통계청 학술연구용역

표본추출 및 관리 매뉴얼

한국보건사회연구원

표본추출 및 관리 매뉴얼

손창균(한국보건사회연구원)
홍기학(동신대학교)
이기성(우석대학교)

제 출 문

통계청장 귀하

본 보고서를 귀 청과 용역계약 한『표본추출 및 관리 매뉴얼 개발』과제의 최종보고서로 제출합니다.

2006년 11월 30일

한국보건사회연구원

원 장 김 용 문

머 리 말

본 매뉴얼을 작성하게 된 주된 목적은 국내의 표본조사관련 연구자 및 실무자들이 올바른 표본조사 설계를 수행하는데 도움을 주고자 하는 것이다. 표본조사설계의 내용과 방법을 모두 이해하기 위해서는 많은 노력과 경험을 필요로 한다. 그러나 현실적으로 표본조사 설계 실무자들에게 그와 같은 기본적인 소양들을 모두 요구하는 것은 무리이며, 가능한 범위에서 그들의 실무에 도움이 되는 방향으로 너무 이론적이지 않으면서, 동시에 최소한의 수준을 유지할 수 있는 방향으로 본 매뉴얼을 작성하고자 노력하였다.

현재 국가적으로 한 나라의 얼굴이 되는 통계의 생산과정에서 기본이 되는 표본조사에 대한 밑그림이 제대로 그려졌는가를 판단하고, 만일 생산된 조사통계들의 신뢰성에 의문이 있다면, 어떤 점에서 문제가 있는 가를 분석해 볼 때, 각 조사통계생산기관들이 적용하고 있는 표본조사 설계의 빈약함 내지는 전문성 결여, 표본설계에 대한 이해부족 등 다양한 측면에서 문제점이 파악되었다. 이를 보다 구체적으로 살펴보면, 다음과 같다.

첫째, 표본수 산정식, 조사대상, 표본추출단위 등의 정의가 모호하거나 기관별로 용어의 통일성이 결여된 것으로 나타났다.

둘째, 모수추정식이 대부분 표본조사 교과서에서 소개되는 단순추정량 식으로 표현되어 무응답들의 문제를 조정할 가중치 적용이 거의 되어 있지 않은 것으로 파악되었다.

셋째, 통계작성 기관별로 표본추출방식에 대한 용어가 서로 다르게 표현되고 있다. 예를 들어, 유선무작위계통추출이란 용어는 통계학에서 층화임의 계통 추출로 통일해야할 것이다.

넷째, 추정을 위한 통계조사와 단순집계만을 위한 통계조사를 구분하여 매뉴얼 작성 작업이 필요한 것으로 파악되었다. 왜냐하면 단순집계에는 추정식이나, 추정오차와 같은 내용이 불필요하기 때문이다.

다섯째, 기업체 부문의 경우 사업체와 기업체 용어의 혼용으로 이용자의 혼란이

가중되고 있다. 따라서 기업체 부분의 경우 해당 통계작성 기준을 명확히 제시할 필요가 있는 것으로 파악되었다.

본 매뉴얼은 다음과 같은 방향으로 실무자 또는 관리자에게 도움이 되고자 한다.

첫째, 표본조사 설계에 대해 기본적인 이론과 방법들에 대한 도움서의 역할을 할 수 있을 것이다.

둘째, 가구단위나 사업체 단위 표본설계를 수행하는 실무자들에게 실제적인 가이드라인을 제공하고자 한다.

셋째, 표본조사의 설계 및 표본의 관리 및 분석 등에 관한 기본적인 지식을 전하고자 한다.

넷째, 단순한 기초지식을 전달하는 교양서적과 같은 수준이 아닌 실무적으로 매우 필요한 가이드북의 역할을 하도록 하였다.

본 매뉴얼의 구성은 크게 표본설계 매뉴얼 부분과 분석을 위한 프로그램, 그리고 사례로 구성되어 있다. 매뉴얼의 내용은 체크리스트를 부여하여 실무자가 각각을 확인함으로서 내용을 구성할 수 있도록 하였다.

본 연구는 손창균 부연구위원의 책임 하에 동신대학교 홍기학 교수와 우석대학교 이기성교수가 공동으로 참여하였다.

본 연구에 대한 검독을 위해 소중한 조언을 해주신 청주대학교 류제복 교수와 연세대학교 김재광 교수, 한국보건사회연구원의 장영식 연구위원, 이연희 부연구위원께 깊은 감사를 드린다.

2006년 11월

한국 보건사회연구원

원 장 김 용 문

목 차

표 목 차

그림 목차

Ⅰ. 표본설계 및 관리 매뉴얼

1. 표본설계란?

표본설계(sampling design)란 추출틀 중에서 모집단을 대표할 수 있는 일부의 단위 집합을 추출하는 과정이다. 표본설계는 작성되는 통계의 품질에 직접적으로 영향을 미치며, 통계가 요구하는 정확도, 예산상의 제약, 표본추출 방법의 이용가능성, 활용 가능한 보조정보의 수준, 이용될 조사기법 등 다양한 요소들을 고려하여 이루어지게 된다. 주어진 여건 하에서 가장 경제적이고 정확성이 높으며 효율적인 표본을 설계하는 것이 표본설계의 목표이다.

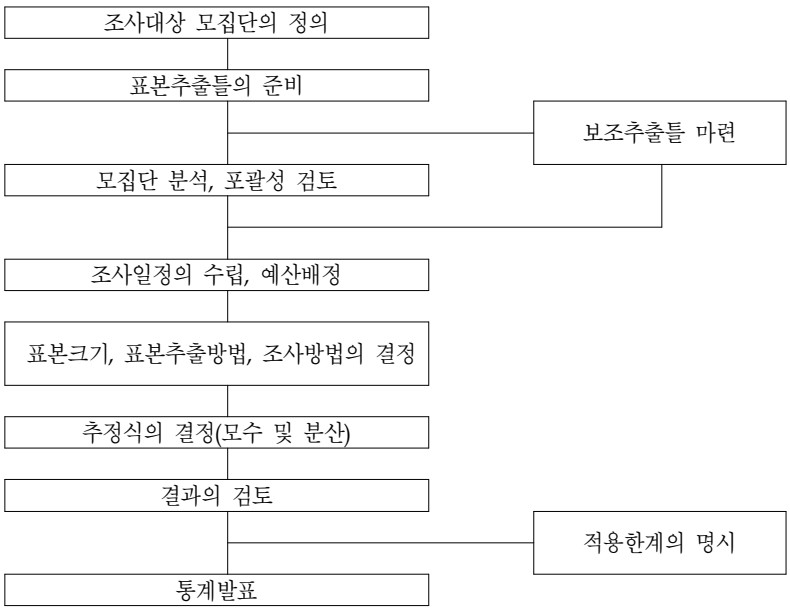
일반적으로 표본설계는 고도로 정교하고 과학적으로 유도된 표본이론을 기초로 하여 이루어진다. 하지만 설계자는 현장에서 조사가 이루어지는 일련의 모든 과정을 염두에 두어 효율성뿐만 아니라 강건성(robustness)까지 고려한 표본설계를 해야 한다. 어떤 설계는 최적의 조건 아래에서는 매우 효율적이지만 예상하지 못한 상황이 생겨 조건에 일부 변화가 생길 경우에는 예상치 못한 문제를 야기하게 되는 경우가 있다. 가령 높은 무응답, 너무 지리적으로 표본이 산재되었을 때 오는 조사상의 어려움 등의 문제가 언제든지 생길 수 있다. 따라서 이런 예상치 못한 상황 아래에서도 충분히 대처 가능한 표본설계가 이루어져야 한다. 이런 점에서 표본설계는 과학적인 면을 지니는 동시에 다른 한편으로 예술적인 면도 지닌다고 할 수 있다.

조사대상이 되는 모집단을 정의하고, 조사 가능한 조사모집단인 표본추출틀을 마련한다. 표본추출틀은 조사목적에 맞으면서 이용 가능한 다양한 추출틀을 확보하여 비교한 후 가장 적합한 추출틀을 정한다. 그리고 조사일정을 수립하고 일정에 따른 비용을 산정한다. 또한 실제로 조사에 필요한 표본크기와 표본추출법, 조사방법을 결정하는 등 조사에 필요한 구체적인 계획들을 마련하고 모수 추정식을 만든다.

표본의 크기는 클수록 조사의 정밀도는 높아지지만 표본의 크기가 너무 커지면 조사비용이나 조사를 위한 노력이 많이 들기 때문에 경우에 따라서는 조사의 질을 떨

어뜨리는 요인이 될 수도 있다. 그러므로 표본의 크기는 조사의 목적에 맞는 목표정도를 정한 후 그것을 만족시키는 범위 내에서 가능한 한 작게 하는 것이 바람직하다. 표본설계 과정을 간략히 순서도로 표현하면 다음과 같다.

<그림 Ⅰ-1-1> 표본설계 및 조사의 흐름도



표본추출법은 각 추출단위가 표본으로 추출될 확률을 사전에 미리 정한 뒤 우연현상에 의해 표본 단위를 선택하는 추출법인 확률추출법을 적용하는 것이 바람직하다. 대표적인 확률추출법으로는 무작위추출법, 층화추출법, 계통추출법, 집락추출법 등이 있다.

자료를 수집하는 방법에는 조사자가 조사단위에 대하여 측정도구를 이용하여 직접 관찰하거나 측정하는 방법과 사회조사에서 많이 이용되는 우편조사, 전화조사, 직접 면접조사, 인터넷 조사 등이 있다. 표본조사에서 어떤 조사방법을 선택할 것인가는 각 조사방법의 특성을 감안하는 동시에 조사목적과 조사의 경비, 시간들을 함께 고려하여 가장 적합한 조사방법을 선택해야 한다.

표본설계 과정에서 중요하게 다루어야 할 몇 가지 중요한 주제로는 모집단과 추출틀, 층화, 표본의 크기 결정 및 표본배분, 추출단계 및 추출법, 추정식 등이 있으며 그밖에도 다목적 표본설계, 비대칭성이 큰 모집단의 문제, 향후 표본보정의 가능성 등을 다루는 것이 필요하다.

2. 표본설계시 검토사항들

(1) 조사목적의 규정

조사주제를 정하고 주제와 관련된 개념들을 정리하여 조사목적을 정하는 단계이다. 새로운 조사를 개발하거나 기존의 조사를 다시 설계할 때에는 조사의 목적을 명확하게 세우는 것이 필요하다. 같은 조사라 하더라도 조사목적이 무엇이냐에 따라 조사의 방향이나 규모, 방법이 달라질 수 있으므로 명확하고 구체적인 조사목적을 세우는 일이 매우 중요하다.

(2) 조사범위 및 조사단위의 결정

조사범위는 조사목적에 의해서 규정되는 것이지만 인원, 비용, 시간 등 제반여건을 감안하여 결정하여야 할 것이다. 조사단위는 조사의 대상을 뜻하는데 통계집단을 구성하는 단위와 반드시 일치하는 것은 아님을 주의해야 한다. 예를 들면 인구주택총조사에서 인구라는 통계집단의 단위는 사람 각자의 단위이지만 실제조사에 있어서는 가구를 조사단위로 하여 조사하는 것과 같다. 또한, 조산단위는 집계단위, 표준분류 적용단위, 표본추출단위 등과도 구별되어야 한다.

(3) 조사사항의 결정

조사사항의 결정은 통계조사기획 중 가장 중요한 과정의 하나이다. 조사사항은 조사목적 특히 이용목적에 따라 결정하되 조사조직의 역량, 조사원 및 응답자의 부담

등을 고려하여 신중하게 결정할 필요가 있는데 조사사항을 결정할 때 주의할 사항은 다음과 같다.

□ 응답자가 사실 그대로 응답할 수 있는 사항인가?
□ 응답자가 쉽게 이해할 수 있는 사항인가?
□ 객관적 파악이 가능한 사항인가?
□ 수량에 관한 것은 응답자가 장부나 기록된 것을 보유하고 있는 사항인가?

이와 같은 사항들을 종합적으로 검토, 판단하여 조사사항을 결정해야 한다.

(4) 조사방법의 선택

통계조사를 하는데 있어서 대상의 전부를 조사하느냐에 따라 전수조사와 표본조사로 구분할 수 있다. 또한 조사표의 배부, 기입 및 수집하는 방법에 따라 타계식조사(면접조사, 전화조사 등)와 자계식조사(배포조사, 우편조사, 인터넷조사 등)로 구분할 수 있으며 이들 중 어떤 방법을 선택할 것인가는 각 조사방법의 특성을 감안하는 동시에 조사목적과 조사의 경비, 시간들을 함께 고려하여 가장 적합한 조사방법을 선택해야 한다.

각 조사방법에 대하여 간략히 정리해 보면 다음과 같다.

① 면접조사

조사자가 응답자(피조사자)를 직·간접적인 방법으로 접촉하여 조사하는 모든 방법은 넓은 의미에서의 면접조사(interview)라 할 수 있다. 따라서 조사자가 응답자를 직접 만나 조사하는 면대면 조사(face-to-face interview)는 물론 전화를 매개체로 하여 조사자가 질문하면 응답자가 응답하는 전화 면접(telephone interview), 온라인상에서 조사자가 응답자에게 질문을 하면 응답자가 온라인으로 응답을 하는 인터넷 조사(internet survey) 등을 모두 망라하기도 한다. 하지만 면접조사는 조사자가 응답자를 직접 만나 조사하는 협의의 개별면접조사(personal interview) 혹은 면대면 조사를 의미한다.

② 전화조사

전화조사(telephone survey)는 전화 보급의 대중화로 오늘날 선거 여론조사나 시장조사 등과 같이 신속을 요하는 조사에서 많이 사용되고 있는 조사방법으로, 조사원이 응답자에게 전화를 걸어 조사내용을 질문하면 응답자가 대답하는 내용을 조사원이 기록하여 자료를 수집하는 방법이다.

③ 우편조사

우편조사(mail survey)란 조사 대상자가 우편으로 보내진 설문지에 응답을 기입하여 다시 우편으로 반송하게 함으로서 자료를 수집하는 방법이다. 그러나 때로는 설문지의 배부와 회수 중에서 어느 한 쪽만 우편을 이용하고 다른 한 쪽은 조사자가 직접 방문하여 배부나 회수를 하는 경우도 있다.

④ 인터넷 조사

인터넷 조사에 대한 정의를 살펴보면 첫째 전산망을 통하여 전산망 가입자에게 직접 질문지 파일을 보내고 응답파일을 받는 형태, 둘째 인터넷상에서 이루어지는 통계조사를 총칭, 셋째 사이버 공간에서 인터넷을 활용하여 조사 자료를 수집하는 방법, 넷째 인터넷 사용자들을 대상으로 웹 또는 전자메일을 이용하여 설문을 진행하고 응답하는 일련의 행위 등 학자들에 따라 다양한 형태를 띄우고 있다.

(5) 조사기준시점, 대상기간, 실시시기의 결정

조사의 기준시점은 「2005년 11월 1일 0시 현재의 인구」등과 같이 파악하고자 하는 정보의 시간적 기준을 말하는 것으로 조사결과의 이용목적에 따라야 한다. 대상기간은 「2006년 1월 1일에서 12월 31일까지의 부가가치 생산액」등과 같이 일정기간으로 정해진 조사대상기간을 말하는데 조사내용을 가장 명확하게 파악할 수 있고 또한 조사결과의 비교가 가능할 수 있도록 정해야 한다. 그리고 조사기간은 실제조사에서 소요되는 일정한 기간을 말하는 것인데 그 기간의 길이는 조사대상수와 조사

내용 및 조사원수에 의하여 결정되어야 하며, 허용되는 예산과 인력의 범위 내에서 최대한 기간을 줄이도록 하는 것이 바람직하다.

(6) 표본설계시 유의사항들

표본설계 과정에서 중요하게 다루어야 할 몇 가지 중요한 주제로는 모집단과 추출틀, 층화, 표본의 크기 결정 및 표본배분, 추출단계 및 추출법, 추정식 등이 있으며 그밖에도 다목적 표본설계, 비대칭성이 큰 모집단의 문제, 향후 표본보정의 가능성 등을 다루는 것이 필요하다. 이를 체크리스트화 하면 다음과 같다.

가. 모집단에 대한 정의

□ 모집단에 대한 정확한 정의가 이루어졌는가?
□ 모집단 정의시 기준년도 등에 관한 설정은 이루어졌는가?
□ 목표모집단과 조사모집단에 대한 조작적 정의가 이루어졌는가?

☞ 가구단위 조사
□ 가구표본의 경우에는 어떠한 모집단을 대상으로 하고 있는가?
□ 표본추출 대상 제외 가구에 대한 명시적 정의가 있는가?
□ 가구표본의 경우에는 표본대상 가구원의 연령 등에 대한 제한점은 고려하고 있는가?

☞ 사업체 또는 기업체 단위 조사
□ 사업체 또는 기업체의 분류단위는 무엇인가?
□ 매출액 또는 종업원 규모별 자료가 모집단자료로 구성되어 있는가?
□ 사업체 또는 기업체 표본의 경우에는 어떠한 모집단을 대상으로 하고 있는가?
□ 사업체 또는 기업체 표본의 경우에는 모집단의 분포 형태를 고려하고 있

는가?

나. 추출틀(표집틀)에 대한 정의

□ 추출틀은 현실적으로 구성이 가능한가?
□ 추출틀의 포괄성은 어느 정도인가?
□ 추출틀의 구성형태는 어떠한가?
□ 이용하고자 하는 추출틀은 리스트 프레임인가? 아니면 Area 프레임인가?

☞ 가구단위 조사

□ 가구단위 조사인 경우 행정 프레임 또는 조사구 프레임을 사용할 것인가?
□ 주 추출틀 이외에 사용가능한 보조프레임은 없는가?

☞ 사업체 또는 기업체 단위 조사

□ 사업체리스트 또는 기업체 리스트 프레임과 그 외 보조프레임은 없는가?
□ 조사구를 추출틀로 사용하고 있는가?

다. 층화의 결정

□ 층별로 추정량의 산출이 필요한가?
□ 층화추출설계가 가능한 모집단인가?
□ 층의 개수는 표본의 크기와 비교해서 적당한가?
□ 층별 표본배분 방법은 어떠한 방법을 사용했는가?
□ 고려한 층별 배분 방법은 적절한가?
□ 중요변수들의 변동계수를 고려하여 층화를 고려하였는가?

☞ 가구단위 표본

□ 가구단위 표본에서는 층화를 어떤 기준에 의해 수행하였는가?

□ **가구단위 표본에 대한 층수는 적절한가?**

☞ **사업체 또는 기업체 단위 표본**

□ **사업체 또는 기업체 단위 표본에서는 어떤 기준에 의해 층화하였는가?**

□ **사업체 규모별로 층화한 경우 전수 층과 표본 층을 구별하여 층화하였는가?**

라. 표본크기의 결정

□ 목표오차를 어느 정도로 고려하였는가?

□ 주어진 예산과 현실을 적절히 반영한 표본규모인가?

□ 표본크기 산정 공식은 어떤 공식을 적용하였는가?

□ 층화 표본설계인 경우 층별로 배분된 표본의 크기는 적절한가?

□ 결정된 표본크기와 더불어 예비표본의 크기까지 함께 고려하였는가?

□ 주요변수의 목표오차를 조정할 수 있도록 표본크기를 고려하였는가?

☞ **가구단위 표본**

□ **표본지역 또는 표본 조사구는 몇 개로 할 것인가?**

□ **지역별 또는 조사구별로 몇 개의 가구를 표본으로 선정할 것인가?**

□ **최종 표본가구 수는 몇 개인가?**

☞ **사업체 또는 기업체 단위 표본**

□ **산업별 또는 규모별 표본사업체 또는 기업체수는 몇 개로 할 것인가?**

□ **최종 표본사업체 또는 기업체는 몇 개인가?**

마. 표본추출방법

□ 모집단을 적절히 대표할 수 있는 표본추출방법인가?

□ 확률 표본추출방법을 사용했는가?
□ 집락추출인 경우 PSU에 대한 정의는 적절한가?
□ 복합 표본추출설계를 고려해야 하는가?
□ 자체가중 표본추출설계는 가능한가?
□ 불균등 확률추출방법을 사용했는가?

☞ 가구단위 표본

□ 가구단위 표본추출의 경우 조사구를 PSU로 고려했는가?
□ PSU의 추출방법은 무엇인가?
□ 최종 표본가구의 추출방법은 무엇인가?

☞ 사업체 또는 기업체 단위 표본

□ 최종 표본사업체 또는 기업체의 추출방법은 무엇인가?
□ 표본추출시에 사용한 보조정보가 있다면 무엇인가?

바. 가중치의 산정

□ 기본가중치는 계산하였는가?
□ 복합표본추출방법을 적용했다면, 각 단계별 가중치는 고려했는가?
□ 100%완전응답 표본이 아니라면, 무응답 가중치는 계산하였는가?
□ 사후층화 가중치는 계산되었는가?
□ 사후층화 가중치 산정시 고려된 방법은 무엇인가?
□ 최종적으로 구한 가중치의 변동을 검토하였는가?
□ 가중치의 효과를 검토하였는가?

☞ 가구단위 표본

□ 지역별 PSU의 추출확률은 고려하였는가?
□ 표본 PSU내의 표본가구의 추출확률은 고려하였는가?

□ **무응답 가구에 대한 가중치 조정 작업은 수행하였는가?**
□ **보조정보를 이용하여 사후층화 가중치 조정은 수행하였는가?**
□ **최종 가구가중치의 변동을 고려하였는가?**

☞ **사업체 또는 기업체 단위 표본**
□ **사업체 또는 기업체 추출확률은 고려하였는가?**
□ **층화 다단계추출인 경우 각 단계별 추출확률은 고려하였는가?**
□ **무응답 가중치 조정은 수행하였는가?**
□ **최종 표본사업체들의 가중치의 변동은 고려하였는가?**

사. 추정산식

□ 적용한 추정식은 표본설계를 적절히 반영하고 있는가?
□ 복합표본추출방법을 적용했다면, 각 단계별 가중치를 고려한 추정산식인가?
□ 사용한 추정량은 비편향성을 지니고 있는가?
□ 추정치는 평균인가? 총합인가? 아니면 비율인가?
□ 주요관심변수들의 추정치는 과거추정치와 시계열성을 유지하고 있는가?
□ 추정치에 대한 표준오차는 계산하였는가?
□ 비선형추정량인 경우 다양한 근사적인 추정방법을 고려하였는가?
□ 분산 추정산식은 적절한가?
□ 추정치의 계산을 위해 사용한 프로그램은 무엇인가?
□ 분산 추정을 위해 사용한 프로그램은 무엇인가?
□ 반복적인 분산 추정방법을 사용하였는가?

☞ **가구단위 표본**
□ **추정치는 가구당 평균 또는 총합인가?**
□ **가구의 추정치에 대한 표준오차는 계산되었는가?**
□ **표준오차의 계산식은 제시되었는가?**

☞ 사업체 또는 기업체 단위 표본
□ 산업별 추정치인가? 아니면 업체별 추정치인가?
□ 각 추정치에 대한 표준오차 값은 제시하였는가?
□ 표준오차의 계산식은 제시되었는가?

3. 표본설계

3.1 모집단의 정의

모집단(population)이란 조사목적에 의하여 규정되는 모든 조사단위의 집단이라고 할 수 있다. 조사결과 작성되는 통계는 모집단을 설명하는 통계가 되므로 모집단을 명확하게 규정하여 정의하는 것이 필요하다.

조사목적에 의하여 개념적으로 규정된 모집단을 목표모집단(target population)이라고 하고 표본추출을 위해 규정된 모집단을 조사모집단(survey population 또는 sampled population)이라고 부른다. 가능한 한 이 두 집단은 일치하는 것이 바람직한데 실제로는 일치하지 않는 경우가 많다. 모집단 정의와 관련하여 Kish(1979)는 모집단을 ① 목표모집단 ② 추출틀 모집단 ③ 조사모집단 ④ 추론모집단 등으로 구분하였다. 이러한 4개의 모집단은 서로 상이할 수 있기 때문에 어떤 조사에서든지 최종결과공표 전에 그 한계를 명시해야 한다.

예를 들어 농가소득조사의 경우 목표모집단을 농업소득이 총소득 중 중요한 부분을 차지하는 모든 가구들의 집합이 되어야 하는데 실제로는 각 가구의 소득을 미리 알 수 없는 관계로 조사모집단은 경지면적 300평 이상을 경작하는 가구의 집합으로 정의된다. 가구조사에서는 조사의 편의를 위해 도서지역의 가구들을 조사대상에서 제외한다. 조사모집단은 목표모집단보다 제한되어 있는 것이 보통이다. 따라서 두 모집단의 차이를 검정하는 작업이 반드시 필요하며 만약 두 집단 간에 차이가 많을 경우 다른 자료에 의해 이 차이를 보충하는 방안을 강구해야 한다.

모집단을 정의할 때 크게 두 가지 사항에 대해 명확히 규정해야 한다. 하나는 모집단의 내용으로 모집단에 포함되는 조사 단위들의 특성과 유형을 명확하게 정해야 한다. 다른 하나는 모집단의 지리적, 공간적, 시간적 범위를 명확히 밝혀야 한다는 점이다. 특히 연속조사인 경우 시간의 흐름에 따른 모집단의 변동을 적절히 반영할 수 있도록 해야 한다.

☞ 체크리스트

□ 모집단에 대한 정확한 정의가 이루어졌는가?
□ 모집단 정의 시 기준년도 등에 관한 설정은 이루어졌는가?
□ 목표모집단과 조사모집단에 대한 조작적 정의가 이루어졌는가?

☞ 가구단위 조사

□ 가구표본의 경우에는 어떠한 모집단을 대상으로 하고 있는가?
□ 표본추출 대상 제외 가구에 대한 명시적 정의가 있는가?
□ 가구표본의 경우에는 표본대상 가구원의 연령 등에 대한 제한점은 고려하고 있는가?

☞ 사업체 또는 기업체 단위 조사

□ 사업체 또는 기업체의 분류단위는 무엇인가?
□ 매출액 또는 종업원 규모별 자료가 모집단자료로 구성되어 있는가?
□ 사업체 또는 기업체 표본의 경우에는 어떠한 모집단을 대상으로 하고 있는가?
□ 사업체 또는 기업체 표본의 경우에는 모집단의 분포 형태를 고려하고 있는가?

☞ 매뉴얼

① 모집단에 포함되는 조사단위의 정의를 명확히 내릴 것.

가령 20세 이상의 성인을 모집단으로 정의한 경우를 생각해보자. 우리나라 사람들의 연령 개념은 다소 모호하여 만 나이도 있고 일반 나이도 있다. 또 호적상의 나이와 실제 나이가 다른 경우도 있다. 따라서 보다 명확한 정의를 내리려면 "호적에 등재된 생년월일이 1983년 6월 31일 이전인 성인"과 같은 식으로 해야 한다.

또 하나 모집단의 지리적, 공간적, 시간적 범위를 명확히 밝혀야 한다. 지리적, 공간적 범위를 밝힌다는 것은 모집단에서 제외되는 지역이 있는 경우 그것을 명확히 밝혀야 한다는 뜻이다. 우리나라 성인들을 대상으로 하는 조사에서 흔히 도서지역을 제외시키는 경우가 있다. 도서지역을 포함시킬 경우 조사를 위한 시간, 비용이 훨씬 많이 드는 까닭이다. 따라서 특정지역을 추출대상에서 제외시킬 경우 그런 부분이 모집단 정의에 명확하게 드러나도록 해야 한다. 한편 시간적 범위란 모집단이 어느 시점을 기준으로 했을 때의 모집단인지를 알리는 것이다.

② 목표모집단과 조사모집단의 차이를 검토할 것.

목표모집단과 조사모집단이 다른 경우 그 차이의 정도가 어느 정도인지를 비교 검토해야 하는데 이를 위해 필요한 자료를 확보하는 것이 필요하다. 두 집단간의 차이가 무시할 수 없을 정도라고 판단될 경우 조사모집단을 확대시키는 것은 검토할 수 있다.

☞ 목표모집단

조사자가 정보를 얻고자 하는 대상 모집단으로서 조사의 목적과 시행조건을 반드시 고려해야 한다.

☞ 조사모집단

특정한 조사조건하에서 실제 조사 가능한 모집단을 말하며, 특정한 조사조건이란 조사의 구성요건, 하부그룹, 조사지침 등 여러 가지가 될 수 있다.
참고로 추출틀 모집단과 조사모집단간의 차이점이 발생하는 요인으로는 ① 추적 불가능한 단위, ② 결측단위, ③ 일시부재, ④ 부재, ⑤ 응답불능, ⑥ 응답거부, ⑦ 기타무응답, ⑧ 중복 또는 외부단위 등이 있다.

