오차를 줄일 수 있는 방법을 고려한다.
- 조사의 보고서를 작성하거나 결과를 보고할 때 측정 오차를 줄이기 위해서 사용한 여러 가지 방법을 설명하는 것이 사용자에게 통계의 품질 수준을 파악하는데 큰 도움을 줄 것이다.

Ⅵ. 자료처리 및 기타 오차

1. 개요

자료처리는 자료수집과정에서 얻은 가공되지 않은 자료를 분석하고 배포할 수 있도록 가공하고 교정한 상태로 변환하는 작업이다. 자료의 처리에는 자료의 입력, 에디팅, 코딩, 대체, 가중 등과 같은 단계들이 있으며, 이러한 일련의 처리 과정을 거쳐 자료의 정확성은 향상된다. 현대에 들어와서 많은 자료 처리 과정들이 자동화되고 서로 통합되어 적은 비용과 인원으로 정확성을 향상시킬 수 있는 경우가 많아졌다. 예를 들어 CAPI(Computer Aided Personal Interview) 방법을 사용하는 조사는 자료의 수집 과정 중에 코딩이 실시되고 논리적 오류를 검사할 수 있는 에디팅 등이 동시에 수행되는 경우가 많다. 또한 자료의 입력도 컴퓨터 입력용 광학 필름 판독장치 등을 이용하여 사람이 입력하는 경우보다 더 정확하게 입력할 수 있는 환경이 구현되었다.

이러한 기술의 발달에 의한 자료처리 과정의 개선에도 불구하고 많은 조사통계의 자료처리 과정에서 많은 오차가 발생하고 있다. 예를 들어 질병에 대한 질문을 표준질병분류에 의해 코딩하거나 또는 직업에 대한 질문을 표준직업분류에 의해 자료를 코딩하는 과정은 전문적인 지식을 필요로 하는 까다로운 작업으로 오차가 발생할 가능성이 많다. 따라서 조사 과정에서 일어나는 자료처리에 대한 오차의 종류와 그 원인을 이해하는 것이 통계의 정확성을 높이기 위한 관리 대책을 세우는 데 매우 중요하다.

자료처리 오차에 대한 연구는 무응답에 의한 오차 또는 측정 오차에 대한 연구에 비해 많지 않다. 다른 오차들에 비해 조사 담당자들의 관심이 적고 또한 전산

기술의 발달과 처리 과정의 통합으로 과거에 비해 오차의 가능성이 줄어든 것은 사실이지만 조사의 전체 비용 중에 큰 부분을 차지하는 자료처리 과정에서 발생하는 오차의 영향을 너무 과소평가하면 안 된다. 특히 자동화와 통합에 의해 전에는 없었던 새로운 오차 발생의 가능성이 있으므로 항상 세심한 주의를 가지고 오차에 대한 관리 대책을 세워야 한다.

여기서는 자료처리에서 나타날 수 있는 자료 입력의 오류, 에디팅 오류, 코딩 오류에 대한 원인과 관리 방법에 대하여 알아보고자 한다. 또한 자료 공개와 잠정치를 생산하는 조사에서 고려되어야 할 정확성 점검 방법도 함께 알아보고자 한다. 대체(imputation)에 대한 내용은 보통 처리 오차에서 다루어 질 수 있지만 본 매뉴얼에서는 대체가 무응답 오차의 관리와 관계가 깊으므로 무응답 오차를 다루는 장에서 언급했다는 점을 밝혀둔다.

2. 자료 입력

자료 입력은 조사용지나 설문지에 기록된 정보를 컴퓨터가 읽을 수 있는 형태로 변환시키는 단계이다. 조사용지에 기록된 정보를 컴퓨터 파일의 형태로 자료를 기록하는 작업은 수작업에 의한 타자 입력, 표시문자 인식(mark character recognition: MCR), 지능형 문자인식(intelligent character recognition; ICR) 등이 있다. 최근에는 CAI(computer aided interview)를 통하여 자료 수집과 동시에 컴퓨터 파일의 형태로 자료 수집 결과가 입력이 되는 통합된 형태가 빈번하게 쓰이고 있다.

수작업을 통해 자료를 입력하는 타자 입력은 입력하는 수단이 사람이기 때문에 다른 방법과 비교할 때 오차가 발생할 가능성이 높으며 입력자에 따라 오차율이 크게 다를 수 있다. 이렇게 상대적으로 높은 오차율을 줄이기 위하여 타자 입력 시에는 많은 경우에 독립적 재검증을 실시한다. 독립적 재검증은 모든 자료 또는 적절하게 선택된 자료의 일부분에 대하여 두 명 이상의 입력자가 독립적으로 자료를 입력하고 서로 확인하는 과정을 말한다. 서로 다르게 입력된 자료가 발생하면 컴퓨터로 검사하여 판정하고 입력 자료를 수정한다. 이 방법은 조사통계에서 오래전부터 시행되어온 대표적인 품질관리 방법이다. 독립적 재검증을 통하여 입력자의 입력 오류율을 파악하여 오차의 원인을 제거하는 여러 가지 전략을 시행할 수 있다. 조사 자료에서 몇 개의 매우 심각한 타자 입력 오류에 대한 우려와 입력이 비교적 저렴한 비용으로 수행될 수 있다는 사실 때문에 많은 기관에서 타자 입력 품질관리와 관련된 활동이 광범위하게 이루어지게 되었다.

타자 입력의 오류율(keying error rates)은 몇 가지 요인에 의해 결정된다. 이 요인들에는 입력자의 경험, 입력자들 간의 오류율 차이, 입력자 변동비율, 입력 과정에 적용되는 품질관리의 양과 종류, 입력할 자료의 해독 난이도, 입력자가 입력해야 할 필드 확인의 어려움 정도, 타자 입력 양식 준비를 위해 수행되는 스캔 에디팅 과정의 양과 질 등이 있다.

오류율을 정의하는 방식이 다양하기 때문에 조사들과 기관들에 따른 타자 입력 오류율은 비교하기 어렵다. 타자에 의한 자료 입력 오차를 측정할 수 있는 지표는 다음과 같이 수량적으로 나타낼 수 있다.

$$\bullet \text{타자 입력 오류율} = \frac{\text{잘못 입력된 글자 수(필드수, 레코드 수)}}{\text{입력된 총 글자수(필드수, 레코드 수)}}$$

독립적 재검증을 실시하면 타자 입력 오류율은 많은 경우에 작은 것으로 나타나지만, 입력된 값이 실제값과 차이가 크게 잘못 입력되면 오류 효과는 상당히 클 수 있다. CAI 조사에서는 입력 오류가 중대한 문제가 아니라는 것이 많은 연구에 의해 입증되었으므로 CAI 방법을 쓰는 것이 타자 입력 오류를 줄이는데 큰 기여를 할 수 있다.

3. 스캔 오류

전산화 도구를 이용하여 자료를 입력 또는 스캔하는 경우, 즉 표시문자 인식, 지능형 문자인식, 음성 인식 등은 타자 입력에 비해서 상대적으로 오차의 발생 가능성이 적지만 입력 오류가 발생하지 않는 것은 아니다. 전산화 도구를 이용하는 경우 교체입력과 거부라는 두 종류의 오류가 발생할 수 있다. 교체입력 오류(substitution error)는 소프트웨어가 문자를 잘못 인식할 때 발생하며 거부 오류는(rejection error) 소프트웨어가 문자를 인식할 수 없어서 거부할 때 발생한다. 거부된 문자들은 수정하여 손으로 시스템에 다시 입력해야 한다. 따라서 거부 오류는 처리에 비용이 많이 들지만 바르게 처리하기만 하면 자료 입력 과정에 아무런 오류가 생기지 않는다.

스캔 오류율(scanning error rates)은 전산 도구 또는 소프트웨어가 자료를 정확하게 읽을 수 있는 성능에 따라 변한다. 따라서 입력 도구의 성능 향상을 위하여 전산화 도구와 소프트웨어의 지속적인 개선이 필요하다. 스캔 오류의 정도를 측정할 수 있는 지표는 다음과 같다.

$$\bullet \text{스캔 오류율} = \frac{\text{스캔이 잘못된 글자 수(필드수, 레코드 수)}}{\text{입력된 총 글자수(필드수, 레코드 수)}}$$

전산화 도구를 이용하여 자료를 입력하는 경우에는 타자 입력에 비해 오차가 발생할 가능성이 매우 적고, 비용이 적게 든다는 장점이 있지만 많은 적용에서 설사 교체입력 오류율이 작다고 할지라도 그 결과로 생기는 오류는 상당히 심각할 수 있다. 예를 들어 스캔에 쓰이는 기계의 광학렌즈에 이물질이 들어가 심각한 체계적인 오류(systematic error)를 발생시킬 수 있다. 따라서 전산화 도구에 대한 지속적인 점검과 성능 향상이 중요하다.

