통 계 연 구

범위오차의 평가와 원인분석

2004년 12월

작성자 : 통계청 통계연구과 이 지 연
Tel. (042)481-2569
(jylee@nso.go.kr)

主 要 內 容

□ 본연구의 목적은 과거 총조사의 범위오차 수준을 평가해보고, 2000년 총조사 사후조사 결과를 이용하여 범위오차 발생의 원인과 유형을 개인, 가구, 조사원의 특성차원에서 분석하는 것임

□ 범위오차는 누락과 조사오차로 구성되며, 조사오차는 다시 중복과 착오에 의한 조사로 세분되는데, 현행 총조사 사후조사방식을 통해서는 중복측정이 불가능함

□ 총조사(1990-2000)의 범위오차 평가결과 범위오차는 최근에 오면서 점차로 안정화되고 있지만, 20-35세 연령층과 이들의 자녀 세대인 5세 이하 연령층은 지속적으로 누락되고 있는 계층임

□ 2000년 총조사 사후조사를 로지스틱 회귀분석한 결과 조사착오를 일으킬 확률은 개인의 특성 변수들과 관계가 깊다면, 누락은 주로 가구차원에서 발생할 확률이 높았으며, 조사원의 연령과 같은 특성도 범위오차에 통계적으로 유의미한 영향을 주고 있었음

통계청
Korea National Statistical Office

범위오차의 평가와 원인분석

I. 센서스의 오차

센서스 자료를 평가하는 두가지 기준은 조사범위(coverage)의 포괄성과 조사의 정확성(accuracy)이다 (UN, 2001). 센서스는 대상 집단의 개개 단위가 일정한 간격으로 동일한 시점에 모두 조사되어야만 한다. 하지만 현실에서 모든 대상이 완전히 조사되기란 어려운 일이기 때문에 모집단의 참값과 조사결과 사이에는 언제나 차이가 발생한다.

센서스에서 발생하는 오차는 크게 범위오차(coverage error)와 내용오차(content error)로 구성된다. 범위오차란 센서스에서 조사된 인구와 실제 인구와의 차이를 의미한다. 내용오차는 조사대상자의 부정확한 응답이나 기록에 의해서 발생하는 자료의 정확성과 관련된 오차이다. 내용오차가 발생하는 경우는 조사항목이나 조사지침 자체가 불분명하거나, 조사대상자가 질문을 잘못 이해하여 응답하거나, 의도적으로 틀리게 응답한 경우, 또는 코딩이나 입력과정에서 오류가 있었을 때이다(UN, 1998).

센서스의 불완전성으로 인해 계획에서 결과공표까지 센서스의 전과정이 범위오차나 내용오차의 발생 가능성으로부터 자유로울 수 없다는 것은 일반적으로 받아들여지고 있는 사실이다. 그러나, 센서스 오차문제의 심각성은 센서스의 불완전성 자체에서 오는 것이 아니라, 그 오차의 수준과 범위를 파악할 수 없을 때 발생한다. 센서스 오차의 객관적인 평가가 중요한 이유는 이용자가 센서스 자료를 통해 측정하고자 하는 현상을 보다 정확하게 해석할 수 있는 정보를 제공하기 때문이다. 또한, 이를 통해서 조사대상의 참값에 근접한 추정치를 만들어 내거나 센서스 수치의 보정작업등의 기초 자료로도 활용될 수 있기 때문이다. 센서스에서 발생하기 쉬운 오차유형이나 원인분석이 이루어진다면 향후 센서스에서 유사한 오차의 발생을 방지하는 장치도 사전에 마련할 수 있을 것이다.

학계와 현장에서 센서스 오차평가의 중요성은 인식되었지만, 지금까지 인구주택총조사(이하 총조사)의 범위오차를 실제적으로 평가해 본 연구들은 없었다[1]. 본 연구에서는 다양한 센서스 범위오차 측정방법들을 소개하고, 지난 총조사의 범위오차를 평가한 후, 2005년 총조사의 범위오차 개선방안을 제시하고자 한다. 먼저 센서스의 범위오차 개념과 구성요인, 주요 오차측정방법들을 살펴보고, 현행 총조사 범위오차의 분류상의 문제점도 함께 논의될 것이다. 다음 장에서는 1990년부터 2000년 까지 총조사의 범위오차를 평가해 볼 것이다. 마지막으로 2000년 총조사 사후조사 결과를 바탕으로 범위오차의 발생 확률을 개인, 가구, 조사원 특성 차원에서 분석하고자 한다.

1) 본 연구에서 '센서스'라는 용어는 전수조사를 의미하는 일반적인 개념으로 사용되고, '총조사'는 한국의 대표적인 센서스인 인구주택총조사를 지칭한다.

II. 센서스 범위오차의 종류

1. 오차의 종류와 현행 분류상의 문제점

일반적으로 범위오차는 조사대상(거처, 가구, 가구원)이 과소집계(undercount)되거나 과대집계(overcount)될 때 발생한다[2]. 과소집계는 주로 조사되었어야 할 대상이 조사되지 않은 경우(누락)로 인해 발생하며, 과대집계는 두번 이상 조사(중복)되었거나, 잘못된 장소나 기간에 착오로 조사된 경우에 발생한다. 예를 들어 센서스 기간 이전 사망자나 혹은 기간 이후 출생자가 조사에 포함된 경우이다 (UN, 2001).

통계청은 1960년부터 총조사의 범위오차를 측정하기 위해서 사후조사를 실시해 왔다. 사후조사는 종속조사 방식(dependent method)으로 진행되는데, 이는 총조사의 표본틀과 진행방식이 사후조사에서도 그대로 활용되는 방식을 말한다 (김민경, 2000). 조사방식에 대한 논의는 뒤에서 보다 자세하게 다루어질 것이다.

현재까지 총조사의 범위오차는 사후조사를 통해서 측정된 중복율과 누락율이라는 두 지표를 통해 평가되어 왔다 (통계청, 2001; 1998). 총오차율은 중복율과 누락율을 합한 값이고, 순누락율은 누락율에서 중복율을 뺀 값이다. 이러한 방식으로 범위오차의 종류를 분류하는 데는 개념상의 문제가 있다. 기존의 사후조사방식으로는 응답자의 중복 여부를 측정할 수가 없다. 엄격히 말해서 사후조사에서 중복을 조사했다기 보다는 착오로 조사된 경우를 조사해서 이를 중복이라고 잘못 지칭해 왔기 때문이다.

이러한 문제가 발생한 데는 두 가지 이유가 있다. 첫 번째는 흔히 총조사의 오차를 누락(omission)과 조사오차(enumeration error)가 아니라 누락과 중복(duplication) 으로 생각해 왔기 때문이다[3]. 어떤 조사가 과대집계를 일으켰다면

---

2) 무응답 대체 (Imputation) 기법이 발달된 국가들에서는 무응답자의 특성을 전혀 알 수 없는 경우에 컴퓨터에 의해 자동으로 조사값이 대체된 경우(whole-person imputation 혹은 non-data defined person) 또한 범주오차의 세 번째 유형으로 분류하기도 한다(Wolfgang, Davis and Stallone, 2001).

3) 이러한 개념상의 문제들이 발생한 이유들 중 하나는 수를 모두 헤아린다는 의미의 enumeration이라는 단어를 번역하는데 적합한 단어가 없었기 때문인 것으로 보인다. Net enumeration rate은 순조사누락율 이라고 종종 번역되면서도, enumeration error는 실제적으로 중복오차를 의미한다고 생각해온 이유도 여기에 있다.

직관적으로 그 원인은 응답자가 중복조사 되었기 때문일 것이라고 생각하기 쉽다. 현재 사후조사에서 중복이라는 개념이 사용되는 이유 또한 A라는 지역에서 조사대상자가 아닌 사람이 조사되었을 지라도, 그 사람이 실제로 조사되었어야 할 B라는 지역에서는 제대로 조사에 포함되었을 것이므로 중복 조사가 발생했을 것이라는 추측에 근거한다. 동일인이 두 번 이상 조사 되어 실제인구보다 조사인구가 많이 집계될 수 도 있지만, 조사대상이 아닌 사람이 조사되었어도 과대집계가 발생할 수 있다. 착오에 의한 조사가 정말 중복조사로 이어졌는지는 경험적인 확인을 통해서 판단할 수 있는 사안일 것이다.

두 번째는, 보다 근본적인 문제인데, 사후조사 자체를 통해서는 중복을 확인할 수가 없다는 사실이 간과되어 온 것이다. 동일인이 두 번 이상 다른 지역에서 중복조사 되었는지는 센서스 자료와 사후조사 자료를 비교해서는 판단할 수 없다. 이는 다른 두 지역의 센서스 자료들중에서 중복이 일어났을 것으로 의심되는 조사표를 상호 비교해 봐야 확인할 수 있는 사안이다. 중복여부는 응답자의 이름이나, 성, 생년월일, 등의 인적사항을 매치시켜보는 과정을 통해서 확인 할 수 있다.