☞ 사례) 모집단의 정의

<표 Ⅰ-3-1> 목표모집단과 조사모집단의 정의

조사의 종류	목표모집단	조사모집단
중소제조업체동향조사	중소제조업 사업체	5인 이상 300인 미만 96,350개 중소업체
가구소비실태조사(한국)	조사시점 당시 전국의 모든 가구	조사시점 당시 <표Ⅰ-3-2>에 해당되는 가구를 제외한 전국 전 가구
가구소비실태조사 (캐나다)	조사시점 당시 캐나다에 거주하고 있는 모든 주민	요양시설 등에 거주하거나 혹은 고정된 주소가 없는 사람들을 제외한 조사시점 당시 캐나다에 거주하고 있는 모든 주민

<표 Ⅰ-3-2> 가계소비실태 조사(한국)의 제외가구

구분	제외가구
2인 이상 가구	- 음식점, 여관, 하숙업을 경영하면서 주거를 겸하는 겸용주택내의 가구 - 영업을 위해 고용한 종업원이 2명 이상 같이 사는 가구 - 직장동료, 친구, 선후배 등이 같이 사는 비혈연 집단가구
1인 가구	- 15세 미만인 사람 - 사회시설에 있는 사람 - 병원에 입원중인 환자 - 2인 이상 가구에서 제외가구 유형에 해당하는 1인 가구

3.2 추출틀

조사모집단이 정의되고 나면 그 모집단을 묘사할 수 있는 틀이 필요한데 이를 추출틀(sampling frame)이라고 한다. 추출틀을 마련하기 위해서는 먼저 추출단위(sampling unit)를 무엇으로 할 것인가가 결정되어야 한다. 추출틀이란 바로 추출단위들의 목록이기 때문이다. 표본설계에서 효과적인 추출틀을 마련하는 일은 매우 중요한 일이다. 추출틀 작성의 단계를 살펴보면 다음과 같다.

1단계) 추출단위의 선택 - 비용, 정보형태, 단위의 안정성, 시간을 고려
2단계) 추출틀의 전용 - 정보의 수집 및 조직화 관리
3단계) 추출틀의 타당도 검토 - 검토범위와 정보의 품질 검토
4단계) 행정조직 - 실사조직
5단계) 유지관리 - 수정 및 보안

통계조사의 추출틀은 조사의 특성에 적합하면서도 모집단에 포함된 조사 단위들의 중복이나 누락을 최소화할 수 있는 추출틀이어야 한다. 추출틀의 설정, 이용, 유지 및 보완 등은 실제 적용이 가능하고 예산상 무리가 따르지 않는 범위 내에서 이루어져야 한다.

추출틀은 표본설계, 자료수집, 추적조사, 추정, 품질평가, 분석 등의 과정에서 고루 이용되므로 높은 수준의 품질이 요구된다. 추출틀에 포함오차가 생기고, 추출틀 내 조사 단위들의 특성을 나타내는 자료들은 오랫동안 갱신되지 않는 낡은 것일 경우 이로 인한 편향이나 오차가 생길 가능성이 커지게 되기 때문이다.

☞ 체크리스트

□ 추출틀은 현실적으로 구성이 가능한가?
□ 추출틀의 포괄성은 어느 정도인가?
□ 추출틀의 구성형태는 어떠한가?

□ 이용하고자 하는 추출틀은 리스트 프레임인가? 아니면 Area 프레임인가?

☞ 가구단위 조사

□ 가구단위 조사인 경우 행정 프레임 또는 조사구 프레임을 사용할 것인가?

□ 주 추출틀 이외에 사용가능한 보조프레임은 없는가?

☞ 사업체 또는 기업체 단위 조사

□ 사업체리스트 또는 기업체 리스트 프레임과 그 외 보조프레임은 없는가?

□ 조사구를 추출틀로 사용하고 있는가?

☞ 매뉴얼

① 조사목적에 적합한 추출틀을 정할 것.

조사목적에 맞으면서 이용 가능한 다양한 추출틀을 확보하여 비교한 후 가장 적합한 추출틀을 마련해야 한다.

한 가지 조사목적을 위해서 사용가능한 추출틀은 여러 가지가 있을 수 있다. 가령 가계조사를 한다고 할 때 주소명부, 인구주택조사구명부, 전화번호부 등이 추출틀로 활용될 수 있다. 이와 같이 사용가능한 다양한 추출틀을 파악한 후 각 추출틀의 적합성과 품질, 경제성 등을 비교, 검토한다. 일반적으로 가장 효과적이면서 널리 활용되는 추출틀로는 통계청에서 수행하는 인구주택총조사 결과를 토대로 작성되는 조사구 추출틀을 들 수 있다.

② 가능한 한 조사 관련 보조정보를 잘 갖춘 추출틀을 마련할 것.

추출틀에 각 조사단위의 특성을 나타내는 보조정보가 많으면 표본설계의 효율을 높일 수 있을 뿐 아니라 더 나아가 추정 과정이나 무응답을 위한 조처를 할 때 보조정보를 적절히 활용할 수 있어서 매우 효과적이다. 따라서 추출틀을 마련할 때 가능한 한 조사단위에 관한 정보를 많이 얻을 수 있는 추출틀을 구하는 것이 바람직하다.

③ 필요에 따라서는 복수의 추출틀을 활용할 수 있음.

다양한 추출틀이 존재하지만 불완전하거나 비용이 많이 소요되는 경우에는 현실적으로 활용가능성이 높은 복수의 추출틀을 동시에 사용할 수도 있다. 이때에는 가능하면 상호보완적인 성격을 지니는 복수의 추출틀을 활용하는 것이 바람직하다. 이 경우 통계의 품질을 보장하기 위해서 전문가의 도움을 받아 이론상으로 문제의 여지가 없는 지를 검토하는 것이 필요하다.

④ 추출틀의 포함범위를 주기적으로 평가하고 보정할 것.

추출틀이 조사모집단을 얼마나 포함하는지를 주기적으로 평가하여 적절한 조치를 취하여야 한다. 추출틀과 조사모집단 사이의 괴리는 추정 값의 편향을 초래하기 때문이다.

추출틀 전체나 그 중 일부 표본을 뽑아 모집단이나 그 하위집단의 비교 가능한 다른 자료와 비교한다. 계속조사일 경우 설계시점의 추출틀을 그대로 둘 것이 아니라 가능한 한 시의적절하게 업데이트하여 모집단의 변화를 반영할 수 있도록 하는 것이 필요하다. 조사관리자는 모집단의 변화를 알 수 있는 행정자료나 가공통계 등을 파악하여 수시로 변화 상황을 점검, 파악하여 변화의 정도가 일정 수준 이상이라고 판단될 때에는 추출틀을 부분적으로 갱신하는 것이 필요하다.

⑤ 추출틀의 품질을 유지, 향상시키기 위한 시스템을 갖출 것.

한 번의 조사로 끝나는 것이 아니라 장기적으로 계속 조사가 이루어지는 조사일 경우 추출틀의 품질은 항상 유지, 관리되어야 한다.

중복이나 누락을 방지하고 새로 발생하거나 소멸하는 단위를 파악하고, 조사 단위들의 특성의 변화를 제대로 갱신할 수 있도록 하는 절차를 구체화하여야 한다. 또한 관련자들에게 추출틀로 인해 생기는 오차의 심각성 및 추출틀 품질 유지의 중요성을 인식시켜야 한다.

⑥ 동일 모집단에 대한 조사에는 가능한 한 동일한 추출틀을 사용하여 일관성을 유지할 것.

계속조사이거나 유사한 다른 조사와 동일한 모집단을 대상으로 조사하는 경우 가능하다면 동일한 추출틀을 사용하여 조사들 간의 일관성을 유지하는 것이 바람직하다.

3.3 층화

표본설계에서 조사목적에 부합되도록 효과적인 층화(stratification)를 하는 것은 무엇보다도 중요한 일이다. 층화란 모집단을 특성에 따라 서로 동질적인 몇 개의 부분집단으로 나누는 과정이다. 모집단을 몇 개의 층으로 나눈 후에 각 층별로 독립적인 표본추출이 이루어지게 된다. 효과적인 층화가 이루어질 경우 추정의 효율을 높일 수 있으며 부분 통계의 생산이 가능하고 경비절감을 가져올 수 있으므로 층화를 잘 하기 위해 노력하는 것은 매우 필요하다.

실제 기업체나 사업체를 대상으로 하는 표본설계를 하는 경우 산업의 종류, 매출액, 종사자, 지역 등이 층화변수로 주로 이용되고 있다.

☞ 체크리스트

□ 층별로 추정량의 산출이 필요한가?
□ 층화추출설계가 가능한 모집단인가?
□ 층의 개수는 표본의 크기와 비교해서 적당한가?
□ 층별 표본배분 방법은 어떠한 방법을 사용했는가?
□ 고려한 층별 배분 방법은 적절한가?
□ 중요변수들의 변동계수를 고려하여 층화를 고려하였는가?

☞ 가구단위 표본

□ 가구단위 표본에서는 층화를 어떤 기준에 의해 수행하였는가?
□ 가구단위 표본에 대한 층수는 적절한가?

☞ 사업체 또는 기업체 단위 표본

□ 사업체 또는 기업체 단위 표본에서는 어떤 기준에 의해 층화하였는가?
□ 사업체 규모별로 층화한 경우 전수 층과 표본 층을 구별하여 층화하였는가?

☞ 매뉴얼

① 설계변수와 밀접한 연관성을 갖는 변수를 선정할 것.

층화변수를 고르는 일반적인 원리로는 설계변수와 밀접한 연관성을 갖는 변수를 선정해야 한다는 점이다. 다목적 조사인데 하나의 특정 변수만을 고려하여 층화를 하다보면 고려되지 않은 다른 변수들에는 부정적인 영향을 끼칠 수 있다. 그러므로 다목적조사에서는 가장 중요하다고 생각되는 복수의 층화변수를 선택하여 고려하는 것이 필요하다. 이 때 좋은 층화변수를 찾는 것이 층화의 핵심적인 사항이다.

② 복잡한 조사, 대규모 조사일 경우 층화 다단추출법에서 1차 추출단위에 대한 층화를 최선으로 고려할 것.

일반적으로 복잡한 조사, 대규모 조사일 경우 층화 다단추출법을 사용하는 것이 보통이다. 이때에는 최초의 추출단위인 1차 추출단위(primary sampling unit : PSU)에 대한 층화를 잘 하는 것이 가장 중요하다.

③ 관심영역에 대한 부분통계의 생산을 원할 때에는 반드시 이를 반영할 수 있는 층화변수를 선정할 것.

④ 층화의 효과를 극대화시키기 위해서는 층을 나눈 후 각 층 내의 모집단 단위들을 관심변수와 가장 관련이 깊은 보조변수의 크기 순으로 정렬한 다음 계통추출(systematic sampling) 방법으로 표본조사 단위들을 추출할 것.

⑤ 층화를 할 때 미리 층의 수를 제한하지 말고 가능한 모든 경우를 다 나눈 후 역으로 합쳐가면서 적절한 층의 개수를 정할 것.

⑥ 층을 나눈 후 모든 충에 대해 획일적인 표본추출방법을 적용시킬 필요는 없으며, 층에 따라서 표본추출의 방법을 달리하는 것도 고려할 것.

⑦ 추출틀에 층화에 필요한 보조정보가 충분히 들어있지 않은 때에는 이중추출(double sampling) 기법을 고려할 것.

이중추출이란 일차로 대규모의 표본을 뽑아 층화변수로 사용할 수 있으면서도 응답이 간편한 변수 값을 구한 후 이를 근거로 1차 표본단위를 층화한 후 각 층에서 일부의 2차 표본을 추출하는 방법이다.

3.4 표본크기의 결정

조사목적과 여건에 맞는 표본크기를 결정하는 단계로 표본의 크기가 커질수록 조

사의 정밀도는 높아진다. 이 점 때문에 일반적으로 조사자들은 가능한 한 표본의 크기를 크게 하려는 경향이 있다. 그러나 표본의 크기가 커지면 조사비용이나 조사를 위한 노력이 많이 들기 때문에 경우에 따라서는 조사의 질을 떨어뜨리는 요인이 될 수도 있다. 그러므로 표본의 크기는 조사목적에 맞는 목표정도(target precision)를 정한 후 그것을 만족시키는 범위 내에서 가능한 작게 하는 것이 바람직하다.

조사 자료를 이용하여 전체적인 추정 외에 세부적인 관심영역(모집단의 부분집합)에 대한 추정도 하는 경우에는 전체 추정값 외에 관심영역에 대한 추정의 목표 정도를 정한 후 이를 만족시킬 수 있는 표본크기를 구하여야 한다.

표본의 크기는 사용할 추정량이 무엇이냐에 따라서도 달라지므로 활용 가능한 보조정보가 있는지 여부들을 파악하여 미리 어떤 추정법을 사용할 것인지를 고려하여 표본의 크기를 정하는 것이 필요하다.

☞ 체크리스트

□ 목표오차를 어느 정도로 고려하였는가?
□ 주어진 예산과 현실을 적절히 반영한 표본규모인가?
□ 표본크기 산정 공식은 어떤 공식을 적용하였는가?
□ 층화 표본설계인 경우 층별로 배분된 표본의 크기는 적절한가?
□ 결정된 표본크기와 더불어 예비표본의 크기까지 함께 고려하였는가?
□ 주요변수의 목표오차를 조정할 수 있도록 표본크기를 고려하였는가?

☞ 가구단위 표본

□ 표본지역 또는 표본 조사구는 몇 개로 할 것인가?
□ 지역별 또는 조사구별로 몇 개의 가구를 표본으로 선정할 것인가?
□ 최종 표본 가구 수는 몇 개로 결정되었는가?

☞ 사업체 또는 기업체 단위 표본

□ 산업별 또는 규모별 표본사업체 또는 기업체수는 몇 개로 할 것인가?

□ 최종 표본사업체 또는 기업체는 몇 개인가?

(1) 표본오차와 표본크기의 관계

표본조사에서는 모집단의 일부인 표본을 조사해서 얻은 결과를 모집단 전체에 대한 것으로 일반화하기 때문에 필연적으로 표본오차가 발생한다. 표본조사를 보면 표본추출과정에서 구체적으로 어떤 조사 단위들이 표본에 포함되는가에 따라 추정값이 변하는데, 표본에 따른 추정값의 변동 정도를 수치로 나타낸 것이 추정량의 표본오차이다. 일반적으로 추정량의 표본오차는 추정량의 표준편차로 설명되는데, 추정량의 표준편차를 표준오차(standard error : SE)라 한다. 확률추출법에서는 추출 방법에 따라 모수를 추정하는 방법 및 표준오차를 계산하는 방법이 달라진다. 여기서는 단순임의추출에서의 표준오차에 대하여 살펴보기로 한다.

단순임의추출법에서는 표본평균 $\bar{y}$를 이용하여 모평균 μ를 추정한다. 따라서 실제 표본에서 얻어진 표본평균과 모평균의 차이, 즉, $|\bar{y}-\mu|$가 추정오차(error of estimation)이다. 추정량의 표준오차는 이런 추정오차가 평균적으로 어느 정도인가를 설명해 주는 값이라고 해석할 수 있다.

우리는 표본조사의 추정오차가 일정한 한계를 넘지 않기를 바라지만, 모평균 μ를 모르기 때문에 단 하나의 추정값에 대한 추정오차는 구할 수 없다. 그러나 확률 $1-\alpha$가 주어지면 근사적으로 다음의 조건을 만족하는 오차의 한계를 구할 수 있다.

$$P(|\bar{y}-\mu| \le e) = 1-\alpha$$

여기서 $1-\alpha$는 표본조사를 반복적으로 시행했을 때 추정오차가 e이하인 경우의 비율을 뜻하고, 그런 의미에서 $(1-\alpha)\times 100\%$를 추정의 신뢰도 또는 신뢰수준이라 한다. 일반적으로 오차의 한계 e는

$$e = (\text{신뢰계수}) \times (\text{추정량의 표준오차})$$

와 같은 식으로 구해지고, 표준오차는 추정량의 분산을 이용하여 구한다. 단순임의추출법에서 표본분산을 s^2이라 할 때 모평균 μ의 추정량 $\bar{y}$의 분산은

$$\widehat{V}(\bar{y}) = \frac{N-n}{N}\frac{s^2}{n}$$

의 공식으로 추정할 수 있다. 따라서 추정량 $\bar{y}$에 대한 오차의 한계는

$$e = z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{s^2}{n}} \approx z_{\alpha/2}\frac{s}{\sqrt{n}}$$

가 된다. 여기서 $z_{\alpha/2}$는 표준정규분포의 제 $100(1-\alpha/2)$ 백분위수를 나타낸다. 예를 들어, $\alpha = 0.05$일 때 표준정규분포의 제 97.5백분위수 $z_{0.025}$는 1.96이다.

한편, 모비율 p는 표본비율 $\hat{p}$를 이용하여 추정하며, 추정량 $\hat{p}$에 대한 오차의 한계는

$$e = z_{\alpha/2}\sqrt{\frac{N-n}{N}\frac{\hat{p}(1-\hat{p})}{n}} \approx z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

로 주어진다.

(2) 표본크기 결정

표본설계에서는 모집단에서 추출할 표본의 크기를 결정해야 한다. 그런데, 표본조사에서는 표본으로 추출된 조사 단위들을 실사하는 비용이 들기 때문에 표본이 필요 이상으로 크면 시간과 인력의 낭비를 초래하며, 반대로 표본의 크기가 너무 작으면 조사결과에 대한 정도나 신뢰도가 떨어진다. 따라서 표본설계에서 적절한 표본의 크기를 결정하는 것은 대단히 중요한 사항이다.

표본의 크기는 목표로 하는 추정오차의 한계, 즉, 목표정도(target precision)와 밀접한 관계가 있다. 일반적으로 표본조사에서 목표로 하는 정도를 설정하고, 그 목표정도를 달성하기 위해 필요한 표본의 크기를 결정하게 된다. 표본조사의 정도는 추정오차의 한계를 이용하여 나타낸다. 조사의 정도를 높이기 위하여 표본의 크기를 증가시키면 현실적으로 그만큼 조사비용이 증가되기 때문에 표본의 크기를 결정할 때는 조사비용도 동시에 고려해야 한다.

추정량의 표준오차는 어떤 확률 추출법을 사용하는가에 따라 달라지기 때문에 원

하는 목표정도를 달성하기 위한 표본의 크기도 추출법에 따라 달라진다. 따라서 표본설계를 할 때는 사전에 설정된 목표정도를 달성하기 위해서 어떤 추출법을 사용하는 것이 효율적인지 검토하고 이를 바탕으로 구체적인 표본의 크기를 정하게 된다. 실제로 표본의 크기를 결정할 때는 표본조사를 통해 얻고자 하는 조사결과, 조사방법, 조사원들의 업무량, 전체적인 조사비용 등을 복합적으로 고려해야 한다.

일반적으로 표본의 크기는 복합표본설계에서 이용 가능한 정보(층별 총합 또는 집락총합)가 없을 경우 먼저 단순임의추출법(simple random sampling)에 근거해서 구한다. 따라서 층화추출법(stratified sampling)이나 집락추출법(cluster sampling) 등에 적합한 표본의 크기를 구하기 위해서는 설계효과(design effect : *deff*)를 곱해줘야 한다. 표본의 크기가 증가하면 증가할수록 표본분산이 감소하기 때문에 추정량의 정도는 좋아진다. 따라서 표본의 크기는 추정량의 정도에 의존한다. 추정량($\hat{\theta}$)의 정도는 추정량의 표본오차, $SE(\hat{\theta})$, 오차의 한계, $z \times SE(\hat{\theta})$, 또는 변동계수, $SE(\hat{\theta})/\hat{\theta}$ 등으로 나타낼 수 있다. 보통 어떤 표본조사에서 표본의 크기를 구한다는 것은 주어진 추정량의 정도를 만족하도록 구하는 것을 의미한다.

표본의 크기를 조정함으로서 표본오차와 무작위무응답으로 인한 비표본오차는 관리할 수 있지만 그밖에 다른 비표본오차는 관리할 수 없다.

☞ 매뉴얼

□ 추정량의 정도 결정시 유의할 점들

조사자는 반드시 추정량에 대하여 자료정리와 분석 그리고 그 추정량을 바탕으로 이루어질 모집단에 대한 판정(결정, 분석) 측면에서 요구되어지는 것이 무엇인지를 살펴보아야 한다.

① 조사로부터 얻어진 추정량에 대하여 허용 가능한 불확실성의 정도는 어느 정도인가?

예를 들어 추정오차의 한계가 95% 신뢰수준에서 ±5%인가 아니면 그 이상인가 이하인가? 를 결정한다.

② 조사모집단에 포함되어 있는 부모집단영역(domain)들에 추정량이 필요한가?

조사모집단 전체를 대상으로 한 추정량에 요구되는 정도와 조사모집단 내 각 부모집단에 대한 추정량에 요구되는 정도는 반드시 구별되어야 한다. 예를 들어 전국을 대상으로 한 조사에서 어떤 추정량에 대한 정도가 전국적으로는 3%일 수 있지만, 도별로는 5%, 군별로는 10%가 될 수 있다.

③ 추정치에 대한 표본분산의 상대적 크기는 어느 정도이어야 하는가?

정도는 추정치의 크기를 고려해서 결정해야 한다. 보통 추정치에 대한 표본분산의 상대적 크기는 추정치의 10 ~20% 정도가 적당하다. 즉, 추정치의 크기가 10일 경우 이에 대한 추정오차의 한계(정도)의 크기는 1 또는 2가 적당하다.

④ 표본의 크기를 늘림으로서 얼마나 정도가 개선되는가?

정도는 표본의 크기를 늘림으로서 개선되지만, 그 개선의 폭은 표본의 크기에 선형 비례하지는 않는다.

만일 모집단의 크기 $N = 100,000$, 모비율 $p = 0.5$인 경우, 단순임의추출 하에서의 추정비율 $\hat{p}$에 대한 표본의 크기 변화에 따른 추정오차의 한계는 95% 신뢰수준에서 다음 표와 같다.

<표 Ⅰ-3-3> 단순임의추출 하에서 표본크기와 오차의 한계

표본의 크기	오차의 한계
50	±.139
100	±.098
500	±.044
1,000	±.031

위 표에서 보는 바와 같이 표본의 크기가 50에서 100으로 2배 증가했다고 해서 오차의 한계가 반으로 줄어들지는 않는다.

⑤ 가장 작은 오차의 한계를 갖는 가능한 한 가장 큰 표본을 선택하는 것이 최선의 방법은 아니다. 이러한 이유로는 부차모집단의 추정량의 정도도 함께 고려해야 하는 경우도 발생하기 때문이다.

□ 표본의 크기에 영향을 미치는 요소들

① 모집단 변동계수(Population coefficient of variation : CV)

조사모집단에서 각 특성들의 변동(variation)은 서로 다르며, 변동의 크기는 표본의 크기에 영향을 준다. 일반적으로 표본조사에서는 서로 다른 변동을 갖는 여러 개의 특성들을 측정한다. 어느 한 특성에 대해 주어진 정도를 만족시키는 충분한 크기의 표본을 구했다 하더라도, 그 보다 더 큰 변동을 갖는 특성에 대해서는 주어진 정도를 만족시키지 못할 수 있다. 따라서 주요한 특성들의 요구정도를 만족시키는 표본의 크기를 구하기 위해서는 가장 큰 변동을 갖는 특성에 맞도록 표본의 수를 결정해야 한다.

② 모집단의 크기(Population Size)

표본의 크기에 영향을 주는 조사모집단의 역할은 그 크기에 따라서 다른데, 조사모집단의 크기가 작을수록 표본의 크기에 큰 영향을 미친다.

예를 들어 모비율 $p = 0.5$인 경우, 단순임의추출 하에서 0.05의 오차의 한계와 95% 신뢰수준에서 모집단 수에 따른 p를 구하는데 필요한 표본의 크기는 다음과 같다.

<표 Ⅰ-3-4> 모집단 크기와 표본크기와의 관계(오차의 한계 : 0.05)

모집단 크기	표본의 크기
50	44
100	80
500	222
1,000	286
5,000	370
10,000	385
100,000	398
1,000,000	400
10,000,000	400

모집단의 크기가 작을수록 모집단 내에서 표본이 차지하는 비율이 커진다. 따라서 작은 크기의 모집단의 경우 표본조사가 아닌 전수조사를 하는 이유가 거기에 있다.

③ 표본설계에 따른 설계효과(design effect)

표본설계와 그에 따른 추정량은 정도에 많은 영향을 미친다. 일반적으로 주어진 정도를 만족시키는 표본의 크기는 층별 총합이나 집락별 총합과 같이 이용가능한 정보가 없을 경우 단순임의추출 하에서의 표본분산을 근거로 구해진다. 따라서 복합설계하에서 같은 정도를 갖는 표본의 크기를 구하기 위해서는 단순임의추출 하에서 구한 표본의 크기에 설계효과(design effect : *deff*)를 곱해 표본설계에 맞는 표본크기를 결정해야한다.

설계효과는 주어진 표본설계에서의 표본분산을 단순임의추출 하에서의 표본분산으로 나눈 값으로서 보통, 주어진 표본설계가 단순임의추출일 경우 *deff*=1, 층화추출일 경우 $deff \leq 1$, 그리고 집락추출일 경우 $deff \geq 1$ 값을 가진다. 설계효과 추정량은 같은 변수들에 대하여 시행된 이전 조사나 비슷한 변수들을 갖는 유사설계 등을 통하여 구한다.

☞ **사례)** 단순임의추출에 의해 뽑힌 $n = 240$ 명의 어린이들을 대상으로 홍역 면역

률을 조사한 결과 $\hat{p}_0 = 160/240 = 0.667$, $\hat{V}(\hat{p}_0) = 0.0009112$로 추정된 반면에 집락추출에 의한 추정치 및 분산추정치가 각각 $\hat{p}_c = 0.667$, $\hat{V}(\hat{p}_c) = 0.002760$으로 추정되었다면,

집락추출의 설계효과는 $deff = \frac{\hat{V}(\hat{p}_c)}{\hat{V}(\hat{p}_0)} = 3.029$,

집락추출의 결과가 단순임의추출과 같은 정도를 갖기 위한 표본의 크기는

$$n_c = 3.029 \times 240 = 726.96 \approx 727$$

가 된다.

---

④ 조사응답률

추정치에 대한 원하는 정도를 얻기 위해서는 표본의 크기를 예상응답률로 조정해주는 작업이 필요하다.

☞ **사례)** 만약 어떤 표본조사에서 주어진 정도를 만족시키는 표본의 크기가 $n = 400$으로 구해졌는데, 이 조사의 예상응답률이 75%라면 조사자는 533개의 표본을 추출해야 한다.

$$n' = \frac{400}{0.75} = 533$$

□ 표본의 크기를 구하는 공식들

다음 공식들은 단순임의표본을 이용해서 구한 모집단 평균이나 비율 추정치들에 대한 주어진 정도를 만족시키는 표본의 크기를 구하는 공식들이다.

① 절대오차한계를 조정하고자 하는 경우(모평균 추정)

[1단계] $\Pr(|\bar{y}-\mu| \geq e) = \alpha$를 정한다.

[2단계] 추정치의 표준오차를 구한다.

$$SE(\bar{y}) = \sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}} \tag{1}$$

[3단계] 얻고자 하는 추정오차의 한계를 설정한다.

$$e = z\sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}} \tag{2}$$

[4단계] 식(2)를 n에 대하여 푼다.

$$\begin{aligned} n &= \frac{z^2 S^2}{e^2 + \frac{z^2 S^2}{N}} \\ &= \frac{NS^2}{N\left(\frac{e}{z}\right)^2 + S^2} \end{aligned} \tag{3}$$

$O = \frac{e}{z}$ 로 놓으면, 식(3)은 다음과 같이 다시 표현된다.

$$\begin{aligned} n &= \frac{N S^2}{NO^2 + S^2} \\ &= \frac{N S^2}{1 + \frac{1}{N}\left(\frac{S}{O}\right)^2} \end{aligned} \tag{4}$$

α : 오차의 한계 e를 초과할 확률(위험)

e : 얻고자 하는 오차의 한계

z : 신뢰계수

N : 모집단의 크기

S^2: 모분산

② 상대오차의 한계(상대오차)를 조정하고자 하는 경우(모평균 추정)

[1단계] $\Pr(\left|\frac{\bar{y}-\mu}{\mu}\right| \geq e) = \Pr(|\bar{y}-\mu| \geq e\mu) = \alpha$를 설정한다.

[2단계] 추정치의 표준오차를 구한다.

$$SE(\bar{y}) = \sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}} \tag{5}$$

[3단계] 얻고자 하는 추정오차의 한계를 설정한다.

$$e\mu = z\sqrt{\left(1-\frac{n}{N}\right)}\frac{S}{\sqrt{n}} \tag{6}$$

[4단계] 식(6)을 n에 대하여 푼다.

$$\begin{aligned} n &= \frac{\left(\frac{zS}{e\mu}\right)^2}{1+\frac{1}{N}\left(\frac{zS}{e\mu}\right)^2} \\ &= \frac{\left(\frac{CV}{O}\right)^2}{1+\frac{1}{N}\left(\frac{CV}{O}\right)^2} \end{aligned} \tag{7}$$

α : 오차의 한계 e를 초과할 확률(위험)
e : 얻고자 하는 오차의 한계
z : 신뢰계수
N : 모집단의 크기
S^2: 모분산
$O = \frac{e}{z}$: 목표오차(허용오차), 추정치의 요구 변동계수
O^2 : 추정치의 허용(요구)상대분산(the desired $(CV)^2$)
CV : 변동계수(실제는 미지임)

상대오차를 조정하고자 하는 경우 표본의 크기는 변동계수의 값에 의존한다. 모분

산과 모평균을 모를 경우 똑같은 성격의 이전 조사의 결과를 이용하거나, 사전조사를 통해 모분산과 표본평균을 추정하고, 전혀 정보가 없을 때는 표본분산과 표본평균으로 대체한다.