4. 에디팅

에디팅(editing)은 통계 산출에 사용되는 각각의 자료에 존재하는 오류와 이상점을 확인하고 만일 필요하면 수정하는 것이다. 즉 자료의 대체와 요약 절차를 수행하기 전에 되도록 많은 잘못된 자료를 수정할 목적으로 잘못되거나 의심스러운 조사 자료를 찾아서 수정하는 모든 절차들을 에디팅이라고 한다. 에디팅은 다음 과정, 즉 통계를 산출하거나 추정값을 계산하는 과정을 가능하게 하는 자료의 정리 단계로 생각할 수 있다. 따라서 에디팅은 조사의 시작 (예를 들어, CAPI에서의 현장 자동 에디팅)부터 마지막 단계(예를 들어, 추정값을 외부 자료와 비교)까지 여러 가지 다양한 형태로 실시되고 있다

에디팅은 현장에서 조사한 자료의 타당성을 여러 가지 체계적인 방법으로 조사하기 때문에 자료의 품질에 대한 유용한 정보를 제공한다. 전체적으로 수집된 자료에서 에디팅 한 자료의 비율이 늘어난다면 이는 통계조사의 품질이 좋지 않다는 것을 의미한다. 또한 전체적인 조사의 품질에 대한 정보뿐만 아니라 면접원별로 수집된 자료의 에디팅 비율을 계산할 수 있으므로 면접원들이 그들의 역할을 어떻게 수행했는지 그리고 응답자들이 정확한 정보를 제공했는지에 대한 품질을 평가할 수 있는 정보를 얻을 수 있다. 더 나아가 지역별, 시간대별 등으로 자료에 대한 다양한 품질 지표를 얻는 것이 가능하다. 한편, 에디팅을 위해 설문지들을 검사하고 컴퓨터 분석을 통하여 자료를 재검증하는 과정을 통해 조사품질과 관련된 유용한 정보를 얻을 수 있으며, 이런 정보는 향후 조사에서 오류를 줄일 수 있는 조사의 개선 전략을 수립하는 데 큰 도움이 될 수 있다. 이런 이유 때문에 조사 설계자들이 자료의 근본적인 오류 원인들을 이해하기 위해 노력하는 것이 중요하며, 이를 통해 자료수집이나 자료처리상의 문제점들을 보완할 수 있다.

자료 에디팅은 조사 자료에 적용되는 규칙으로 구성되어 있다. 예를 들어, 어떤 값은 파일의 특정한 위치에 항상 존재해야 하고, 변수에 따라 취할 수 있는 값이 제한적이며, 특정 변수들에 대응되는 값들의 조합에도 제한이 있고, 합계에 해당하

는 것은 그것을 구성하는 값들의 합과 일치해야 한다. 또한 특정한 변수의 값은 미리 설정된 구간 안에 반드시 포함되어야 한다. 이런 에디팅들을 결정적 에디팅(deterministic edits)이라고 부르며, 만일 이것들이 지켜지지 않으면 틀림없이 오류가 있다는 것을 나타낸다. 결정적 에디팅은 자료가 사용 가능하기 위하여 반드시 수정되어야 하는 잘못된 값들을 찾아내는 중대한 에디팅(critical editing), 의심스러운 값들을 확인하는 의문 에디팅(query editing), 총괄(aggregate) 자료(예를 들어, 평균이나 총계)를 대상으로 수행되는 또는 기록 전체를 대상으로 하는 검사인 매크로 에디팅(macro editing) 등이 있다. 매크로 에디팅은 규칙에 근거하여 총계가 의심스럽다고 생각되면 비정상적인 총계를 산출하게 된 원인을 한 개나 몇 개의 잘못된 레코드에서 찾을 수 있는지 확인하기 위해 해당 총계를 구성하는 각각의 레코드를 검사한다. 그렇지만 만일 총계가 의심스럽지 않으면, 그 총계를 구성하는 각각의 레코드들은 에디팅 과정을 통과한 것으로 본다.

최근 연구에서 얻어진 에디팅에 대한 중요한 교훈은 자료에 있는 모든 오류를 고치기는 것보다 선택적 에디팅(selective editing)을 수행해야 한다는 것이다. 선택적 에디팅에서는 의심스러운 항목을 100% 재검토하는 대신 의문 에디팅 대상 중에서 일부를 표본단위의 중요성, 연구대상 변수의 중요성, 오류의 심각성 그리고 의심스러운 항목을 자세히 조사하는데 드는 비용 등을 근거로 선택하는 것이다. 표본단위의 중요성은 단위의 크기나 추정 과정에서 단위에 부여되는 가중값을 근거로 한다. 최근 들어 매크로 에디팅과 선택적 에디팅에 대한 관심이 높아지고 있다.

통계조사에서 각 항목에 대한 에디팅률(editing rates)을 계산하면 해당하는 항목에 대한 측정 가능한 품질 지표를 얻을 수 있다.

$$\bullet \text{ 에디팅률} = \frac{\text{에디팅을 통해 수정된 조사단위의 수}}{\text{항목에 응답해야 되는 총 조사단위의 수}}$$

총 조사예산 중 많은 부분이 에디팅에 투여된다. 에디팅을 통해 얻은 이득을 최대한 활용하기 위하여 에디팅 단계에서 얻은 결과들을 조사 과정을 개선하는데 사용해야 한다. 다시 말해, 설문지, 면접원 교육, 또는 에디팅 실패의 원인이 될 수 있는 다른 조사 과정들을 개선하기 위하여 에디팅에서 얻은 정보를 적극 활용해야 한다. 또한 잠재적인 오류를 더 효과적으로 찾아내는 방법을 찾아서 에디팅 과정 자체를 개선할 필요가 있다.

5. 코딩

코딩이 모든 조사에 필요한 것은 아니지만, 조사결과에 나쁜 영향을 주는 잠재적

인 오류 발생 원인이면서 동시에 많은 조사에서 매우 중요한 작업이다. 코딩은 흔히 서술형 질문에 대한 응답으로 얻어지는 원시 자료를 분류하는 작업으로, 추정, 표 작성, 분석 등을 하기에 적절하도록 코드번호나 범주값을 부여하는 분류 과정이다. 코딩은 작업자나 코딩 담당자에 의해 수작업으로 실시되거나 특수하게 설계된 코딩 소프트웨어로 자동으로 실시될 수 있다.

코딩은 응답한 자료를 각 변수에 대하여 미리 정해진 특정한 코드번호를 부여하는 작업이며 코딩은 미리 정해진 지침이나 규칙에 따라 수행된다. 코딩 과정에서 많은 문제들이 발생할 수 있는데 이를 인지하고 관리하는 것은 매우 어렵다. 코딩 규칙들은 항상 바르게 적용되지 않고 코딩 규칙 자체가 불완전하기 때문에 오차가 발생한다. 예를 들어, 가능한 모든 서술형 응답을 처리할 수 있게 명명법을 구성할 수 없고, 매우 노련한 코딩 담당자들도 알맞은 코드번호에 대하여 의견이 일치하지 않는 경우가 많다. 코딩은 때로 매우 주관적인 행동이기 때문에 우수한 코딩 방법을 개발하는 것은 어렵다. 어떤 서술형 응답들은 코드번호를 명확하게 부여하기에 적당하지 않을 수 있기 때문에 담당자들은 응답결과를 코딩하기 위하여 주관적인 판단을 해야 한다.

전형적으로 코딩을 필요로 하는 변수들에는 산업 및 직업 분류, 대학 전공 분야, 업무 장소 등이 있다. 이런 변수에 대해서는 몇 백 개의 코드번호나 범주가 명명법에 주어져 있을 수 있다. 각 요소 및 대상 변수에 대하여 올바른 코드번호가 항상 존재한다고 가정한다. 코딩 오류는 요소에 올바른 코드번호가 아닌 다른 것을 부여했을 때 발생 한다. 흔히 응답이나 코딩할 요소의 애매함 때문에 올바른 코드번호를 결정하기 어렵다. 더 나아가 코딩 오류라기보다 오히려 응답자에 의해 발생한 측정 오차에 해당하는 경우도 많다. 자세한 응답이 주어졌다고 할지라도 정확한 코드번호를 부여하는데 문제가 발생할 수 있다. 코딩 전문가들 사이에 변동이 상당히 있을 수 있고, 그 결과 정확한 코드번호는 운영상의 규칙을 통해서 정의해야 한다. 이런 운영상의 규칙은 다수결의 원칙이다. 예를 들면 두 명 이상의 코딩 전문가가 독립적으로 응답을 코딩한다고 하자. 그러면 정확한 코드번호는 전문가들 중 다수가 부여한 것으로 정한다. 과반수에 이르지 못한 경우에는 추가 규칙들이 사용되어야 한다.

코딩 오류율은 다양한 방법으로 계산될 수 있다. 특정 조사 변수, 특정 코드번호나 코드번호 단계(자리 수), 개별 코딩 담당자들에 대한 오류율 등이 계산될 수 있다. 다음은 일반적은 코딩 오류율(coding error rates)을 계산하는 식이다.

$$\bullet \text{ 코딩오류율} = \frac{\text{코딩 오류의 수}}{\text{코딩된 전체 단위의 수}}$$

코딩오류율의 정확한 계산은 매우 어려우며 코딩된 총 단위들을 재조사하는 방

법 등으로 추정할 수 있다.