지금까지는 총조사에서 착오에 의한 조사와 중복의 개념이 명확하게 분류되지 않아도 별다른 문제가 되지 않았다. 대부분 방문조사라는 단일한 조사 방법에 의해서 조사표가 작성되었기 때문에, 막대한 시간과 비용을 들여 조사표의 중복여부를 확인할 필요성 적었다. 그러나, 센서스 뿐만 아니라 기타 여러 조사에서 응답자의 편의를 위해서 다양한 조사 방법을 동원하고 있는 것은 이미 국제적인 추세이고, 이 때 조사의 중복 여부의 확인은 매우 중요한 문제가 되고 있다 (Diffendal, 2001). 개인의 사생활 보호에 대한 인식이 점차 확산되고 있고, 다양한 주거형태와 생활습관으로 인해 조사원이 조사대상자를 대면하는 것 조차 점점 힘들어지고 있다. 전통적인 조사방법이나 단일한 응답방법에만 의존할 경우에 나타나는 응답율 저조현상을 우려하는 목소리는 여러 나라에서 공통적으로 발견되는 현상이다 (De Heer, 1999).

미 센서국이 2000년 센서스에서 우편, 방문, 컴퓨터, 전화등 다양한 조사방법 동원이 가능했던 이유도, "주택 중복방지 (Housing Unit Unduplication Operation)“와 같이 주택과 개인차원에서 대대적으로 중복을 확인하는 방법을 개발해 냈기 때문이었다 (Nash, 2000). 앞으로 총조사의 응답율을 높이기 위해 다양한 방식으로 응답이 가능하게 한다면 범위오차를 측정하는데 있어서도 착오로 조사된 경우와 중복 여부를 분리해서 평가할 필요성은 대두될 것이다.

III. 총조사 범위오차의 평가, 1990-2000

한 센서스에서 총인구나 성과 연령별 인구의 누락이나 중복이란 실제 인구보다 조사된 사람이 적거나 더 많은 경우를 의미한다. 제1차 UN의 인구센서스 권고안(1998)에 따르면, 센서스의 범위오차를 측정하는 방법은 크게 세 가지로 분류되는데, 내적일관성 체크, 인구분석방법, 사후조사 방법이 있다. 인구분석방법은 범위오차를 추정하는데 다양한 방식으로 활용되고 있는데, 다음의 네 가지 형태로 구분할 수 있다. i) 인구균형방정식: 직전 센서스결과를 바탕으로 센서스간의 출생, 사망, 인구이동의 동태자료를 고려하여 작성된 기대인구와 현 센서스 결과를 비교하는 방법, ii) 센서스간 추정방법: 직전 센서스 결과와 출산력, 사망력, 인구이동 결과를 코호트 조성법을 이용해서 작성한 인구추계치와 현 센서스 결과치를 비교하는 방법, iii) 센서스간 코호트 생잔율법: 센서스간 코호트 생잔율에 기초해서 두 센서스간의 연령분포를 비교하는 방법, iv) 코호트 회귀생잔계수 방법등이다.

이 장에서는 자주 사용되는 평가방법인 인구균형방정식, 인구추계치와의 비교, 코호트 생잔율법에 의한 평가 및 사후조사 결과를 이용하여 1990년부터 2000년까지의 총조사 결과를 중심으로 범위오차의 수준을 평가해보고자 한다.

1. 인구균형방정식과 보정인구

한 센서스에서 총인구나 성·연령별 인구가 과대 혹은 과소집계 되었다는 것은 실제 인구보다 조사된 인구가 더 많거나 더 적은 경우를 의미하는데, 이것은 다른 인구 동태자료에 실제 인구로 기록된 숫자와 비교해 보면 측정 가능하다. 인구균형방정식에 의해 산출된 2000년 11월 1일자 인구는 다음과 같다 (단위: 천명):

2000년 인구 = 1995년 인구 + (출생 - 사망) - 국제이동

47,130 = 45,255 + (3,244 - 1,251) - 117

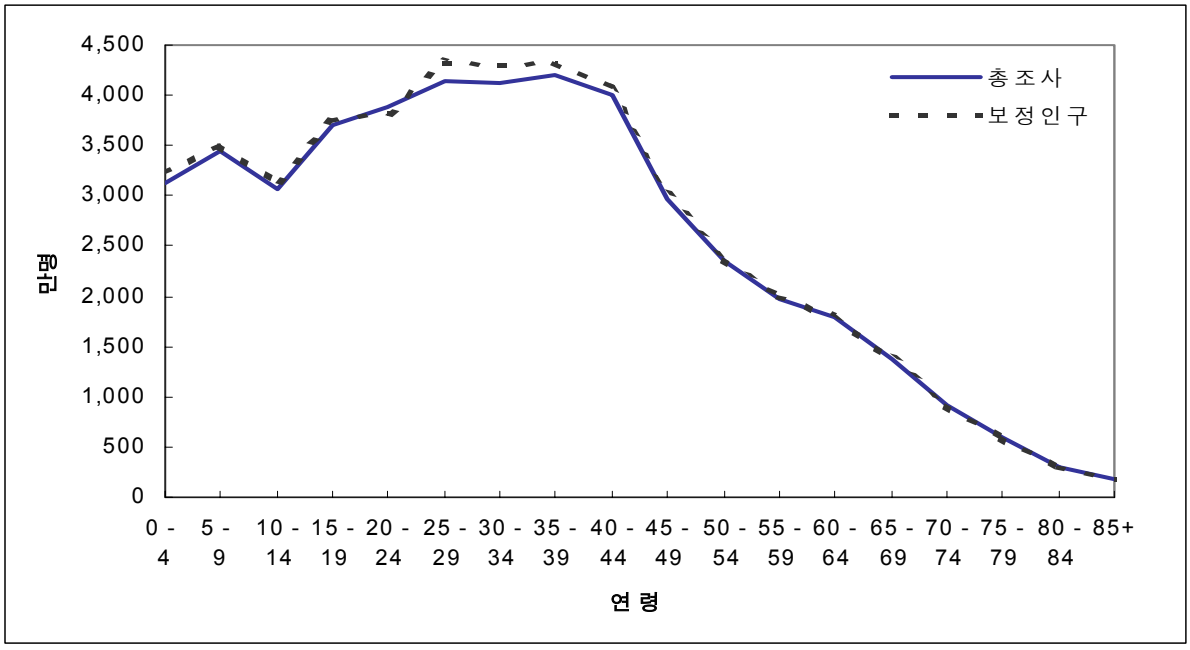
인구균형방정식에 의한 결과와 센서스 결과의 차이는 누락을 의미한다. 2000년 총조사에서 집계된 인구는 46,136천명으로 인구분석 방법에 의한 인구 47,130

명보다 990천명이 적어, 누락율은 2.1%로 추정된다.

그림1은 2000년 총조사 인구와 인구균형방정식에 의해 보정된 인구의 연령별 분포이다. 보정인구는 인구동태통계의 연도별 출생아수, 주민등록인구, 국제이동, 생잔율등을 감안하여 연령별 분포를 보정한 인구이다. 연령을 비교해 보면, 총조사에서는 젊은 연령층의 누락율이 높은 것을 알 수 있다(통계청, 2001). 특히, 25-29세 (4.4%), 30-34세(4.29%), 0-4세(2.96%) 집단은 동태자료를 활용한 인구분석방법에 의해 보정된 수치에 비해 총조사 인구가 훨씬 적게 나타났다.

인구분석방법은 특정시점에 인구에 관한 전국치를 제공하기 때문에 성과 연령에 따른 총조사의 범위오차를 측정하는 벤치마크로서 꾸준히 활용되어 왔다. 하지만 이방법의 단점은 총인구, 성과 연령별 인구의 전국치만이 산출 가능하고, 국내이동을 고려하지 못하기 때문에, 지역단위 누락율은 측정할 수 없다는 점이다 (White and Rust, 1997). 또한, 각 행정자료 자체가 가지는 오차에 노출되어 있으며, 센서스간 인구동태 기록이 없거나 부정확한 경우 결과를 신뢰할 수 없다는 단점이 있다 (Fosu, 2001).

그림 1. 2000년 총조사 인구와 인구분석방법에 의해 보정된 인구 (성·연령별)



자료: 통계청(2002b), 『인구추계작성방법』 *(내부자료).*

2. 인구추계치 비교

총조사의 누락을 측정하는 다른 한 방법은 한 시점의 총조사 결과와 그 보다 앞선 시점에 작성된 추계 결과를 비교해 보는 것이다. 만약 이전의 조사 결과와 연령별 사망률이 정확하다면, 추계된 인구와 이후에 조사된 인구의 차이는 조사 누락을 의미한다. 추계치와 센서스 실측치간의 차이를 나타내는 오차율은 다음과 같이 계산 된다.