---

③ 완전응답을 가정한 모비율의 추정에 필요한 표본크기

[1단계] 분산추정치 s^2을 구한다.

$$s^2 = \hat{p}(1-\hat{p})$$

[2단계] s^2을 앞의 식(2)에 대입해서 푼다.

$$n = \frac{z^2\hat{p}(1-\hat{p})}{e^2 + \frac{z^2\hat{p}(1-\hat{p})}{N}}$$

□ 표본크기를 결정하는 일반적인 방법

표본크기 결정방법에서 통상적으로 사용되는 방법을 단계별로 설명하면 다음과 같다.

① 절대오차의 한계를 조정하는 경우

[1단계] 식(2)에서 유한모집단수정항을 무시한 추정오차의 한계를 설정한다.

$$e = z\frac{S}{\sqrt{n}}$$

[2단계] 위 식을 n에 대하여 푼다.

$$n_1 = \frac{z^2 S^2}{e^2}$$

[3단계] n_1을 다음과 같이 조정해 준다.

$$n_2 = n_1\frac{N}{N+n_1} = \frac{n_1}{1+\frac{n_1}{N}}$$

위 결과는 n_1을 식(3) 대입해서 푼 것으로서, 이 때 n_2는 식(3)의 n과 같다.
즉, n_1은 복원추출에서의 표본의 크기이고, 식(3)의 n과 n_2는 비복원추출에서의 표본의 크기이다.

[4단계] 단순임의표본에 의한 표본이 아닐 경우 설계효과를 곱해준다.

$$n_3 = deff \times n_2$$

[5단계] 완전응답을 보장 못할 경우, 예상응답률(r)을 적합한 최종 표본의 크기를 구한다.

$$n = \frac{n_3}{r} \quad (8)$$

② 상대오차의 한계(상대오차)를 조정하는 경우

[1단계] 식(6)에서 유한모집단수정항을 무시한 추정오차의 한계를 설정한다.

$$e_{\mu} = z\frac{S}{\sqrt{n}}$$

[2단계] 위 식을 n에 대하여 푼다.

$$n_1 = \left(\frac{zS}{e_{\mu}}\right)^2 = \frac{1}{O^2}\left(\frac{S}{\mu}\right)^2 = \left(\frac{CV}{O}\right)^2$$

[3단계] n_1을 다음과 같이 조정해 준다.

$$n_2 = n_1\frac{N}{N+n_1} = \frac{n_1}{1+\frac{n_1}{N}}$$

위 결과는 n_1을 식(7) 대입해서 푼 것으로서, 이 때 n_2는 식(7)의 n과 같다.
즉, n_1은 복원추출에서의 표본의 크기이고, 식 (7)의 n과 n_2는 비복원추출에서의 표본의 크기이다.

[4단계] 단순임의표본에 의한 표본이 아닐 경우 설계효과를 곱해준다.

$$n_3 = deff \times n_2$$

[5단계] 완전응답을 보장 못할 경우, 예상응답률(r)을 적합한 최종 표본의 크기를 구한다.

$$n = \frac{n_3}{r}$$

3.5 표본의 배분

표본설계시에 주로 층화를 하게 되는데, 이 때 고려해야 할 또 하나의 중요한 문제는 각 층에 표본을 배분(allocation)하는 방법이다. 표본을 구성할 때 각 층에서 추

출할 표본의 크기를 얼마로 하는 것이 효율적인가를 결정해야 한다. 일정한 비용 하에서 추정량의 정도를 최대로 할 수 있는 최적의 표본배분 방법은 다음과 같은 요인들의 영향을 받는다.

① 각 층을 구성하는 총 조사단위의 수 - 층의 크기(N_h)
② 각 층을 구성하는 조사 단위들 간의 변동 - 층내 분산(S_h^2)
③ 각 층에서 조사단위당 실사 비용(c_h)

표본의 배본문제는 층화임의추출의 경우에 발생한다. 층화임의추출의 효율성은 표본의 배분방법에 크게 의존한다.

모집단을 서로 겹치지 않는 L개의 층으로 나눴다고 가정할 때 층화임의추출에서 모집단의 크기와 표본은 다음과 같이 나타낼 수 있다.

모집단의 크기 : $N = N_1 + N_2 + \cdots + N_L = \sum_{h=1}^{L} N_h$

표본의 크기 : $n = n_1 + n_2 + \cdots + n_L = \sum_{h=1}^{L} n_h$

(1) 배분기준

크기 n인 표본을 L개의 층으로 배분하는 기준은 크게 2가지가 있는데, 고정표본 기준(fixed sample size)과 고정변동계수 기준(fixed CV)이다.

① 고정표본 기준

주어진 정도를 만족시키는 전체 표본의 크기 n을 구한 다음 이를 각 층 들에 적당한 비율로 배분한다.

$$n_h = n \times a_h \ , \quad \left(\sum_{h=1}^{L} a_h = 1 \right)$$

층화임의표본을 이용하여 주어진 변동계수의 한도 내에서 모집단 총합을 구하는데 필요한 표본의 크기를 구하면 다음과 같다.

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^2} \quad (9)$$

N_h : 모집단의 h층의 크기

S_h^2 : h층의 모분산

a_h : h층에 할당된 표본비율

CV : 모집단 총합에 대한 지정된 변동계수

Y : 모집단 총합

위 식에 $n_h = n \times a_h$를 대입해서 정리하면,

$$n_h = a_h \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^2} \quad (10)$$

가 된다.

따라서 각 층에 대한 a_h의 값을 결정한 다음, 각 층의 표본의 크기 n_h를 계산한다.

② 고정변동계수 기준

주어진 정도를 만족시키는 표본의 크기를 각 층별로 구한 다음 이를 합쳐서 전체 표본의 크기를 구한다. 이 방법의 장점은 각 층 별로 요구되는 정도를 만족시켜주기 때문에 자연스럽게 추정치에 대한 전체 정도를 만족시켜준다는 것이다. 단점은 계산이 복잡하고, 표본의 크기가 처음 계획했던 것보다 커져서 비용이 많이 들어갈 수 있다는 것이다.

(2) 배분방법

① 비례배분 방법(proportional allocation)

각 층별로 모집단의 구성비만큼 표본을 배분하는 방법으로서 표본층의 구성비를

모집단층의 구성비에 맞추어 주는 방법으로 모집단에 대한 이용가능한 정보가 거의 없을 경우 활용되는 배분방법이다. 비례배분의 장점으로는 추정량의 식이 자체가중 추정량의 식으로 변환되어 추정식이 간단해지며, 단순임의 추출에 비해 분산이 작다.

$$n_h = \frac{N_h}{N} n$$

$$a_h = \frac{n_h}{n} = \frac{N_h}{N}$$

② 모집단 층합 비례배분 방법(Y-proportional allocation)

이 방법은 관심변수들의 분포가 왜도(skewness)되기 쉬운 기업체조사 등에 많이 이용된다. 예를 들어 제조업분야에서의 고용에 관한 조사나 도·소매업 동태조사 등에서 대표적으로 이용되는 표본배분방법이다.

$$a_h = \frac{Y_h}{Y}$$

③ $\sqrt{N}$ 비례배분 방법

이 방법은 조사에서 전체 추정치에 대한 정도보다는 각 층별로 추정치에 대한 정도를 개별적으로 관리하고자 할 때 유용하게 쓰인다.

$$a_h = \frac{\sqrt{N_h}}{\sum_{h=1}^{L} \sqrt{N_h}}$$

④ $\sqrt{Y}$ 비례배분 방법(역배분)

$\sqrt{N}$ 비례배분 방법과 마찬가지로 전체 추정치에 대한 정도보다는 각 층별로 추정치에 대한 정도를 따로 관리하고자 할 때 유용하게 쓰인다.

$$a_h = \frac{\sqrt{Y_h}}{\sum_{h=1}^{L} \sqrt{Y_h}}$$

⑤ 최적배분 방법(optimal allocation)

이 방법은 단위당 조사비용이 서로 다르고, 층별 분산간의 변동이 존재할 때 이를 감안해서 표본을 배분할 때 유용하게 쓰인다. 최적배분 방법을 이용하기 위해서는 조사비용을 모형화한 다음과 같은 비용함수기 필요하다.

$$C = c_0 + \sum_{h=1}^{L} c_h n_h$$

최적배분 방법을 이용해서 구한 할당모수(allocation parameter) a_h는 다음과 같다.

$$a_h = \frac{N_h S_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h S_h / \sqrt{c_h}}$$

⑥ 네이만(Neyman)배분 방법

단위당 조사비용이 모두 같다고 가정했을 때 쓸 수 있는 방법이다.

$$a_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h}$$

⑦ 각 층별 분산이 같을 경우 배분 방법

층별 분산이 모두 같다고 가정했을 때 쓸 수 있는 방법이다.

$$a_h = \frac{N_h / \sqrt{c_h}}{\sum_{h=1}^{L} N_h / \sqrt{c_h}}$$

(3) 층화추출에서 표본의 크기에 대한 일반적인 방법

이 방법은 조사자가 추정치에 대한 분산 값을 미리 한정하고 이를 만족시켜주는 표본의 크기를 구하는 방법이다.

① 일반적인 방법

식(9)는 다음과 같이 다시 쓸 수 있다.

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^2} = \frac{\sum_{h=1}^{L} W_h^2 S_h^2 / a_h}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2}$$

- 각 층에서 복원추출을 할 경우(유한모집단수정항을 무시할 경우)

$$n_1 = \frac{1}{V}\sum_{h=1}^{L} \frac{W_h^2 S_h^2}{a_h}$$

- 각 층에서 비복원추출을 할 경우(유한모집단수정항을 무시할 수 없을 경우)

$$n_2 = \frac{n_1}{1 + \frac{1}{NV}\sum_{h=1}^{L} W_h S_h^{\,2}}$$

② 네이만(Neyman)과 비례배분 방법인 경우

만약 $a_h \approx N_h S_h$, 즉, $a_h = \frac{N_h S_h}{\sum_{h=1}^{L} N_h S_h} \left(= \frac{W_h S_h}{\sum_{h=1}^{L} W_h S_h} \right)$ 라고 가정하면, 식(9)는 다음과 같이 쓸 수 있다.

$$n = \frac{\sum_{h=1}^{L} N_h^2 S_h^2 / a_h}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^2}$$

$$= \frac{(\sum_{h=1}^{L} N_h S_h)^2}{CV^2 Y^2 + \sum_{h=1}^{L} N_h S_h^{\,2}} \qquad (11)$$

$$= \frac{(\sum_{h=1}^{L} W_h S_h)^2}{CV^2 \mu^2 + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^{\,2}}$$

$$= \frac{(\sum_{h=1}^{L} W_h S_h)^2}{V + \frac{1}{N}\sum_{h=1}^{L} W_h S_h^2} \quad (12)$$

- 각 층에서 복원추출을 할 경우(유한모집단수정항을 무시할 경우)

$$n_1 = \frac{(\sum_{h=1}^{L} W_h S_h)^2}{V}$$

- 각 층에서 비복원추출을 할 경우(유한모집단수정항을 무시할 수 없을 경우)

$$n_2 = \frac{n_1}{1 + \frac{n_1}{N}}$$

위 식에서 V는 조사자가 원하는 추정치의 정도를 얻기 위하여 한정한 분산 값이다. V 대신 오차의 한계(e)를 이용한다면 $V = (e/z)^2$이 된다. 모평균에 대한 추정치일 경우 식(9) 또는 식(11)을 이용하는데 결과는 같다. 이 때 CV와 V는 각각 평균추정량에 대한 변동계수와 분산을 의미한다.

☞ **매뉴얼**

① 모집단 총합 비례배분과 $\sqrt{Y}$ 비례배분 방법을 이용할 경우 보조정보를 이용할 것.

② $n_h \geq N_h$일 경우에는 해당 층을 전수조사 할 것.

이 경우 전체 표본의 크기는 계획된 표본의 크기 n보다 작을 수 있다. 이럴 경우에는 표본이 작게 배분된 층의 표본의 크기를 좀 더 늘려 전체 표본의 크기를 맞춘다.

③ 각 층에 배분되는 조사단위의 수는 적어도 2개 이상이 되도록 할 것.

즉, $n_h \geq 2$인 배분방법에 따라 각 층에 표본을 할당하다보면 조사단위가 전혀 배분되지 않는 층이 생길 수 있다. 이럴 경우 해결책 중의 하나는 각 층에 각각 크기 2인 표본을 강제 배분하고 나머지를 배분방법에 의해 배분하는 것이다.

④ 여러 개의 변수에 근거해서 표본을 배분하는 방법을 고려할 것.

⑤ 층별 모분산, S_h^2을 모를 경우 표본분산, s_h^2으로 대체할 것.

3.6 표본추출법과 표본추출단계

표본추출법은 확률추출법(probability sampling method)과 비확률추출법(nonprobability sampling method)으로 구분되는 데 대부분의 표본설계에서는 추출틀 내의 각 추출단위가 표본으로 추출될 확률을 사전에 미리 정한 뒤 우연 현상에 의해 표본단위를 선택하는 추출법인 확률추출법을 사용한다. 다시 확률 추출법은 균등 확률추출법(equal probability sampling)과 불균등 확률추출법(unequal probability sampling)으로 구분된다.

표본에 입각한 모든 추정이론은 확률추출법으로 표본이 추출될 때라야 적용 가능한 것이므로 오늘날 과학적 표본조사라고 하면 당연히 확률추출법을 근간으로 한다. 대표적인 확률추출법으로는 단순임의추출(simple random sampling), 층화추출(stratified sampling), 계통추출(systematic sampling), 집락추출(cluster sampling) 등이 있으며 일반적으로는 이러한 추출법들을 복잡하게 결합하여 사용하는 것이 보통이다. 이와 더불어 비확률 추출법으로는 판단추출법(judgement sampling), 편의추출법(convenience sampling), 할당추출법(quota sampling), 눈덩이추출법(snowball sampling) 등이 있다.

표본추출방법을 결정하는 일은 통계전문가의 영역으로 매우 중요하다고 할 수 있다. 관심변수에 대한 사전 정보를 얻을 수 있는 경우에는 이를 층화변수로 적용하는 것이 효율적이다. 하지만 층화추출법을 적용하는 데 있어서 층의 크기나 층의 경계를 정하는 일은 간단하지만은 않다. 그리고 층화 2단계추출법 또는 층화 다단추출법을 적용해야할 것인지는 추출단위의 결정과 연계하여 고려해야 할 사안이다. 또한 층화추출법을 적용하더라도 추출단위에 대한 실제 추출과정에서는 단순임의추출법을 사용할 것인지 아니면 과거 조사결과나 외부정보를 이용하여 확률비례추출법을 사용할 것인지도 고려하여야 한다.

☞ **체크리스트**

□ 모집단을 적절히 대표할 수 있는 표본추출방법인가?
□ 확률 표본추출 방법을 사용했는가?
□ 집락추출인 경우 1차 추출단위(PSU)에 대한 정의는 적절한가?
□ 복합 표본추출설계를 고려해야 하는가?
□ 자체가중 표본추출설계는 가능한가?
□ 불균등 확률추출방법을 사용했는가?

☞ **가구단위 표본**

□ **가구단위 표본추출의 경우 조사구를 1차 추출단위(PSU)로 고려했는가?**
□ **1차 추출단위(PSU) 의 추출방법은 무엇인가?**
□ **최종 표본가구의 추출방법은 무엇인가?**

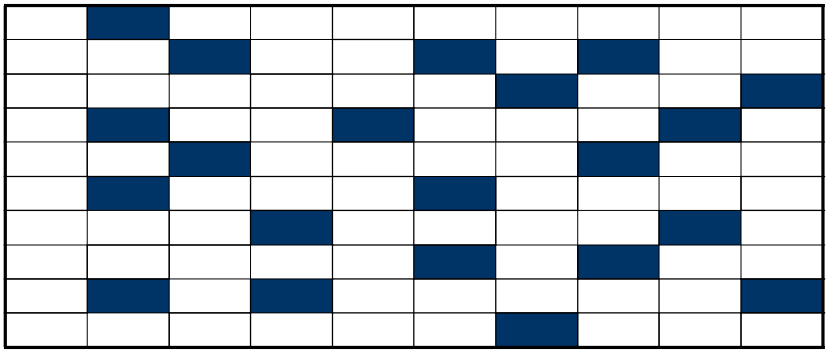
☞ **사업체 또는 기업체 단위 표본**

□ **최종 표본사업체 또는 기업체의 추출방법은 무엇인가?**
□ **표본추출시에 사용한 보조정보가 있다면 무엇인가?**

(1) 단순임의추출법

① 정의 및 개념

단순임의추출법(simple random sampling)은 가장 간단한 확률표본추출법이지만 다른 확률추출법의 기초가 되므로 가장 기본이 되는 추출법이다. 단순임의추출법이란 크기 N인 모집단에서 n개의 추출단위를 뽑을 수 있는 모든 가능한 경우의 각각에 표본으로 추출될 가능성을 동일하게 부여하여 표본을 추출하는 방법이다. 그리고 이런 방법에 의하여 얻어진 표본을 단순임의표본이라 한다. 랜덤(random)이란 추출단위에 동일한 추출확률을 부여하는 것인데, "기회 균등(equally likely)"의 의미로 해석할 수 있다. <그림 Ⅰ-3-1>은 $N=100$인 모집단에서 임의로 $n=20$개의 표본을 추출한 결과를 나타내고 있다. 이때 추출하는 과정은 난수를 사용하거나, 특정한 표본추출프로그램을 사용하여 조사단위를 추출하게 되는 데, 각 단위가 표본으로 추출될 가능성은 모두 1/50로 동일하게 된다.



<그림 Ⅰ-3-1> 단순임의추출($n=20$)

② 한계점

단순임의표본을 추출하려면 모집단에 속하는 모든 추출단위들의 목록이 필요하다. 그러나 추출틀을 구하는 것이 현실적으로 쉽지 않고, 또, 추출틀을 구했더라도 추출된 조사 단위들이 지리적으로 넓게 퍼져 있다면 실사비용이 많이 들 수 있으며, 더 낮은 비용으로 동일한 정도를 보장하는 다른 표본추출법들이 많이 있기 때문에, 실

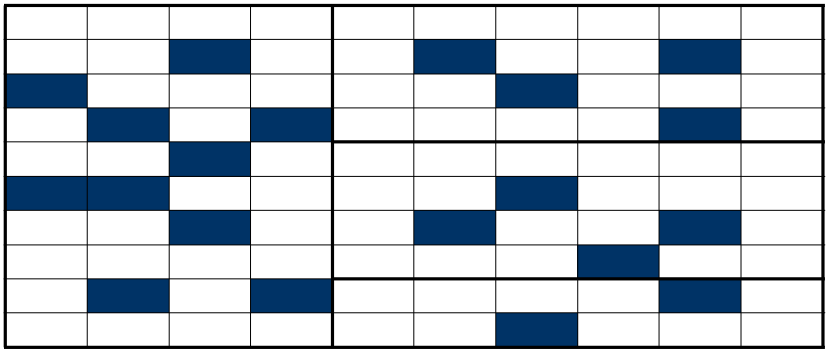
제 표본설계에서 단순임의추출법이 단독으로 사용되는 경우는 거의 없다.

(2) 층화추출법

① 정의 및 개념

표본조사에서 관심변수와 관련된 보조변수(auxiliary variable)에 대한 자료를 쉽게 얻을 수 있다면 효율적인 표본설계를 위하여 그 보조정보를 이용하여야 한다. 보조정보는 추정의 정도를 높이기 위하여 비추정량(ratio estimator), 회귀 추정량 등과 같은 모수의 추정량에 이용하거나, 표본설계 및 추출단계에 이용할 수 있다. 층화확률추출법은 표본설계에서 보조정보를 이용하는 표본추출법이다.

층화임의추출법(stratified random sampling)이란 모집단을 보조변수의 값이 유사한 추출단위들을 묶어서 만든 층(strata)들로 분할하고 각 층에서 단순임의추출법으로 표본을 추출하는 방법이다. 층화추출법으로 표본을 추출하려면 우선 모집단을 서로 겹치지 않는 부분집단으로 나누어야 한다. 일반적으로, 동질적인 추출단위들의 묶음이 되도록 모집단을 분할하는데 때에 따라서는 관심을 갖고 통계를 산출해내려는 집단별로 분할하기도 한다. 이렇게 분할된 부차모집단(sub-population) 각각을 층(stratum)이라고 부른다. <그림 Ⅰ-3-2>는 모집단이 $h=1,2,3,4$인 4개의 층으로 구분되고, 각 셀은 단위를 나타내는 것으로 $N=100$개로 구성된다. 그러면 이 모집단으로부터 $n=20$개의 표본을 추출하기 위해 첫 번째 층에서는 $N_1=40$으로부터 $n_1=10$개를 추출하고, 두 번째와 세 번째 층은 $N_2=N_3=24$인 모집단 층으로부터 각각 $n_2=4$, $n_3=4$ 인 표본을 추출하고, 네 번째 층은 $N_4=10$에서 $n_4=2$인 표본을 추출한 결과를 나타내고 있다.



<그림 Ⅰ-3-2> 층화추출($n_1 = 10$, $n_2 = n_3 = 4$, $n_4 = 2$)

층화추출법에서 모집단을 층화하는 기준으로 사용되는 변수를 층화변수(stratification variable)라고 한다. 층화임의추출법이 효율적이기 위해서는 전체 추출단위에 대하여 사전에 정보를 확보할 수 있는 적절한 층화변수를 선택하는 작업이 대단히 중요하다.

② 층화추출법이 널리 사용되는 이유

층화추출법이 실제 표본설계에서 널리 사용되는 이유를 살펴보면, 다음과 같다.

첫째, 필요에 따라 집단별(층별) 추정값을 얻을 수 있다. 집단별 평균 또는 비율에 관심이 있는 경우에 각 층과 관심 집단이 일치하도록 층화 작업을 하고 각 층별로 적절한 표본의 크기를 사전에 배정하면 어느 정도 신뢰할 수 있는 각 관심 집단별 추정값들을 얻을 수 있다. 이런 경우 층화 변수 및 각 층의 분류기준은 조사목적에 의해 자연스럽게 결정된다. 우리나라 통계청에서 실시되는 대부분의 조사들을 보면 전국을 광역시 및 도별로 층화하고 표본을 추출하여 전국에 대한 추정값 및 각 시도별 추정값을 산출해내고 있다.

둘째, 적절한 층화 변수를 확보하여 모집단을 층화하면 단순임의추출에 비하여 동일한 비용 하에서 표본조사의 정도(precision)를 높일 수 있다. 즉, 층화 변수의 기준에 따라 동질적(homogeneous)인 추출단위들끼리 하나의 층으로 묶어주는 층화 작업에 의하여 층 내의 변동이 적어지게 되므로, 추정의 정도가 높아진다.

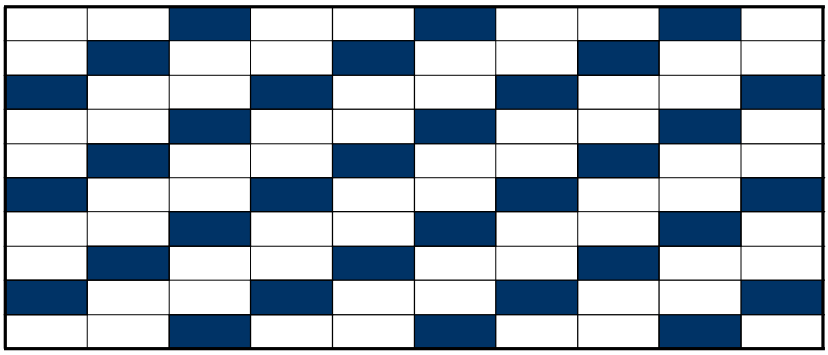
셋째, 단순임의추출에 비해 조사비용을 절감할 수 있다. 예를 들어, 지역별로 층을 구성하는 경우나 여론조사를 위해 성별-연령별로 층을 구성하는 경우에 유사한 속성을 지닌 개체들로 구성된 각 층별로 전반적인 조사관리를 할 수 있어서 조사관리가 용이하고 비용을 절감할 수 있다. 따라서 동일한 비용으로 단순임의추출에서 보다 더 많은 조사단위를 표본에 포함시킬 수 있어서 결과적으로 추정의 정도를 높일 수 있다.

(3) 계통추출법

① 정의 및 개념

단순임의추출법으로 표본을 추출하는 과정을 보면, 추출틀에 속하는 모든 추출단위에 고유의 표지번호를 부여해야 하는데 실제 표본설계에서 추출단위에 번호를 부여하는 작업이 현실적으로 어려운 경우가 많이 있다. 예를 들면, 전화번호부에 등재되어 있는 각 전화번호에 일련번호를 부여하는 작업이 그렇게 간단하지 않음을 알 수 있다. 이와 같은 단순임의추출에서의 어려움을 해결하기 위한 목적으로 표본추출과정을 보다 단순하고 편리하게 변형한 추출법이 계통추출법(systematic sampling)이다.

계통추출법은 실제 조사현장에서 단순임의추출 대신으로 매우 폭넓게 활용되고 있는 추출법이다. 계통추출이란 추출틀에 수록된 처음 k개의 추출단위들에서 하나를 랜덤하게 뽑고, 그 다음부터는 매 k번째에 해당되는 추출단위를 뽑는 추출법이다. 이러한 추출법으로 선정된 표본을 $1/k$ 계통표본이라 한다. <그림 Ⅰ-3-3>는 $N=100$인 모집단에서 추출간격이 3인 계통 표본 $n=33$을 추출한 결과를 나타낸 것이다. 즉 매 3번째 단위를 표본으로 추출한 결과 즉, 1/3 계통 표본을 내타내고 있다.



<그림 Ⅰ-3-3> 계통추출($n=33$)

결국, 시간적이나 공간적으로 일정한 간격을 두고 추출단위를 뽑는 방법이 계통추출법인데, 방법이 간편하고 모집단 전체에서 추출단위가 골고루 뽑히는 장점이 있다.

추출틀에 추출단위들이 랜덤하게 나열되어 있다면 계통추출법에 의해 추출된 표본은 단순임의표본과 거의 동일하다. 이럴 경우 계통표본은 단순임의표본과 이론적으로 거의 동일하게 취급되며 계통표본에서 얻어진 자료를 통계적으로 분석하는 과정에서도 단순임의표본에 적용되는 방법이 그대로 사용된다.

② 계통추출법이 널리 사용되는 이유

계통추출법은 다음과 같은 몇 가지 이유로 인하여 단순임의추출법 대신으로 사용되고 있는 유용한 표본추출방법이라 할 것이다.

첫째, 계통추출은 실제 조사 현장에서 적용하기 편리한 방법이다. 뿐만 아니라 적절한 추출틀이 마련되어 있지 않은 경우에도 조사자의 주관인 선택에 의해 발생하는 선택오차(selection errors)를 줄일 수 있다는 장점을 지닌다.

둘째, 모집단의 크기를 사전에 파악할 수 없어도 원하는 추출률에 따른 표본을 얻을 수 있다.

셋째, 관심변수와 관련된 특성에 따라 순서대로 나열된 추출틀을 확보할 수 있을 때는 계통추출법을 사용하는 것이 단순임의추출에 의한 것보다 모집단을 더 잘 대표할 수 있는 표본을 얻을 수 있다. 실제로 통계청 등에서 정부공식통계를 생산하기

위한 표본설계를 보면 표본의 대표성 제고를 위해 이런 형태의 계통추출법이 폭 넓게 활용되고 있다.

넷째, 계통추출은 단위비용 당 얻을 수 있는 정보의 양이라는 측면에서 단순임의추출보다 더 많은 정보를 제공하는 추출법이다. 다시 말해서, 계통추출은 추출과정이 간편하므로 비용 및 시간 측면에서 단순임의추출에 비해 효율적이다.

한편, 계통추출법을 사용하는 경우 주의할 사항이 있다. 우선, 이론적으로 계통추출은 단순임의추출과 상당한 차이가 있다. 예를 들어, 10% 계통추출법에서 표본으로 추출될 수 있는 가능한 경우는 정확히 열 가지이고, 랜덤하게 선정된 그 중 하나가 표본이기 때문에 이론적으로 추정량의 분산을 추정할 수 없다. 그래서 추출단위들이 랜덤하게 나열되었다는 가정 하에 계통표본을 단순임의표본인 것처럼 간주하여 추정량의 분산을 추정한다. 이 경우 일반적으로 표본오차가 실제보다 과대 추정되는 경향을 보인다. 또한, 추출단위들이 관심변수와 관련된 특성에 따른 주기성을 갖고 나열되어 있는 경우에 주기와 추출간격이 일치하게 되면 표본의 대표성에 심각한 문제가 생길 수 있다. 따라서 주기를 갖는 추출틀에서 표본을 추출해야 할 때는 다른 추출법을 사용하거나 주기와 추출간격이 일치하지 않도록 한 계통추출법을 사용해야 한다.