코딩의 방법은 크게 수동 코딩과 자동화 코딩으로 나뉜다. 수동 코딩을 관리하기 위해 사용되는 기본적인 두 가지 방법으로 종속검증과 독립검증이 있다. 종속검증(dependent verification)에서는 코딩 담당자 A가 한 요소를 코딩한다. 이렇게 입력된 코드번호는 검증자인 B에 의해 재검토 된다. B는 코드번호를 면밀히 살피고 그것이 정확한지를 확인한다. 만일 그것이 맞는다고 생각되면 그대로 두고, 그렇지 않으면 그가 맞는다고 생각하는 것으로 코드번호를 바꾼다. 독립검증(independent verification)에서는 검증자 입장에서 최초 코딩을 기초로 한 판단을 할 필요가 없다(즉, 검증자는 처음 부여된 코드번호를 알 수 없다). 독립검증에서는 코딩 담당자 A가 어떤 요소에 코드번호 X를 부여하고, 같은 요소를 두 번째 코딩 담당자인 B가 다시 코딩하는데, 코드번호 Y를 부여한다. 두 코딩 담당자들이 독립적으로 코딩을 하기 때문에 부여한 코드번호(X, Y)를 서로 모른다. 두 코드번호를 비교하여 최종 코드번호를 결정하기 위하여 결정 규칙을 정하고 적용한다. 코드번호와 관련된 단어나 단어의 일부를 포함하는 사전이나 데이터베이스가 컴퓨터에 저장되어 있어야 한다. 이것이 수동 코딩에서 사용되는 명명법 설명서에 상응하는 역할을 한다.

자동화 코딩은 응답이 온라인으로 또는 스캐닝이나 타자 입력과 같은 방법으로 컴퓨터에 입력된다. 응답들은 사전의 해설들과 대응시키고, 그 대응에 수반하는 결정 규칙을 근거로 코딩 번호를 부여하거나 응답들을 코딩을 하지 않은 채로 놓아둔다. 자동화 코딩(automated coding)은 일괄처리방식 또는 컴퓨터 보조방식으로 수행될 수 있다. 처리되지 못하고 남겨진 모든 코딩은 수동으로 처리한다.

6. 자료 공개와 잠정치

모든 국가 통계기관들과 많은 조사기관들은 매크로 자료와 마이크로 자료를 외부로 공개하는 것과 관련된 지침을 갖고 있다. 매크로 자료(macro data)는 통계표, 총계, 빈도 등을 포함하는 파일을 말하고, 마이크로 자료(micro data)는 사람, 가구, 사업체 등 개별 조사 단위들에 대한 자료를 담고 있는 레코드들의 파일을 가리킨다. 노출방지(disclosure avoidance)라는 용어는 모집단에 있는 특정 단위가 표본에 있는 단위와 매칭이 되어 일반적으로 알려지지 않은 특정 단위에 대한 정보가 노출될 수 있을 때, 자료 공개에 따른 노출 위험을 줄이려는 다각적인 노력을 포괄적으로 의미한다. 따라서 통계표나 마이크로 자료를 공개하는 경우에 항상 노출 위험 허용 수준을 사전에 검토해 개별 단위에 대한 정보가 노출되지 않도록 해야 한다.

노출방지를 위해 적용된 기법들은 정확하게 기술되어야 하고, 기법이 적용된 뒤에는 품질관리 차원에서 다음과 같은 내용들이 점검되어야 한다.

- 노출방지 기법들이 적용된 자료에서 원시자료(original data)에 나타난 항목들의 관계가 계속 유지되는가를 점검해야 한다. 이는 변수들의 상관관계(correlation coefficient)를 비교하거나 여러 가지 통계적 측도를 사용해 점검할 수 있다.
- 노출방지 기법들이 적용된 자료에서 산출된 총합 또는 평균들이 원시자료의 총합 또는 평균과 크게 차이가 나는지 점검해야 한다. 이는 가공된 자료의 총합이나 평균이 실제 값과 얼마나 다른지를 나타내는 상대 오차 등으로 점검할 수 있다.
- 노출방지 기법들이 적용된 테이블의 도수 또는 분포가 원시자료의 도수 또는 분포와 크게 차이가 나는지 점검해야 한다. 이는 가공된 자료의 테이블에서 계산된 카이제곱 통계량을 실제 테이블에서 계산된 값과 얼마나 다른지 비교하는 방법으로 점검할 수 있다.

많은 경제관련 통계에서는 사용자의 시급한 요구에 부응하기 위하여 초기에는 잠정치(provisional statistics)를 발표하고, 일정 기간이 경과한 후에 최종치(final statistics)를 발표하는 경우가 많다. 특별히 월별 또는 분기별로 발표되는 경제 지표는 잠정치의 발표가 매우 흔한 일이다. 잠정치와 최종치의 차이는 일종의 오차로 볼 수 있으며, 이런 오차를 줄이기 위해서는 잠정치와 최종치에 대한 지속적인 비교를 통해 그 차이를 줄일 수 있도록 노력해야 한다. 실제로 잠정치와 최종치의 차이는 잠정치를 발표할 때 자료의 일부분이 결측된 상태이기 때문에 적절한 방법으로 결측치를 대체하기 때문에 발생하는 경우가 많다. 잠정치의 발표 이후 결측되었던 자료들이 대부분 회복이 되면 최종치를 발표하게 되는데, 이런 과정에서 잠정치와 최종치의 차이, 즉 오차가 어떤 요인에 의해 생기는지 파악하여 오차를 줄일 수 있는 대체 방법을 개발하는 것이 바람직하다. 잠정치를 생산하는 조사에서 통계품질을 향상시키기 위해서는 일정 기간 동안의 잠정치와 최종치간의 평균절대차이(absolute mean revision) 등을 지속적으로 계산하여 관리하는 것이 필요하다.

$$\bullet \text{ 평균절대차이 } = \frac{1}{n}\sum_{i=1}^{n}|P_i - F_i|$$

여기서 n은 분석대상 기간이며(예, n=12 개월) P_i와 F_i 는 각각 해당하는 기간의 잠정치와 최종치이다.

부록 A. 사용자 가이드라인

A.0 개요

사용자 가이드라인에서는 사용자가 본 매뉴얼에서 설명한 정확성 향상을 위해 고려해야 하는 중요한 주제들을 스스로 점검할 수 있는 방법을 요약하여 제공한다. 주어진 주제별로 정확성에 대한 요인의 개념과 사용자가 고려해야할 사항을 간략하게 기술하고 측정 가능한 지표를 제시함으로서 사용자가 담당 통계의 정확성을 향상하려는 업무에 도움을 주려고 한다. 유의할 점은 가이드라인에 제시된 모든 지표가 항상 계산 가능한 것은 아니므로, 측정 가능한 지표의 선택은 조사의 종류나 성격에 따라서 통계 작성자가 판단하여 정한다.

A.1 표본추출오차

(1) [표본추출] 표본추출방법

표본추출방법에는 확률표본추출과 비확률표본추출이 있다. 단순확률추출, 계통추출, 집락추출, 층화추출 등이 대표적인 확률추출법에 해당하고 대규모 표본조사의 경우 흔히 이런 추출법들을 단계별로 적용한 좀 더 복잡한 형태의 복합표본설계를 사용한다. 추정량의 계산은 표본추출방법에 의해 정해지므로 사용자에게 표본설계 내용을 정확하게 전달해야한다.

(2) [표본추출] 추정량에 대한 표준오차 또는 상대표준오차

통계 추정량의 표준오차(standard error)는 조사에서 얻어진 추정값이 실제값과 얼마나 가까운지를 나타내는 측도이기 때문에 통계조사의 정확성을 파악하는데 가장 중요한 지표이다. 표준오차는 추정량의 분산의 제곱근이며 표본추출 방법에 따라 달라진다. 표준오차 대신에 상대표준오차를 의미하는 변동계수(coefficient of variation)를 계산하여 추정결과의 정확성을 나타낼 수도 있다. 보고서에는 표준오차나 변동계수를 구하는 방법도 기술해야한다.

(3)[표본추출] 변동 추정에 대한 표준오차 또는 상대표준오차

시간에 따른 추정값의 변동은 매우 중요한 관심대상이다. 따라서 주요변수들의 절대 변동이나 상대 변동에 대한 표준오차를 구한다.

$$\text{절대 변동} : var(\widehat{Y}_2 - \widehat{Y}_1) = var(\widehat{Y}_2) + var(\widehat{Y}_1) - 2cov(\widehat{Y}_1, \widehat{Y}_2)$$

$$\text{상대 변동} : var(\frac{\widehat{Y}_2}{\widehat{Y}_1}) \approx (\frac{\widehat{Y}_2}{\widehat{Y}_1})^2 [\frac{var(\widehat{Y}_1)}{\widehat{Y}_1^2} + \frac{var(\widehat{Y}_2)}{\widehat{Y}_2^2} - \frac{2cov(\widehat{Y}_1, \widehat{Y}_2)}{\widehat{Y}_1 \widehat{Y}_2}]$$

(4) [표본추출] 표본으로 추출된 단위들의 개수와 그에 대한 정보

표본조사에서는 추출된 조사 단위들의 개수가 정확성을 측정하는데 매우 중요한 요소이다. 따라서 표본의 개수를 정하는 방법을 자세히 기술하고 조사에서 얻고자 하는 목표정도의 형태와 값에 대하여 정확히 기술한다. 예를 들어 추정량의 오차가 95% 신뢰수준에서 오차의 한계 B를 넘지 않도록 하는 것을 목표로 표본의 개수가 n으로 결정된 것을 사용자에게 전달해야 한다.