오차율= [(추계인구- 총조사 인구)/ 총조사 인구] * 100

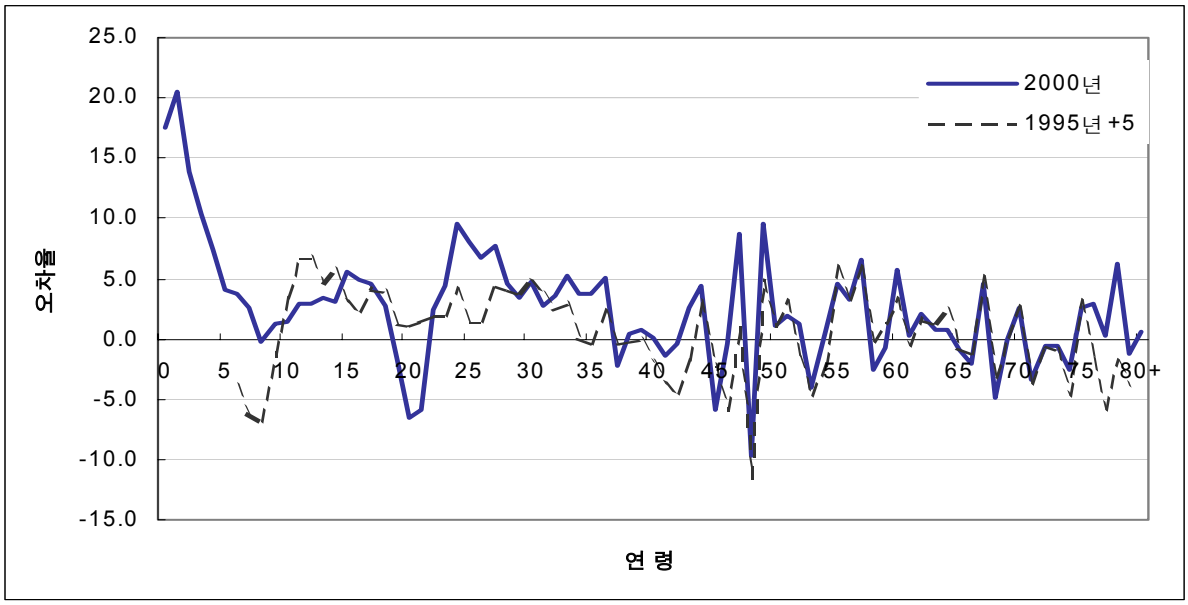
오차율이 양의 값을 갖는 것은 추계치보다 총조사에서 조사된 인구가 적었음을 의미한다. 추계치와 센서스 결과치가 유의미한 차이를 보인다면 두 센서스중 적어도 한 센서스에 문제가 있었거나, 인구추계시 출생, 사망, 이동에 대한 가정이 부정확했기 때문에 발생한다.

그림2는 1996년에 작성된 추계치를 바탕으로 2000년 총조사의 오차율을 연령별로 추정해 본 것이다. 1996년 추계에 따르면 2000년 연앙인구는 47,274천명으로 예상되었다. 2000년 총조사 인구를 7월 1일자로 환산해 보면 총조사는 추계인구에 비해서 전체적으로 약 2.2%의 인구가 과소집계된 것을 알 수 있다. 추계인구를 기준으로 보면 총조사에서 저연령층(특히 35세이하 집단)은 과소집계되고 고연령층은 상대적으로 과대집계된 경향을 발견할 수 있다.

사실 이러한 경향은 한국 뿐만 아니라 대부분의 국가에서 추계인구와 센서스 인구를 비교할 때 자주 발생한다. 센서스 범위오차율을 측정하는 방법과는 분모/분자가 상반된 경우가 인구추계가 얼마만큼 정확한지를 측정하기 위해서 실측치인 센서스 인구와 비교해 본 경우들이다. Bongaarts와 Bulatao (2000)가 세계 여러 나라에서 행해지고 있는 추계의 정확성을 종합적으로 분석한 결과 공통적으로 나타나는 현상으로 추계인구가 저연령층은 너무 높게, 고연령층은 너무 낮게 추계하는 경향이 있다는 점을 발견했다. 환언하자면, 추계인구를 기준으로 센서스의 오차를 측정하면, 정도의 차이는 있겠지만 센서스에서는 저연령층이 과소집계되고, 고

연령층은 과대집계되는 것은 일반적인 현상임을 알 수 있다.

그림 2. 추계인구와 총조사 인구와의 차이



자료: 통계청 (1996, 2001b) 『장래인구추계』.

다만 추계인구를 기준으로 총조사 오차를 측정할 때 주의할 점은 추계인구가 개인의 보고에 의존하는 동태자료에 영향을 받기 때문에, 특정 연령코호트에서는 불안정성이 발생한다는 사실이다. 2000년 경우 특히 문제가 되는 연령은 19-21세 집단이다. 일반적으로 이 연령층은 총조사에서 누락율이 가장 높은 집단으로 여겨지고 있고, 이는 2000년 사후조사 결과에서도 확인된 사실이다. 하지만 이 연령이 추계치보다 총조사에서 더 많이 집계된 원인은 사실 명확하지 않다.

그림2에서 '1995년 +5'는 1991년에 작성된 1995년 추계인구를 기준으로 1995년 총조사의 오차율을 계산한 수치에 5세씩을 더해서 2000년 총조사 오차율의 연령별 코호트와 일치 시킨 것이다. 48세와 52세 코호트처럼 '2000년'과 '1995년+5'가 일치한다는 것은 특정 연령 코호트에서는 반복적인 과대·과소집계나 혹은 과대·과소추계 발생하고 있다는 것을 의미한다. 총조사간의 연령별 범위오차는 유사한 패턴을 보이지만 완전히 일치하기 어렵다는 점을 감안한다면, 이것은 지속적인 과대·과소추계가 발생하고 있음을 의미한다.

3. 코호트 생잔율법

지난 총조사에서 조사된 코호트가 이번 총조사에서는 얼마나 조사되었는 지를 파악해 보면 이것을 통해 총조사 오차를 측정해 볼 수 있다. 다음의 그림3-1과 3-2는 1990-1995년, 1995-2000년 총조사간의 남녀별 코호트 생잔율이다. 두 코호트 생잔율을 비교해 보면, 1995-2000년 생잔율이 일반적인 센서스 코호트 생잔율 곡선에 더 가까워 지고 있고, 연령간의 변동 폭도 더 적게 나타나는 등 안정적인 패턴을 보여주고 있다. 이는 1990년과 1995년 총조사 보다는 1995년, 2000년 총조사간의 범위오차가 유사한 수준에서 안정화되고 있음을 의미한다. 예를 들어, 0-10세 인구는 두 생잔율 곡선이 특히 상이한 패턴을 보여주고 있다. 이것은 1990년, 1995년, 2000년 총조사간의 오차율의 차이가 반영된 것이다. 1990년 조사는 과대집계율(특히 저연령층에서)이 역대 총조사 중 가장 높은 3.46%였던 반면, 2000년 총조사의 과대집계율은 1.74%에 그쳤기 때문이다.

남자와 여자의 생잔율을 비교해 보면 여자의 생잔율이 전 연령에 걸쳐서 보다 고르게 나타난다. 이는 여자보다는 남자에게서 누락 혹은 조사착오가 더 많이 일어나고 있다는 것을 의미한다. 남자는 특히 젊은 연령층에서 주로 급격한 변동이 나타나는데 이러한 경향은 두 생잔율 곡선에서 공통적으로 발견되는 현상이다.

연령별로 생잔율을 살펴보면 1990년부터 2000년까지의 총조사 모두 0-4세 인구 보다는 5-9세 인구가, 20-24세 인구보다는 15-19세 인구가, 20대 인구보다는 30대가 총조사에서 더 완전하게 집계되는 계층임을 알 수 있다. 이것은 남자와 여자에게 공통적으로 발견되는 현상이지만 남성의 경우 연령에 따른 생잔율의 차이가 더욱 크게 나타난다.