계통추출법은 다양한 분야에서 실제로 널리 활용되는 표본추출법으로 우리나라 인구주택 총조사에서는 표본조사를 실시할 가구를 선정할 때 조사구를 기준으로 10개 조사구에서 한 조사구를 표본조사구로 추출하는 10% 계통추출법을 사용하고 있으며, 품질관리를 위한 표본검사 또는 시장조사를 위한 표본조사에서도 계통추출이 많이 활용되고 있다.

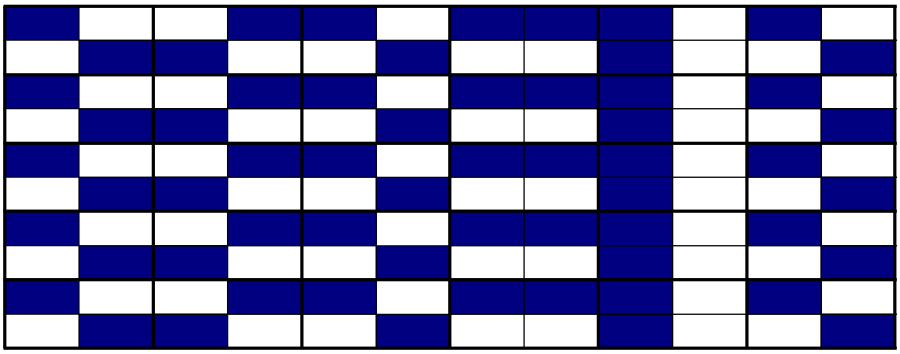
(4) 집락추출법

① 정의 및 개념

단순임의추출법이나 층화임의추출법을 사용하기 위해서는 모집단의 조사 단위들을 모두 포함한 추출틀을 작성해야 한다. 전국의 가구 집단과 같은 대규모 모집단을 대

상으로 하는 표본설계에서는 전국 가구의 목록과 같은 추출틀을 작성하는데 엄청난 비용이 들거나 또는 현실적으로 확보가 불가능할 수도 있다. 또한 방대한 규모의 추출틀이 작성되었다고 하더라도 그 추출틀에서 직접 가구들을 단순임의추출하면 조사대상 가구들이 전국적으로 흩어져서 산재하기 때문에 표본가구들을 실사하는 데에는 막대한 비용 및 시간이 소요된다. 이러한 문제를 극복하기 위해서 우선 전국의 가구를 지역적으로 인접한 것끼리 묶은 집단(집락)을 구성한 다음, 집락(cluster)들의 집합에서 일정한 수의 집락을 추출하고, 선정된 집락 내의 일부 또는 모든 가구를 조사하는 방법을 사용하게 되는데, 이런 추출법을 집락추출법이라고 한다.

집락추출법(cluster sampling)이란 서로 인접한 조사 단위들을 묶어서 집락 또는 조사구를 만든 다음, 집락들 중에서 일부의 집락을 추출하고 추출된 집락에 속한 조사 단위들의 일부 또는 전부를 표본으로 추출하는 확률추출법이다. 즉, 집락추출법에서는 조사 단위들의 집합인 집락을 추출단위로 사용한다. <그림 Ⅰ-3-4>에서는 $N=120$개의 조사단위가 $M=30$개의 집락(PSU)으로 구성된 모집단으로부터 집락당 $n=2$개씩의 표본을 추출하는 과정을 나타내고 있다. 즉, 집락을 조사구로 고려하면 30개의 조사구에서 각각 2가구씩을 표본으로 추출하여 총 60가구를 최종표본으로 추출하는 과정을 나타낸 것이다.



<그림 Ⅰ-3-4> 집락추출($n=60$)

집락추출법은 모집단의 모든 조사 단위들을 망라한 좋은 추출틀이 없거나, 추출틀

을 얻는데 많은 비용이 드는 반면에 집락에 대한 추출틀은 쉽게 얻을 수 있는 경우 또는 조사 단위들이 멀리 떨어져 있을수록 실사 비용이 크게 증가하는 경우에 최소의 비용으로 최대의 정보를 확보하기 위한 효과적인 표본추출법이다.

한 예로 우리나라 통계청에서 사용하는 조사구가 가장 대표적인 집락이다. 정부공식통계를 생산하기 위한 표본설계에서는 대부분 조사구를 집락으로 이용하는 집락추출법을 사용하고 있다. 그 밖에도 조사단위가 사람들인 경우에 가구는 그 자체가 집락이 된다.

② 집락추출법을 사용할 때 유의사항

집락추출법은 여러 종류의 표본조사에서 활용되고 있는 표본추출법이지만 이 추출법을 사용할 때는 반드시 유의해야 할 사항이 있다. 일반적으로 집락은 인접한 조사단위들을 묶어서 구성되기 때문에 집락 내의 조사 단위들은 여러 가지 측면에서 상당히 유사한 특성을 갖게 된다. 예를 들어, 지리적으로 인접한 가구들을 묶어서 집락을 만들면 한 집락에 속해 있는 가구들은 소득, 교육수준, 정치적 성향 등에 있어서 상당히 동질적일 가능성이 높다. 즉, 집락 내에서 얻어지는 관측값들 사이에 상당히 높은 상관관계가 존재할 수 있는데, 이런 경우에는 집락 내에서 많은 가구들을 조사하는 것은 비효율적이다. 이때에는 일차적으로 선정된 집락에서 다시 일부의 조사단위만을 추출하여 조사하는 2단계집락추출법을 사용하는 것이 효율적이다.

③ 층화추출법과 집락추출법의 차이점

층화추출법과 집락추출법은 모집단을 서로 중복되지 않는 그룹으로 분할하여 구성한다는 측면에서는 유사하다. 하지만 그룹의 구성 방법이나 사용목적 등을 보면 층화추출법의 층과 집락추출법의 집락 사이에는 큰 차이가 있다. 이들의 주요 차이점을 살펴보면, 첫째, 층화추출법에서는 가능한 한 동질적인 조사 단위들로 층을 구성하여 층 사이에는 가급적 서로 이질적인 층이 되도록 하는 것이 바람직하다. 그러나 집락추출법에서는 집락들을 구성하고 있는 조사 단위들이 이질적일수록 집락간의 변동이 줄어들게 되어서 동일한 조사비용으로 보다 효율적인 추정이 가능해진다.

둘째, 층화추출의 경우 모든 층에서 표본을 추출하여 정보를 확보하지만, 집락추출

에서는 전체 집락에서 표본으로 추출된 일부 집락에서만 다시 표본을 추출하여 정보를 얻기 때문에 추정방법 등에 있어서 본질적으로 큰 차이가 있다.

집락추출에서 표본의 크기는 집락의 크기와 표본으로 추출될 집락의 개수라는 두 가지 요소를 동시에 고려하여 결정해야 한다. 이단집락추출에서는 추출된 집락에서 뽑게 될 최종 조사단위의 수도 결정해야 한다. 이러한 결정은 집락 내 조사 단위들의 동질성과 매우 밀접한 관계를 갖고 있다. 집락 내 조사 단위들의 동질성 정도는 급내상관계수(intra-class correlation coefficient :ICC)로 표현된다.

④ 집락 설계효과

어떤 표본의 집락 효과는 부분적으로 설계효과(design effect; $deff$)에 의해 측정된다. 그러나 $deff$는 또한 층화 효과를 반영한다. 대부분의 표본설계에서는 비용을 최소로 하며, 정도를 최대로 하는 표본설계를 하고자 한다. 정도를 최대화 하고자 할 때에는 가능한 한 설계효과를 최소로 하거나 아니면, 통제해야 한다. 이 값을 어떻게 최소화 하거나 통제할 수 있는지를 보여주기 위해 $deff$의 집락부분에 대한 수학적인 정의는 다음과 같다.

$$deff \approx 1 + \rho(\tilde{n} - 1)$$

여기서 ρ는 집락내 성관계수를 나타내며, 만일 이 값이 크면 집락 안에서 단위들 간의 동질성이 높음을 의미한다. $\tilde{n}$은 집락에서 목표모집단의 단위들의 수를 나타낸다.

설계효과는 급내상관계수(ρ)와 집락의 크기($\tilde{n}$)가 증가할수록 증가한다. 설계효과에 대한 다양한 해석이 존재하지만, 한 가지는 조사에서 사용한 실제 표본설계의 표본분산이 동일한 표본크기의 단순임의 설계의 표본분산보다 얼마나 큰지를 나타내는 인자로 사용된다. 둘째로는 실제 표본설계가 정도(precision) 면에서 단순임의 추출설계보다 얼마나 떨어지는가를 나타내는 척도로 사용된다. 셋째로는 동일한 표본분산을 얻기 위해 단순임의 추출에 비해 얼마나 많은 표본이 필요한가를 나타내는 척도로 해석된다. 예를 들어 $deff = 2.0$은 단순임의 표본과 동일한 신뢰를 얻기 위해서는 2배의 표본이 요구된다고 할 수 있다. 항상 성립하는 것은 아니지만, 대체로 $deff$ 값이 2.5~3.0이면, 표본설계는 바람직하지 못하다고 해석된다.

설계효과는 통상적으로 표본조사가 완료된 후에 구할 수 있는 값이다. 표본으로부

터 분산을 추정하고, 동일한 크기의 표본으로부터 단순임의 분산을 구한 후 이들의 비를 구하여 설계효과를 구한다.

⑤ 집락 수의 결정

집락표본추출에서 최소한의 비용 하에서 집락의 수를 적절히 조정할 수 있는 방법 중의 하나가 설계효과($deff$) 가 최소가 되도록 하는 표본수를 결정하는 것이다.

다음의 표는 다양한 급내상관계수와 표본수에 대해 설계효과 값을 나타낸 것으로 이를 이용하여 적절한 집락수를 결정할 수 있을 것이다.

예를 들어 집락의 크기가 20일 경우 급내상관계수가 작지 않으면 표본설계에서 받아들이기 어려운 설계효과를 가지게 된다. <표 Ⅰ-3-5>로부터 $\tilde{n}$은 일반적으로 집락 내 목표모집단에 있는 단위의 수를 나타내는 것이지 최종 표본수를 나타내는 것은 아니다. 즉, 가구원수를 나타낼 경우 $\tilde{n}$에 집락당 평균 가구원수를 곱해야 하며, 만일 가구 수를 나타낼 경우에는 $\tilde{n}$이 최종 표본수가 될 수도 있다. 만일 사업체 조사인 경우라면 집락내 표본사업체수가 될 수 도 있다.

<표 Ⅰ-3-5> 급내상관계수와 표본의 크기에 따른 설계효과($deff$)비교

$\tilde{n}$	$ICC(\rho)$						
	0.02	0.05	0.10	0.15	0.20	0.35	0.5
5	1.08	1.2	1.4	1.6	1.8	2.4	3
10	1.18	1.45	1.9	2.35	2.8	4.15	5.5
20	1.38	1.95	2.9	3.85	4.8	6.65	10.5
30	1.58	2.45	3.9	5.35	6.8	11.15	15.5
50	1.98	3.45	5.9	8.35	10.8	18.15	25.5
75	2.48	4.7	8.4	12.1	15.8	26.19	38

☞ **사례)** 목표모집단을 전체 인구라고 가정하고, 건강조사에서 급만성 질환자를 추정하고자 한다. 또한 조사에서는 10가구 단위의 집락을 사용하고자 한다. 그러면 $\tilde{n}$은 집락 당 평균 가구 수에 10을 곱한 값이 된다. 만일 집락 당 평균 가구 수가 5라면 $\tilde{n}$은 50이 되며, 이때 설계효과를 찾으면 된다. 이 경우 $\rho = 0.02$일 때를 제외하고는 나머지에 대해서는 매우 큰 설계효과 값을 가진다.

(5) 확률비례추출법(PPS : Sampling with probability proportional to size)

모집단을 구성하고 있는 집락의 규모가 심하게 차이가 날 경우 각 집락을 추출할 확률을 동일하게 뽑지 않고, 불균등 확률로 뽑는 추출방법으로 규모가 큰 집락의 추출확률을 높게 하고, 규모가 작은 집락의 추출확률을 낮추어 주는 역할을 한다.

PPS 추출을 이용할 경우 표본설계자는 집락추출에서 최종 표본크기를 조정하기를 원한다. 집락들이 모두 동일한 크기이거나 거의 같을 때에는 PPS 표본추출이 무의미하다. 예를 들어 어떤 도시의 모든 블록(blocks)은 정확하게 100가구씩으로 구성되어 있다고 하고, 50개의 표본 블록에서 1,000개 가구를 추출하려고 한다. 가장 확실한 방법은 50개의 블록을 단순임의 추출하는 것이다. 즉, 동일한 확률(epsem)로 각 블록에서 1/5 표본 가구를 선정하면 된다. 즉, 각 블록마다 20개 가구씩 50개 블록에 대해 표본이 추출됨으로 총 1,000가구를 표본으로 선정하게 된다. 이 관계를 확률 식으로 표현하면 다음과 같다.

$$P = (50/M) \cdot (1/5) = 10/M$$

여기서 P는 하나의 가구를 표본으로 선택할 확률이며, (50/M)는 하나의 블록을 추출할 확률, (1/5)은 표본 블록 내에서 하나의 가구를 선택할 확률이다.

실제 조사상황에서는 집락의 크기는 매우 다양함으로 epsem을 이용한다는 것은 우연히 동일한 크기를 가진 집락이 표본으로 추출되지 않는 한 거의 불가능하다. 2단계 표본추출의 경우에 추출확률은 다음과 같이 정의된다.

$$P(\alpha\beta) = P(\alpha)P(\beta|\alpha)$$

여기서 $P(\alpha\beta)$는 집락 α에서 가구 β의 추출확률, $P(\alpha)$는 집락 α의 추출확률, $P(\beta|\alpha)$는 1단계에서 집락 α가 추출되고, 집락 α의 조건하에서 2단계로 가구 β가 추출될 조건부 확률을 나타낸다.

가구 수의 측면에서 전체 표본크기를 고정하기위해 모집단의 N개의 가구에서 n개의 표본가구를 등확률(epsem)로 추출하길 원한다. 이 경우 전체 추출률은 n/N이며 이는 $P(\alpha\beta)$와 같게 된다. 더욱이 만일 표본으로 추출될 집락의 수가 a개로 결정되고 추출된 집락의 크기에는 무관하게 각 집락으로부터 b개의 가구를 선택하고

자 할 때가 있다. 만일 i번째 집락의의 크기를 m_i라 하면, $P(\beta|\alpha)$는 b/m_i와 같게 된다. 따라서

$$P(\alpha\beta) = P(\alpha)(b/m_i)$$

이고, $n = ab$이므로

$$ab/N = P(\alpha)(b/m_i),\ N = \sum m_i$$

이다. 이를 $P(\alpha)$에 대해 정리하면, 다음과 같다.

$$P(\alpha) = (a)\, m_i/N$$

그러므로 집락을 추출할 확률은 해당 집락의 크기에 비례하게 된다. 1단계 추출단위를 PPS 로 추출했음에도 최종단위의 추출확률은 균등확률 추출이 된다.

$$P(\alpha\beta) = \left(\frac{a \cdot m_i}{\sum m_i}\right)\left(\frac{b}{m_i}\right) = \left(\frac{ab}{\sum m_i}\right)$$

즉, 표본추출설계가 PPS로 설계되었지만, 결과적으로 자체가중설계의 형태가 되었으며, 최종 식으로부터 각 원소들은 고정된 상수 값들이다.

(6) 추정된 집락의 크기에 비례한 확률추출법(PPES : Sampling with probability proportional to estimated size)

PPS 표본추출방법은 다소 이상적인 추출방법이다. 왜냐하면 1단계에서 집락의 추출확률을 결정하는데 사용한 크기척도(MOS : measure of size)값이 2단계에서 최종 추출단위의 크기척도가 아니기 때문이다.

특히 가구조사의 경우 일반적으로 PSU의 1단계 추출에서 적용되는 MOS는 가장 최근의 센서스로부터 얻은 모집단의 가구 수가 되는 경우가 많다. 심지어 센서스가 최근에 수행된 경우라면, 조사당시의 실제 가구 수는 차이가 나는 것이 일반적이며, 만일 전 단계에서 사용했던 MOS를 이용하여 2단계에서 가구를 동일 프레임에서 직접 추출한 경우에는 별 차이가 없다.

대부분의 가구단위 조사는 이전의 센서스 조사를 기반으로 이루어지며, 이러한 센서스 자료는 이미 상당기간 시간이 흐른 다음 사용되는 경우가 대부분이기 때문에 추출단계에서 사용되는 가구프레임은 가장 최근 프레임을 준비하는 것이 바람직하다. 즉, 프레임을 갱신하는 작업이 요구된다.

표본을 PPES에 의해 추출할 때에는 다음과 같은 확률에 의해 최종 단위가 추출된다.

$$P(\alpha\beta) = \left(\frac{a \cdot m_i}{\sum m_i}\right)\left(\frac{b}{m_i'}\right)$$

여기서 m_i'는 listing 과정에 따른 가구의 수이다.

위 식으로부터 m_i와 m_i'은 다르기 때문에 추출확률 또한 다르게 계산된다. 센서스와 MOS 간의 차이를 보상상하기 위해 정확한 가중치를 사용해야 하며, 결과적인 추정량은 비편향성을 갖도록 해야 한다.

☞ **매뉴얼**

① 적용 가능한 다양한 표본추출법들을 고려하여 가장 효율적인 추출법을 선택할 것.

② 가능한 한 단순한 추출법을 사용할 것.

다단추출법이나 집락추출법 등 다소 복잡한 추출법은 그 추출법을 사용함으로 인해 명백한 이점이 있는 경우에 한해 적용하도록 한다. 추출법이 복잡해질수록 추정식이나 관리가 까다로워진다는 사실을 명심하여야 한다.

③ 추출틀과 확보 가능한 보조정보를 고려하여 거기에 맞는 적합한 추출법을 결정할 것.

추출틀 마련이 어려운 경우 1차로 추출틀 마련이 용이한 규모의 집락을 형성하는 다단추출법이 바람직하다. 이 때 집락을 어떻게 구성하는 것이 최적인지에 대한 연구가 필요하다.

④ 표본추출방법을 결정할 때 조사의 용이성, 조사비용 등을 함께 고려할 것.

예를 들어 표본으로 뽑힌 조사 단위들이 지리적으로 너무 떨어지지 않고 인접되도록 하려면 지리적으로 인접하는 마을이나 지역을 하나의 집락으로 하는 추출법을 생각할 수 있다. 또한 조사 단위들 간의 조사비용이 서로 다를 경우 조사비용을 고려하는 표본추출이 이루어져야 한다.

⑤ 가능하다면 자체가중설계(self-weighting design)가 되도록 표본을 추출할 것.

자체가중설계란 모집단에 속하는 최종추출단위(ultimate sampling unit)들의 추출확률을 동일하게 하는 방법이다. 자체가중설계인 경우 각 조사 값들의 가중 값이 같아지므로 나중에 분석을 할 대 매우 편리하다. 일반적인 통계 소프트웨어에서 제공되는 여러 분석 방법들은 조사 값들의 추출확률이 같은 것으로 고려하기 때문에 자체가중설계가 아닌 경우 분석 과정에서 편향이 생길 여지가 많이 생기기 때문이다. 하지만 자체가중설계를 하는 것이 어려운 경우도 있으므로 반드시 해야만 하는 것은 아니다.

⑥ 자체가중설계가 이루어지지 않은 때에는 표본추출 방법에 따른 설계가중값을 반드시 명시하여 추정에 반영할 것.

⑦ 조사가 주기적으로 반복되는 계속조사일 때에는 가급적 융통성 있게 표본 추출 설계를 고려할 것.

향후 모집단이나 표본의 상황에 변동이 생길 수 있으므로 표본크기의 변화, 재층화, 추출확률의 수정 등이 가능하도록 한다. 또한 주기적 조사일 경우 응답자의 응답부담도 사전에 고려하여 이를 반영할 수 있는 추출법을 선택하는 것이 필요하다.

⑧ 계속조사의 경우 표본의 품질을 지속적으로 모니터 할 수 있는 절차를 개발할 것.

품질이 크게 저하된 층이 생길 경우 수정, 보완, 또는 재설계를 위한 전략을 수립한다. 이를 위해 가능하면 모집단의 변동을 감지할 수 있도록 하는 방안을 강구하는 것이 바람직하다.

3.7 가중

표본조사의 주된 목적은 모집단으로부터 확률 추출된 표본자료를 통해 미지의 모집단의 특성을 추측(inference)하는데 있다. 이때 표본을 추출하는 방법은 앞서 언급한데로 다양한 확률표본추출 방법을 적용할 수 있을 것이다. 이와 같이 표본으로 추출된 추출단위들은 모집단에 있는 표본으로 추출되지 않은 단위들을 잘 대표해야 한다. 즉, 표본자료는 모집단에 대한 대표성을 확보해야 하며, 이때 고려할 수 있는 것이 표본자료에 대한 가중(weighting)이다. 즉, 모집단으로부터 크기 n의 확률표본을 단순임의 추출할 경우 모집단 총합 τ_y에 대한 추정량 $\hat{\tau}_y$는 $N/n\sum y_i$로서 i번째 단위에 대한 승수가 N/n이 되며 이를 i번째 단위에 대한 가중치라 한다.

결과적으로 가중의 주된 목적은 표본자료의 모집단에 대한 대표성확보라 할 수 있다.

이러한 가중은 첫째, 분산과 비용을 감소시킬 수 있으며, 둘째 영역별로 서로 다른 추출률을 할당하여 작은 영역의 추출률을 높임으로서 전체적인 비용을 감소시킬 수 있다. 셋째, 중복, 비포괄성 등의 불완전 추출틀 문제를 해결할 수 있다. 넷째, 조사

에서 발생하는 무응답문제를 적절히 해결할 수 있다. 그런데, 이와 같은 가중 절차에서 추출단위에 항상 상수의 가중치를 고려할 수 있는데, 모집단 단위가 표본으로 뽑힐 확률이 0이 아닌 값을 갖는 경우로서 이와 같은 가중치를 "자체가중(self weighting)"이라 한다. 이러한 자체가중은 단순임의추출, 층화추출, 확률비례추출, 다단계 추출 등에서 적용할 수 있다. 자체가중방법은 첫째, 조사의 복잡성을 감소시키며, 둘째 잘못된 가중치의 문제를 해결할 수 있으며, 셋째, 동일한 표본이 다양한 목적과 다른 조사에서 사용될 수 있는 유연성과 편리성을 가지며, 넷째, 다양한 표본설계에 로버스트(robust)하며, 다섯째 일반적인 가중절차에 비해 이해하기가 쉽다는 장점을 가진다.

일반적으로 조사단위에 대해 가중의 절차는 (1) 추출가중치 또는 기본가중치 (2) 무응답가중치 (3) 사후층화가중치의 단계로 이루어진다.

☞ 체크리스트

□ 기본가중치는 계산하였는가?
□ 복합표본추출방법을 적용했다면, 각 단계별 가중치는 고려했는가?
□ 100%완전응답 표본이 아니라면, 무응답 가중치는 계산하였는가?
□ 사후층화 가중치는 계산되었는가?
□ 사후층화 가중치 산정시 고려된 방법은 무엇인가?
□ 최종적으로 구한 가중치의 변동을 검토하였는가?
□ 가중치의 효과를 검토하였는가?

☞ 가구단위 표본

□ 지역별 1차추출단위(PSU)의 추출확률은 고려하였는가?
□ 표본 1차추출단위(PSU)내의 표본가구의 추출확률은 고려하였는가?
□ 무응답 가구에 대한 가중치 조정 작업은 수행하였는가?
□ 보조정보를 이용한 사후층화 가중치 조정은 수행하였는가?
□ 최종 가구가중치의 변동을 고려하였는가?

☞ **사업체 또는 기업체 단위 표본**

☐ **사업체 또는 기업체 추출확률은 고려하였는가?**

☐ **층화 다단계추출인 경우 각 단계별 추출확률은 고려하였는가?**

☐ **무응답 가중치 조정은 수행하였는가?**

☐ **최종 표본사업체들의 가중치의 변동은 고려하였는가?**

(1) 기본가중치

기본가중치(base weight)는 표본추출 설계로부터 직접적으로 얻어지는 값이다. 임의의 모집단으로부터 적절한 크기의 표본을 추출할 때 단위가 표본에 포함될 확률의 역수로 기본가중치를 계산할 수 있다. 이때 포함확률은 기지(known)의 값이다. 다음의 표본 각 추출설계에 따른 기본가중치를 산정한 것이다.

<표 Ⅰ-3-6> 추출설계와 가중치

추출설계	가중치
단순임의 추출(비복원)	$w_i = N/n$
단순계통추출(추출간격이 r인 경우)	$w_i = r$
층화임의 추출(전체 L개의 층)	$w_{hi} = N_h/n_h$
계통 추출(불균등확률)	$w_i = M/(nM_i)$
2단계 추출 $p_i^{(1)}$: i번째 PSU의 추출확률 $p_{ij}^{(2)}$: i번째 PSU에서 j번째 SSU의 조건부추출확률	$w_{ij} = 1/(p_i^{(1)} p_{ij}^{(2)})$

(2) 무응답 조정가중치

무응답 조정가중치(nonresponse weight)의 중요한 역할은 조사로부터 발생한 무응답으로 인한 무응답 편향을 제거하기 위한 것이다. 무응답 편향은 무응답 자들이 응

답자들과 차이가 있을 때 발생하며, 모집단에 대해 매우 높은 비율로 추정치에 영향을 줄때 발생한다. 무응답 조정가중치를 계산하는 방법으로는 이용 가능한 정보의 근원에 따라 "표본에 기초한 무응답조정 방법"과 "외부정보를 이용한 무응답 조정방법"이 있다.

① 표본에 기초한 무응답 조정방법

이용가능한 정보가 표본으로 한정되며, 전체 모집단에 대한 정보는 알 수 없고, 무응답 단위들의 기본가중치를 표본응답자들에게 배분하여 응답단위들에 대해 조정된 가중치의 합이 전체표본단위들에 대한 기본가중치의 합이 된다.

$$F_c = \frac{\sum_{i=1}^{n_1} w_i MOS_i + \sum_{i=1}^{n_2} w_i MOS_i}{\sum_{i=1}^{n_1} w_i MOS_i}$$

여기서 n_1은 표본응답자들의 수이고, n_2는 무응답 자들의 수이다. w_i는 기본가중치이고 MOS_i는 표본추출설계에 의해 결정되는 적절한 크기척도를 나타낸다.

그러면 조정값 F_c를 표본응답자들의 기본가중치를 곱하여 각 단위의 가중치로 고려하면 무응답 조정 가중값 $w_i^{(a)}$를 계산할 수 있다.

$$w_i^{(a)} = \begin{cases} F_c w_i, & \text{그룹1} \\ 0, & \text{그룹2} \\ w_i, & \text{그룹3} \end{cases}$$

이때 그룹1은 응답그룹을 그룹2는 무응답그룹을 그룹3은 부적절한 표본그룹을 나타낸다.

② 외부정보를 이용한 무응답 조정방법

표본에 기초한 무응답 조정 가중치를 계산한 후 외부자료를 이용하여 사후층화, raking, 또는 calibration등의 방법으로 이 가중치를 조정한다. 외부자료를 이용한 경우 조정인자 F_c가 다음과 같이 수정된다.

$$F_c^* = \frac{MOS_c}{\sum_{i=1}^{n_1} w_i MOS_i}$$

여기서 MOS_c는 계급 c에 대해 외부데이터로부터 얻는 크기척도이다.

이때 유일한 외부정보는 표본추출틀로부터 구할 수 있으며 이를 가중계급을 구성하는데 이용한다. 가중계급을 구성하기위해 이용 가능한 방법으로는 경험적인 방법, 분류소프트웨어(CHAID), 성향점수모형, 다중가법회귀나무(MART)등이 있다.

(3) 사후층화 가중치

① 사후층화 조정

사후층화(post-stratification) 조정은 추출틀의 불완전으로 인한 포괄성의 차이, 표본의 불균형 또는 비대표성, 무응답에 의한 차이 등을 조정하기 위해 광범위하게 이용되는 방법이다. 즉, 표본응답자들의 가중치를 조정함으로서 가중된 표본분포가 기지의 모집단분포와 같아지도록 하는 방법이다. 이러한 사후층화 조정을 실시하는 주된 이유는 첫째, 추정치의 정도를 개선할 수 있으며, 둘째, 추정분산을 줄이기 위해 층화와 비추정을 사용할 수 있고, 셋째, 부차모집단간의 포괄성과 무응답에 따른 추정치의 편향을 감소시키며, 넷째, 모집단의 다양한 그룹에 대해 추정치의 일치성을 보장한다.

□ 사후층화를 위한 초기 셀 구성의 원칙

▷ 모집단 수는 부차 데이터로부터 이용 가능해야 하며 이들의 특성 값들은 관심변수와 양의 상관관계를 가져야 한다.
▷ 모집단 수는 정확히 목표모집단 수가 되어야 한다.
▷ 모집단 수는 가중된 조사추정치보다 더욱 신뢰할 수 있어야 한다.
▷ 하나의 표본응답자는 하나의 셀로 분류될 수 있어야 한다.
▷ 표본에 대한 사후층의 정의와 기지의 모집단은 서로 일치해야 한다.
▷ 사후층 셀들은 충분히 커야 한다.