표본의 개수, 목표정도(신뢰수준, 오차의 한계), 표본추출률, 유효표본의 크기 등이 측정 가능한 지표이다.

(5) [표본추출] 주요 변수들의 설계효과

설계효과는 조사에서 사용된 표본추출방법이 단순확률추출법보다 얼마나 좋은지를 분산의 비율을 통하여 나타내는 측도이다. 설계효과는 집락추출 등에서 표본의 크기를 결정할 때 사용된다.

$$\text{설계효과}(Deff) = \frac{\text{주어진 표본추출법에서 } Var(\hat{Y})}{\text{동일한 표본크기의 } SRS\text{에서 } Var(\hat{Y})}$$

(6) [표본추출] 추정값을 계산할 때 사용된 가중치와 그 계산방법

표본조사에서 편향이 없는 추정결과를 얻기 위해서는 적절한 가중치를 반영한 추정식을 사용하는 것이 필요하다. 가중치를 산출한 방법에 대한 내용은 사용자에게 매우 중요하며, 이는 조사의 품질을 판단하는데 중요한 역할을 한다. 따라서 가중치를 정할 때 고려되어진 사항들, 즉 추출확률의 산출, 무응답 가중치 조정, 사후층화 등과 관련된 사항들을 정확하게 기술한다.

A.2 추출틀오차

(1) [추출틀오차] 미포함에 의한 누락

목표모집단의 구성원이 추출틀에 누락되는 문제가 추출틀 결함 중 가장 빈번하게 발생하면서 가장 문제가 심각한 유형이라고 볼 수 있다. 미포함에 의한 누락의 원인을 파악하고 그 비율을 추정한다. 또한 조사의 정확성에 미치는 영향에 대하여 기술하며, 가능하면 미포함에 의하여 발생되는 추정량의 편향에 대한 추정값을 구한다.

$$\frac{\text{추출틀에 누락된 단위의 수}}{\text{목표모집단 전체 단위의 수}}$$

(2) [추출틀오차] 과다포함

목표모집단의 구성원이 아닌 조사단위들이 추출틀에 포함되는 것을 과다포함이라고 하며, 이런 과다포함 오차를 제거하지 않으면 총계를 과대 추정할 가능성이 높다. 과다포함의 원인을 파악하고 조사의 정확성에 미치는 영향을 기술한다.

(3) [추출틀오차] 중복

추출틀상에 있는 두 개 이상의 단위가 목표모집단의 동일한 단위와 대응이 되는 경우 중복에 의해 오차가 발생할 수 있다. 이러한 중복의 원인을 파악하고 그 비율을 추정한다. 또한 조사의 정확성에 미치는 영향에 대하여 기술한다.

$$\frac{\text{추출틀에 중복 기재된 단위의 수}}{\text{추출틀에 등재된 전체 단위의 수}}$$

(4) [추출틀오차] 부적격단위

추출틀상에 부적격단위가 존재하면 과다 포함과 유사한 오차가 발생할 수 있다. 이러한 부적격단위의 포함 원인을 파악하고 그 비율을 추정한다. 또한 조사의 정확성에 미치는 영향에 대하여 기술한다.

$$\frac{\text{추출틀에 등재된 부적격 단위의 수}}{\text{추출틀에 등재된 전체 단위의 수}}$$

(5) [추출틀오차] 오분류

보조변수를 사용해 단위를 분류할 때 부정확한 정보에 의해 분류 오차가 발생할 수 있다. 이러한 오분류 비율을 추정하고 그 원인을 분석한다. 또한 조사의 정확성에 미치는 영향에 대하여 기술한다.

$$\frac{\text{추출틀에 등재된 적격 단위 중 오분류된 단위의 수}}{\text{추출틀에 등재된 적격 단위의 수}}$$

A.3 무응답오차

(1) [무응답오차] 단위무응답에 대한 응답률 (가구조사의 경우)

무응답의 유형 중 추출단위에 대한 조사를 통해 응답거절 등의 이유로 어떤 응답도 얻지 못한 경우를 단위무응답이라고 한다. 응답률은 통계조사의 품질을 파악할 수 있는 측정 가능한 지표 중에 가장 중요한 지표 중 하나이며 가구조사에 대한 응답률에는 다음과 같은 것들이 있다.

$$\text{전체응답률} = \frac{I+P}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

$$\text{완전응답률} = \frac{I}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

$$\text{협조율} = \frac{I+P}{(I+P)+(R+O)+e_c UC}$$

$$\text{접촉률} = \frac{(I+P)+R+O+e_c UC}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

$$\text{거부율} = \frac{R}{(I+P)+(R+NC+O)+e_c UC+e_n UN}$$

I = 완전응답(complete interview)
P = 부분응답(partial interview)
R = 응답거부(refusal)
NC = 접촉불능 (noncontact)
O = 그 밖의 모든 무응답 (other nonresponse)
UC = 무적격단위인지 판단이 불가능 - 접촉 가능한 경우
UN = 무적격단위인지 판단이 불가능 - 접촉 불가능한 경우
e_c = 무적격단위인지 판단이 불가능(접촉 가능한 경우) 중 적격단위의 비율
e_n = 무적격단위인지 판단이 불가능(접촉 불가능한 경우) 중 적격단위의 비율

(2) [무응답오차] 단위무응답에 대한 응답률 (경기조사의 경우)

무응답의 유형 중 추출단위에 대한 조사를 통해 응답거절 등의 이유로 어떤 응답도 얻지 못한 경우를 단위무응답이라고 한다. 응답률은 통계조사의 품질을 파악할 수 있는 측정 가능한 지표 중에 가장 중요한 지표 중 하나이며 경기조사에 대한 응답률에는 다음과 같은 것들이 있다.

$$\text{전체응답률} = \frac{FC+FP+PC+PP}{(FC+FP+PC+PP)+RNU+NR+e(U)}$$

$$\text{완전응답률} = \frac{FC}{(FC+FP+PC+PP)+RNU+NR+e(U)}$$

FC = 모든 기간에 대한 완전응답 (Full period return with complete data)
FP = 모든 기간에 대한 부분응답 (Full period return with partial data)
PC = 부분 기간에 대한 완전응답 (part period return with complete data)
PP = 부분 기간에 대한 부분응답 (part period return with partial data)
RNU = 응답하였으나 쓰이지 못하는 경우(returned but not used)
NR = 무응답 (non-response)
U = 무적격단위인지 판단이 불가능 (cases of unknown eligibility)
e = 무적격단위인지 판단이 불가능한 단위중 적격인 비율

(3) [무응답오차] 주요 변수들에 대한 항목 응답률

완성된 설문지를 얻었더라도 설문지에 있는 모든 항목이 완성되지 않았을 수 있다. 이와 같이 일부 항목에 대한 응답이 누락된 경우를 항목무응답이라고 한다. 응답률은 통계조사의 품질을 파악할 수 있는 측정 가능한 지표 중에 가장 중요한 지표 중 하나이며 항목 응답률은 다음과 같이 계산할 수 있다.

$$\frac{\text{항목에 응답한 조사단위의 수}}{\text{항목에 응답해야 하는 조사단위의 수}}$$

(4) [무응답오차] 무응답에 의한 편향과 분산

무응답에 의한 편향은 응답자와 무응답자간의 체계적인 차이에 의해 생기는 편향이다. 무응답이 많아지면 무응답에 의한 편향이 생길 가능성도 많아지며 목표한 추정량의 정도(precision)를 만족시킬 수 없게 된다. 즉 무응답이 발생함으로 인하여 표본의 실제 크기는 감소하며 추정량의 분산은 일반적으로 증가한다.

무응답에 의한 편향이 어느 정도 발생하는지 또는 무응답 발생에 의해 추정량의 정확성이 얼마나 감소하는지 등은 무응답의 특성, 즉 실제값을 조사할 수 없다는 무응답 특성 때문에 수량적으로 측정하는 것은 현실적으로 매우 어렵다. 하지만 응답자와 무응답자의 차이를 여러 가지 다양한 방법으로 분석해 보면 그 차이점이 파악되고 그 차이의 성격과 크기에 따라 무응답에 의한 편향과 추정량의 정확성에 미치는 영향을 어느 정도를 추정할 수 있는 경우도 있다.

(5) [무응답오차] 대체율

결측치 대체는 무응답으로 발생한 결측 값을 다양한 방법을 통해 특정한 값으로 대체하는 것을 말한다. 무응답 단위와 유사한 속성을 갖는 응답 단위 중에서 하나를 선택하여 무응답을 대체하는 핫덱대체법 또는 최근방대체법 등이 있다. 핫덱대체법이나 최근방대체법은 단위 무응답인 경우에도 효과적으로 사용할 수 있다. 또한 결측된 자료를 응답자들의 평균으로 대체하는 평균대체법, 두 개의 관련된 변수의 비(ratio)를 사용하는 비대체법, 관련 변수들을 이용한 회귀모형을 만들어 그 추정값으로 대체하는 회귀대체법 등이 있다.