그림 3-1. 총조사간 생잔율 비교 (남자): 1990 - 2000

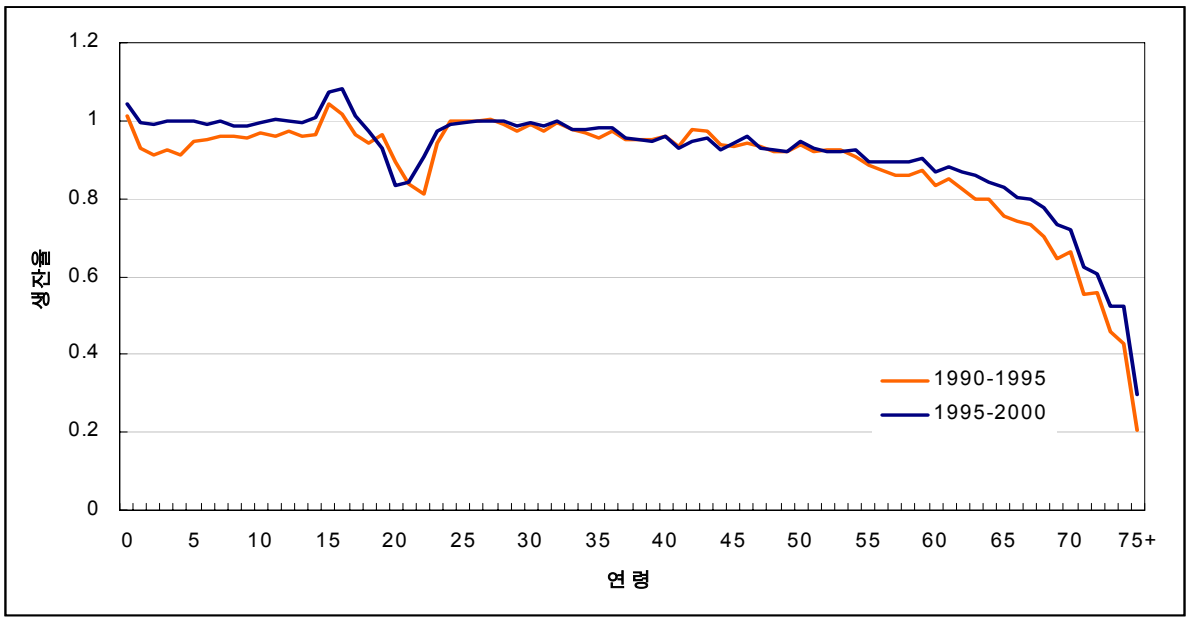
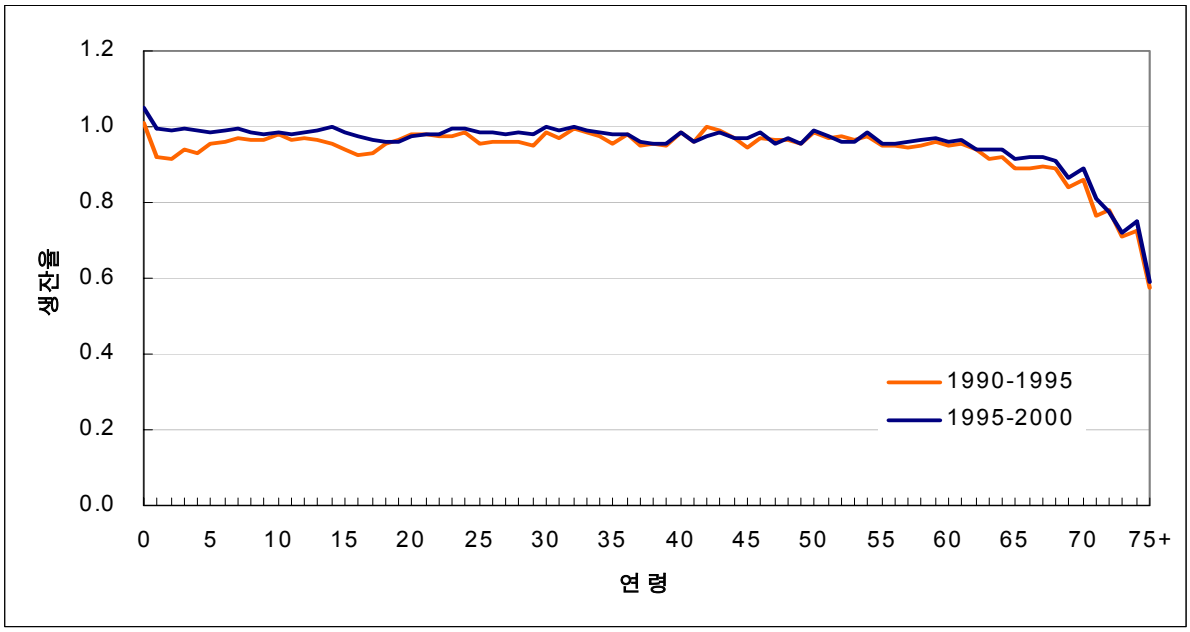


그림 3-2. 총조사간 생잔율 비교 (여자): 1990 - 2000



자료: 통계청 (1993, 1997, 2002a) 『인구주택총조사 보고서』

1995-2000년 총조사에서 14-17세 코호트와 같이 생잔율이 1.0 이상인 경우는 1995년 총조사에서 9- 12세 연령이 누락되었거나, 2000년 총조사에서 14-17세 연령이 과대집계 되었음을 의미한다. 일반적으로 10세 전후의 취학연령층이 생애 가장 이동성이 적기 때문에 오차율이 낮은 집단임을 고려한다면, 2000년 총조

사에서 14-17세 집단은 과대집계된 것으로 보인다. 이것은 아마도 이 연령대가 학업으로 인해 외지로 유학할 가능성이 높고, 이로 인해 조사착오나 누락이 발생할 가능성이 높기 때문인 것으로 보인다.

1990-1995년과 1995-2000년과 남자의 생잔율을 비교해 보면 누락이 가장 높은 연령이 22살에서 20살로 낮아지고 있음을 알 수 있다. 이러한 생잔율의 변화는 한국 사회의 인구이동이 점차적으로 낮은 연령대에 까지 활발하게 이루어 지고 있음을 의미한다. 반면에, 50세 이상 인구의 집계는 최근에 올수록 더 완전해 지고 있는 것을 알 수 있다.

4. 사후조사

많은 국가들이 사후조사를 통해 센서스자료의 범위오차와 내용오차를 평가하고 있다. 센서스와는 독립적으로 사후조사를 실시할 경우 누락된 사례와 착오로 조사된 사례를 모두 조사할 수 있기 때문에 순오차율을 측정할 수 있다는 점은 사후조사의 장점이다.

통계청은 1960년부터 사후조사를 통해서 총조사의 순누락율을 추정해 왔다. 2000년 사후조사는 총조사 실시 후 한달 후인 12월 1일에 600개 표본조사구 약 36천 가구를 대상으로 8일간 실시되었다. 원칙적으로 단수추정방식을 사용하기 때문에 총조사에서 사용된 조사방법이 사후조사에도 동일하게 적용되었다. 사후조사 요원은 총조사 기간에 작성된 조사구 지도와 가구명부를 바탕으로 해당가구를 방문하여 총조사에서 조사된 내용을 확인하고 누락이나 중복(실제 착오로 조사된 경우를 말함), 전입가구 사항등을 조사했다.

2000년 총조사 사후조사 결과를 보면 누락율은 3.30%, 조사오차율 1.74%로서 순오차율은 1.56%이고 총오차율은 5.04%이다 (통계청, 2001a). 순오차율 수준을 비슷한 시기에 2000년 센서스 사후조사를 실시한 주변국들과 비교해 보면 일본의 1.11% 보다는 높고, 중국의 1.81% 보다는 낮은 수치이다(Zhang and Cui 2004, Takami 2003).

총조사의 순오차율은 1975년부터 1990년 까지 점차로 감소하였다가 다시 증가하는 추세이다. 그러나 순오차율은 몇가지 측면에서 총조사의 정확성과 신뢰성을 평가하는데 문제가 있다. 만약, 과대집계율과 과소집계율이 둘다 동일하게 높을 경우, 서로의 영향이 상쇄되어 순오차율이 제로에 가깝게 나타난다는 문제점이 있다. 실제로 1990년 총조사의 경우 누락이 3.46%이고, 조사오차가 3.43%로 총오차율은 6.89%였지만, 순오차율 -0.04%로 역대 조사 중 가장 낮게 나타난다.

이러한 현상은 연령별로도 나타난다. 20-29세 연령은 누락(5.42%)과 조사착오(4.24)가 가장 많이 발생한다. 총오차는 9.66%로 가장 높지만, 순오차율은 1.19%로 기타 연령에 비해 가장 낮다. 사후조사에서 발견되는 오차유형별 특성은 뒷장에서 보다 자세하게 논의될 것이다. 현행 사후조사의 근본적인 문제는 센서스에서의 누락이 무작위적으로 분포된 것이 아니고, 개인의 특성에 따라 차별적으로 누락된 것이라면, 이러한 특성을 가진 사람들은 사후조사에서도 누락될 가능성이 높다는 데 있다.