② 래킹비조정

래킹비조정(raking ratio adjustment)은 전수조사 자료와 표본조사 자료간의 일치성을 확보하기 위해 1940년에 미국에서 Deming과 Stephan에 의해 처음 제안된 방법으로 2차원 분류표상의 각 셀 값을 반복적으로 조정해 가는 방법이다. 래킹비 조정 절차는 기본가중치를 하나의 주변분포를 이용하여 조정한 후, 두 번째 주변분포를 재차 이용하여 가중치를 조정한다. 이러한 과정을 특정한 수렴조건을 만족할 때까지 반복적으로 수행한다.

래킹비 조정 절차의 단점은 첫째, 반복적으로 셀 값을 조정해가는 과정에서 수렴성을 보장할 수 없다는 점이다. 즉, 래킹해야 할 차원을 잘못 선택할 경우 래킹 과정이 수렴하지 않을 수 있다. 둘째, MSE 계산이 매우 어렵다는 점이다. 왜냐하면 래킹 추정량의 구조가 복잡하기 때문에 직접적으로 MSE를 계산하기 어렵다는 문제가 있다. 셋째, 무응답과 같은 상황에 바로 적용하기란 쉽지 않다.

③ 보정

보정(calibration)은 추출설계에 따른 기본가중치 또는 추출가중치와 보정된 새로운 가중치 간의 차이를 나타내는 일종의 거리함수(distance function)를 주어진 조건을 만족하도록 최소로 하는 가중값을 구한 방법이다. 이때 제한조건을 나타내는 식을 보정방정식(calibration equation)이라 하며 다음과 같이 정의한다.

$$\sum_{s} w_i x_i = \sum_{U} x_i = \tau_x$$

여기서 x_i는 관심변수 y_i와 강한 상관이 있는 보조변수이다.

보정방정식의 조건하에서 특정한 거리함수를 최소로 하는 새로운 가중치 w_i를 구하면 된다.

즉, 선형 거리함수 $\sum G(w,d) = \sum(w_k - d_i^2)/2d_i$를 최소로 하는 새로운 w_i를 구하면 $w_i = d_i(1 + \sum d_i x_i)(\sum x_i x_i')^{-1}(\tau_x - \hat{\tau}_x)$이며 이를 총합추정량의 식에 대입하면 다음과 같은 일반화 회귀추정량(generalized regression estimator : GREG)이 된다.

$$\hat{\tau}_{yGREG} = \hat{\tau}_{yHT} + \hat{B}'(\tau_x - \hat{\tau}_{xHT})$$

이와 같은 회귀추정량의 장점은 연속형 보조정보를 사용할 수 있고, 유연성을 가지며, 분산을 정의할 수 있다는 점이다.

(4) 가중의 효과

표본의 추출률과 조사에서 발생하는 무응답, 그리고 프레임의 불완전성으로 인한 비포괄성 문제 등을 해결하기 위해서는 반드시 가중치를 고려해야 한다. 만일 가중을 하지 않으면, 추정치의 편향이 증가하게 되며 따라서 추정치의 왜곡을 피할 수 없게 된다. 그러나 가중치를 고려하게 되면 추정치의 편향은 감소하지만 가중의 효과로 인해 분산이 증가하게 된다. 이러한 이유 중의 하나는 각 단위에 부과되는 가중치의 변동이 매우 클 경우에 발생한다.

모집단 평균을 추정함에 있어 분산의 증가분에 기여한 가중치의 효과는 다음과 같은 인자에 의해 측정된다.

$$L = n \times \frac{\sum_h n_h w_h^2}{(\sum_h n_h w_h)^2}$$

여기서 $n = \sum_h n_h$로서 실제 표본크기를 나타내며, w_h는 최종가중치, n_h는 h층의 실제 표본크기이다.

이식은 다음과 같이 가중치의 변동계수(CV)의 식으로 재표현이 가능하다.

$$L = n \times \frac{\sum_j w_j^2}{(\sum_j w_j)^2} = 1 + CV^2(w_j)$$

이때 $CV^2(w_j) = \frac{n}{(\sum_j w_j)^2}\left(\sum_j w_j^2 - \frac{1}{n}(\sum_j w_j)^2\right)$이다.

(5) 가중치의 절단

일단 가중치가 계산되어 불완전성을 보상하기 위해 조정된 다음, 조정된 가중치의 분포를 파악하는 것이 바람직하다. 극히 작은 표본에 의해 극단적으로 큰 가중치는 추정치의 분산을 증가시키는 요인이 되기 때문이다. 그러므로 가중치의 변동을 고려하여 최대값 수준에서 극단가중치를 절단하는 것이 필요하다. 가중치의 절단(trimming of weights)은 대체로 무응답에 대한 조정 후에 수행하는 것이 일반적이다.

가중치의 절단 작업은 분산을 축소시키는 효과가 있지만, 다른 한편으로는 추정치의 편향을 야기하게 된다. 따라서 매우 큰 가중치에 대해 절단 작업을 수행하면, 분산의 크기는 줄일 수 있지만, 편향의 크기가 상대적으로 증가하게 되는 문제가 있다. 따라서 가중치의 절단은 절단을 함으로서 발생하는 분산의 감소분 보다 총 MSE상의 영향이 적도록 가중치를 절단하는 것이 바람직하다.

층화 추출설계에 대해, 가중치 절단과정은 각 층 내에서 수행되는 것이 이상적이다. 먼저 원 가중치들에 대한 상한(upper bound)를 정의하고, 전체 가중치들에 대해 절단된 가중치들의 합이 원가중치들의 합과 같아지도록 조정한다. w_{hi}를 h층의 i번째 단위에 대한 최종 가중치라 하고, w_{hB}를 h층에 대해 정해진 가중치들의 상한이라 하자. 그러면, h층의 i번째 표본단위에 대한 절단 가중치는 다음과 같다.

$$w_{i(T)} = \begin{cases} w_{hi}, & w_{hi} \leq w_{hB} \\ w_{nB}, & w_{hi} \geq w_{hB} \end{cases}$$

그러면 전체 표본에 대해 절단된 가중치들의 합이 원 가중치의 합과 같아지도록 조정되어야 한다. 층 내에서 상수인 가중치를 가정하고, F_T를 원 가중치들의 합과 절단된 가중치들의 합의 비라 하자.

$$F_T = \frac{\sum_h n_h w_h}{\sum_h n_h w_{h(T)}}$$

만일 h층에 대해 조정된 절단가중치를 $w^*_{h(T)}$이라 하면 이는 다음과 같다.

$$w^*_{h(T)} = F_T \times w_{h(T)}$$

이때 $w^*_{h(T)}$는 $\sum_h n_h w^*_{h(T)} = \sum_h n_h w_h$를 만족한다.

☞ 매뉴얼

① 표본설계시 자체가중설계를 지향할 것.

자체가중설계의 장점은 이해가 쉽고 다른 추출설계에 직접적으로 적용이 가능하다는 점이다. 따라서 표본추출 설계시 가능하면 자체가중설계를 고려함으로서 표본설계의 유연성을 확보할 수 있다.

② 자체가중 설계가 어렵다면 가중치 조정단계에 따라 가중값을 부여할 것.

만일 자체가중 설계가 표본설계과정에서 고려되지 않았거나 고려하기 어려운 경우라면, 기본가중치와 무응답가중치 그리고 사후가중치 조정단계에 따라 가중치를 계산하여 추정치 산정에 반영해야 할 것이다. 또한 가중치 조정단계에서 이용 가능한 정보를 적극 활용하는 것도 고려해야 한다.

③ 가중치 조정과정에서 이용 가능한 보조정보를 적극 활용할 것.

이용 가능한 정보는 표본과 표본 외부에 모두 존재할 수 있으며, 만일 외부정보를 활용할 수 있다면 이를 적극적으로 활용하는 것이 바람직하다. 사후층화 조정시에 이러한 외부정보를 사용함으로서 표본과 모집단간의 일치성을 보장할 수 있을 것이다.

④ 가중치를 적용했다면 가중치 효과를 반드시 계산할 것.

통상적으로 가중치를 사용하지 않으면, 추정량의 편향이 발생하게 되는 반면에 가중치를 사용하면 추정량의 편향은 줄일 수 있으나, 분산이 증가하는 문제가 있다. 따라서 무조건 가중치를 적용할 것이 아니라 먼저 가중효과를 분산팽창인자를 이용하여 산정해봄으로서 적절한 가중치를 적용해야 할 것이다.

⑤ 가중치의 변동을 고려하여 극단 가중치는 추정치에 큰 영향을 줌으로 개관적인 방법으로 절단하여 사용할 것.

극단 가중치는 분산을 크게 함으로서 추정치의 정도를 떨어뜨리는 역할을 한다. 따라서 적절한 기준 하에서 극단가중치를 조정할 필요가 있다. 이는 총 MSE 차원에서 다루는 것이 합당한데, 왜냐하면 가중치를 조정하여 분산감소의 효과를 볼 수 있는 반면 추정치의 편향이 발생하기 때문이다.

3.8 추정량과 추정식

일반적으로 추정단계에서 가중치를 이용하면 모집단에 대한 특성치인 모수에 대한 비편향추정량(unbiased estimator)을 얻을 수 있다. 만약 통계분석 과정에서 가중치를 무시하고 분석한 추정치는 심각한 편향(bias)이 발생할 수 있다. 표본의 크기가 큰 대규모 조사에서 문제가 되는 것은 추정량의 편향이기 때문에 추정과정에서 반드시 가중치를 이용해야 한다. 일반적으로 복합표본조사(complex sample survey)의 가중치는 설계 가중치, 무응답에 대한 조정, 사후층화에 대한 조정 등의 세 가지 요인을 통합하여 산정된다.

복합표본조사 데이터를 분석할 때 가중치를 무시하고 분석하면 모수 추정에 심각한 편향(bias)이 발생할 수 있고, 추정량의 분산이 과소평가되어 문제가 된다. 따라서 기업체나 사업체 조사에서 모집단의 특성치에 대한 추정은 가중치를 이용한 통계치를 이용해야 하고, 만약 단순표본평균을 사용하면 추정치에 편향이 발생할 수 있다. 모집단의 특성치에 대한 추정은 가중치를 이용한 가중표본평균을 사용한다.

☞ 체크리스트

□ 적용한 추정식은 표본설계를 적절히 반영하고 있는가?
□ 복합표본추출방법을 적용했다면, 각 단계별 가중치를 고려한 추정산식인가?

□ 사용한 추정량은 비편향성을 지니고 있는가?
□ 추정치는 평균인가? 총합인가? 아니면 비율인가?
□ 주요관심변수들의 추정치는 과거추정치와 시계열성을 유지하고 있는가?
□ 추정치에 대한 표준오차는 계산하였는가?
□ 비선형추정량인 경우 다양한 근사적인 추정방법을 고려하였는가?
□ 분산 추정산식은 적절한가?
□ 추정치의 계산을 위해 사용한 프로그램은 무엇인가?
□ 분산 추정을 위해 사용한 프로그램은 무엇인가?
□ 반복적인 분산 추정방법을 사용하였는가?

☞ 가구단위 표본

□ 추정치는 가구당 평균 또는 총합인가?
□ 가구의 추정치에 대한 표준오차는 계산되었는가?
□ 표준오차의 계산식은 제시되었는가?

☞ 사업체 또는 기업체 단위 표본

□ 산업별 추정치인가? 아니면 업체별 추정치인가?
□ 각 추정치에 대한 표준오차는 제시하였는가?
□ 표준오차의 계산식은 제시되었는가?

(1) 전통적인 평균, 총계, 비율의 추정과 분산추정량

① 단순임의추출 하에서의 추정량

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

$\bar{y}$의 분산 : $V(\bar{y}) = \frac{N-n}{N}\frac{S^2}{n}$, $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i-\mu)^2$

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

$\bar{y}$의 표준오차 : $\widehat{SE}(\bar{y}) = \sqrt{\hat{V}(\bar{y})}$

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_y = N\bar{y}$

$\hat{\tau}_y$의 분산 : $V(\hat{\tau}_y) = N^2\frac{N-n}{N}\frac{S^2}{n}$

$\hat{\tau}_y$의 추정분산 : $\hat{V}(\hat{\tau}) = N^2\frac{N-n}{N}\frac{s^2}{n}$

$\hat{\tau}_y$의 표준오차 : $\widehat{SE}(\hat{\tau}_y) = \sqrt{\hat{V}(\hat{\tau}_y)}$

□ 비율 추정

모비율 p의 추정량 : $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} y_i$, 이때, y_i는 0 또는 1의 값을 갖는다.

$\hat{p}$의 분산 : $V(\hat{p}) = \frac{N-n}{N-1}\frac{pq}{n}$

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

$\hat{p}$의 표준오차 : $\widehat{SE}(\hat{p}) = \sqrt{\hat{V}(\hat{p})}$

② 층화추출 하에서의 추정량

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \bar{y}_h$, $\bar{y}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$

$\bar{y}_{st}$의 분산 : $V(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h}$,

$$S_h^2 = \frac{1}{N_h - 1}\sum_{i=1}^{N_h}(y_{hi} - \mu_h)^2$$

$\bar{y}_{st}$의 추정분산 : $\hat{V}(\bar{y}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h}$,

$$s_h^2 = \frac{1}{n_h - 1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2$$

$\bar{y}_{st}$의 표준오차 : $\widehat{SE}(\bar{y}_{st}) = \sqrt{\hat{V}(\bar{y}_{st})}$

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{st} = N\bar{y}_{st}$

$\hat{\tau}_{st}$의 분산 : $V(\hat{\tau}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h}\frac{S_h^2}{n_h}$

$\hat{\tau}_{st}$의 추정분산 : $\hat{V}(\hat{\tau}_{st}) = \sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h}\frac{s_h^2}{n_h}$

$\hat{\tau}_{st}$의 표준오차 : $\widehat{SE}(\hat{\tau}_{st}) = \sqrt{\hat{V}(\hat{\tau}_{st})}$

□ 비율 추정

모비율 p의 추정량 : $\hat{p}_{st} = \frac{1}{N}\sum_{h=1}^{L} N_h \hat{p}_h$, $\hat{p}_h = \frac{1}{n_h}\sum_{i=1}^{n_h} y_{hi}$

$\hat{p}_{st}$의 분산 : $V(\hat{p}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h - 1}\frac{p_h q_h}{n_h}$

$\hat{p}_{st}$의 추정분산 : $\hat{V}(\hat{p}_{st}) = \frac{1}{N^2}\sum_{h=1}^{L} N_h^2 \frac{N_h - n_h}{N_h}\frac{\hat{p}_h \hat{q}_h}{n_h - 1}$

$\hat{p}_{st}$의 표준오차 : $\widehat{SE}(\hat{p}_{st}) = \sqrt{\hat{V}(\hat{p}_{st})}$

③ 계통추출 하에서의 추정량

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y}_{sy} = \bar{y}_i = \frac{1}{n}\sum_{j=1}^{n} y_{ij}, \quad i = 1, 2, \cdots, k$

$\bar{y}_{sy}$의 분산 : $V(\bar{y}_{sy}) = \frac{1}{k}\sum_{i=1}^{k}(\bar{y}_i - \mu_y)^2$

$\bar{y}_{sy}$의 추정분산 : $\hat{V}(\bar{y}_{sy}) = \frac{N-n}{N}\frac{s^2}{n}$, $s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_{ij} - \bar{y}_{sy})^2$

$\bar{y}_{sy}$의 표준오차 : $\widehat{SE}(\bar{y}_{sy}) = \sqrt{\hat{V}(\bar{y}_{sy})}$

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{sy} = N\bar{y}_{sy}$

$\hat{\tau}_{sy}$의 분산 : $V(\hat{\tau}_{sy}) = \frac{N^2}{k}\sum_{i=1}^{k}(\bar{y}_i - \mu_y)^2$

$\hat{\tau}_{sy}$의 추정분산 : $\hat{V}(\hat{\tau}_{sy}) = N^2\hat{V}(\bar{y}_{sy}) = N^2\frac{N-n}{N}\frac{s^2}{n}$

$\hat{\tau}_{sy}$의 표준오차 : $\widehat{SE}(\hat{\tau}_{sy}) = \sqrt{\hat{V}(\hat{\tau}_{sy})}$

□ 비율 추정

모비율 p의 추정량 : $\hat{p}_{sy} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$, y_{ij}는 0 또는 1의 값을 가진다.

$\hat{p}_{sy}$의 분산 : $V(\hat{p}_{sy}) = \frac{N-n}{N-1}\frac{pq}{n}$

$\hat{p}_{sy}$의 추정분산 : $\hat{V}(\hat{p}_{sy}) = \frac{N-n}{N}\frac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}$

$\hat{p}_{sy}$의 표준오차 : $\widehat{SE}(\hat{p}_{sy}) = \sqrt{\hat{V}(\hat{p}_{sy})}$

④ 집락추출 하에서의 추정량

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y}_{cl} = \dfrac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} N_i}$

$\bar{y}_{cl}$의 분산 : $V(\bar{y}_{cl}) = \dfrac{M-m}{Mm\bar{N}^2} S_c^2$, $S_c^2 = \dfrac{\sum_{i=1}^{M} (y_i - \mu N_i)^2}{M-1}$

$\bar{y}_{cl}$의 추정분산 : $\hat{V}(\bar{y}_{cl}) = \dfrac{M-m}{Mm\bar{N}^2} s_c^2$, $s_c^2 = \dfrac{\sum_{i=1}^{m} (y_i - \bar{y}_{cl} N_i)^2}{m-1}$,

여기서 M : 모집단에 있는 집락 수, n : 단순임의추출로 뽑은 집락 수,

N_i : 집락 i에 있는 조사단위의 수, $i = 1, 2, \cdots, M$

$\bar{n} = \dfrac{1}{m} \sum_{i=1}^{m} N_i$: 표본에 대한 평균 집락 크기

$N = \sum_{i=1}^{M} N_i$: 모집단에 있는 조사단위의 수

$\bar{N} = N/M$: 모집단에 대한 평균 집락 크기

y_i 또는 a_i : i번째 집락에 있는 조사단위의 합계

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{cl} = N \dfrac{\sum_{i=1}^{m} y_i}{\sum_{i=1}^{m} N_i}$

$\hat{\tau}_d$의 분산 : $V(\hat{\tau}_{cl}) = M^2 \dfrac{M-m}{Mm} S_c^2$

$\hat{\tau}_d$의 추정분산 : $\hat{V}(\hat{\tau}_{cl}) = M^2 \dfrac{M-m}{Mm} s_c^2$

$\hat{\tau}_d$의 표준오차 : $\widehat{SE}(\hat{\tau}_{cl}) = \sqrt{\hat{V}(\hat{\tau}_{cl})}$

□ 비율 추정

모비율 p 의 추정량 : $\hat{p}_{cl} = \frac{\sum_{i=1}^{m} a_i}{\sum_{i=1}^{m} N_i}$

$\hat{p}_{cl}$의 분산 : $V(\hat{p}_{cl}) = \frac{M-m}{Mm\overline{N}^2} S_p^2, \quad S_p^2 = \frac{\sum_{i=1}^{M} (a_i - pN_i)^2}{M-1}$

$\hat{p}_{cl}$의 추정분산 : $\hat{V}(\hat{p}_{cl}) = \frac{M-m}{Mm\overline{N}^2} s_p^2, \quad s_p^2 = \frac{\sum_{i=1}^{m} (a_i - \hat{p}_{cl} N_i)^2}{m-1}$

$\hat{p}_{cl}$의 표준오차 : $\widehat{SE}(\hat{p}_{cl}) = \sqrt{\hat{V}(\hat{p}_{cl})}$

⑤ 확률비례집락추출 하에서의 추정량

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{pps} = \frac{1}{m} \sum_{i=1}^{m} \frac{y_i}{p_i}$,

p_i는 i번째 집락이 표본으로 추출될 확률.

$\hat{\tau}_{pps}$의 분산 : $V(\hat{\tau}_{pps}) = \frac{1}{m} \sum_{i=1}^{M} p_i \left(\frac{y_i}{p_i} - \tau \right)^2 = \frac{1}{m} \left(\sum_{i=1}^{M} \frac{y_i^2}{p_i} - \tau^2 \right)$

$\hat{\tau}_{pps}$의 분산추정 : $\hat{V}(\hat{\tau}_{pps}) = \frac{1}{m(m-1)} \sum_{i=1}^{m} \left(\frac{y_i}{p_i} - \hat{\tau}_{pps} \right)^2$

$\hat{\tau}_{pps}$의 표준오차 : $\widehat{SE}(\hat{\tau}_{pps}) = \sqrt{\hat{V}(\hat{\tau}_{pps})}$

□ **총계 추정**

모총계 τ_y의 추정량 : $\hat{\tau} = \frac{M}{m}\sum_{i=1}^{m} N_i \bar{y}_i = \frac{M}{m}\sum_{i=1}^{m} y_i$

$\hat{\tau}$의 분산 : $V(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\frac{S_b^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{S_{wi}^2}{n_i}$

$S_b^2 = \frac{1}{M-1}\sum_{i=1}^{M}(\tau_i-\mu_1)^2$, $S_{wi}^2 = \frac{1}{N_i-1}\sum_{i=1}^{N_1}(y_{ij}-\mu_i)^2$, $\mu_1 = \tau/M$

$\hat{\tau}$의 분산추정 : $\hat{V}(\hat{\tau}) = M^2\left(1-\frac{m}{M}\right)\frac{s_b^2}{m} + \frac{M}{m}\sum_{i=1}^{m} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_{wi}^2}{n_i}$

$s_b^2 = \frac{1}{m-1}\sum_{i=1}^{m}(\hat{\tau}_i - \hat{\mu}_1)^2$, $s_{wi}^2 = \frac{1}{n_i-1}\sum_{i=1}^{N_1}(y_{ij}-\bar{y}_i)^2$,

$\hat{\mu}_1 = \hat{\tau}/M$

$\hat{\tau}$의 표준오차 : $\widehat{SE}(\hat{\tau}) = \sqrt{\hat{V}(\hat{\tau})}$

⑥ 2단계 집락추출 하에서의 추정량

□ **평균 추정** (N을 알 때)

모평균 μ_y의 추정량 : $\hat{\mu} = \frac{\hat{\tau}}{N} = \frac{1}{N}\left(\frac{M}{m}\sum_{i=1}^{m} N_i \bar{y}_i\right)$

$\hat{\mu}$의 분산 : $V(\hat{\mu}) = \frac{1}{N^2}M^2\left(1-\frac{m}{M}\right)\frac{S_b^2}{m} + \frac{M}{m}\sum_{i=1}^{M} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{S_{wi}^2}{n_i}$

$\hat{\mu}$의 분산추정 :

$$\hat{V}(\hat{\mu}) = \frac{1}{N^2}\left[M^2\left(1-\frac{m}{M}\right)\frac{s_b^2}{m} + \frac{M}{m}\sum_{i=1}^{m} N_i^2\left(1-\frac{n_i}{N_i}\right)\frac{s_{wi}^2}{n_i}\right]$$

$\hat{\mu}$의 표준오차 : $\widehat{SE}(\hat{\mu}) = \sqrt{\hat{V}(\hat{\mu})}$

□ 평균 추정 (N을 모를 때)

모평균 μ_y의 추정량 :

$$\hat{\mu}_r = \frac{\hat{\tau}}{\frac{M}{m}\sum_{i=1}^{m} N_i} = \frac{\frac{M}{m}\sum_{i=1}^{m} N_i \bar{y}_i}{\frac{M}{m}\sum_{i=1}^{m} N_i} = \frac{\sum_{i=1}^{m} N_i \bar{y}_i}{\sum_{i=1}^{m} N_i} \; (= \bar{y}_{2clu})$$

$\hat{\mu}_r$의 분산 :

$$V(\hat{\mu}_r) = \left(1 - \frac{m}{M}\right)\frac{1}{\bar{N}^2}\frac{S^2}{m} + \frac{1}{mM\bar{N}^2}\sum_{i=1}^{M} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{S_{wi}^2}{n_i}$$

$$\bar{N} = \sum_{i=1}^{M}\frac{N_i}{M}, \; S^2 = \frac{1}{M-1}\sum_{i=1}^{M} N_i^2(\mu_i - \mu)^2,$$

$$S_{wi}^2 = \frac{1}{N_i - 1}\sum_{j=1}^{N_i}(y_{ij} - \mu_i)^2$$

$\hat{\mu}_r$의 분산추정 :

$$\hat{V}(\hat{\mu}_r) = \left(1 - \frac{m}{M}\right)\frac{1}{\bar{n}^2}\frac{s^2}{m} + \frac{1}{mM\bar{n}^2}\sum_{i=1}^{m} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{s_{wi}^2}{n_i}$$

$$\bar{n} = \frac{1}{m}\sum_{i=1}^{m} N_i, \; s^2 = \frac{1}{m-1}\sum_{i=1}^{m} N_i^2(\bar{y}_i - \hat{\mu})^2,$$

$$s_{wi}^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2$$

$\hat{\mu}_r$의 표준오차 : $\widehat{SE}(\hat{\mu}_r) = \sqrt{\hat{V}(\hat{\mu}_r)}$

□ 비율 추정

모비율 p의 추정량 : $\hat{p} = \frac{\sum_{i=1}^{m} N_i \hat{p}_i}{\sum_{i=1}^{m} N_i}$, $\hat{p}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$, y_{ij}는 1 또는 0

$\hat{p}$의 분산추정 :

$$\hat{V}(\hat{p}) = \left(1 - \frac{m}{M}\right)\frac{1}{\bar{n}^2}\frac{s^2}{m} + \frac{1}{mM\bar{n}^2}\sum_{i=1}^{m} N_i^2\left(1 - \frac{n_i}{N_i}\right)\frac{\hat{p}_i\hat{q}_i}{n_i - 1},$$

$$s^2 = \frac{1}{m-1}\sum_{i=1}^{m} N_i^2(\hat{p}_i - \hat{p})^2$$

$\hat{p}$의 표준오차 : $\widehat{SE}(\hat{p}) = \sqrt{\hat{V}(\hat{p})}$

⑦ 2단계 층화 집락 추출 하에서의 추정량

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y}_{st} = \frac{\sum N_h M_h \bar{y}_h}{\sum N_h M_h} = \sum W_h \bar{y}_h$, $W_h = \frac{N_h M_h}{\sum N_h M_h}$,

$$\bar{y}_h = \frac{1}{n_h}\sum_{i}^{n_h} y_{hi} = \frac{1}{n_h m_h}\sum_{i}^{n_h}\sum_{j}^{m_h} y_{ij}$$

$\bar{y}_{st}$의 분산 : $V(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_{1h}}{n_h} S_{1h}^2 + \frac{1-f_{2h}}{n_h m_h} S_{2h}^2 \right)$,

$$f_{1h} = \frac{n_h}{N_h},\ f_{2h} = \frac{m_h}{M_h}$$

$\bar{y}_{st}$의 추정분산 : $\hat{V}(\bar{y}_{st}) = \sum_{h=1}^{L} W_h^2 \left(\frac{1-f_{1h}}{n_h} s_{1h}^2 + \frac{1-f_{2h}}{n_h m_h} s_{2h}^2 \right)$

$\bar{y}_{st}$의 표준오차 : $\widehat{SE}(\bar{y}_{st}) = \sqrt{\hat{V}(\bar{y}_{st})}$

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_y = \sum N_h M_h \bar{y}_h$

$\hat{\tau}_y$의 분산 : $V(\hat{\tau}_y) = \sum_{h=1}^{L} (N_h M_h)^2 \left(\frac{1-f_{1h}}{n_h} S_{1h}^2 + \frac{1-f_{2h}}{n_h m_h} S_{2h}^2 \right)$

$\hat{\tau}_y$의 추정분산 : $\hat{V}(\hat{\tau}_y) = \sum_{h=1}^{L} (N_h M_h)^2 \left(\frac{1-f_{1h}}{n_h} s_{1h}^2 + \frac{1-f_{2h}}{n_h m_h} s_{2h}^2 \right)$

$\hat{\tau}_y$의 표준오차 : $\widehat{SE}(\hat{\tau}_y) = \sqrt{\hat{V}(\hat{\tau}_y)}$

⑧ HH(Hansen-Hurwitz) 추정량

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{HH} = \frac{1}{n}\sum_{i}^{n}\frac{y_i}{p_i}$

$p_i = M_i / M_0$: i번째 단위의 추출확률, $M_0 = \sum M_i$

$\hat{\tau}_{HH}$의 분산 : $V(\hat{\tau}_{HH}) = \frac{1}{n}\sum_{i}^{N}\left(\frac{y_i}{p_i} - \tau\right)^2 p_i$

$\hat{\tau}_{HH}$의 분산추정 : $\widehat{V}(\hat{\tau}_{HH}) = \frac{1}{n(n-1)}\sum_{i}^{n}\left(\frac{y_i}{p_i} - \hat{\tau}_{HH}\right)^2$

$\hat{\tau}_{HH}$의 표준오차 : $\widehat{SE}(\hat{\tau}_{HH}) = \sqrt{\widehat{V}(\hat{\tau}_{HH})}$

□ 평균 추정

모평균 μ_y의 추정량 : $\bar{y}_{HH} = \frac{1}{nN}\sum_{i=1}^{n}\frac{y_i}{p_i}$

$\bar{y}_{HH}$의 분산 : $V(\bar{y}_{HH}) = \frac{1}{n}\sum_{i}^{N}\left(\frac{y_i}{Np_i} - \mu_y\right)^2 = \frac{1}{n}\left(\frac{1}{N^2}\sum_{i=1}^{N}\frac{y_i^2}{p_i} - \mu_y^2\right)$

$\bar{y}_{HH}$의 분산추정 : $\widehat{V}(\bar{y}_{HH}) = \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(\frac{y_i}{Np_i} - \bar{y}_{HH}\right)^2$

$\bar{y}_{HH}$의 표준오차 : $\widehat{SE}(\bar{y}_{HH}) = \sqrt{\widehat{V}(\bar{y}_{HH})}$

⑨ HT(Horvitz-Thompson) 추정량

□ 총계 추정

모총계 τ_y의 추정량 : $\hat{\tau}_{HT} = \sum_{i=1}^{n} \frac{y_i}{\pi_i}$

π_i : i단위표본에 포함될 확률

$\hat{\tau}_{HT}$의 분산 : $V(\hat{\tau}_{HT}) = \sum_{i}^{N} \frac{(1-\pi_i)}{\pi_i} y_i^2 + 2\sum_{i}^{N}\sum_{j>i}^{N} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j} y_i y_j$

π_{ij} : i단위와 j단위가 동시에 표본에 포함될 확률

$\hat{\tau}_{HT}$의 분산추정 : $\hat{V}_{HT}(\hat{\tau}_{HT}) = \sum_{i}^{n} \frac{(1-\pi_i)}{\pi_i^2} y_i^2 + 2\sum_{i}^{n}\sum_{j>i}^{n} \frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_i\pi_j\pi_{ij}} y_i y_j$

$\hat{\tau}_{HT}$의 표준오차 : $\widehat{SE}(\hat{\tau}_{HT}) = \sqrt{\hat{V}(\hat{\tau}_{HT})}$

□ 평균 추정

모평균 μ의 추정량 : $\bar{y}_{HT} = \frac{1}{N}\sum_{i=1}^{d} \frac{y_i}{\pi_i}$

$\bar{y}_{HT}$의 분산 : $V(\bar{y}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{i<j}^{N} (\pi_i\pi_j - \pi_{ij})\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$

$\bar{y}_{HT}$의 분산추정 : $\hat{V}_{YG}(\bar{y}_{HT}) = \frac{1}{N^2}\sum_{i=1}^{d}\sum_{i<j}^{d} \frac{(\pi_i\pi_j - \pi_{ij})}{\pi_{ij}}\left(\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right)^2$

$\bar{y}_{HT}$의 표준오차 : $\widehat{SE}(\bar{y}_{HT}) = \sqrt{\hat{V}(\bar{y}_{HT})}$

(2) 가중치를 고려한 평균, 총합 추정량

① 평균의 추정

표본설계를 층화2단계추출에 근거했다고 가정할 때 관심변수인 모평균 μ의 추정량 및 분산추정량은 다음과 같다.

$$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}}{w_{\cdots}}$$

$$\hat{V}(\bar{y}) = \sum_{h=1}^{L}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(e_{hi.}-\bar{e}_{h..})^2$$

여기서 $w_{\cdots} = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}$, $e_{hi.} = \left(\sum_{j=1}^{m_{hi}} w_{hij}(y_{hij}-\bar{y})\right)/w_{\cdots}$,

$\bar{e}_{h..} = \left(\sum_{i=1}^{n_h} e_{hi.}\right)/n_h$ 이고, w_{hij}는 변수값 y_{hij}에 부여된 가중치이다.