단위 무응답과 항목 무응답의 처리 방법 또는 결측치 대체법은 반드시 사용자에게 설명되어져야 한다. 또한 결측치 처리에 대한 측정 가능한 지표를 제공하는 것이 바람직하다.

- 결측치 대체로 대체된 단위의 비율

$$\frac{\text{결측치 대체된 단위의 수}}{\text{전체 추출 단위의 수}}$$

- 중요 항목에서 결측치가 대체된 비율

$$\frac{\text{결측치 대체된 단위의 수}}{\text{항목에 응답해야 하는 단위의 수}}$$

- 중요 항목의 추정에서 결측치 대체가 차지하는 비율

$$\frac{\text{결측치 대체된 자료에 대한 가중치 적용 합계}}{\text{모든 단위의 자료에 대한 가중치 적용 합계}}$$

A.4 측정 오차

(1) [측정오차] 측정 오차 감소를 위한 대책

통계조사에서 측정 오차의 요인을 파악하고 오차를 줄일 수 있는 개선 방안을 알아내기 위하여 측정 오차에 대한 지속적인 관리가 필요하다. 조사에서 채택된 측정 오차를 줄이기 위한 방안이 있다면 사용자에게 그 목적, 방법과 결과 등을 제공하는 것이 바람직하다.

(2) [측정오차] 면접원 변동

면접원 변동은 같은 질문에 대하여 같은 대상에게 면접하여 응답을 받았을 때 면접원에 따른 차이를 말한다. 예를 들어 같은 답을 가진 응답자가 다른 면접원에게 상이한 응답을 하는 경우에 발생한다. 예를 들어 개방형 질문에서는 면접원의 태도와 생각에 따라 응답이 달라질 가능성이 높다.

면접원 변동이 크다고 판단될 때는 변동의 측정을 위한 소규모 연구를 진행 할 수 있다. 서로 다른 응답자들이 같은 면접원에 대해 어떻게 다르게 반응하는지에 대한 연구를 진행할 수 있다.

(3) [측정오차] 자료수집방법

자료수집방법은 조사 대상자를 접촉하고 질문에 대한 응답을 얻는데 사용되는 수단을 가리킨다. 현재 사용되는 자료수집방법은 응답자와 접촉의 정도, 자료수집자인 면접원의 참여 정도, 자료 수집 도구 등에 의해 면접조사, 전화조사, 우편조

사, 일지조사, 직접관측 등으로 분류될 수 있다. 자료수집방법의 선택에 따른 오차의 발생 가능성과 조사 정확성에 대한 영향을 고려하여 적절한 방법을 선택한다.

(4) [무응답오차] 설문지

설문지의 설계는 통계조사의 품질을 결정하는 중요한 요소이다. 설문지 설계단계부터 정확한 목표를 설정하고 지속적으로 관리하는 것이 필요하다. 시험조사에서 중요 항목에 대하여 응답자에게 응답을 받고 질문이 이해하기 쉬웠는지 등에 대한 질문을 통해 질문에 대한 이해 정도를 파악할 수 있다.

A.5 자료처리 오차

(1) [자료처리 오차] 자료처리 방법과 품질 관리

자료처리는 자료수집과정에서 얻은 가공되지 않은 자료를 분석하고 배포할 수 있도록 가공하고 교정한 상태로 변환하는 작업이다. 자료의 처리에는 자료의 입력, 에디팅, 코딩, 대체, 가중 등과 같은 단계들이 있으며, 이러한 일련의 처리 과정을 거쳐 자료의 정확성은 향상된다. 또한 정확성 향상을 위해 채택된 품질 관리 기법들을 각 단계에 맞추어 개발하여 적용해야 한다.

(2) [자료처리 오차] 타자입력 오류율

수작업을 통해 자료를 입력하는 타자입력은 입력하는 수단이 사람이기 때문에 다른 방법과 비교할 때 오차가 발생할 가능성이 높다. 독립적 재검증을 이용하여 오류를 줄일 수 있지만 오류의 영향은 상당히 클 수 있다.

$$\text{타자 입력 오류율} = \frac{\text{잘못 입력된 글자 수(필드수, 레코드 수)}}{\text{입력된 총 글자수(필드수, 레코드 수)}}$$

(3) [자료처리 오차] 스캔 오류율

스캔 오류율은 전산 도구 또는 소프트웨어가 자료를 정확하게 읽을 수 있는 성능에 따라 변한다. 따라서 입력 도구의 성능 향상을 위하여 전산화 도구와 소프트웨어의 지속적인 개선이 필요하다. 스캔 오류의 정도를 측정할 수 있는 지표는 다음과 같다.

$$\text{스캔 오류율} = \frac{\text{스캔이 잘못된 글자 수(필드수, 레코드 수)}}{\text{입력된 총 글자수(필드수, 레코드 수)}}$$

(4) [자료처리 오차] 에디팅률

에디팅(editing)은 통계 산출에 사용되는 각각의 자료에 존재하는 오류와 이상점을 확인하고 만일 필요하면 수정하는 것이다. 즉 자료의 대체와 요약 절차를 수행하기 전에 되도록 많은 잘못된 자료를 수정할 목적으로 잘못되거나 의심스러운 조사 자료를 찾아서 수정하는 모든 절차들을 에디팅이라고 한다.

에디팅은 현장에서 조사한 자료의 타당성을 여러 가지 체계적인 방법으로 검증하기 때문에 자료의 품질에 대한 유용한 정보를 제공한다. 통계조사에서 각 항목에 대한 에디팅률(editing rates)을 계산하면 해당하는 항목에 대한 측정 가능한 품질 지표를 얻을 수 있다.

$$\text{에디팅률} = \frac{\text{에디팅을 통해 수정된 조사단위의 수}}{\text{항목에 응답해야 되는 총 조사단위의 수}}$$

(5) [자료처리 오차] 코딩오류율

코딩은 흔히 서술형 질문에 대한 응답으로 얻어지는 원시 자료를 분류하는 작업으로, 추정, 표 작성, 분석 등을 하기에 적절하도록 코드번호나 범주값을 부여하는 분류 과정이다. 코딩은 응답한 자료를 각 변수에 대하여 미리 정해진 특정한 코드번호를 부여하는 작업이며 코딩은 미리 정해진 지침이나 규칙에 따라 수행된다.

코딩 오류율은 다양한 방법으로 계산될 수 있다. 특정 조사 변수, 특정 코드번호나 코드번호 단계(자리 수), 개별 코딩 담당자들에 대한 오류율 등이 계산될 수 있다. 코딩오류율의 정확한 계산은 매우 어려우며 코딩된 총 단위들을 재조사하는 방법 등으로 추정할 수 있다. 다음은 일반적은 코딩 오류율(coding error rates)을 계산하는 식이다.

$$\text{코딩오류율} = \frac{\text{코딩 오류의 수}}{\text{코딩된 전체 단위의 수}}$$

A.6 기타 오차

(1) [기타 오차] 정보공개

노출방지를 위해 적용된 기법들은 정확하게 기술되어야 하고, 기법이 적용된 뒤에는 품질관리 차원에서 다음과 같은 내용들이 점검되어야 한다.

- 노출방지 기법들이 적용된 자료에서 원시자료에 나타난 항목들의 관계가 계속 유지되는가를 점검해야 한다.
- 노출방지 기법들이 적용된 자료에서 산출된 총합 또는 평균들이 원시자료의 총합 또는 평균과 크게 차이가 나는지 점검해야 한다.
- 노출방지 기법들이 적용된 테이블의 도수 또는 분포가 원시자료의 도수 또는 분포와 크게 차이가 나는지 점검해야 한다.

(2) [기타 오차] 잠정치

많은 경제 관련 통계에서는 초기에는 잠정치(provisional statistics)를 발표하고 일정 시간이 지나서 최종치(final statistics)를 발표하는 경우가 많다. 이러한 경우 잠정치와 최종치에 대한 지속적인 비교를 통해 그 차이를 줄일 수 있도록 노력해야 한다.

잠정치를 생산하는 조사에서 통계품질을 향상시키기 위해서는 일정 기간 동안의 잠정치와 최종치간의 평균절대차이(absolute mean revision)를 지속적으로 계산하여 관리하는 것이 필요하다.

$$\text{평균절대차이} = \frac{1}{n}\sum_{i=1}^{n}|P_i - F_i|$$

여기서 n은 분석대상 기간이며, P_i와 F_i는 각각 해당하는 기간의 잠정치와 최종치이다.