Ⅳ. 총조사 범위오차의 원인과 유형분석: 사후조사 결과분석

1. 범위오차의 원인

센서스에서 조사된 인구가 실제 인구보다 적거나 더 많은 것으로 나타났다면, 그 오차가 어디서 발생되었는지를 이해하는 것은 매우 중요한 문제이다. UN(2001)에 의하면 센서스에서 과소집계와 과대집계가 발생하는 원인은 i) 부정확하거나 불완전한 조사지도나 조사명부, 지리적으로 접근이 어려운 경우, ii) 이동중이거나 거처가 불명확해 조사되기 어려운 경우, iii) 센서스 요건이 공중에게 잘못 전파된 경우, iv) 조사 요원들이 조사의 정의와 절차를 잘못 이해한 경우, v) 관리 감독의 부재로 인해 조사활동에 대한 품질확신이 부족한 경우 등 다섯 가지로 분류하고 있다.

가. 과대집계

과대집계를 발생시키는 원인은 대부분 중복된 조사표 작성과 착오로 조사에 포함된 경우에 발생한다. 최근 미국의 센서스 기록상 가장 오차율이 높았던 1990년 센서스를 사례로 과대집계의 유형을 살펴보자[4]. 사후조사(Post Enumeration Survey)와 주택단위 커버리지 연구(Housing Units Coverage Study)결과를 분석한 Griffin과 Moriarity(1992)의 연구에 따르면, 센서스 우편조사에서 조사오차의 약 87%는 중복이나 착오로 조사에 포함된 경우였다. 나머지 오류는 주소지가 잘못 조사되었거나, 허위로 작성된 경우였다. 중복이 발생하는 이유는 응답자가 우편으로 배달받은 조사표를 작성한 상태에서, 무응답 팔로우업 기간에 조사원 방문시 다시 그 해당가구의 조사가 진행되었기 때문에 발생한다. 착오로 조사에 포함된 경우가 과대집계의 50%를 차지하는데, 대부분 센서스의 상주자 원칙을 잘못 이해한 경우에 발생한다. 이는 조사표 설계상의 문제점과 연관되는데, 누가 조사에 포함되어야 하는지에 대한 명확한 정의가 조사표에 제시되었어야 함을 의미한다.

4) 미국의 센서스는 4월 1일 센서스 데이를 기점으로 우편조사방식을 주로 하여, 원거리 지역이나 주소지가 불명확한 경우에는 조사원 면접방식으로 진행된다. 센서스 기간 직후 조사원 면접방식에 의한 무응답 팔로우 업이 실시된다. 이때 센서스 트랙 전부를 조사원이 방문 하는 경우도 있고, 표본에 선정된 가구만 방문하는 경우들이 있다. 표본 선정방식은 센서스 트랙별로 주소지 숫자 대비 응답가구율을 파악하여, 90%의 응답률을 목표로 표본수를 결정한다. 예를 들어 60퍼센트 인 지역에서는 조사대상가구 4개당 3개를 무응답 팔로우업의 대상으로, 85% 이상인 경우는 2개당 1개를 방문하는 방식이다 (Dimirtry and Treat, 2000)..

과대집계의 유형은 조사표 작성자, 가구규모, 주택의 특성, 조사방식에 따라서 다르게 나타난다. 먼저 우편조사의 경우, 착오로 조사된 사례를 응답자 특성에 따라 분류해 볼 때 조사표가 가구원에 의해서 작성된 경우는 약 3%의 오차를 보이는 반면, 주변사람이나 집주인등에 의해서 작성되는 경우는 7%가 넘는 오차를 보였다. 주변사람에 의해서 작성될 경우 그 주소지에서 조사되어서는 안 될 사람이 착오로 조사되는 경우가 많다는 것을 의미한다.

가구의 규모도 과대집계와 연관이 있는데, 1인 가구 오차율은 2.9%인 반면, 5인 이상의 가구는 오차율이 3.7%에 달한다. 다가구가 살고 있는 주택의 경우 오차율이 4.0%인 반면, 단독가구는 2.8%에 달한다. 주인이 점유하는 주택의 오차는 2.8%인 반면, 임차인 점유 주택의 경우 3.9%의 오차율이 있다.

조사원 인터뷰 방식으로 조사표가 작성된 경우, 가구원에 의해서 응답이 이루어진 경우는 오차가 7.7%인 반면, 친척이나 집주인등에 의해서 작성된 경우는 13.4%에 달한다. 이는 우편조사와 마찬가지로 가구원에 의해서 조사가 완료될 때 오차가 낮아 질 수 있음을 의미한다. 반면에 조사원 인터뷰 방식에서 1인가구의 오차율(10.6%)은 5인 이상 가구보다(8.5%) 더 높게 나타났다. 이것은 조사원이 만나기 힘든 대상의 정보를 가공으로 만들어 내는 경우가 있음을 의미한다.

요약하자면, 과대집계의 주요 원인은 중복이나 착오에 의한 조사이다. 센서스의 상주자 원칙을 조사대상자 혹은 조사원이 명확하게 이해하지 못한 경우, 가구원 이외의 사람이 응답한 경우, 다가구가 점유한 주택, 세든 가구일수록 오차율이 높다. 또한 우편조사는 5인 이상의 규모가 클수록 오차율이 높은 반면, 조사원면접조사 경우는 1인가구가 중복이나 착오로 조사에 포함되기 쉽다.

나. 과소집계

조사되었어야할 사람이 조사에서 빠진 경우인 과소집계도 무작위로 발생하는 것이 아니라 일정한 패턴을 가지고 있다. Simpson과 Middleton(1997)이 미국, 캐나다, 호주, 영국의 센서스 결과를 비교한 후 누락의 유형을 다음과 같이 정리 하였다. 연령에 따라 누락률이 차이가 나는데 유년 인구는 어릴수록 조사되기 어려운

반면, 장년인구에서는 나이가 많을수록 보다 쉽게 조사된다. 누락의 전형적인 사례가 연령에 따른 성비의 차이이다. 센서스에서는 여자보다는 남자가 누락되는 경우가 많은데, 단기간 이동자이거나 독신인 경우에 많이 발생한다.

다른 연구 결과에 따르면 미혼이거나, 45세 이하이거나, 19-44세의 남성이거나, 가구주와 관련되지 않은 친척의 경우 누락될 확률이 높다. 정리하자면, 누락은 주로, 사회에서 주변적인 위치이거나, 지리적 이동성이 높거나, 거주지나 동거가구와의 관계가 일반적이지 않은 환경에 살고 있는 경우 발생한다.

과소집계는 과대집계와는 정반대의 경우이지만 그 원인별 유형을 살펴보면 공통점을 발견할 수 있다. 접촉하기 어렵거나 센서스의 상주자 원칙적 적용이 모호한 경우라는 점이다. 미국의 1990년 센서스의 예로 돌아가자. 가구 내에서 가구원이 누락되는 경우가 1.8%, 가구 전체가 누락되는 경우가 2.0%, 거처자체가 누락되는 경우가 1.8%이다.

오차사례 중 9%는 응답자의 비전형적인 거주형태 때문에 발생하며, 이중 5%는 가구대표와 친척이 아닌 기타 가구원 사이의 상주개념 차이로 인한 것이다. 가구대표가 기타 동거인을 가구원으로 보고하지 않을 확률은 반대 경우에 비해 3배 이상이나 높기 때문이다 (Martin and Griffin, 1994).

가구원을 정확히 기록하지 않는 이유는 가구원에 대한 정의가 불명확하거나, 조사원이나 응답자가 가구원 개념을 정확히 적용하지 못하고 있거나, 응답자의 거주형태가 복잡해서 판단하기 어렵거나, 응답자가 자신의 신상공개를 원치 않기 때문이다.

2. 총조사 범위오차 발생의 상대적 위험도 분석

가. 조사대상 특성

표1. 사후조사 대상자 특성

전체사례수	114,666
가구수	35,667
성	
남자	48.9%
여자	51.1%
연령	
0-9	13.2%
10-19	14.8%
20-29	14.9%
30-39	18.1%
40-49	16.8%
50-59	9.6%
60+	12.5%
거주지	
읍면	20.3%
동	79.7%

총조사에서 발생하는 오차의 유형을 분석하기 위해, 2000년 인구주택총조사 사후조사 자료를 이용하였다. 표1은 사후조사의 조사대상자의 특성을 정리한 것이다. 전체사례 114,666 명, 35,667 가구를 대상으로 분석이 실시되였다. 여자가 51.1%로 남자에 비해 약간 많았으며, 30대가 (18.1%) 가장 많고, 50대(9.6%)가 가장 적었다. 일반적으로 도시지역으로 분류되는 동지역 거주자가 전체 응답자의 79.9% 차지하고 있다.