② 총합의 추정

표본설계를 층화2단계추출에 근거했다고 가정할 때 관심변수인 모집단 총합 τ의 추정량 및 분산추정량은 다음과 같다.

$$\hat{\tau}_y = \sum_{h=1}^{L}\sum_{i=1}^{n_h}\sum_{j=1}^{m_{hi}} w_{hij}y_{hij}$$

$$\hat{V}(\hat{\tau}_y) = \sum_{h=1}^{L}\frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_h}(y_{hi.}-\bar{y}_{h..})^2$$

여기서 $y_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij}y_{hij}$, $\bar{y}_{h..} = \left(\sum_{i=1}^{n_h} y_{hi.}\right)/n_h$이다.

☞ **참고 1 :**

위에서 다룬 분산추정식이 성립하기 위해서는 다음과 같은 조건이 필요하다.

① 서로 다른 층으로부터 뽑힌 표본들은 서로 독립이다.
② 각 층 내에서 PSU는 복원추출방법에 의해 뽑는다.
③ PSU의 개수는 2개 이상이다.

(3) 분산추정량

표본조사자료로부터 추정치의 분산이나 표본오차를 추정하기위한 방법에는 여러 가지가 있다. 예를 들면 가장 단순하게 전통적인 표본추출방식에 따른 추정량의 분산추정치를 계산할 수 있을 것이다. 이러한 측면에서 다음과 같이 4가지로 분산 추정방법을 분류할 수 있다.

□ 정확한 방법(Exact Method)
□ 최종 집락 방법(Ultimate Cluster Method)
□ 선형화 방법(Linearization Method)
□ 반복적인 방법(Replication Methods)

① 정확한 방법

이 방법이 앞에서 언급한 추정량과 분산추정식을 사용하여 분산 추정을 위한 가장 최선의 방법이다. 그러나 표본의 추출과정에서 복잡한 형태의 인자들(무응답, 포괄오차)이 개입됨으로서 실제적으로 적용하는 데는 한계가 있다. 첫째, 대부분의 표본추출설계는 단순임의 추출이 아닌 매우 복잡한 형태로 이루어져 있다. 둘째, 관심변수의 추정량은 단순하게 관측 자료들의 선형결합의 형태로 나타나지 않는 경우가 있으며, 따라서 분산추정식 또한 단순임의 추출이나 층화 추출에서의 추정식과 같이 완전한 수식의 형태로 나타나지 않는 경우가 있다. 더욱이 이 추정방법은 표본추출설계에 의존하기 때문에 대부분 가중치를 고려한 방법으로 관심변수의 추정식을 사용하고 있다.

② 최종 집락 방법

분산 추정에 있어서 집락 방법은 복합표본설계에 의한 표본에 근거한 분산 추정방법이다. 이 방법에 의해 집락은 다단계 설계에 의해 부차적인 표본추출이 이루어졌음에도 불구하고, 1개의 PSU가 전체 표본으로 구성된다. 분산 추장치는 각 단계별 추출에서 분산성분을 계산하지 않고 PSU 총합들 간의 변동만을 계산한다.

즉, 모집단의 h층으로부터 크기가 n_h인 PSU를 표본으로 추출한다면, h층의 모총합에 대한 추정식은

$$\hat{Y}_h = \sum_{i=1}^{n_h} \hat{Y}_{hi}$$

이며, 이때 $\hat{Y}_{hi} = \sum_{j=1}^{m_i} W_{hijk} Y_{hjik}$이다.

PSU 차원에서의 추정량 $\hat{Y}_{hi}$는 $\hat{Y}_h / n_h$의 추정치이다. 그러므로 개개의 PSU 차원의 추정치의 분산은 다음과 같다.

$$\hat{V}(\hat{Y}_{hi}) = \frac{1}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2$$

또한 h층에 대해 모총합의 크기가 n_h인 단순임의 표본에 의해 추정량인 $\hat{Y}_{hi}$의 총합인 $\hat{Y}_h$의 분산은 다음과 같다.

$$\hat{V}(\hat{Y}_h) = \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} \left(\hat{Y}_{hi} - \frac{\hat{Y}_h}{n_h} \right)^2$$

최종적으로 각 층별로 독립적으로 표본추출이 되므로 전체 모집단 총합의 분산 추정치는 층화 차원에서 다음과 같이 계산된다.

$$\hat{V}(\hat{Y}) = \sum_{h=1}^{L} \hat{V}(\hat{Y}_h)$$

③ 선형화 방법

대부분의 복합표본설계로 부터의 추정량의 형태는 비선형 추정량의 형태를 가진다. 선형화 방법에서는 비선형추정량을 테일러 전개를 이용하여 근사적인 선형화 형태로 만들어 추정량으로 사용한다. 즉, 비선형 함수를 테일러 전개하여 선형화 한 후 선형화된 함수의 1차 항만을 고려하여 이를 정확한 방법의 분산추정식에 대입하여 최종 분산 추정치를 구하게 된다.

선형화 과정을 간단히 설명하면, 우선 모수 Θ의 추정치인 $\hat{\Theta}$의 분산을 구하기

위해 $\hat{\Theta}$를 $Y_1, Y_{2,}\cdots, Y_m$의 관찰치인 $y_1, y_2, \cdots, y_m$의 비선형 함수라 하자.

$$\hat{\Theta} = f(y_1, y_2, \cdots, y_m)$$

그러면 $\hat{\Theta}$가 Θ에 근사하다고 가정하고, $\hat{\Theta}$을 테일러 전개한 후 $\hat{\Theta} - \Theta$의 1차 항을 고려하면

$$\hat{\Theta} = \Theta + \sum_{i=1}^{m} d_i(y_i - Y_i)$$

이며, 이때 $d_i = \partial\ \hat{\Theta}_i / \partial y_i$이다.

따라서 $\hat{\Theta}$의 분산은 위의 선형함수의 분산으로 나타낼 수 있기 때문에, 실제적인 분산 추정치를 계산할 수 있다.

$$\hat{V}(\ \hat{\Theta}) = \hat{V}(\sum_i d_i y_i) = \sum_{i=1}^{m} d_i^2\, \hat{V}(\ y_i) + \sum_{i \neq j} d_i d_j cov(y_i, y_j)$$

이러한 선형화 방법은 실제적인 분산 추정에 많이 사용되는 방법으로서 어떤 통계량에도 선형화가 가능하다면, 우리가 알고 있는 전형적인 분산추정량으로 사용이 가능하다.

□ 선형화 방법의 장점

분산 선형화 방법은 이미 오래전부터 사용해 오던 방법으로 이론적으로 매우 발전된 방법이며, 반복적인 방법(replication method) 보다 다양한 표본설계에 적용이 가능하다. 만일 선형화 과정에서 편미분 값만 안다면, 선형화로부터 얻은 분산 추정치는 비(ratio)나 회귀계수와 같이 비선형 추정량에 대해 모든 선형추정량의 분산으로 적용이 가능하다.

□ 선형화 방법의 한계

선형화 방법의 장점 중에서 편미분 값이 알려지고 높은 차원(2차원 이상)의 값이 맞는다면, 매우 좋은 추정방법이다. 그러나 이러한 점들이 만족되지 않는다면, 매우 심각한 편향을 가지게 된다. 또한 가중치를 포함하고 있는 복합함수에 이를 적용하기란 일반적으로 어렵다. 이러한 문제를 해결하기 위해서는 특별한 프로그램을 사용

해야 한다는 측면도 한계점 중의 하나이다.

또한 무응답이나 비포괄성의 상황에 선형화를 적용하기란 매우 어려운 측면도 존재한다. 이는 표본설계, 관심추정치, 가중치 과정 등에 좌우되는 방법이다.

④ 반복적인 방법

반복적인 방법의 기본은 표본데이터로부터 반복적으로 부차표본(subsample)를 반복적으로 추출하여 각 반복별로 가중 추정치를 새롭게 계산한 후 전체 표본을 이용한 추정치와 반복추정치간의 변동을 분산으로 계산하는 방법이다.

반복적인 방법에는 다음과 같은 방법들이 있다.

○ 임의 그룹 방법(random group)
○ 균형 반복방법(Balanced Repeated Replication: BRR)
○ 잭나이프 방법(Jackknife Replication : JK1 , JK2, JKn)
○ 붓스트랩 (Bootstrap)

반복과정을 단계별로 정리하면 다음과 같다.

- **1단계** : 반복 표본을 만들기 위해 부차표본과 전체 표본간의 차이를 제거한다.
- **2단계** : 각각의 반복표본에 대해 추정과정을 반복하여 반복가중치를 생성한다.
- **3단계** : 전체표본으로부터 추정치를 구하고 반복가중치의 집합으로부터 추정치를 구한다.
- **4단계** : 반복 추정치와 전체 표본으로부터 구한 추정치간의 차이의 제곱으로 추정치의 분산을 계산한다.

이를 식으로 표현하면, k개의 반복으로 부터 얻어진 추정량을 각각 $\hat{\theta}_{1,}\ \hat{\theta}_{2,}\cdots,\ \hat{\theta}_{k}$ 라하고, 전체 표본으로부터 구한 추정량을 $\hat{\theta}_{0}$라 하면 반복분산 추정량은 다음과 같다.

$$\widehat{V}(\hat{\theta}) = \frac{1}{c}\sum_{r=1}^{k}(\hat{\theta}_r - \hat{\theta}_0)^2$$

이다.

상수 c는 반복 방법에 따라 다르게 표현되는데 이는 다음과 같다.

<표 Ⅰ-3-7 > 반복적인 방법의 분산항의 c값

방법	c값
임의그룹	$k(k-1)$
BRR	k
JK1	1
JK2	2
JKn	$k/(k-1)$
Bootstrap	$k-1$

☞ 사례) ICT 투자규모 표본설계(한국전자거래진흥원, 2006)

2006년 한국전자거래진흥원에서는 ICT 투자규모 예측을 위한 층화 표본설계를 하였다.

① 전체추정량

관심변수에 대한 평균($\bar{y}$)은 다음과 같다.

$$\bar{y} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}y_{hi}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}}$$

여기서, $w_{hi} = \frac{N_h}{n_h}$ 는 h층(산업별 종사자 규모별)의 i번째 표본 기업체의 가중치이며, y_{hi}는 h층의 i번째 표본 기업체로부터 얻은 변수값이고, L은 층의 총수를 나타낸다.

그리고 전체 산업의 관심변수에 대한 평균의 분산추정량은 다음과 같다.

$$\widehat{V}(\bar{y}) = \sum_{h=1}^{L} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_{h.})^2$$

여기서, $e_{hi} = w_{hi}(y_{hi} - \bar{y})/w_{..}$, $\bar{e}_{h.} = \left(\sum_{i=1}^{n_h} e_{hi}\right)/n_h$, $w_{..} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}$이고,

$f_h = \frac{n_h}{N_h}$ 는 추출률이다.

또한, 전체 산업의 관심변수에 대한 평균의 상대표준오차는 다음의 식을 통해서 계산한다.

$$\widehat{RSE}(\bar{y}) = \frac{\sqrt{\widehat{V}(\bar{y})}}{\bar{y}} \times 100(\%)$$

한편, h층에서의 표본기업체 n_h개 중에서 r_h개의 기업체만이 응답한다고 하면, $\bar{y}^*$는 다음과 같이 계산이 된다.

$$\bar{y}^* = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}^* y_{hi}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}^*}$$

여기서, $w^*_{hi} = w_{hi} \times \frac{n_h}{r_h}$ 는 무응답 기업체가 발생한 경우 최종적으로 부여되는 가중치이다.

그리고 전체 산업의 관심변수에 대한 평균의 분산추정량은 다음과 같다.

$$\widehat{V}(\bar{y}^*) = \sum_{h=1}^{L} \frac{r_h(1-f_h^*)}{r_h-1} \sum_{i=1}^{r_h} (e_{hi} - \bar{e}_{h.})^2$$

여기서, $e_{hi} = w_{hi}^*(y_{hi} - \bar{y}^*)/w_{..}^*$, $\bar{e}_{h.} = \left(\sum_{i=1}^{r_h} e_{hi}\right)/r_h$, $w_{..}^* = \sum_{h=1}^{L}\sum_{i=1}^{r_h} w_{hi}^*$,

$f_h^* = \frac{r_h}{N_h}$ 이다.

또한, 전체 산업의 관심변수에 대한 평균의 상대표준오차는 다음의 식을 통해서

계산한다.

$$\widehat{RSE}(\bar{y}^*) = \frac{\sqrt{\widehat{V}(\bar{y}^*)}}{\bar{y}^*} \times 100(\%)$$

② 층별 추정량

기업체조사에서 각 산업별 기업체 규모별 관심변수에 대한 평균($\bar{y}_h$)은 다음과 같다.

$$\bar{y}_h = \frac{\sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{i=1}^{n_h} w_{hi}}$$

그리고 각 산업별 기업체 규모별 관심변수에 대한 평균의 분산추정량은 다음과 같다.

$$\widehat{V}(\bar{y}_h) = \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_{h.})^2$$

여 기 서 ,

$e_{hi} = w_{hi}(y_{hi} - \bar{y})/w_{..}$, $\bar{e}_{h.} = \left(\sum_{i=1}^{n_h} e_{hi}\right)/n_h$, $w_{..} = \sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi}$, $f_h = \frac{n_h}{N_h}$

이다.

또한, 각 산업별 기업체 규모별 관심변수에 대한 평균의 상대표준오차는 다음의 식을 통해서 계산한다.

$$\widehat{RSE}(\bar{y}_h) = \frac{\sqrt{\widehat{V}(\bar{y}_h)}}{\bar{y}_h} \times 100(\%)$$

☞ 매뉴얼

① 표본설계에 알맞은 추정식을 결정할 것.

각종 모수의 추정은 층화, 추출률, 추출방법을 고려하여, 표본설계에 알맞은 가중치 작성 방법과 모수 추정방법을 제시한다. 모집단의 보조정보를 가중치 작성에 반영하여 모수를 추정할 수 있으며 정확한 보조정보가 있는 경우에 보조정보를 추정단계에서 이용하면 추정의 정도는 상당히 향상될 수 있다.

② 통상적으로 복합표본설계에 대한 추정식은 가중치를 고려한 추정식이므로 자체 가중설계가 아닌 경우에는 가중 추정식을 사용할 것.

표본설계의 내용이 복합표본추출설계인 경우 단순한 추정식을 사용하게 되면, 추정량의 편향과 분산추정의 문제가 발생하게 된다. 따라서 표본설계의 내용이 복합표본설계인 경우 반드시 추출단계별 가중치를 함께 고려한 추정식을 사용하는 것이 바람직하다.

③ 상대표준오차를 제시할 것.

발표된 통계의 신뢰도에 대한 지표로서 추정량의 분산 혹은 상대표준오차가 사용된다. 또한 사업체나 기업체 통계조사에서 발표된 통계자료가 이후 새로운 분석의 기초 자료로 사용될 경우에 정확한 분산은 더욱 필수적이다. 사업체나 기업체 통계조사는 대개 복합표본설계에 의해서 표본이 추출되고, 추정단계에서 가중치를 이용한 추정방법이 사용된다. 따라서 분산 추정에도 복합표본설계와 가중치를 반영하여 각 항목에 대해 추정치의 통계적 정확성을 평가할 수 있어야 한다.

4. 표본의 사후관리

일회성 조사가 아닌 계속 조사인 경우 처음 조사를 기획했던 때의 개념이나 품질, 수준이 시간이 흘러도 계속 유지될 수 있도록 체계적으로 관리되어야 한다. 그렇지 않을 경우 처음 시작 때에 아무리 많은 수고를 했을지라도 나중에는 신뢰하기 어려

운 조사가 될 가능성이 높다. 조사담당자나 조사기관 조직의 변화, 모집단이나 표본 상황의 변화 등은 조사의 수준이나 품질이 변화되도록 하는데 결정적인 영향을 미칠 수 있기 때문이다. 우리나라의 경우 상대적으로 조사의 사후관리에 소홀하게 다루어 온 감이 없지 않아 있다.

조사관리의 핵심은 처음 조사를 기획, 설계할 때의 개념, 원칙, 방법을 늘 동일한 수준으로 유지하는데 있다. 담당자가 바뀌는 상황일 경우 새로운 담당자가 조사의 여러 국면을 이해하는 데에는 시간이 필요한데 그러는 과정에서 조사의 개념이 달라질 수 있기 때문이다. 그러므로 언제 누가 그 일을 맡아도 동일한 원칙을 유지할 수 있도록 하기 위해서는 조사의 전 과정을 상세하게 체계화한 조사시스템을 구축하는 것이 필요하다.

조사관리에서 고려해야 할 또 다른 부분은 모집단 및 표본을 관리하는 작업이다. 시간이 지남에 따라 모집단의 상황이 달라질 수 있다. 조사 단위들이 생성, 소멸되기도 하며 조사단위의 속성이 변하기도 한다. 조사불응 등이 생겨 표본에 변화가 생기기도 한다. 따라서 모집단 상황을 나타내는 추출틀의 관리, 보완, 표본의 대체 또는 보정 등의 작업이 필요하다.

조사가 진행되면 계속 데이터들이 쌓여가게 된다. 방대한 양의 데이터와 작성된 통계들을 어떻게 어떤 수준으로 보관, 관리할 것인지를 고려해야 하는 데이터베이스 관리부분도 사후관리의 측면에서 고려되어야 할 중요한 부분이다.

4.1 조사시스템 구축

체계적인 조사관리를 위해서는 매번 조사할 때마다 최선의 노력을 기울이는 것 이상으로 늘 일정한 수준의 조사품질을 담보할 수 있는 조사시스템을 구축하는 것이 필요하다. 조사시스템의 핵심적인 내용으로는 예산과 조직과 인력, 구체적인 조사관련 업무를 규정하는 조사매뉴얼, 장비와 소프트웨어 등이 될 것이다.

(1) 예산, 조직, 인력

수준 높은 조사가 이루어지기 위해서는 조사에 합당한 예산과 조직 및 인력이 확보되어야 한다는 것은 가장 기본적인 요구이다. 그러기 위해서는 조사의 전 과정을 세밀하게 분석하여 업무량을 계산하고 이를 가장 효율적으로 성취할 수 있는 조직을 구성하고 인원을 배분하는 것이 필요하다.

☞ 매뉴얼

① 기대하는 품질수준의 통계를 생산하기 위한 예산과 인력을 배분할 것.

통계 생산에서 적정한 수준의 예산과 인력이 배분되지 않을 경우 그 영향은 반드시 조사의 질 저하로 나타나기 마련이다. 신뢰할 수 없는 통계를 생산하는 것은 그 자체가 예산과 인력의 낭비라고 할 수 있을 것이다. 그러므로 무조건 예산과 인력을 줄이거나, 그 반대로 늘이는 것이 능사가 아니고 해당 통계 생산을 위해 가장 적정한 수준의 예산과 인력을 확보하는 것이 필요하다.

② 조사의 전 과정을 상세하게 이해하고 관리할 수 있는 전문적 조사관리자를 확보할 것.

해당 조사에 정통한 전문적 조사관리자를 키우지 않을 경우 처음 조사를 기획하고 설계하는 단계에서 가장 중요하게 고려되었던 부분들이 시간이 갈수록 잊혀져서 원래 의도와 다른 모습으로 관리될 우려가 있다. 뿐만 아니라, 세월이 지나면서 모집단이나 표본에 새로운 상황이 발생하게 될 경우 이를 원래 설계 시의 의도에 부합하게 대처하지 못하게 되어 편향을 초래할 가능성도 있다. 담당자가 바뀔 경우라고 해도 미리 대비를 하여 조사의 전 과정을 꿰뚫어 볼 수 있는 관리자가 있게 하는 것이 필요하다.

(2) 조사매뉴얼

다음으로는 조사의 모든 과정을 상세히 담고 있는 조사매뉴얼을 갖추는 일이 필요

하다. 조사매뉴얼은 조사의 지속적인 관리를 위해 조사 당사자의 입장에서 조사의 모든 과정을 자세히 이해할 수 있도록 기록한 것이다. 조사의 기획, 표본설계, 조사표설계, 조사원 훈련 및 자료수집, 조사 결과 분석, 이용자 서비스, 사후관리 등의 전 과정에서 고려된 사항들을 상세하게 기록하여야 한다. 누구든지 이 문서를 주의 깊게 읽으면 같은 개념으로 동일한 수준의 조사를 수행할 수 있을 정도로 조사의 과정을 명확하게 설명하는 것이 필요하다.

☞ 매뉴얼

① 각 조사 단계별로 최종적인 결정사항들을 명확히 기술하고 아울러 필요한 참조사항들을 자세히 기록할 것.

처음 조사를 기획, 설계하는 과정에서 여러 가지 결정들을 내리게 된다. 이러한 결정을 내리는 과정에서 여러 상황들을 고려하여 특정한 결정을 내리는 것이 보통인데 그런 결정을 내리게 된 이유들이 있을 것이다.

가령 표본의 크기를 전국에서 1,000가구를 조사하기로 결정을 내리게 되었다고 하자. 그런 결정이 내려지게 된 배경에는 조사목적, 조사예산의 제약, 목표정도의 크기 등이 종합적으로 고려되었을 것이다. 최종적인 결정사항은 표본의 크기가 1,000가구라는 점이며, 필요한 참조사항이란 이 크기가 정해지는데 영향을 미친 여러 고려사항들을 의미하는데 이런 점들을 처음 보는 사람도 이해할 수 있는 수준으로 자세히 기록하는 것이 필요하다. 만일 친절한 조사 매뉴얼이 기록되지 않으면 시간이 지날수록 결정사항만 남게 되고 그런 결정이 내려지게 된 배경은 잊혀지기 쉽다. 그럴 경우 담당자가 바뀐다든지 하면서 조사의 개념이 변질될 우려가 많아진다.

② 조사관리자, 감독자, 조사원 훈련을 위한 매뉴얼을 마련할 것.

조사는 결국 사람에 의해 이루어지므로 조사 관련 담당자들이 표준화된 개념, 방법으로 조사에 임하여야 한다. 이를 위해서는 그들을 체계적으로 훈련시키는 것이 필요하다 조사매뉴얼에는 조사관련 담당자들의 훈련을 위한 지침도 포함되어야 한다.

③ 예비조사, 본조사, 편집, 자료분석 등 조사수행 과정에 필요한 운영사항들을 문서화할 것.

표준화된 조사, 표준화된 자료처리 및 분석을 위해 필요한 운영사항들을 상세하게 문서화하는 것이 필요하다. 매번 조사 때마다 새로운 사항들이 발견되면 이를 계속 보완해 가는 것이 필요하다.

(3) 장비 및 소프트웨어

마지막으로 조사를 하는데 필요한 장비나 소프트웨어들을 확보하여 관리하는 것이 필요하다. 조사 및 조사점검, 자료입력, 자료전송, 편집과 대체, 추정, 데이터베이스 관리 등을 위한시스템을 마련하는 것 등이 그 예가 된다.

☞ 매뉴얼

① 각각의 시스템에 오류가 없는지 철저히 점검할 것.

모든 시스템은 초기에 많은 오류가 발견된다. 따라서 충분한 경험이 쌓여지기까지는 시스템을 무작정 믿지 않고 철저히 점검하여 수정, 보완하는 것이 필요하다. 관련 전문가들에게 거듭 검토를 받는 것 또한 필요하다.

② 전문적인 시스템의 도입이 어려운 경우 범용 소프트웨어를 활용할 것.

큰 규모이면서 중요한 조사일 경우 전문적인 시스템을 도입하는 것이 바람직하다. 하지만 큰 규모의 조사가 아니고 예산상으로도 여유가 없는 조사일 경우에는 널리 일반화된 범용 소프트웨어를 사용하여 시스템을 구출할 수도 있는데 이때에는 개발 비용이나 시간 등을 많이 줄일 수 있다.

4.2 추출틀 관리 및 표본 관리

처음 표본설계를 할 때에는 대체로 그 당시 시점의 모집단을 비교적 잘 포함하는 추출틀이 마련되는 것이 일반적이다. 그러나 시간이 지남에 따라 모집단에도 새로운 변화가 생긴다. 새로운 조사단위가 생성되기도 하고 기존의 조사단위가 소멸되기도 하며 일부 조사단위는 그 특성이 변화되기도 한다. 이 때 모집단의 이런 변화를 반영하기 위해 매번 새로운 표본설계를 하는 것은 현실적으로 불가능한 경우가 대부분이다. 대개는 이런 때에 모집단의 변화를 추출틀에 계속 반영시켜 적절히 관리하고 그에 맞게 표본을 보완해 주는 조치를 취하게 된다. 만일 이런 조치를 취하지 않으면 표본의 대표성에 문제가 생겨 편향이 발생한다.

(1) 추출틀 관리

통계청에서 5년에 한 번씩 실시하는 인구주택총조사 결과는 많은 표본조사의 추출틀로 활용되고 있는데 매우 방대한 조사이기 때문에 매년 실시하기 못하고 5년에 한 번 실시하는 실정이다. 이런 경우 모집단의 변화를 매번 관찰하여 추출틀을 보완하는 것은 많은 경우 방대한 작업이며 현실적으로 어려울 수 있다. 이런 어려움으로 인해 우리나라의 많은 조사를 보면 표본설계를 새로 할 때 이외에는 추출틀 관리를 전혀 하지 않는 경우가 허다하다. 그러나 추출틀을 보완하지 않을 경우 시간이 지날수록 표본의 대표성이 떨어져서 통계의 질이 문제가 될 수 있다.