부록 B. 주요 용어 설명

• **가중치 (weight)**
표본조사에서 편향이 없는 추정결과를 얻기 위해서는 적절한 가중치를 반영한 추정식을 사용하는 것이 필요하며, 다양한 통계모형을 기초로 한 분석을 하는 경우에도 가중치를 적용함으로써 올바른 분석결과를 얻을 수 있음. 일반적으로 조사 자료의 분석에서는 추출확률을 기초로 한 설계가중치, 무응답 조정 가중치, 사후층화 가중치 등을 종합해 얻은 가중치를 사용함

• **개방형 질문 (open-ended question)**
주어진 몇 가지 지문에서 선택하는 것이 아니라 응답자의 생각을 자유롭게 기술하도록 하는 질문 양식. 서술형 질문이라고 부르기도 함. 개방형 질문과 달리 주어진 몇 개의 지문 중에서 해당하는 지문을 선택하도록 하는 질문 양식을 폐쇄형 질문 또는 선지형 질문이라고 함

• **거부율 (refusal rate)**
가구나 사업체 모집단에서 성공적으로 접촉한 사람들의 일부가 필요한 정보 제공을 거절하는 경우, 전체 조사대상자 중 조사를 거부한 사람의 비율. 거절한 단위의 수를 원래 정보를 얻고자 하는 전체 표본 단위의 수로 나누어 산출함

• **결측치 대체 (imputation)**
결측치 대체는 일반적으로 항목 무응답일 때 사용하며, 무응답의 편향을 줄이고 결측치가 없는 완벽한 데이터를 얻기 위해 사용함. 결측치 대체는 자료수집 또는 데이터 에디팅 과정에서 파악된 결측, 논리적으로 불가능한 응답, 다른 항목과 불일치하는 응답 등 자료가 갖고 있는 다양한 문제를 해결하기 위해 활용됨. 대체는 결

측치를 그럴듯한 값이나 일관성 있는 방법을 적용해 채우는 작업을 말함

• 과대포함 (over-coverage)

과대포함은 목표모집단에 속하지 않은 단위가 추출틀에 포함되거나 동일한 단위가 한 번 이상 추출틀에 포함된 경우에 발생함. 예를 들어 사업체조사에서 명부 작성 시 사업 중이던 사업체가 조사 당시에는 폐업하여 존재하지 않는 경우, 또는 동일 사업체가 두 번 이상 중복되어 추출틀에 기재된 경우 이에 해당함

• 과소추정 (under estimation)

표본조사 자료를 이용해 모수를 추정할 때 참값보다 추정값이 체계적으로 작아지는 경향이 있는 경우를 말함. 추출틀에 목표모집단에 속하는 모든 단위가 포함되지 않는 경우, 추출틀에 기재된 정보를 기초로 총계를 추정하게 되면 과소추정이 흔히 발생함.

• 과소포함 (under-coverage)

과소포함은 목표 모집단에 속하는 일부 단위가 추출틀에 누락된 경우에 발생하며, 사업체 조사의 경우 조사당시 신규 사업체는 사업체 명부 작성 시 누락됨으로서 실제 모집단에는 포함되지만 추출틀에는 존재하지 않는 경우 이에 해당함

• 구역추출 (area sampling)

전체 지역을 체계적인 방법으로 구역으로 분할하고, 설정된 구역을 추출단위로 사용한 추출법. 가구 대상 조사를 위해 인구주택총조사의 조사구를 추출단위로 사용하는 경우가 대표적인 구역추출법에 해당함

• 다단계추출 (multi-stage sampling)

확률추출의 한 방법. 모집단 지역이 광범위할 때는 먼저 몇 개의 집락(조사구, 학교, 병원 등)을 추출하고 다시 표본 집락 내에서 그 다음 단계의 추출단위를 추출하는 방법. 조사의 편의 또는 추출틀 확보를 위해 표본을 몇 단계로 나누어 추출하는 방법

• 단순확률추출 (simple random sampling)

크기가 N인 모집단에서 크기가 n인 표본을 추출할 때 모든 가능한 표본들이 추출될 확률이 동일하도록 해주는 추출법. 이 경우 각 단위가 표본에 포함된 확률은 추출률에 해당하는 n/N 으로 일정함. 흔히 단순임의추출, 단순무작위추출, 단순랜덤추출이라고 부르기도 함

• **단위무응답 (unit nonresponse)**
무응답의 유형 중 여러 가지 이유로 조사를 통해 어떤 응답도 얻지 못한 경우를 단위무응답이라고 함. 단위무응답은 주로 조사단위와 접촉할 수 없거나 접촉에 성공했지만 그들이 조사에 참여하기를 거부하기 때문에 발생하며, 단위무응답과는 달리 일부 항목에 대해서만 응답을 얻지 못한 경우는 항목무응답이라고 함. 단위무응답의 경우 흔히 가중치조정이나 사후층화 등의 방법으로 단위무응답에 의한 편향을 보정함

• **단위응답률 (unit response rate)**
표본으로 추출된 조사대상 단위 중에서 응답을 완료한 단위의 비율. 일반적으로 가중치를 반영해 응답률을 산출함

• **대화식 면접 (conversational interviewing)**
질문자와 응답자가 서로 대화를 통해 내용을 주고받는 조사 형식이며 상황에 따라 면접원이 재량껏 면접과정을 조정할 수 있음. 표준화 면접에 비해 면접원과 응답자 사이에 훨씬 높은 수준의 상호작용을 요구하는 면접 방식으로 면접원은 설문지에 있는 질문의 의미를 좀 더 명확히 하고 그 질문이 응답자의 특정한 상황에 어떻게 적용되는지 확실히 하기 위해 질문의 표현을 바꿀 수 있음

• **마이크로 에디팅(micro-editing)**
개별적인 조사 자료를 검사함으로써 오류를 찾아내고 오류를 수정하는 작업으로 각 조사단위로부터 얻은 레코드나 질문 항목 수준에서 수행하는 데이터 에디팅 작업을 의미함

• **매크로 에디팅 (macro-editing)**
매크로에디팅은 평균이나 총계와 같은 총괄(aggregate) 자료나 레코드 전체에 대해 점검을 통해 개별적인 오류를 찾아내는 과정을 말함. 규칙에 근거하여 총계가 의심스럽다고 생각되면 비정상적인 총계를 산출하게 된 원인을 한 개나 몇 개의 잘못된 레코드에서 찾을 수 있는지 확인하기 위해 해당 총계를 구성하는 각각의 레코드를 검사하는 방식으로 에디팅 작업을 함

• **모집단 (population)**
모집단은 표본조사를 통해 특성을 파악하고자 하는 연구대상 전체 집단을 말하며, 구체적으로 보면 관심대상이 되는 모든 기본단위들의 집합을 모집단이라고 함. 연구목적에 따라 개념적으로 정의된 모집단을 목표모집단이라고 하고, 실제 표본추출

을 하는 과정에서 현실적인 제약을 반영하여 구체적으로 표본추출 대상이 될 수 있는 기본단위들로 구성한 모집단을 조사모집단 또는 조사가능모집단이라고 함

• 무응답 (nonresponse)

통계조사에서 추출된 단위(개인, 가구, 기업체 등)와 접촉이 불가능하거나 응답 자체를 거부한 경우, 또는 조사항목 중 일부 항목에 대해서는 응답하지 않은 경우를 모두 무응답이라고 부름. 일바적으로 무응답은 단위무응답과 항목무응답으로 구분함

• 무응답 가중치 조정 (nonresponse weighting adjustment)

무응답에 의한 편향을 줄이기 위해 응답한 개체의 속성을 분석해 무응답에 따른 결손을 응답 개체의 가중치를 확대함으로써 보상하는 방법. 무응답 조정층을 구성하거나 응답성향점수를 이용해 무응답 조정 상수를 산출해 가중치를 조정함

• 무응답 오차 (nonresponse error)

무응답오차는 조사에서 한 개 또는 여러 개의 항목에 대해 응답을 받지 못함으로써 발생하는 오차를 의미함. 무응답이 발생하면 유효한 표본크기가 줄어 결과적으로 사전에 설정한 목표정도를 달성할 수 없게 됨. 일반적으로 무응답자들과 응답자들은 여러 가지 속성이 서로 다르기 때문에 편향을 발생시키는 주요 원인이 되며, 무응답 오차는 통계조사의 전반적 품질을 떨어뜨리는 중요한 요인 중 하나

• 변동계수 (coefficient of variation; CV)

변동계수는 표준오차를 해당 추정값으로 나누어 백분율(%)로 표시한 것으로, 상대표준오차를 의미하며 추정값 대비 상대적인 변동을 설명함. 표본조사 전문가들은 표준오차에 비해 변동계수를 표본추출오차를 나타내는 척도로 더 많이 사용하며, 추정량의 정도(precision)를 평가하거나 목표오차를 만족하는 표본크기를 산출하는데 사용함

• 부적격단위 (ineligible units)

조사대상 기간 동안 목표 모집단에 속하지 않기 때문에 추출틀에 포함되어서는 안되는 단위로서 만일 이러한 단위가 추출틀에 있어서 조사에 포함되면 총계 추정에 있어서 과다포함에 따른 편향이 흔히 발생함

• 분산 (variance)

통계 추정량의 확률분포가 평균을 중심으로 퍼진 정도를 나타내는 척도. 분산추정을 통해 추정량의 정도를 평가할 수 있으며 신뢰구간을 계산하고 통계적 추론을

하는 데 핵심적인 역할을 함. 표본분산은 표본조사와 추정의 품질에 대한 주요 지표중 하나이며 표본설계 및 추정단계에서 매우 주요한 정보로 사용됨

• 비표본추출오차 (non sampling error)