나. 조사착오율과 누락율의 분포

표2는 사후조사 결과에서 나타난 조사착오율과 누락율의 분포를 개인특성별로 살펴본 것이다. 사후조사의 범위 오차는 성과 연령에 따라 다르게 나타난다. 이동성이 높은 남자가 주로 누락을 일으킬 것이라는 예상과 달리 남자와 여자의 누락율은 차이가 없지만, 조사 착오율은 여자보다 높았다. 누락율은 연령별로 차이를 나타내는데, 5세 이하 유년 인구의 누락율이 높고, 취학연령기에 낮아졌다가 점차

높아져서 20대에서 가장 높고 (4.4%), 다시 낮아지다가 60세 이후(2.5%) 급격히 상승한다. 젊은층의 누락은 학업이나 취업과 관련된 지리적 이동성이 높아서 발생한다면, 고연령층의 경우는 친척집이나 외지를 방문하고 있는 경우가 많아서 생기는 누락으로 보인다.

표2. 개인 특성별 총조사 조사착오율과 누락율 분포, 2000

변수	총오차율	조사착오율	누락률
성	4.4	1.8	2.6
남자		2.0	2.6
여자		1.5	2.6
연령	4.3	1.7	2.6
10세 이하		0.8	2.5
10-19세		1.8	2.1
20-29세		4.2	4.4
30-39세		1.6	2.5
40-49세		1.2	2.1
50-59세		0.9	1.9
60세 이상		0.9	2.5
결혼상태	4.7	2.0	2.7
미혼		4.2	4.1
기혼		1.0	1.8
사별		1.7	3.7
이혼		3.4	6.3
지역	4.2	1.7	2.5
서울		2.0	2.4
대구		1.9	2.8
광주		1.3	2.7
전남		3.6	4.3
충남		1.5	4.2
전북		2.2	2.9
경남		2.5	2.5
강원		1.9	3.0

결혼 상태에 따라서도 오차율이 다르게 나타나는데, 전반적으로 유배우자 보다는 무배우자의 오차율이 높게 나타난다. 이러한 경향은 다른 국가들에서도 공통적으로 발생하는 현상인데, 기혼자의 안정적인 주거형태 때문이다. 무배우자 집단 내에서도 는 오차 유형의 차이가 발생하는데, 미혼자는 조사착오율와 누락율이 모두 비슷하게 높은 반면, 이혼자의 경우는 누락율이 조사착오율보다 두 배 가까이 높

다. 이러한 차이의 일부분은 결혼상태와 연령의 상관관계에서 오는 것으로 보인다. 미혼자의 88%가 조사착오율과 누락율이 모두 높은 30세 이하인 반면, 이혼자는 누락율이 높은 40대에 가장 많기 때문이다.

범위오차 발생율을 지역별로 살펴보면, 특광역시 지역에서는 대구가 4.7%로 가장 높았고, 도지역에서는 전남이 7.9%로 높게 나타났다. 오차율의 차이의 원인은 대구의 경우는 타지역 비해 상대적으로 60세 이상의 인구에서 발생하는 오차율이 높았고, 전남의 경우는 젊은 연령층에서의 오차 발생률이 높았기 때문이다.

범위오차의 발생은 가구의 특성과도 연관되는데, 가구차원에서 발생하는 오차의 대부분은 1인 가구에서 발생한다. 이것은 1인 가구의 불안정한 주거 형태 때문이다. 1인 가구에서 발생한 오차율은 12.0% 인데, 2000년 당시 인구를 기준으로 이 오차율을 적용해 보면, 전국적으로 26만의 1인 가구가 누락 또는 중복등의 오차를 일으킨 것으로 추정할 수 있다. 1인 가구의 대부분은 조사원이 낮 시간에 만나기 힘든 젊은 계층 (19.1%)이고, 다가구 주택(57.0%)에 거주하는 경우가 많다. 오차의 원인을 세분화하여 살펴보면 1인 가구에서 발생한 중복오차의 59.7%가 가공의 인물 조사로 인해 발생했고, 1인 가구 누락오차의 55.3%가 세 들어 사는 가구가 차지한다. 응답방식을 다양화 시켜서 1인 가구의 조사참여율을 높이지 않는 한 오차율은 계속 높아 질 것으로 보인다.

표3. 가구특성별 총조사 조사착오율과 누락율 분포, 2000

변수	총오차율	조사착오율	누락률
가구규모	4.0	1.4	2.6
1인		4.2	8.0
2인		1.5	3.2
3인		1.1	1.6
4인		0.6	1.0
5인 이상		0.4	1.1
6인 이상		0.9	1.4
가구유형	4.5	3.5	1.0
가족		0.3	1.5
가족+가족이외		0.7	2.8
1인 가구		4.2	7.8
5인 이하 남남		0.3	13.1
6인 이상 남남		-	18.2
가구주와의 관계	4.4	1.8	2.6
가구주		1.4	2.6
배우자		0.8	1.7
자녀		2.3	1.9
자녀배우자		2.0	5.9
부모		2.0	5.0
기타친인척		1.8	10.9
기타동거인		2.3	18.5
거처종류	3.9	1.3	2.6
일반		1.7	2.3
다가구		2.0	3.9
아파트		0.7	1.1
연립		1.4	1.5
영업용건물		1.6	6.8
거주기간	3.0	0.4	2.6
1년 이하		0.5	5.9
1-2년		0.5	3.3
2-3년		0.4	2.5
3-5년		0.4	2.0
5-10년		0.3	1.2
10년 이상		0.4	1.1

가구주와의 관계를 살펴보면 일반적으로는 가구주와의 관계가 직접적이지 않고 멀수록 오차율이 높아진다. 자녀보다는 부모의 오차율이 높고, 기타친인척 보다는 기타동거인이 오차율이 높다.

범위오차는 거처의 특성과도 연관되는데, 전체 사례 수에서 차지하는 상대적으

로 낮긴 하지만 영업용 건물 내 주택(8.4%)과 같이 일반적인 주택이외의 거처에서 발생하는 오차율이 가장 높았다. 아파트(1.8%)에 비해 세들어 사는 가구가 많아 가구 전체가 누락되기 쉬운 다가구 주택(5.9%)이나 일반가구(4.0%)의 오차율도 높게 나타난다. 가구의 거주기간은 조사착오율과는 관계없지만 누락율과는 마이너스 상관관계를 갖는 것으로 보인다. 거주기간이 1년 미만인 경우 누락율은 5.9%이지만, 5년이상인 경우 누락율은 1.2%로 떨어진다.

조사원의 특성도 범위오차에 영향을 준다. 조사원이 남자일 경우 대체적으로 오차율이 여자에 비해 높지만, 특히 누락율이 높게 나타난다. 60대 이상과 같이 연령이 많거나 20대 이하와 같이 어린 경우는 조사착오율이 다른 연령에 비해 상대적으로 높았고, 20대와 50대의 경우는 누락율이 높았다. 조사원의 직업이 대학생일 경우 오차율이 가장 높았으며, 통반장일 경우는 조사착오율과 누락율이 모두 낮게 나타났다.

표4. 조사원 특성별 총조사 조사착오율과 누락율 분포, 2000

변수		평균	조사착오율	누락율
조사원 성		3.9	1.3	2.6
	남자		1.6	3.4
	여자		1.3	2.4
연령		3.9	1.3	2.6
	20대 이하		3.1	1.0
	20-29세		1.6	3.6
	40-49세		1.2	2.4
	50-59세		1.3	3.6
	60세이상		3.1	2.6
직업		3.9	1.3	2.6
	반장		1.4	2.2
	주부		1.4	2.5
	부녀회원		0.8	3
	대학생		2.1	5
	퇴직자		1	3

다. 범위오차 발생의 상대적 위험도

개인과 가구, 조사원의 특성 차이가 범위오차를 일으킬 확률에 미치는 상대적인 영향력을 파악하기 위해서 로지스틱 분석이 실행되었다. 결과는 표5,6,7에 제시되어 있다. 회귀계수가 1 이하 이면 괄호안의 준거 그룹에 비해 비교그룹에서 오차가 발생할 상대적인 확률이 낮은 것을 의미하고, 계수가 1 이상이면 비교그룹에서 오차를 일으킬 확률이 높은 것을 의미 한다.

(1) 개인차원

표5에서 개인의 특성 변수들이 조사착오 혹은 누락을 일으킬 확률을 얼마만큼 잘 설명하고 있는 지는 Model Chi-Square 값을 통해 검증해 볼 수 있다. 두 값을 비교해 보면 개인차원의 변수들은 조사착오(1,374) 보다는 누락(1,525)을 일으킬 확률과 더 잘 설명하고 있다는 것을 알 수 있다.