모집단이 단시간에 크게 변하는 경우는 흔하지 않으므로 현실적으로 합리적인 주기를 정하여 정기적으로 보완하는 것이 하나의 방법이 될 수 있다. 조사와 관련된 다른 행정통계 등을 이용하여 예상외로 수월하게 추출틀을 보완할 수 있는 경우가 많으므로 이에 대한 관심을 기울이는 것이 필요하다.

☞ 매뉴얼

① 추출틀 보완주기를 미리 결정할 것.

추출틀 보완주기는 모집단의 변화의 양상에 따라 달라져야 한다. 변화가 극심한 경우에는 주기가 짧아야 하고 변화가 적은 편이면 상대적으로 주기를 길게 해도 된다. 조사의 관리 체계를 세울 때 추출틀의 보완도 미리 결정해 두는 것이 바람직하다.

② 해당 조사의 추출틀 보완을 위해 참조할 관련 통계를 찾을 것.

해당 조사의 추출틀을 무엇으로 하였건 간에 이것과 유사한 관련 통계를 찾는 것이 가능할 때가 있다. 정부의 여러 부서에서 매년 생산하는 행정통계 등이 있으므로 해당조사의 추출틀 보완에 도움이 되는 통계들을 발견해 두는 것이 필요하다. 추출틀 보완이 여의치 않은 상황에서는 모집단에 뚜렷한 변화가 생겼을 때 해당 부분에 대한 추출틀을 별도로 마련할 수 있다.

③ 대표성 확보를 위해 여러 개의 추출틀을 동시에 활용할 수도 있음.

추출틀이 동시에 여러 개 활용하여 모집단에 대한 포함률을 높이는 표본설계에 관한 연구들이 있으므로 추출틀 보완을 위해 여러 개의 추출틀을 활용하는 방안도 생각할 수 있다.

(2) 표본 관리

계속조사에서 시간이 지날수록 표본에도 여러 문제가 생길 수 있다. 가령 표본단위의 소멸, 응답불응 등이 그 예이다. 추출틀 보완에 따라 새로이 표본이 추가되거나 감소되는 경우가 생길 수가 있다. 이런 경우 매 조사마다 일정한 표본크기를 유지하는 것이 현실적으로 불가능할 때도 있다.

일반적인 표본 관리의 방법으로는 표본의 대치(substitution), 추가(addition), 삭제(deletion), 표본개편 등이 있다. 표본의 대치는 조사단위에 대한 조사가 더 이상 개의치 않을 때에 유사한 다른 조사단위로 대치하는 것을 의미한다. 추가나 삭제는 모집단의 변동을 표본에 반영하기 위해 고려하게 된다. 표본개편은 표본의 모집단에 대한 대표성이 떨어졌다고 판단될 때 실시하는데 인구주택총조사 주기에 따라 5년에 한번은 표본을 개편하는 것이 일반적이다.

☞ 매뉴얼

① 표본 조사단위를 대치해야 할 때와 삭제해야 할 때를 구분하여 조치할 것.

만일 조사단위가 존재하는데 불응이나 장기부재 등으로 인해 조사가 어려울 때에는 대치를 하는 것이 좋다. 그렇지 않고 조사단위 자체가 소멸되는 경우에는 대치를 하기보다 표본에서 삭제하는 것이 바람직하다.

일반적으로 표본 조사단위에 이상이 생기면 무조건 대치하는 것이 일반적인데 이는 잘못이다. 대치를 하는 이유는 표본의 크기를 일정하게 유지하려 하기 때문이다. 표본에서의 소멸은 모집단의 변동 상황을 일정 부분 반영하는 것으로 볼 수 있기 때문이다. 그렇지 않고 모든 경우에 대해 대치를 하면 표본이 왜곡될 가능성이 많다.

어업기본통계조사 같은 경우 대치로 인해 문제가 생긴 구체적인 예이다. 우리나라의 경우 지난 여러 해 동안 매년 어가가 감소하는 추세였다. 그런데 표본에서 어가가 소멸될 경우 다른 어가로 대치하는 바람에 설계로부터 4-5년이 지났을 대 표본을

통해 추정한 어가 수나 어가인구수는 실제보다 과대추정 되어 문제가 생겼다.

② 표본 조사단위가 추가 또는 삭제될 때에는 관련된 기록을 남길 것.

기존에 조사되던 표본이 추가되거나 삭제될 때에는 해당 기록을 남겨 두면 추후에 모집단 변동에 대한 정보로 활용할 수 있다.

③ 표본의 추가나 삭제가 일어날 경우 이를 추정에 적절히 반영할 것.

표본의 추가나 삭제가 일어날 경우 이에 따른 가중값 조정 등의 조치가 취해져야 하며 이는 추정식의 수정을 야기하게 되므로 추정 과정에서 이를 적절히 반영해주어야 한다. 그렇지 않으면 추정에서 편향이 초래될 수도 있다.

④ 일정한 주기가 되면 표본을 전면적으로 개편할 것.

계속조사에서 처음 표본이 아무리 잘 설계되었다고 해도 시간이 경과함에 따라 표본의 모집단에 대한 대표성은 떨어질 수밖에 없다. 모집단의 변동 상황을 파악하여 이를 표본에 반영하는 것은 일반적으로 매우 어려운 일이다. 따라서 일정 기간이 경과하면 표본을 전면적으로 재설계하여 개편하는 것이 필요하다. 표본개편을 위해서는 좋은 추출틀의 마련이 전제되어야 하므로 일반적으로 인구주택총조사 등과 같은 총조사 시행 주기에 맞추어서 표본을 개편하는 것이 바람직하다.

4.3 데이터베이스 관리

계속조사에서 얻어지는 조사 데이터는 시계열 자료(time series data)가 되어 여러 가지 입체적인 분석을 위한 자료로 활용될 여지가 많다. 처음 조사된 자료를 통해서는 단순한 추정밖에 할 수 없다고 해도 시계열 자료가 모여지면 보다 다양하고 방대한 분석이 가능해지지 때문이다. 따라서 조사된 데이터를 어떤 양식으로 저장, 관리할 것인가 하는 점도 중요한 문제가 된다. 가능한 한 앞으로 이 데이터를 이용해서 할 수 있는 다양한 종류의 분석들을 미리 고려하여 보다 효과적으로 활용될 수 있는

데이터베이스를 관리하는 것이 필요하다.

☞ 매뉴얼

① 다양한 가능성을 충분히 고려하여 데이터베이스 설계를 할 것.

가장 단순한 형태의 데이터베이스 관리는 조사 수행 후 입력된 원본 데이터 파일을 그대로 보관하는 것이며 이것은 당연히 해야 할 일이다. 그러나 경우에 따라서는 원본 데이터뿐만 아니라 보고서에 발표된 통계들을 보관·관리하는 것이 필요하기도 하다.

일반적인 이용자들은 원본 데이터가 너무 방대하기 때문에 그것보다는 일차 가공된 형태의 데이터베이스를 요구하게 된다. 따라서 다양한 이용자들의 요구와 앞으로 예상할 수 있는 다양한 형태의 분석을 감안하여 적절히 데이터베이스를 관리하는 것이 필요하다. 굳이 한 가지 형태의 데이터파일이 아니라 여러 가지 다양한 데이터파일을 관리하는 것이 필요하다.

② 새로운 통계수요를 대비할 것.

사회가 변화해감에 따라 통계의 수요도 달라져간다. 그러므로 현재의 수요뿐만 아니라 가능하다면 장래 요구될 가능성이 많은 통계수요도 감안하여 데이터베이스를 관리하는 것이 바람직하다.

가령 현재는 지역별 통계만 작성되지만 장차 직업군별 통계가 필요할 것으로 생각된다면 응답자의 직업을 나타내는 필드를 보관하는 것이 바람직하다.

4.4 무응답 대책

조사과정에서 모든 조사관계자들이 아무리 노력한다고 해도 무응답 사례는 생기게 마련이다. 이러한 무응답은 전체 조사의 일정 및 조사의 질에 영향을 미친다. 따라서 사전에 미리 무응답에 대한 대책을 세워두는 것은 조사원 품질을 일정 수준 이상으로 유지하고 관리하기 위해 매우 필요하다.

무응답은 크게 단위무응답(unit nonresponse)과 항목무응답(item nonresponse)으로 구분된다. 단위무응답이란 응답자가 조사 자체에 불응한 경우에 생기는 것이고 항목무응답은 전체 조사항목 중 일부 조사항목에 대해 응답을 않은 경우이다. 무응답은 표본의 크기를 원래 목표한 것보다 작아지게 함으로 조사의 효율에 영향을 미치고, 무응답이 어떤 경향성을 띠게 되는 경우 추정값의 편향을 초래할 수 있다.

가능한 한 무응답이 발생하지 않도록 하는 것이 일차적인 관심이어야 하지만 부득불 무응답이 발생하였을 경우 이에 대해 적절히 대처하는 것 또한 중요하다.

<표 Ⅰ-4-1> 무응답 유형과 처리방법

무응답 유형	처리방법
단위무응답 (Unit Nonresponse)	가중셀 조정(Weighting adjustment)
	래킹비 조정(Raking ratio adjustment)
	보정방법(Calibration)
항목무응답 (Item Nonresponse)	대체(Imputation)

☞ 매뉴얼

① 조사의 전 과정에서 응답률을 극대화시킬 수 있는 방안을 마련할 것.

일반적으로 응답률에 큰 영향을 미치는 요소로는 조사방법, 조사원의 능력, 조사원의 업무량, 조사주제, 응답부담, 조사표의 길이와 복잡성, 응답자 인센티브 등이 있다.

② 가능하다면 무응답에 대해 재조사(callback)를 실시할 것.

무응답자에 대한 재조사는 응답률을 높이는 데 기여하는 동시에 무응답층의 특징을 파악하는데 도움이 된다. 무한정 재조사를 실시할 수는 없으므로 재조사를 몇 회까지 실시할 것인지, 전체를 재조사할 것인지 아니면 일부만 재조사할 것인지에 대해 구체적인 지침을 마련하는 것이 필요하다. 재조사를 할 때 무응답으로 인한 편향이 클 것으로 생각되는 조사단위에 우선순위를 두는 것이 좋다.

③ 무응답의 원인을 기록하고 모니터 할 것.

매번 조사 시마다 응답거부, 부재, 기타 무응답이 발생한 원인을 체계적으로 기록하여 관리한다. 이는 추후에 무응답에 대한 종합적인 분석 및 대책 마련을 할 때 중요한 정보가 될 수 있다.

④ 무응답 데이터에 대해 가중값 조정 또는 대체 등 적절한 조치를 취할 것.

무응답 데이터를 삭제하고 그 영향을 고려하여 가중값을 제거하거나 아니면 보정방법을 사용하여 무응답을 보정하는 조치를 취한다. 무응답에 대해 취한 조치에 따라 나중에 추정 과정에서도 이를 반영해주어야 하며 이를 명확히 밝혀야 한다. 특히 보정을 하는 경우 데이터 세트에서 보정값 여부를 나타내는 표시(flag)를 반드시 해주어야 한다.

⑤ 응답률 데이터를 공표할 것.

모든 조사단위를 응답과 무응답으로 분류하여 표시하고 각 조사의 응답률을 공표하여 조사가 지니는 한계를 밝히는 것이 필요하다.

⑥ 무응답에 관한 정보들을 축적하고 체계적인 연구를 할 것.

무응답에 관한 정보들이 축적되면 이를 이용하여 여러 유용한 정보들을 얻을 수 있으므로 조사과정에서 무응답과 관련된 정보들을 체계적으로 수집해가는 것이 필요하다. 응답자와 무응답자 사이에 특성 차이 등이 밝혀지면 무응답으로 인한 편향 등을 추측하는데 큰 도움이 된다.

Ⅱ. 표본설계 사례

1. 사업체 표본설계의 사례

1) 도・소매업 통계조사를 위한 표본설계(통계청)

(1) 조사목적

전국의 도·소매업 및 숙박·음식점업에 대한 경영실태 및 구조변화를 파악하여 각종 정책수립과 연구·분석을 위한 기초자료 제공하기 위함이다.

(2) 조사대상

한국표준산업분류상(제6차 개정 1991년 9월, 제8차 개정 2000년 1월)의 대분류 G, H 업종의 사업체 중 표본으로 선정된 약 29,000개 사업체를 조사대상으로 한다.

(3) 조사주기

○ 매년

(4) 조사기준시점 및 조사대상기간

○ 조사기준시점 : 조사대상년도 12. 31.
○ 조사대상기간 : 조사대상년도 1. 1. ~ 12. 31.

(5) 표본설계(2003년 기준 조사)

가. 추출틀

2002년 기준 사업체기초통계조사 결과 중 산업중분류 50(자동차판매 및 차량연료소매업), 51(도매업 및 상품중개업), 52(소매업 ; 자동차제외), 55(숙박 및 음식점업)에 해당하는 모든 사업체 중에서 52820(노점 및 유사 이동판매업), 52899(기타 무점포소매업) 및 55223(이동음식점 업)에 해당되는 사업체는 제외하였다.

나. 표본의 구성

① 전수조사 업종
- ○ 백화점(52111), 기타종합소매업(52119), 호텔업(55101)
- ○ 산업 세분류 및 시·도별 모집단 사업체수가 11개 미만인 업종

② 표본조사 업종
- ○ 전수층 : 종사자수가 일정규모(절사점) 이상인 모든 사업체
- ○ 표본층 : 사업체수가 많아 전수조사가 불가능하여 사업체 단위로 표본사업체를 추출하여 일부만 조사하는 업종

다. 층화

- ○ 산업세분류, 16개 시·도별 및 종사자 수 순으로 층화한 후 표본사업체를 선정하였다.

① 산업세분류

- 50 : 자동차판매, 수리 및 차량연료소매업(5개 산업세분류)
- 51 : 도매업(24개 산업세분류)
- 52 : 소매업 ; 자동차 제외 (25개 산업세분류)
- 55 : 숙박 및 음식점업 (6개 산업세분류)

② 시·도별 : 16개 시·도

③ 재층화

○ 표본조사업종의 경우 절사점을 기준으로 전수층과 표본층으로 재층화

○ 전수층 : 종사자수가 일정규모(절사점) 이상인 사업체는 모두 표본으로 선정

○ 표본층 : 종사자수가 일정규모(절사점) 미만인 사업체는 계통추출방법으로 표본선정

- 표본층에서 매출액이 100억 이상이거나 업종평균 매출액의 100배 이상인 사업체는 전수층으로 모두 선정

※ 절사점 : 시·도별로 종사자수가 큰 사업체들의 누적비와 변동계수 및 상대허용오차와 신뢰계수에 의해 결정하며 전수층과 표본층을 나누는 경계점

라. 표본규모 결정

○ 신뢰도 68%와 상대허용오차 14%로 표본규모를 결정, 서울의 경우는 신뢰도 68%와 상대허용오차 15%로 표본규모를 결정

마. 표본규모 계산

① 특성치 : 매출액(x)

○ 총 표본규모 : $n_{hi} = {}_{c}n_{hi} + {}_{s}n_{hi}$

○ 표본층 표본규모 : $${}_{s}n_{hi} = \frac{\frac{z^2 \cdot (Q_{hi} \cdot CV_{hi})^2}{e^2}}{1 + \frac{z^2 \cdot (Q_{hi} \cdot CV_{hi})^2}{{}_{s}N_{hi} \cdot e^2}}$$

○ 첨자 - h : 산업세분류
i : 시·도
c : 전수층
s : 표본층

○ 변수 - n : 표본수
N : 모집단수
Q : 종사자 총합 중 표본층이 차지하는 비율
CV : 표본층 변동계수
S : 표본층 표준편차
e : 허용상대오차
z : 신뢰계수

바. 표본추출

3개 전수조사 업종에 대해서는 해당되는 모든 사업체를 조사하고, 표본조사 업종에 대해서는 사업체단위로 절사법 표본설계 방법을 응용하여 표본 추출를 실시하였다.

○ 전수층 : 절사점 이상인 사업체는 모두 선정
○ 표본층 : 절사점 미만인 사업체는 주어진 표본규모에 따라 계통추출방법으로 표본을 선정

사. 표본오차

① 업종별·시도별 총량 추정시

○ 분산 : $V(\ \widehat{X}_{hi}) = \ {}_sN_{hi}^{\ 2}\left(\frac{{}_sN_{hi} - {}_sn_{hi}}{{}_sN_{hi}}\right) \cdot \frac{{}_ss_{hi}^{\ 2}}{{}_sn_{hi}}$

여기서 ${}_s\ s_{hi}^{\ 2} = \frac{1}{{}_sn_{hi} - 1}\left(\sum_j^{{}_sn_{hi}} {}_sx_{hij}^{\ 2} - \frac{(\ \sum_j^{{}_sn_{hi}} {}_sx_{hij})^2}{{}_sn_{hi}}\right)$,

j : 개별사업체 를 나타냄.

○ 표준오차 : $SE(\ \widehat{X}_{hi}) = \sqrt{\ V(\ \widehat{X}_{hi})}$

○ 상대표준오차 : $RSE(\ \widehat{X}_{hi}) = \frac{SE(\ \widehat{X}_{hi})}{\widehat{X}_{hi}} \cdot 100$

② 업종별 전국 총량 추정시

○ 분산 : $V(\ \widehat{X}_h) = \sum_i V(\ \widehat{X}_{hi})$

○ 표준오차 : $SE(\ \widehat{X}_h) = \sqrt{\ V(\ \widehat{X}_h)}$

○ 상대표준오차 : $RSE(\ \widehat{X}_h) = \frac{SE(\ \widehat{X}_h)}{\widehat{X}_h} \cdot 100$

아. 모수추정

○ 표본조사 결과 계산된 업종별·시·도별 합계에 2002년 기준 사업체기초통계조사결과의 해당 사업체수 또는 종사자수를 기준으로 승수를 주어 총량 추정

① 업종별·시도별 총량 :

$$\widehat{X}_{hi} = \sum_{j}^{c n_{hi}} x_{hij} + w_{hi} \sum_{j}^{s n_{hi}} x_{hij} \ , \quad w_{hi} = {}_{s}N_{hi} / {}_{s}n_{hi}$$

② 업종별 전국 총량 : $\widehat{X}_{h} = \sum_{i=1}^{16} \widehat{X}_{hi}$ (첨자 $\widehat{X}$는 추정치)

2) 기업체노동비용조사 표본설계(노동부)

(1) 조사의 목적

기업체노동비용조사는 상용근로자 10인 이상 기업체를 대상으로 근로자의 고용에 따른 제반 노동비용 파악을 위한 통계조사이다. 기업체노동비용조사의 목적은 기업에 대하여 사용자가 근로자를 고용하여 발생하는 모든 비용의 종류 및 금액 등을 종합적으로 조사 파악하여 노동정책 입안자료는 물론 기업의 근로자 복지후생 증진을 위한 기초 자료를 제공하는데 있다.

(2) 조사의 대상

이 조사의 대상은 농업, 수렵업, 임업 및 어업 등을 제외한 한국표준산업분류 상의 전 산업(단, 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관은 제외)에 대하여 상용근로자 10인 이상을 고용하고 있는 기업체이다.

이 조사의 단위는 기업단위로 단독법인 기업이나 본사가 조사대상이 되며, 본사의 경우에는 기업에 속하여 있는 전체의 사업체(본사 이외의 지점, 공장, 영업소 등)분을 일괄하여 조사하고 있다. 한편, 이 조사는 표본기업체에 고용된 전체 상용근로자를 대상으로 한다.

이 조사의 조사기준기간은 노동비용에 대해서는 매년 1월부터 동년 12월 말까지 1년간으로 한다. 다만, 회계연도가 조사기준 기간(1.1. ~ 12. 31)과 다른 기업체에 대

해서는 회계연도를 기준으로 조사되고 있다.

(3) 자료 산출 범위

각 산업 중분류와 기업체 규모에 따라서 기업체의 노동비용 구조에 대한 제반 통계 생산을 원칙으로 한다.

(4) 현행 표본설계의 개요

현행 기업체노동비용조사는 기업에 대하여 사용자가 근로자를 고용하여 발생하는 모든 비용의 종류 및 금액 등을 조사 파악하여 노동정책 입안 자료는 물론 기업의 근로자 복지후생증진을 위한 기초 자료를 제공하는데 목적이 있다. 현행 조사의 조사범위는 농업, 수렵업, 임업 및 어업 부문을 제외한 전산업(단, 국가 또는 지방행정기관, 군 · 경찰 및 국 · 공립 교육기관은 제외)이다.

2004년도 기업체노동비용조사는 상용근로자 10인 이상 기업체 중 통계적 방법에 의하여 추출된 2,438개 표본기업체를 조사대상으로 한다. 2004년도 조사의 조사대상 기간은 2003년 회계연도(1년간)이고, 조사 실시 기간은 3월 말 이전 결산법인에 대해서는 2004. 5. 15부터 6. 5까지 진행되었고, 4~6월말 결산법인은 2004. 6. 25부터 7. 5까지 진행되었다.

조사내용은 크게 기업체 속성에 관한 사항과 노동비용에 관한 사항으로 구분된다.

가. 기업체 속성에 관한 사항

기업체명, 소재지, 대표자 성명, 상용근로자 수, 주요 생산품목 또는 영업종목, 노조결성 유무 등

나. 노동비용에 관한 사항

급여지급 연인원, 현금급여 총액(정액 및 초과급여, 상여금 등 특별급여), 현물지급의 비용, 퇴직금의 비용, 모집비, 교육훈련비, 법정복리비(건강보험료, 산재보험료, 국

민연금, 고용보험료, 장애인고용부담금, 기타 법정복리비), 법정외복리비(주거, 의료·보건, 식사, 문화·체육·오락, 보험료지원금, 경조, 저축장려금, 학비보조, 사내 근로복지기금 출연금, 보육비지원금, 근로자휴양, 종업원지주제도 지원금, 기타 법정 외 복리비), 기타 노동비용(작업복 비용, 전근비용, 사보에 관한 비용, 표창 비용 등)

조사방법은 조사대상 기업의 조사표 기입담당자를 정하여 기업체에서 직접 조사표를 기입하여 관할 노동부 지방관서에 제출하는 자계식 조사를 원칙으로 하였다. 그러나 조사표 기재상 중대한 착오, 누락 등이 있어 간접적인 지도로는 정정할 수 없거나 조사표 제출이 과도하게 지연된 경우에는 조사원이 기업체를 방문하여 조사표의 작성을 지도하거나 자료제출을 요구하여 조사표를 작성하였다.

<표 II-1-1> 산업대분류별 기업체 규모별 표본 수 현황

(단위: 개)

규 모 / 산업대분류	~29인	30~99인	100~299인	300~499인	500인 이상	합 계
C. 광업	10	12	5	2	2	31
D. 제조업	290	307	253	95	176	1,121
E. 전기, 가스 및 수도사업	3	7	6	2	6	24
F. 건설업	86	41	14	5	19	165
G. 도매 및 소매업	98	31	22	7	16	174
H. 숙박 및 음식업	8	4	5	7	9	33
I. 운수업	36	28	76	12	27	179
J. 통신업	1	4	2	0	5	12
K. 금융 및 보험업	30	41	33	13	31	148
L. 부동산업 및 임대업	53	28	7	4	6	98
M. 사업 서비스업	71	38	49	32	42	232
O. 교육 서비스업	13	19	16	3	12	63
P. 보건 및 사회 복지사업	21	20	27	8	18	94
R. 기타, 개인 서비스업	25	21	17	8	14	85
합 계	745	601	532	198	383	2,459

(자료 : 노동부, 「기업체노동비용조사 결과」, 2003)

2003년도에 조사된 기업체노동비용실태조사의 전체 표본 수는 2,459개 기업체이고, 2004년도 조사의 표본기업체 수는 2,438개 기업체이다. 이들 표본기업체의 산업대분류별, 기업체 규모별 표본크기 현황은 다음의 <표 Ⅱ-1-1>과 <표 Ⅱ-1-2>와 같다. 전체적으로 제조업체에 해당하는 표본기업체가 약 1100여개로 전체 표본크기의 절반을

차지하고 있다. 한편, 표본기업체의 기업체 규모별 분포를 보면 산업분류별 분포에 비해서 상대적으로 규모별 표본크기의 편차는 적은 것을 알 수 있다.

<표 Ⅱ-1-2> 산업대분류별 기업체 규모별 표본 수 현황

(단위: 개)

규 모 / 산업대분류	~29명	30~99명	100~299명	300~499명	500명 이상	합 계
C. 광업	10	11	5	2	2	30
D. 제조업	299	296	245	96	178	1,114
E. 전기, 가스 및 수도사업	3	7	6	2	6	24
F. 건설업	85	34	19	6	19	163
G. 도매 및 소매업	93	35	23	4	19	174
H. 숙박 및 음식업	6	7	6	6	9	34
I. 운수업	33	36	69	9	27	174
J. 통신업	2	2	3	0	5	12
K. 금융 및 보험업	32	38	31	13	30	144
L. 부동산업 및 임대업	56	27	6	3	7	99
M. 사업 서비스업	64	43	48	34	41	230
O. 교육 서비스업	15	17	16	3	12	63
P. 보건 및 사회 복지사업	20	19	28	5	20	92
R. 기타, 개인 서비스업	25	18	17	11	14	85
합 계	743	590	522	194	389	2,438

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004)

(4) 현행 조사데이터에 대한 분석

가. 직접노동비용 분석

다음 <표 Ⅱ-3-3>은 2004년도 기업체노동비용조사 결과를 분석한 것으로 산업대분류별 기업체 규모별 직접노동비용의 평균과 그 상대표준오차를 정리한 결과이다. 산업대분류 및 기업체 규모 구분에 따라서 평균 직접노동비용 추정값에 편차가 크게 나타나고 있다. 한편, 산업대분류별 또는 기업체 규모별 통계의 상대표준오차는 대체로 안정적인 결과를 나타내고 있지만, 산업대분류 내의 기업체 규모별 평균 직접노동비용 추정액의 상대표준오차는 편차가 대단히 큰 것으로 나타났다. 특히, 10~29인, 30~99인, 100~299인 규모에서 상대표준오차가 크게 나타나는 경우가 많았다.

<표 Ⅱ-1-3> 산업대분류별 기업체 규모별 평균 직접노동비용 및 상대표준오차

(단위 : 억원, %)

산업대분류 \ 규모		10~29명	30~99명	100~299명	300~499명	500~999명	1,000명이상	합 계
C.광 업	평균	1663.05	1734.56	2152.70	2667.56	-	2375.34	2059.84
	상대표준오차	10.14	6.27	9.40	12.80	-	3.72	3.80
D. 제조업	평균	1140.07	1227.21	1371.33	1638.31	1764.01	2303.58	1687.15
	상대표준오차	1.85	1.87	1.65	2.29	2.47	2.71	3.42
E. 전기, 가스 및 수도	평균	1345.52	2597.92	3349.59	3752.93	3524.16	3683.00	3578.81
	상대표준오차	19.33	11.86	5.56	5.06	0.00	0.00	0.51
F. 건설업	평균	1324.23	1430.31	1771.90	1689.33	1873.73	2713.28	1567.98
	상대표준오차	2.59	4.92	5.77	6.47	6.03	4.69	2.33
G. 도매 및 소매업	평균	1616.26	1775.74	2404.40	1736.57	2311.39	2160.38	1891.04
	상대표준오차	4.22	6.17	6.98	14.79	8.92	9.54	2.89
H. 숙박 및 음식업	평균	833.72	936.65	1028.16	1555.32	1855.55	1639.88	1349.93
	상대표준오차	6.56	4.54	16.05	4.51	7.25	12.09	5.81
I. 운수업	평균	1016.65	1175.88	1224.61	1557.78	1556.77	2246.82	1483.85
	상대표준오차	10.25	5.75	3.30	6.17	4.39	3.12	4.12
J. 통신업	평균	1076.78	1642.94	1852.48	-	1790.11	2633.13	2514.64
	상대표준오차	20.37	5.34	15.12	-	0.00	1.24	2.94
K. 금융 및 보험업	평균	2181.94	2810.71	2830.11	3205.43	3688.60	3434.05	3367.04
	상대표준오차	6.28	10.89	5.51	0.00	5.96	3.83	3.22
L. 부동산업 및 임대업	평균	1101.62	976.20	1868.87	2022.57	2194.36	2907.36	1852.75
	상대표준오차	6.11	12.74	21.02	37.41	17.35	3.27	9.06
M. 사업 서비스업	평균	1507.38	1671.65	1715.64	2149.16	2124.97	2139.07	1830.65
	상대표준오차	5.39	15.48	8.29	8.56	8.15	9.49	4.66
O. 교육 서비스업	평균	1459.08	2362.32	2464.96	2705.94	2787.79	3019.99	2515.05
	상대표준오차	14.19	5.86	2.92	3.27	2.11	5.61	3.22
P. 보건 및 사회 복지	평균	1331.52	1524.14	1726.78	1769.71	2172.00	2309.17	2034.90
	상대표준오차	2.44	3.54	2.24	2.22	0.83	3.80	1.30
R. 기타, 개인 서비스업	평균	1320.15	1711.15	2125.81	2123.74	2493.26	2554.04	1829.41
	상대표준오차	7.56	15.64	9.55	0.00	7.20	5.72	6.11
전 체	평균	1301.10	1377.43	1512.71	1786.30	1970.90	2501.11	1834.31
	상대표준오차	1.48	2.69	1.81	2.56	2.17	1.71	1.87

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004)

나. 간접노동비용 분석

다음 <표 Ⅱ-1-4>는 2004년도 기업체노동비용조사 결과를 분석한 것으로 산업대분

류별 기업체 규모별 간접노동비용의 평균과 그 상대표준오차를 정리한 결과이다. 산업대분류 및 기업체 규모 구분에 따라서 평균 간접노동비용 추정액도 직접노동비용과 마찬가지로 편차가 크게 나타나고 있음을 살펴 볼 수 있다. 한편, 간접노동비용 추정에 대한 상대표준오차는 직접노동비용 추정의 경우에 비해서 크게 나타났고, 통계작성 단위별 편차도 크다.