표본조사에서 표본의 추출에 따른 변동으로 인해 발생하는 오차이외의 다양한 원인에 의해 발생하는 오차. 추출틀의 결함, 무응답, 측정상의 문제, 자료처리상의 오류 등에 의해 발생함. 비표본추출오차에 의해 보통 체계적인 오차에 해당하는 편향이 발생할 소지가 많기 때문에 조사의 정확성을 확보하기 위해서는 적극적으로 비표본추출오차를 관리해야 함

• 비확률표본추출 (non-probability sampling)

확률표본추출의 반대 개념으로 특정 표본이 추출될 확률을 알 수 없고 주관적으로 표본을 추출하는 경우를 말함. 편의추출, 판단추출, 할당추출, 눈덩이 추출 등이 대표적인 비확률표본추출법에 해당함. 비확률표본추출은 확률적인 이론에 근거한 추출법이 아니기 때문에 결과를 분석하는 데 있어서 통계적인 이론을 적용할 수 없다는 한계를 갖고 있음

• 사후층화 (post-stratification)

알려진 모집단 구성비율과 표본의 구성 비율이 일치하도록 가중치를 조정하는 방법. 단순평균 대신에 실제 적용된 추출법과 상관없이 층화추출법에서의 추정공식으로 사용되는 가중 평균을 사용하여 평균이나 총계를 추정하는 것으로 볼 수 있음

• 상대표준오차 (relative standard error)

상대표준오차는 추정량의 표본추출오차를 나타내는 척도임. 상대표준오차는 추정값의 표준오차를 추정값 자체로 나누어서 산출함. 흔히 백분율로 표시하며 변동계수(CV)라고 부르기도 함

• 설계효과 (design effect)

단순확률추출을 기준으로 실제 적용된 추출방법의 효율성을 나타내는 척도. 해당 추출법을 적용할 때 얻어지는 추정량의 분산을 동일한 표본크기를 갖는 단순확률추출을 적용했을 때의 추정량의 분산으로 나눈 값으로 1보다 크면 적용된 추출법이 단순확률추출법에 비해 효율이 떨어진다는 것을 의미함

• 스캔 오류 (scan error)

자료를 자동화 기계를 이용하여 입력하여 컴퓨터 파일로 변환시킬 때 관측자료와 변환자료가 다르게 입력되는 오류

• 시험조사 (pilot study)

예비조사라고도 하며 본 조사에 앞서 실제 조사 현장에서 발생할 수 있는 다양한 문제점을 사전에 확인하기 위해 수행하는 소규모 조사. 연구목적을 달성하기 위해서는 어떤 질문을 사용해야 하는지 검토하고, 응답자의 반응을 관찰해 조사 질문의 적절성을 평가하며, 자료 수집 방법의 타당성을 검증하는 등 표본조사 전반적인 과정에서 발생할 수 있는 다양한 문제점을 사전에 파악하고 개선방안을 수립하기 위해서 수행하는 조사를 말함

• 에디팅 (editing)

데이터 에디팅은 통계 산출에 사용되는 각각의 자료에 존재하는 오류와 이상점을 확인하고, 불일치 자료를 확인해 수정하는 전반적인 과정을 의미함. 즉 자료의 분석과정을 수행하기 전에 잘못된 자료를 수정할 목적으로 잘못되거나 의심스러운 조사 자료를 찾아서 수정하는 모든 절차들을 에디팅이라고 함.

• 오분류 (misclassification)

오분류란 설정된 분류기준에서 어긋나게 통계단위를 잘못 분류하는 경우를 말함. 예를 들면 사업체의 산업분류나 근로자의 직업분류를 잘못 하는 경우가 이에 해당함

• 오차 (error)

일반적으로 하나의 실수 또는 일상적인 오류를 의미하며, 보다 제한적인 의미에서 오차는 통계학에서 사용되는 개념으로 추정된 값과 실제 참값과의 차이를 의미함. 흔히 평균제곱오차로 조사에서 발생하는 전체 오차를 설명함

• 유의추출 (purposive sampling)

조사를 실시하는 전문가의 경험이나 주관적인 판단에 의해 모집단을 잘 대표 할 수 있다고 생각되는 대상들을 표본으로 추출하는 방법. 전형적인 비확률추출법으로 판단추출 또는 전문가 선택에 의한 추출법이라고 부르기도 함

• 자료처리 오차 (processing error)

자료처리 오차는 자료수집과정에서 얻은 가공되지 않은 자료를 분석하고 배포할 수 있도록 가공하고 교정한 상태로 변환하는 작업 중 발생하는 오차를 의미함. 자료의 처리에는 자료의 입력, 에디팅, 코딩, 표의 작성, 추정 등과 같은 단계들이 있으며, 이러한 일련의 처리 과정을 거쳐 자료의 정확성은 향상될 수 있음

• **잠정치 (provisional statistics)**
모집단의 모수에 대한 최종적인 추정값이 나오기 전에 사용자의 시급한 요구에 부응하기 위해 잠정적으로 산출한 추정값

• **재가중 (reweighting)**
재가중은 추정값을 산출할 때 원래 가중치를 수정하여 주는 작업 또는 그 값을 의미함. 재가중은 주로 추출틀이 갖고 있는 결함이나 단위무응답에 따른 편향 등을 줄이기 위해 보조정보를 사용해 원래의 가중치를 조정해 주는 작업을 말함. 이를 위해 사후층화, 보정(calibration), 응답성향모형을 이용하는 방법 등이 흔히 사용됨

• **접촉률 (contact rate)**
전체 조사대상 단위들 중 접촉 가능한 단위들의 비율을 나타냄

• **조사설계 (survey design)**
조사설계에서는 주어진 조건 내에서 모든 실현 가능한 조사방법 중 가장 효율적인 방법을 선택하는 과정을 의미함. 조사기획, 조사원 선정 및 훈련, 설문지 작성, 표본추출, 추정, 자료처리 등 조사 전반의 계획을 수립하는 전반적인 과정을 말함. 표본설계, 표본추출계획 등과 유사함 의미로 사용되기도 함.

• **조사원 오차 (interviewer error)**
조사원 오차는 조사원들이 동일한 조사를 다른 방식으로 수행함으로써 응답자의 답변에 영향을 주기 때문에 발생하는 오차임. 질문을 잘못 읽거나, 응답자가 답을 선택하는데 영향을 줄만한 행동이나 설명을 하는 경우, 또는 응답자의 답변을 조사원이 잘못 기입하는 경우 등 조사원에 의해 발생하는 다양한 형태의 오차를 말함

• **집락추출 (cluster sampling)**
조사단위들의 집합 또는 집락을 추출단위로 사용하는 확률추출법을 말함. 흔히 단순확률추출을 하면 조사단위들이 너무 산재되어 조사에 어려움이 예상되거나 직접 조사단위를 추출하는 경우 표본추출을 위한 추출틀을 확보하기 어려운 경우 집락추출법을 흔히 사용함. 학교, 병원, 조사구 등이 집락추출에서 주로 사용되는 집락의 형태임

• **참값 (true value)**
완벽한 조사를 수행한다고 가정하는 경우, 다시 말해 어떤 형태의 오류도 범하지

않는 경우 얻을 수 있는 실제 모집단의 값을 의미함

• 최근방 대체 (nearest neighbor imputation)
결측치가 있을 때 거리 척도(distance measure)를 정의하고 거리 척도를 기준으로 가장 근접한 단위의 응답값으로 결측치를 대체하는 방법. 다시 말해 거리 척도의 기준이 되는 특정 속성을 기준으로 결측된 단위와 가장 유사한 단위의 응답값으로 결측치를 대체하는 방법

• 추정값 (estimate)
관심 모수를 추정하기 위해 표본조사 결과를 통해 얻은 자료를 기초로 추정공식에 해당하는 추정량을 이용해 산출한 값을 의미함

• 추정량 (estimator)
추정량은 모집단의 특성을 나타내는 모수를 추정하는 공식 또는 방법을 말함. 보통은 표본 관측치의 함수로 표현되며, 추정량의 정확성을 통해 조사과정 전체의 정확성이 설명되기 때문에 조사통계의 정확성을 평가하는데 매우 중요한 역할을 하는 요소

• 추출틀 오차 (sampling frame error)
표본조사에서 표본추출틀이 가지고 있는 결함 때문에 발생하는 오차, 추출틀에 모집단 구성원이 아닌 단위가 포함되거나, 포함되어야 할 단위가 누락 또는 중복되는 경우 추출틀 오차가 발생함

• 측정 오차 (measurement error)
다양한 원인에 의해 조사단위가 갖고 있는 참값을 측정하지 못하기 때문에 발생하는 오차. 사회조사의 경우 측정도구에 해당하는 설문지가 정교하지 못한 경우 측정오차가 발생할 수 있으며, 응답자가 여러 가지 이유로 정확하지 않은 응답을 하는 경우도 측정오차로 볼 수 있음

• 층화추출 (stratified sampling)
모집단을 서로 겹치지 않는 그룹들, 층으로 나누고 각 층에서 확률추출법으로 표본을 추출하는 방법. 모집단을 상호배타적인 동질적인 단위들로 이루어진 층(부차모집단)으로 나누고, 각 층으로부터 전체 분산을 줄이거나 각 층별 통계를 효과적으로 산출할 수 있도록 층별로 표본을 배분해 표본을 추출하는 방법