조사착오를 일으킬 상대적 확률은 여자가 남자의 86.1% 수준 정도로 낮았다. 이는 남자가 비전형적인 거주지에 거주하거나 동거가구와의 관계가 일반적이지 않은 환경에서 생활하고 있는 경우가 여자 보다 많기 때문에 조사 당시 상주자 원칙을 명확하게 적용하기 어려웠기 때문인 것으로 보인다. 하지만 누락을 일으킬 확률에 있어서 남자와 여자의 차이는 통계적으로 무의미한 것으로 나타났다.

연령에 따라 범위오차가 발생할 확률도 차이가 난다. 조사착오를 일으킬 확율은 20대가 10대에 비해 2.3배 이상으로 가장 높다가 연령이 증가하면서 점차 낮아지지만 어떤 연령층이든 10대 보다는 조사착오 확률이 높다. 누락을 일으킬 확률은 이와는 약간 다른 유형을 나타낸다. 누락을 일으킬 가능성은 조사착오의 경우와 마찬가지로 20대에서 가장 높지만, 40대 이후부터는 10대보다 누락될 확률이 낮아진다. 60세 이상의 누락확률은 10대의 58% 수준까지 떨어진다.

결혼상태의 차이는 조사착오와 누락을 일으킬 확률 양쪽 모두 유사한 방향의 영향력을 행사하고 있다. 기혼자가 조사착오를 일으킬 확률은 미혼자의 45% 수준이며, 누락확률도 32% 수준 밖에는 되지 않았다. 사별한 경우도 미혼자보다는 범위오차을 일으킬 확률이 낮았다. 하지만 이혼자의 경우는 미혼자와의 차이가 통계

적으로 유의미하지 않은 것으로 나타났다.

가구주와의 관계에 따라서도 범위오차 발생확률이 차이가 난다. 조사착오나 누락을 일으킬 확률 모두 가구주의 배우자가 가장 낮았다. 가구주보다 배우자의 오차 발생확률이 낮은 이유는 가구주에 누락과 착오가 가장 많이 일어난 계층인 1인 가구가 포함되어 있기 때문인 것으로 보인다. 또한 실제 조사시 가구주의 배우자인 여성들이 주로 가구원의 상태를 응답하기 때문에 다른 가구원에 비해 가구주와의 관계를 잘못 보고할 가능성이 낮기 때문인 것으로 보인다. 조사착오를 일으킬 확률은 가구주와의 관계가 직접적이지 않을수록 높았다. 다른 가구원에 비해 가구주의 직계존속의 조사착오 확률이 낮았고, 기타친인척이 기타동거인보다 잘못 조사될 확률이 더 높았다. 이와는 달리 누락의 경우는 가구주와의 관계가 가장 먼 기타친인척이나 기타동거인 보다 가구주의 손자녀나 형제등의 누락확률이 더 높았다.

도시에 거주하는 사람들이 농촌지역 거주자에 비해서 범위오차가 발생할 가능성이 낮았다. 읍면 지역 거주자에 비해 동지역 거주자의 조사착오 확률은 65.8%, 누락은 60.7% 수준이었다. 지역별로도 차이가 나는데, 특광역시들 중에서는 서울이 조사착오와 누락이 발생할 확률 모두 가장 높았다. 전남을 제외하면 대체적으로 한 지역의 조사착오 확률이 높으면, 누락을 일으킬 확률은 상대적으로 낮았다. 조사착오 확률은 전남이 서울의 약 1.85배로 가장 높았고, 누락 확률은 충남이 1.3배로 가장 높았다. 대전과 경기는 두 확률 모두 서울보다 낮았다.

표5. 개인 특성별 범위오차 발생의 상대적 위험도(Relative risk)

변수	Exp(b)	
	조사착오	누락
상수	.025	.108
성		
(남자)		
여자	.861*	1.079
연령		
(10-19세)		
20-29세	2.335**	1.426**
30-39세	2.026**	1.006
40-49세	1.782**	.919
50-59세	1.308	.745*
60세이상	1.233	.579**
결혼상태		
(미혼)		
기혼	.454**	.319**
사별	.695*	.421**
이혼	1.176	1.092
가구주와의 관계		
(가구주)		
배우자	.609**	.770**
자녀	1.783**	.272**
자녀배우자	1.485	2.188**
부모	1.575**	2.846**
배우자부모	2.132	2.846**
손자녀	1.745	5.992**
증손자녀	.057	.879
조부모	.062	.009
형제	2.487	7.018**
형제자녀	.954	1.246*
기타친인척	3.626**	3.454**
기타동거인	2.175**	1.913**
행정단위		2.904**
(읍면)		
동	.658**	.607**
지역		
(서울)		
부산	.452**	.689**
대구	.931	1.109
인천	.524**	.972
광주	.631**	1.116
대전	.572**	.716**
울산	.617**	.897
경기	.589**	.672**
강원	.981	.989
충북	1.365**	.765*
충남	.754*	1.303*
전북	1.137	.997
전남	1.855**	1.258*
경북	.824	.964
경남	1.298*	.895
제주	.964	.765
Model Chi-Square	1,374	1,525
자유도	37	37

* p<= .05 ** p<= .01

(2) 가구차원

표6은 가구차원의 특성 변수에 따라 범위오차를 일으킬 확률이 얼마나 차이가 나는지를 보여준다. Model Chi-Square 값을 비교해 보면, 누락은 주로 가구차원에서 발생하는 문제임을 알 수 있다. 가구변수들의 누락 확률 설명력(1,294)은 조사착오 확률 설명력(362)의 3배가 넘기 때문이다.

조사대상자가 일반적인 가족의 일원인 경우 범위오차를 발생시킬 확률이 가장 낮다. 가족에 비해 1인가구는 조사착오 (3.6배)나 누락을 일으킬 확률(2.7배)이 모두 높았다. 가족이외의 남남으로 이루어진 가구에서는 조사착오 보다는 누락이 심각한 문제가 되고 있음을 알 수 있다.

표6. 가구 특성별 범위오차 발생의 상대적 위험도(Relative risk)

항목	Exp(b)	
	조사착오	누락
상수	.009	.052
가구유형		
(가족)		
가족+ 가족이외	1.091	2.078**
1인가구	3.607**	2.762**
5인이하 남남	1.019	5.288**
6인이상 남남	1.128	6.392**
거처종류		
(일반)	.774**	1.030
다가구	.821*	.375**
아파트	.579**	.482**
연립	.502**	.583**
다세대	1.072	2.209**
영업용	.940	.863
오피스텔	4.353	1.424
호텔	.024	3.118*
기숙사	.886	7.248**
거주기간		
(1년이하)		
1-2년	1.008	.705**
2-3년	.851	.609**
3-5년	1.284*	.581**
5-10년	1.577**	.505**
10년이상	2.174**	.380**
Model Chi-Square	362	1,294
자유도	18	18

* p<= .05 ** p<= .01

일반주택 보다는 공동주택에서 범위오차가 발생할 확률이 낮았다. 대표적인 공동주택인 아파트의 경우 일반주택 거주자에 비해서 조사착오 확률은 87%, 누락 확률은 32% 수준이었다. 조사착오는 거처보다는 주택에서 주로 발생할 확률이 높았고, 누락은 기숙사(7.2배)이나 다세대주택(7.2배)과 같이 공동으로 많은 사람들이 거주할 경우에 발생할 확률이 높았다.

거주기간에 따라서도 범위오차 발생 확률이 차이가 난다. 거주기간과 누락은 부의 상관관계를 나타내고 있는데, 그 거처에 10년 이상 거주한 경우 누락 확률은 거주기간 1년이하의 최근 이주한 사람의 약 1/3순준으로 떨어진다. 반면에 조사착오를 일으킬 확률은 거주기간이 길수록 높아진다.

(3) 조사원차원

표7의 Model Chi-Square 값을 살펴보면 조사원의 특성 변수는 범위오차 발생 확률을 설명하는데 있어서 다른 차원의 변수들에 비해 영향력이 적다는 것을 알 수 있다. 하지만 조사원의 성이나 직업에 따라서 범위오차의 발생확률은 유의미한 차이가 있다.

조사원이 여자인 경우 조사착오(77%)나 누락(71%)을 일으킬 확률 모두 상대적으로 낮았다. 조사원의 연령이 20대 이하인 경우와 비교할 때 50대인 경우는 2.3배 이상으로 가장 높은 누락 확률을 나타냈다. 통계적으로 유의미하지는 않았지만 30대 연령층의 조사원이 누락과 착오를 일으킬 확률에 있어서 가장 안정적인 것으로 보인다.