• **코딩 (coding)**
코딩은 구두로 말한 정보를 수치나 다른 기호로 전환하는 기술적인 절차로서 이를 통해 계산과 통계표 작성이 용이해짐

• **코딩 오차 (coding error)**
조사응답에 대해 잘못된 코드가 부여된 경우에 발생하는 오차를 말함

• **크기비례확률추출법 (probability proportional to size sampling)**
각 단위에 대해 사전에 크기척도(size measure)를 정의하고, 단위별 크기척도에 비례하는 확률로 표본을 추출하는 경우 이를 크기비례확률추출이라고 함. 집락추출에서 집락의 크기에 비례하는 확률로 집락을 추출하는 경우가 대표적인 크기비례확률추출법에 해당함

• **편의추출 (convenience sampling)**
조사자가 손쉽게 접촉할 수 있는 조사대상을 표본으로 선정하는 방법으로 가장 쉽게 접촉할 수 있고, 쉽게 응답을 얻을 수 있는 사람을 표본으로 선택하는 추출법. 자의적인 선택에 의한 표본추출법이기 때문에 대표성 있는 표본을 확보할 수 없음

• **편향 (bias)**
통계적 추정결과가 체계적으로 한 쪽으로 치우치는 경향을 보임으로써 발생하는 오차. 일반적으로 추정결과가 크거나 작아짐에 따라 발생하는 변동오차(variation error)와는 달리 추정결과가 체계적으로 한 쪽 방향으로 치우치는 경향을 보이는 오차를 의미함

• **평균제곱오차 (mean squared error)**
추정에 대한 평균제곱오차는 편향의 제곱에 분산을 더한 것. 표본추출오차, 추출틀오차, 무응답오차, 측정오차, 자료처리오차 등이 표본조사에서 평균제곱오차의 편향과 분산에 영향을 주는 주요 요소

• **포함오차 (coverage error)**
포함오차는 표본추출에 사용된 추출틀에 목표모집단의 일부 단위가 포함되지 않았거나 중복 기재되는 경우 또는 모집단에 해당하지 않는 단위가 포함됨으로써 발생하는 오차를 말함. 흔히 과소포함과 과대포함으로 구분되며, 추출틀에서 발생하는 대표적인 오차 발생 원인

• 포함확률 (inclusion probability)

모집단의 특정 단위가 표본에 포함될 확률. 확률추출법에서 포함확률을 계산할 수 있으면 Horvitz-Thomson(HT) 추정량을 이용해 비편향성을 만족하는 추정결과를 항상 얻을 수 있음

• 표본설계 (sample design, sampling plan)

표본설계는 전반적인 조사방법을 기획하는 과정을 의미하며, 조사설계와 동일한 의미로 사용하기도 함. 특히 표본추출이론을 기초로 연구목적에 따른 표본크기를 정하고 주어진 추출틀에서 효율적으로 표본을 추출하는 동시에 모수를 추정하는 효율적인 방법을 중점적으로 다루는 일련의 과정을 의미함

• 표본추출오차 (sampling error)

전수조사 대신 확률추출법(probability sampling)에 의해 모집단에서 추출된 표본으로부터 모수를 추정하기 때문에 발생하는 추정결과와 모수간의 차이를 의미함. 확률표본추출의 경우 추정량의 분산, 표준오차 또는 변동계수 등으로 표본추출오차를 설명함

• 표본추출틀 (sampling frame)

모집단에서 실제 표본을 추출하기 위해서 사용되는 모든 추출단위가 나열된 명부 또는 목록을 말함. 특정 표본이 추출될 확률을 계산할 수 있는 과학적인 확률추출법을 구현하기 위해서는 정확한 표본추출틀을 확보하는 것이 필수적임. 사업체 명부, 전화번호부, 조사구 명부 등이 흔히 표본추출틀로 사용되며 표본추출틀에는 해당 단위를 접촉할 수 있는 정보가 수록되어 있어야 함

• 표준오차 (standard error)

표본조사에서 사용되는 추정량의 분산에 대한 제곱근을 말함. 표본추출오차를 나타내는 척도로 흔히 사용됨

• 표준화 면접 (standardized interviewing)

사전에 일정하게 표준화된 면접, 질문지를 만들어서 모든 응답자에게 동일한 질문과 동일한 질문 순서에 따라서 면접을 수행하는 방법. 면접원과 응답자사이의 상호작용을 표준화 하는 것으로 표준화 면접의 목적은 응답이 면접원에 의해 조금도 영향을 받지 않게 하기 위하여 모든 응답자에게 같은 질문을 정확히 같은 방법으로 하도록 만드는 것

• **할당추출 (quota sampling)**
관심 변수에 영향을 주는 주요 특성에 대한 모집단 비율이 표본에 그대로 유지되도록 하는 추출법을 말함. 여기서 구성 비율을 제어하는 것은 조사원에 의한 선택편향을 제어하기 위한 것이고, 미리 정해진 기준에 따라 전체 표본을 여러 집단으로 구분하고 집단별로 할당된 수의 대상을 표본으로 추출하는 방법

• **핫덱 대체 (hot-deck imputation)**
동일한 조사에서 다른 응답자로부터 얻은 자료를 이용해 결측치를 대체하는 방법을 포괄적인 의미로 핫덱이라고 말함. 흔히 자료의 입력순서에 따라 바로 앞에 응답 결과로 결측치를 대체하거나, 유사한 단위들로 대체군(imputation class)를 구성하고, 대체군내에서 다른 단위의 응답값으로 결측치를 대체하는 방법을 의미함. 이과 대조되는 방법으로 동일한 조사 자료가 아닌 다른 자료를 사용해 결측치를 대체하는 경우 콜드덱 대체(cold-deck imputation)라고 부름.

• **항목무응답 (item nonresponse)**
어떤 조사단위가 조사에 참여해 대부분의 항목에 대해서는 응답을 하였으나 일부 항목에 대해 응답을 하지 않은 경우. 항목무응답의 경우 응답된 항목들로부터 해당 단위에 대한 유익한 정보를 얻을 수 있기 때문에 응답 항목으로부터 얻은 정보를 활용한 대체방법을 적용하는 것이 효과적임

• **항목응답률 (item response rate)**
항목응답률은 특정한 항목에 대하여 응답을 얻어야 하는 전체 단위 중에서 해당 항목에 응답결과를 얻은 단위의 비율을 의미함

• **협조율 (co-operation rate)**
전체 조사대상 단위 중 조사에 응답한 단위들의 비율을 나타냄

• **확률표본추출법 (probability sampling)**
모든 추출단위들이 표본으로 추출될 확률을 계산할 수 있는 표본추출법을 말함. 확률표본추출법에 의한 표본을 통해 얻은 추정결과에 대해서는 통계적인 이론에 따른 해석이 가능함. 단순확률추출법, 계통추출법, 층화확률추출법, 집락추출법 등이 대표적인 확률추출법에 해당함

▌참고문헌▐

강형철 · 한상태 · 김지연 · 정용찬 · 허명회 (2008). “RDD 전화조사와 주요결과.” 조사연구 9권 1호: 1-22.

김영원 · 류제복 · 박진우 · 홍기학 (2006).「표본조사의 이해와 활용 (6판)」 톰슨 코리아.

노동부 (2008).「임금구조기본통계조사와 사업체근로실태조사 통합을 위한 표본설계 최종보고서」

박홍래 (2000).「통계조사론 (개정판)」 영지문화사.

통계청 (2003).「가구부문 표본개편보고서」 통계청 내부 보고서.

통계청 (2005).「2005 인구주택총조사 조사지침서」

한국고용정보원 (2006).「2005년 청년패널 기초분석보고서」

한국노동연구원 (2007).「2006년 고령화연구패널조사 1차기본조사 사용자안내서」

한국고용정보원 (2008).「2006 대졸자 직업이동 경로조사 기초분석보고서」

한국통계학회 조사통계연구회 (2005).「조사방법의 이해 (개정판)」 교우사.

해양수산부 (2007).「어업생산통계조사의 표본개편설계 연구보고서」

Biemer, P. P. and Lyberg, L. E. (2003). *Introduction to Survey Quality*. John Wiley and Sons.

Cochran, W. G. (1977). *Sampling Technique, 3rd edition*. John Wiley & Sons, New York.

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

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press.

Office for National Statistics (2007). *Guidelines for measuring statistical quality*. Office for National Statistics, UK

Schuman, H. and Presser, S. (1981). *Questions and Answers in Attitude Survey*. Academic Press.

The American Association for Public Opinion Research (2008). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*.

집필진

연구책임자	김영원	숙명여자대학교 통계학과 교수
공동연구원	이용희	서울시립대학교 통계학과 교수

조사통계의 정확성지표 품질관리 매뉴얼

발행일 : 2008년 12월

발행처 : 통계청

대전광역시 서구 둔산동 920 정부대전청사

T. 042 – 481–2499

홈페이지(www.nso.go.kr, http://quality.nso.go.kr)

발간등록번호 11-1240000-000471-14, ISBN 978-89-92372-21-3 부가기호 93330 (비매품)