표7. 조사원 특성별 범위오차 발생의 상대적 위험도(Relative risk)

항목	Exp(b)	
	조사착오	누락
상수	.025	.016
조사원 성		
(남자)		
여자	.772**	.711**
연령		
(20대 이하)		
20-29세	.595	1.678
30-39세	.591	1.488
40-49세	.648	1.557
50-59세	1.014	2.282*
60세이상	1.630	1.806
직업		
(반장)		
주부	1.390**	1.397**
부녀회원	1.107	1.165
대학생	1.610**	1.865**
퇴직자	1.791**	1.222
기타	1.206*	1.462**
Model Chi-Square	135	132
자유도	11	11

* p<= .05 ** p<= .01

조사원이 반장인 경우 기타 다른 직업에 비해서 조사착오나 누락을 일으킬 확률이 가장 낮았다. 반면에 대학생인 경우에 조사착오를 일으킬 확률은 반장인 경우에 비해 1.6배, 누락은 1.8배이상 높았다.

V. 결론

지금까지 본 연구는 1990년부터 2000년 총조사까지 발생해온 범위오차를 다양한 평가방법을 통해 측정해 보고, 2000년 총조사 사후조사 결과를 이용하여 범위오차 발생의 원인과 유형을 개인, 가구, 조사원의 특성차원에서 파악해 보았다.

논의된 결과는 다음의 네가지로 요약될 수 있다.

첫 번째, 기존의 범위오차의 개념을 재정의 해야 한다는 점이다. 범위오차는 총조사에서 조사되었어야 할 사람이 빠졌거나, 조사되어서는 안될 사람이 조사될 때 발생한다. 후자는 조사표가 중복으로 작성된 경우와 상주자 원칙이 잘못 적용되어 조사된 경우를 포함하지만 이전까지는 이를 모두 중복이라는 개념으로 포괄적으로 지칭해 왔다. 현재까지는 총조사가 대부분 조사원 면접 방식이라는 단일한 조사방식으로 진행되어 왔다. 따라서 조사표의 중복작성 여부를 확인할 필요성이 적었기 때문에 중복과 착오에 의한 조사를 분리해서 측정하지 않았다. 하지만 응답율을 높이기 위해서 인터넷 조사등의 새로운 응답방식이 총조사에 활용될 경우 중복과 상주자 원칙의 잘못된 적용을 명확히 분류해서 생각해야할 필요성이 대두될 것이다. 따라서 본 연구에서는 범위오차는 누락과 조사오차로 구성되며, 조사오차는 중복과 착오에 의한 조사를 포함하는 포괄적인 개념으로 재정의 할 것을 제안하고자 한다.

두 번째, 과거 총조사의 범위오차를 인구균형방정식, 인구 추계치 비교, 코호트 생잔율법 및 사후조사를 통해 평가해 보았다. 평가결과를 종합해 보면 총조사의 범위오차는 차수마다 증감이 있지만 최근에 오면서 점차로 안정화된 패턴을 찾아 가고 있다. 하지만 이동성이 높은 젊은 연령층인 20-35세 연령층과 이들의 자녀 세대인 5세 이하 연령의 누락은 지속적으로 발견되고 있다. 또한 여자보다는 남자가 누락되는 경우가 많다는 사실을 알 수 있다.

향후 총조사의 커버리지를 향상시키기 위해서는 지금까지 어느 정도 수준에서 얼마만큼의 범위오차가 발생해 왔는지에 대한 평가 뿐만 아니라 왜 그리고 어떠한 유형으로 오차가 발생해 왔는지를 파악하는 것도 중요하다. 2000년 총조사 사후조

사를 분석한 결과 조사착오를 일으킬 확률은 개인의 특성 변수들과 관계가 깊다면, 누락은 주로 가구차원에서 발생할 확률이 높은 것으로 나타났다. 조사원의 연령과 같은 특성도 범위오차에 통계적으로 유의미한 영향을 주고 있다는 점이 확인되었다. 앞으로 조사방법과 관련된 오차에 관해서도 심도 있는 논의가 필요할 것으로 보인다.

지금까지 총조사의 오차측정은 총조사의 자료의 질을 평가하고, 조사과정이 당초 기획한 데로 이루어 졌는지를 점검하는 사후적인 수단으로서만 사용되어 왔다. 그러나 앞으로 오차의 평가는 단순히 오차수준과 내용을 지적하는데 그치는 것이 아니라, 총조사 이후의 인구추정과 추계과정에 있어서 총조사 수치를 보다 참값에 근접하게 보정하기 위한 기초자료로도 활용될 수 있는 만큼 그 유용성을 극대화시킬 수 있는 방안에 대한 검토가 필요한 것으로 보인다.

참고문헌

Bongaarts, J. and R. A. Bulatao, Eds. 2000. *Beyond Six Billion*. Washington, DC: National Academy Press.

De Heer, Wim. 1999. "International Response Trends: Results of an International Survey." *Journal of Official Statistics* 15:129-142.

Diffendal, Gregg, 2001, "The Hard-To-Interview in the American Community Survey." *American Statistical Association Proceedings 2001.* Alexandria, VA: American Statistical Association.

Dimitri, Robert C. James B. Treat. 2000. "Examination of Census2000 Initial Response Rates and the '90 PLUS FIVE PROJECT.

Fosu, Gabriel B., 2001, "Evaluation of Population Census Data through Demographic Analysis." Symposium on Global Review of 2000 Round of Population and Housing Censuses: Mid-Decade Assessment and Future Prospects, UNSD, New York.

Griffin, Deborah. H. and Christopher Moriarity. 1992. "Characteristics of Census Errors." *American Statistical Association Proceedings 1992.* Alexandria, VA: American Statistical Association.

Hogan, Howard. 2000, "Accuracy and Coverage Evaluation: Theory and Application." Dual System Estimation Workshop of the National Academy of Sciences Panel to Review the 2000 Census. U.S. Bureau of the Census, Washington D.C..

Martin, Elizabeth A. and Griffin, Deborah H. 1994. “The Role of Questionnaire Design in Reducing Census Coverage Error.” *American Statistical Association Proceedings 1994.* Alexandria, VA: American Statistical Association.

Moriarity, Christopher and Childers, Danny R. 1993. "Analysis of Census Omissions: Preliminary Results." *American Statistical Association Proceedings 1993.* Alexandria, VA: American Statistical Association.

Nash, Fay F. 2000. "*Overview of the Duplicate Housing Unit Operation.*" Census 2000 Informational Memorandoum No.78. U.S. Bureau of the Census.

Robinson, J.G., K. West, and A. Adlakha. 2002. "Coverage of the Population in Census 2000: Results From Demographic Analysis." *Population Research and Policy Review* 21:19-38.

Robinson, Gregory J. 2001. "Accuracy and Coverage Evaluation: Demographic Analysis Results." *DSSD Census 2000 Procedures and Operations Memorandum Series B-4*. U.S. Bureau of the Census.

Simpson, Stephen and Elizabeth Middleton. 1997. "Who is Missed by a National Census? A Review of Empirical Results from Australia, Britain, Canada, and the

USA." Working Paper 2. Manchester: Centre for Census and Survey Research, University of Manchester.

Takami, Akira. 2003. "Evaluation of Accuracy of the 2000 Population Census of Japan." The 21st Population Census Conference. Statistics Bureau of Japan.

UN, 2001, *Handbook on census Management for Population and Housing Censuses*, United Nations Department of Economic and Social Affairs Statistics Division: NewYork.

------. 1998. *Principles and Recommendations for Population and Housing Censuses (Revision 1)*, United Nations Department of Economic and Social Affairs Statistics Division: NewYork.

U.S. Bureau of the Census. 2001. "Our Homes, Our Neighborhoods," American Houing Brief. http://www.census.gov/mp/www/pub/con/mscho19a.html

White, Andrew A. and Rust Keith, F. 1997. *Preparing For the 2000 Census: Interim Report II.* Report of the Panel to Evaluate Alternative Census Methodologies. Washington, D.C.: National Academy Press.

Wolfgang Glenn, Peter P. Davis, and Phawn Stallone, 2001. "Accuracy and Coverage Evaluation Persons Not Matched in Census 2000," *American Statistical Association Proceedings 2001.* Alexandria, VA: American Statistical Association.

Zhang, Weimin and Hongyan Cui, 2004. "An Evaluation on the Accuracy of China's 2000 Census." International Seminar on China's 2000 Population and Housing Census. National Bureau of Statistics China.

김민경, 2000. 『인구센서스의 이해』, 통계청

통계청. 2004. KOSIS. http://Kosis.nso.go.kr/.

--------. 2002a. 『인구주택총조사보고서』. 통계청.

--------. 2002b. 『인구추계작성방법』. 내부자료. 통계청

--------. 2001a. 『2000 인구주택총조사 사후조사 분석결과』 내부자료. 통계청

--------. 2001b. 『장래인구추계』. 통계청.

--------. 1997. 『1995년 인구주택충조사 보고서』. 통계청.

--------. 1996. 『장래인구추계』. 통계청.

--------. 1993. 『1990년 인구주택총조사 보고서』. 통계청.