<표 II-1-4> 산업대분류별 기업체 규모별 평균 간접노동비용 및 상대표준오차

(단위 : 억원, %)

산업대분류 \ 규모		10~29명	30~99명	100~299명	300~499인	500~999명	1,000명이상	합 계
C. 광 업	평균	635.43	958.14	1448.60	1347.72	-	1512.99	1158.61
	상대표준오차	12.24	23.06	15.27	25.24	-	2.07	5.35
D. 제조업	평균	386.49	439.24	612.03	747.28	747.95	954.04	685.66
	상대표준오차	5.00	4.10	7.28	5.42	4.55	3.13	3.42
E. 전기, 가스 및 수도업	평균	223.89	715.82	1058.23	828.81	806.55	1029.76	983.62
	상대표준오차	48.60	23.35	16.78	13.66	0.00	0.00	1.52
F. 건설업	평균	407.33	506.57	818.66	953.74	824.82	1258.16	591.78
	상대표준오차	7.20	8.13	12.78	18.02	10.03	12.55	4.73
G. 도매 및 소매업	평균	362.00	424.86	750.00	389.41	701.91	649.47	495.64
	상대표준오차	6.63	9.80	21.86	18.88	12.11	16.14	6.49
H. 숙박 및 음식업	평균	208.20	337.45	354.50	493.61	747.06	630.21	490.28
	상대표준오차	19.94	14.68	23.99	6.59	7.20	15.66	8.09
I. 운수업	평균	275.76	487.00	475.07	545.66	660.03	1454.71	712.55
	상대표준오차	16.79	11.76	5.55	9.93	9.13	24.03	14.24
J. 통신업	평균	296.61	445.86	627.61	-	1503.89	2486.21	2279.82
	상대표준오차	2.46	26.93	22.96	-	0.00	8.89	13.62
K. 금융 및 보험업	평균	688.47	773.44	981.05	870.18	1039.68	766.05	785.73
	상대표준오차	15.37	16.97	11.11	0.00	6.94	12.59	10.66
L. 부동산업 및 임대업	평균	213.49	306.75	340.78	534.21	664.75	889.63	484.07
	상대표준오차	7.00	25.76	16.10	39.62	27.35	29.93	13.86
M. 사업 서비스업	평균	346.00	482.95	450.86	584.33	471.27	671.42	499.65
	상대표준오차	7.55	22.06	12.51	12.37	10.19	15.75	7.64
O. 교육 서비스업	평균	257.12	251.67	209.26	210.35	261.84	297.42	265.31
	상대표준오차	8.24	10.48	5.20	5.50	0.50	7.89	4.20
P. 보건 및 사회복지사업	평균	323.72	282.64	363.74	339.18	396.26	497.29	405.83
	상대표준오차	3.62	7.56	1.99	9.91	4.04	7.48	3.28
R. 기타, 개인 서비스업	평균	341.41	612.43	1106.30	596.36	707.47	1241.14	692.26
	상대표준오차	10.67	31.32	21.68	0.00	7.26	9.54	11.90
전 체	평균	370.44	459.95	584.94	670.01	701.65	1046.20	694.67
	상대표준오차	2.95	4.24	4.62	4.05	3.23	8.74	5.44

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004)

또한 산업대분류 내의 기업체 규모별 평균 간접노동비용 추정에 대한 상대표준오차는 편차가 대단히 큰 것으로 나타났다. 특히, 30~99명, 100~299명, 1,000명 이상 규모에서 상대표준오차가 크게 나타나는 경우가 많았다.

다. 총노동비용 분석

다음 <표 Ⅱ-1-5>는 2004년 기업체노동비용조사 결과 중에서 산업대분류별 기업체 규모별 총 노동비용의 평균과 그 상대표준오차를 정리한 결과이다. 산업대분류 및 기업체 규모 구분에 따라서 평균 간접노동비용 추정액도 다른 노동비용과 마찬가지로 편차가 크게 나타나고 있음을 살펴 볼 수 있다. 한편, 총 노동비용 추정액에 대한 상대표준오차는 직접노동비용 추정의 경우에 비해서 상대표준오차가 크게 나타났고, 상대적인 편차도 크다. 또한 산업대분류 내의 기업체 규모별 평균 총노동비용 추정에 대한 상대표준오차는 편차가 대단히 큰 것으로 나타났다. 특히, 30~99명, 300~499명, 1,000명 이상 규모에서 상대표준오차가 크게 나타나는 경우가 많았다.

<표 Ⅱ-1-5> 산업대분류별 기업체 규모별 평균 총 노동비용 및 상대표준오차

(단위 : 억원, %)

산업대분류 \ 규모		10~29명	30~99명	100~299명	300~499명	500~999명	1,000명이상	합 계
C.광 업	평균	2298.48	2692.70	3601.30	4015.28	-	3888.34	3218.44
	상대표준오차	8.25	8.76	9.89	16.98	-	1.47	3.50
D. 제조업	평균	1526.56	1666.45	1983.36	2385.59	2511.97	3257.62	2372.81
	상대표준오차	2.20	2.12	2.85	2.71	2.67	2.42	3.30
E. 전기, 가스 및 수도업	평균	1569.41	3313.73	4407.82	4581.73	4330.71	4712.76	4562.44
	상대표준오차	13.52	14.01	6.55	6.62	0.00	0.00	0.64
F. 건설업	평균	1731.56	1936.88	2590.56	2643.06	2698.55	3971.44	2159.76
	상대표준오차	3.24	5.14	6.38	10.32	6.90	6.05	2.64
G. 도매 및 소매업	평균	1978.26	2200.59	3154.40	2125.98	3013.29	2809.85	2386.69
	상대표준오차	4.31	6.64	7.58	15.46	9.05	10.33	3.13
H. 숙박 및 음식업	평균	1041.92	1274.10	1382.66	2048.93	2602.61	2270.09	1840.21
	상대표준오차	8.53	4.95	17.58	4.40	6.65	13.00	6.29
I. 운수업	평균	1292.42	1662.88	1699.68	2103.44	2216.79	3701.53	2196.41
	상대표준오차	11.09	6.71	3.45	6.82	5.56	10.48	6.68
J. 통신업	평균	1373.40	2088.80	2480.08	-	3294.00	5119.34	4794.46
	상대표준오차	15.44	1.55	17.08	-	0.00	3.96	7.67
K. 금융 및 보험업	평균	2870.42	3584.15	3811.16	4075.61	4728.28	4200.10	4152.77
	상대표준오차	7.02	10.96	6.06	0.00	4.29	4.64	3.88
L. 부동산업 및 임대업	평균	1315.10	1282.94	2209.65	2556.77	2859.11	3796.99	2336.82
	상대표준오차	5.85	15.14	20.17	37.79	19.25	9.52	9.34
M. 사업 서비스업	평균	1853.38	2154.60	2166.50	2733.49	2596.24	2810.49	2330.30
	상대표준오차	5.31	16.06	8.70	8.71	8.19	10.66	5.06
O. 교육 서비스업	평균	1716.20	2613.99	2674.22	2916.29	3049.63	3317.41	2780.36
	상대표준오차	11.72	4.95	2.71	3.39	1.90	5.34	2.91
P. 보건 및 사회 복지사업	평균	1655.24	1806.78	2090.52	2108.89	2568.27	2806.46	2440.73
	상대표준오차	2.20	3.86	2.01	1.29	1.09	3.18	1.17
R. 기타 개인서비스업	평균	1661.57	2323.58	3232.11	2720.10	3200.73	3795.18	2521.67
	상대표준오차	6.34	19.63	11.64	0.00	6.84	6.90	7.45
전 체	평균	1671.54	1837.38	2097.65	2456.31	2672.55	3547.31	2528.98
	상대표준오차	1.55	2.86	2.11	2.62	2.15	2.95	2.50

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004)

라. 연도별 노동비용 추정액 및 상대표준오차 분석

다음 <표 Ⅱ-1-6>과 <표 Ⅱ-1-7>은 2003년과 2004년 기업체노동비용조사 결과를 분석한 것으로 각각 산업대분류 및 기업체 규모별 평균 노동비용 추정액 및 상대표

준오차의 연도별 변화를 살펴보기 위한 것이다. 먼저 <표 Ⅱ-1-6>을 살펴보면 각 산업대분류별 노동비용 추정액의 상대표준오차에 대하여 연도별 변화는 대체로 안정적인 것으로 나타나고 있음을 확인할 수 있다. 한편, <표 Ⅱ-1-7>을 살펴보면 각 기업체 규모별 노동비용 추정액의 상대표준오차에 대하여 연도별 변화는 대체로 안정적이지만 간접노동비용 추정의 경우에는 연도별로 편차가 다소 크게 나타나고 있다.

<표 Ⅱ-1-6> 연도별 산업대분류별 평균 노동비용 및 상대표준오차

(단위 : 억원, %)

산업대분류 \ 연도		직접노동비용		간접노동비용		총 노동비용	
		2003년	2004년	2003년	2004년	2003년	2004년
C.광 업	평균	1834.47	2059.84	983.99	1158.61	2818.46	3218.44
	상대표준오차	3.29	3.80	6.39	5.35	3.72	3.50
D. 제조업	평균	1596.80	1687.15	657.52	685.66	2254.31	2372.81
	상대표준오차	3.78	3.42	7.42	3.42	3.82	3.30
E. 전기, 가스 및 수도사업	평균	3173.16	3578.81	860.29	983.62	4033.45	4562.44
	상대표준오차	0.56	0.51	1.02	1.52	0.64	0.64
F. 건설업	평균	1537.43	1567.98	638.17	591.78	2175.60	2159.76
	상대표준오차	2.04	2.33	5.58	4.73	2.60	2.64
G. 도매 및 소매업	평균	1685.22	1891.04	541.99	495.64	2227.21	2386.69
	상대표준오차	3.02	2.89	8.60	6.49	3.82	3.13
H. 숙박 및 음식업	평균	1262.82	1349.93	558.18	490.28	1821.00	1840.21
	상대표준오차	5.61	5.81	9.24	8.09	6.15	6.29
I. 운수업	평균	1186.19	1483.85	622.13	712.55	1808.32	2196.41
	상대표준오차	5.43	4.12	12.90	14.24	7.09	6.68
J. 통신업	평균	2403.06	2514.64	832.95	2279.82	3236.02	4794.46
	상대표준오차	3.45	2.94	11.75	13.62	1.66	7.67
K. 금융 및 보험업	평균	2992.92	3367.04	707.74	785.73	3700.66	4152.77
	상대표준오차	3.66	3.22	6.40	10.66	3.58	3.88
L. 부동산업 및 임대업	평균	1693.34	1852.75	516.50	484.07	2209.85	2336.82
	상대표준오차	8.55	9.06	18.13	13.86	9.16	9.34
M. 사업 서비스업	평균	1610.69	1830.65	455.40	499.65	2066.09	2330.30
	상대표준오차	4.66	4.66	9.26	7.64	5.23	5.06
O. 교육 서비스업	평균	2308.30	2515.05	283.39	265.31	2591.69	2780.36
	상대표준오차	2.68	3.22	7.87	4.20	2.55	2.91
P. 보건 및 사회 복지사업	평균	1942.22	2034.90	462.35	405.83	2404.57	2440.73
	상대표준오차	1.84	1.30	6.15	3.28	1.92	1.17
R. 기타, 개인 서비스업	평균	1526.50	1829.41	701.02	692.26	2227.51	2521.67
	상대표준오차	5.46	6.11	11.63	11.90	6.37	7.45
전 체	평균	1662.96	1834.31	626.60	694.67	2289.56	2528.98
	상대표준오차	2.16	1.87	4.42	5.44	2.21	2.50

(자료 : 노동부, 「기업체노동비용조사 결과」, 2003, 2004)

<표 II-3-7> 연도별 기업체 규모별 평균 노동비용 및 상대표준오차

(단위 : 억원, %)

규모 \ 구분		직접노동비용		간접노동비용		총 노동비용	
		2003년	2004년	2003년	2004년	2003년	2004년
10~29명	평균	1214.39	1301.10	375.31	370.44	1589.70	1671.54
	상대표준오차	2.26	1.48	3.90	2.95	2.37	1.55
30~99명	평균	1266.74	1377.43	440.86	459.95	1707.60	1837.38
	상대표준오차	2.43	2.69	5.33	4.24	2.82	2.86
100~299명	평균	1417.70	1512.71	686.53	584.94	2104.23	2097.65
	상대표준오차	2.21	1.81	19.63	4.62	7.46	2.11
300~499명	평균	1585.94	1786.30	567.74	670.01	2153.68	2456.31
	상대표준오차	2.62	2.56	3.65	4.05	2.63	2.62
500~999명	평균	1700.25	1970.90	702.69	701.65	2402.94	2672.55
	상대표준오차	3.74	2.17	7.01	3.23	4.17	2.15
1,000명 이상	평균	2262.75	2501.11	825.98	1046.20	3088.72	3547.31
	상대표준오차	2.26	1.71	3.68	8.74	1.85	2.95
전 체	평균	1662.96	1834.31	626.60	694.67	2289.56	2528.98
	상대표준오차	2.16	1.87	4.42	5.44	2.21	2.50

(자료 : 노동부, 「기업체노동비용조사 결과」, 2003, 2004)

마. 각종 노동비용 추정에서 가중 추정방법과 비가중 추정방법의 비교

다음 <표 Ⅱ-1-8>과 <표 Ⅱ-1-9>는 2004년 기업체노동비용조사 결과를 분석한 것으로 각각 산업대분류 및 기업체 규모별 평균 노동비용 추정액에 대해서 가중값을 준 경우와 그렇지 않은 경우를 비교하여 정리한 것이다. 산업대분류별 구분에 따른 노동비용관련 통계를 살펴보면 가중치 부여하여 구한 경우와 그렇지 않은 경우의 추정값에 상당한 차이가 있음을 확인할 수 있다. 현재 가중치는 산업대분류와 기업체 규모를 구분하여 주고 있는데, 기업체 규모에 따라서 추출률의 차이가 크기 때문에 산업대분류별 노동비용 통계의 추정값은 가중치를 준 경우와 그렇지 않은 경우에 차이가 크게 발생하게 된다. 그러나 <표 Ⅱ-1-9>를 보면 기업체 규모 구분에 따라서 노동비용 추정값을 살펴보면 상대적으로 가중값을 준 경우와 그렇지 않은 경우에 차이가 거의 없음을 확인할 수 있다.

<표 II-1-8> 산업대분류별 가중 및 비가중 평균 노동비용과 상대표준오차

(단위 : 억원, %)

산업대분류 \ 연도		직접노동비용(2004)		간접노동비용(2004)		총 노동비용(2004)	
		비가중	가중	비가중	가중	비가중	가중
C.광 업	평균	2298.51	2059.84	1410.41	1158.61	3708.92	3218.44
	상대표준오차	3.62	3.80	3.99	5.35	3.09	3.50
D. 제조업	평균	2065.06	1687.15	862.81	685.66	2928.87	2372.81
	상대표준오차	3.18	3.42	2.92	3.42	2.88	3.30
E. 전기, 가스 및 수도사업	평균	3645.95	3578.81	1005.02	983.62	4605.97	4562.44
	상대표준오차	0.22	0.51	0.70	1.52	0.28	0.64
F. 건설업	평균	2265.13	1567.98	1035.82	591.78	3300.95	2159.76
	상대표준오차	4.21	2.33	8.55	4.73	4.73	2.64
G. 도매 및 소매업	평균	2159.81	1891.04	637.94	495.64	2797.74	2386.69
	상대표준오차	5.69	2.89	9.92	6.49	6.13	3.13
H. 숙박 및 음식업	평균	1609.48	1349.93	604.57	490.28	2214.05	1840.21
	상대표준오차	7.04	5.81	9.23	8.09	7.55	6.29
I. 운수업	평균	1940.88	1483.85	1136.13	712.55	3077.01	2196.41
	상대표준오차	4.39	4.12	20.98	14.24	9.44	6.68
J. 통신업	평균	2612.39	2514.64	2450.61	2279.82	5062.99	4794.46
	상대표준오차	1.13	2.94	9.62	13.62	4.56	7.67
K. 금융 및 보험업	평균	3391.13	3367.04	791.95	785.73	4183.08	4152.77
	상대표준오차	3.22	3.22	10.69	10.66	3.91	3.88
L. 부동산업 및 임대업	평균	2219.44	1852.75	641.76	484.07	2861.19	2336.82
	상대표준오차	8.78	9.06	17.64	13.86	9.65	9.34
M. 사업 서비스업	평균	2069.35	1830.65	584.50	499.65	2653.84	2330.30
	상대표준오차	5.45	4.66	9.93	7.64	6.20	5.06
O. 교육 서비스업	평균	2828.24	2515.05	272.52	265.31	3110.77	2780.36
	상대표준오차	3.48	3.22	5.06	4.20	3.30	2.91
P. 보건 및 사회 복지사업	평균	2061.45	2034.90	975.15	405.83	2473.10	2440.73
	상대표준오차	1.37	1.30	8.14	3.28	1.24	1.17
R. 기타, 개인 서비스업	평균	2385.76	1829.41	701.02	692.26	3360.91	2521.67
	상대표준오차	3.75	6.11	11.63	11.90	4.82	7.45
전 체	평균	2293.35	1834.31	912.62	694.67	3205.97	2528.98
	상대표준오차	1.48	1.87	6.70	5.44	2.43	2.50

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004)

<표 Ⅱ-1-9> 기업체 규모별 가중 및 비가중 평균 노동비용과 상대표준오차

(단위 : 억원, %)

규모 \ 구분		직접노동비용(2004)		간접노동비용(2004)		총 노동비용(2004)	
		비가중	가중	비가중	가중	비가중	가중
10~29인	평균	1305.52	1301.10	372.45	370.44	1677.97	1671.54
	상대표준오차	1.41	1.48	2.94	2.95	2.43	1.55
30~99인	평균	1464.83	1377.43	468.64	459.95	1933.47	1837.38
	상대표준오차	2.32	2.69	3.76	4.24	2.41	2.86
100~299인	평균	1631.20	1512.71	608.42	584.94	2239.63	2097.65
	상대표준오차	1.50	1.81	4.24	4.62	1.82	2.11
300~499인	평균	1916.37	1786.30	683.96	670.01	2600.33	2456.31
	상대표준오차	2.10	2.56	3.66	4.05	2.18	2.62
500~999인	평균	1998.99	1970.90	691.42	701.65	2690.41	2672.55
	상대표준오차	1.91	2.17	3.01	3.23	1.92	2.15
1000이상	평균	2532.01	2501.11	1049.93	1046.20	3581.94	3547.31
	상대표준오차	1.58	1.71	8.25	8.74	2.74	2.95
정 체	평균	2293.35	1834.31	913.62	694.67	3205.97	2528.98
	상대표준오차	1.48	1.87	6.70	5.44	2.43	2.50

(자료 : 노동부, 「기업체노동비용조사 결과」, 2004.)

(5) 모집단 분석

가. 통계청 기업체·사업체명부 작성 개요

새로운 표본설계의 모집단 추출틀은 통계청의 2003년 기준 기업체 명부 작성 결과를 이용한다. 기업체 명부는 정확한 기업체단위 모집단 자료 구축 및 관리를 통하여 기업의 구조 및 변동파악은 물론 기업체 단위의 조사를 위한 모집단 명부를 제공하는 것을 목적으로 한다.

기업체 명부의 작성범위는 회사법인(주식회사, 유한회사, 합자회사, 합명회사)이고, 사업체기초통계조사 결과를 기초 자료로 하여 사업체구분(단독기업, 본사, 지사(본사명, 소재지)), 법인등록번호, 사업자등록번호 등을 고유번호로 하여 사업체 단위를 기업체 단위로 연계하여 작성되었다. 기업체 명부의 구체적인 작성과정은 다음과 같다.

- 기초자료 정리(회사법인 확인, 기업체 고유번호 부여)
- 본·지사 연계(사업체기초통계조사 이용)

- 행정자료 보완(연계가 어려운 사업체, 누락사업체)
  · 본사확인 및 누락보완(금감원 기업공시자료, 전년도 기업체명부)
  · 40대 다기업 집단 연계(상공회의소 기업정보 자료)
  · 1000대 기업 적정성 확인(상공회의소 기업정보 자료)
  · 기업변동 자료 보완(법원 행정처)
  · 기업의 산업분류(상공회의소 기업정보 자료)
- 타 통계자료 보완(광업·제조업통계조사, 도소매 및 서비스업통계조사)

나. 기업체 현황 자료 분석

현행 조사에서 기업체노동비용조사의 대상은 농업, 수렵업, 임업 및 어업 등을 제외한 한국표준산업분류 상의 전 산업(단, 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관은 제외)에 대하여 상용근로자 10명 이상을 고용하고 있는 기업체이다.

<표 II-1-10> 산업대분류별 기업체 규모별 현황

(단위 : 개)

규모 / 대분류	10~29명	30~99명	100~299명	300~499명	500~999명	1,000명이상	합 계
C. 광업	213	66	9	3	0	1	292
D. 제조업	18,222	9,989	2,497	397	278	178	31,561
E. 전기, 가스 및 수도사업	30	30	13	5	2	4	84
F. 건설업	8,184	1,268	170	26	19	10	9,677
G. 도매 및 소매업	7975	1,555	255	35	28	25	9,873
H. 숙박 및 음식점업	467	248	80	19	17	12	843
I. 운수업	2,316	1,446	898	83	46	10	4,799
J. 통신업	206	90	22	9	5	8	340
K. 금융 및 보험업	387	213	61	13	20	56	750
L. 부동산업 및 임대업	1,042	282	111	24	15	11	1,485
M. 사업 서비스업	5,574	2,091	585	117	75	37	8,479
O. 교육 서비스업	373	181	13	1	1	0	569
P. 보건 및 사회복지사업	24	12	3	0	0	0	39
Q. 오락, 문화 및 운동관련업	477	312	78	7	4	5	883
R. 기타, 개인 서비스업	1,546	519	43	4	2	2	2,116
전 체	47,036	18,302	4,838	743	512	359	71,790

(자료 : 통계청, 「기업체 실태현황」, 2003년 12월 현재)

<표 Ⅱ-1-10>은 2003년 기준의 통계청 기업체 현황 자료를 분석한 결과이다. 우리나라에서 상용근로자 10인 이상을 고용하고 있는 기업체는 모두 71,790 개소로 나타났다. 전체 기업체 중 47,036개소는 상용근로자 10～29명을 고용하고 있는 소규모 기업체에 속하는 것으로 나타났다.

<표 Ⅱ-1-11> 지역별 기업체 규모별 기업체 현황

(단위 : 개)

규모 / 지역	10～29명	30～99명	100～299명	300～499명	500～999명	1,000명 이상	합 계
서 울	15,922	5,310	1,675	347	279	252	23,785
부 산	2,913	1,141	362	56	26	5	4,503
대 구	1,582	794	200	23	9	2	2,610
인 천	2,425	1,085	266	28	23	8	3,835
광 주	1,268	404	78	9	5	6	1,770
대 전	932	346	76	11	9	4	1,378
울 산	712	349	91	19	6	4	1,181
경 기	10,265	4,257	945	126	68	42	15,703
강 원	1,030	321	48	10	6	2	1,417
충 북	1,380	615	135	14	16	0	2,160
충 남	1,444	645	182	26	10	6	2,313
전 북	1,100	361	96	11	7	2	1,577
전 남	1,182	369	78	4	3	3	1,639
경 북	1,921	911	275	25	18	9	3,159
경 남	2,551	1,286	308	32	26	14	4,217
제 주	409	108	23	2	1	0	543
전 체	47,036	18,302	4,838	743	512	359	7,1790

(자료 : 통계청, 「기업체 실태현황」, 2003년 12월 말 기준)

한편, 상용근로자 1,000명 이상을 고용하고 있는 대규모 기업체에 해당하는 경우는 359개소인 것으로 나타났다. 산업대분류별 현황을 보면 제조업에 속한 기업체가 31,561개소로 가장 많은 것으로 나타났고, Q. 보건 및 사회복지 업에 속한 기업체는 39개소로 나타났다. 이와 같이 산업대분류 구분에 따라서 기업체 수는 그 편차가 대단히 큰 것으로 나타났다. 한편, <표 Ⅱ-1-11>은 지역 및 규모 구분에 따른 기업체 현황을 정리한 것이다. 전체적으로 지역에 따라서 기업체 수에 편차가 큰 것으로 나타났다.

<표 Ⅱ-1-12>는 상용근로자 1,000명 이상인 기업체에 대해서 기업체 규모를 세분하

여 기업체 현황을 정리한 것이다. 상용근로자 3,000명 이상 종사하는 기업체 수는 전체 97개 기업체인 것으로 나타났다. 상용근로자 3,000명 이상 종사하는 기업체의 대부분은 제조업인 것으로 나타났다.

<표 II-1-12> 상용근로자 1,000명이상 산업대분류별 기업체 규모별 현황

(단위 : 개)

규모 / 대분류	1,000~2,999명	3,000~4,999명	5,000~10,000명	10,000명 이상	합계
C. 광업	1	0	0	0	1
D. 제조업	125	30	13	10	178
E. 전기, 가스 및 수도사업	3	1	0	0	4
F. 건설업	8	2	0	0	10
G. 도매 및 소매업	16	6	2	1	25
H. 숙박 및 음식점업	11	1	0	0	12
I. 운수업	6	2	1	1	10
J. 통신업	4	2	1	1	8
K. 금융 및 보험업	37	9	8	2	56
L. 부동산업 및 임대업	11	0	0	0	11
M. 사업 서비스업	34	2	1	0	37
O. 교육 서비스업	0	0	0	0	0
P. 보건 및 사회복지사업	0	0	0	0	0
Q. 오락, 문화 및 운동관련업	4	1	0	0	5
R. 기타 , 개인 서비스업	2	0	0	0	2
전 체	262	56	26	15	359

(자료 : 통계청, 「기업체 실태현황」, 2003년 12월 말 기준.)

한편, 기업체노동비용조사에서 주요 조사항목은 전체 근로자에 대한 내용이 아니라 상용근로자에 대한 것인데, 상용근로자는 다른 노동통계조사와 마찬가지의 정의를 사용하고 있다. 상용근로자는 다음의 네 가지 경우 중에서 하나에 해당하는 자를 말한다.

- ▸ 기간을 정하지 않거나 1개월을 초과하는 기간을 정하여 고용된 자
- ▸ 임시 또는 일용근로자로서 조사기준 이전 3개월을 통산하여 45일 이상 고용된 자
- ▸ 중역, 이사 등의 임원으로서 기업에서 일정한 직무에 종사하고 임원보수 이외에 일반 근로자와 동일한 급여규칙 또는 동일 기준으로 매월급여가 산정되고 있는

자

▸ 사업주의 가족으로서 상시 근무하고 일반근로자와 동일한 급여규칙 또는 동일 기준으로 매월급여가 산정되고 있는 자

다음 <표 Ⅱ-1-13>은 산업대분류별 기업체 규모별 상용근로자 현황을 정리한 것이다. 기업체노동비용조사의 전체 조사대상 기업체에 종사하고 있는 상용근로자 총수는 4,246,081명인 것으로 나타났다. 전체적인 상용근로자 수 분포 현황을 보면 기업체 분포 현황과 유사한 것으로 나타났다.

<표 Ⅱ-1-13> 산업대분류별 기업체 규모별 상용근로자 수

(단위 : 개)

규모 / 대분류	10~29명	30~99명	100~299명	300~499명	500~999명	1,000명이상	합 계
C. 광업	3,875	2,927	1,479	1,127	0	1,204	1,0612
D. 제조업	30,9124	506,943	406,542	149,471	192,555	657,375	2,222,010
E. 전기, 가스 및 수도사업	579	1,572	2,342	1,955	1,492	7,347	15,287
F. 건설업	12,2746	59,787	25,917	9,661	12,712	19,100	249,923
G. 도매 및 소매업	12,1139	74,923	40,992	13,229	18,832	76,833	345,948
H. 숙박 및 음식점업	7,994	13,056	13,784	7,208	12,425	23,526	77,993
I. 운수업	38,282	81,888	140,808	31,344	32,260	34,792	359,374
J. 통신업	3,473	4,261	3,428	3,583	3,706	50,086	68,537
K. 금융 및 보험업	6,415	10,884	10,960	5,108	13,939	177,133	224,439
L. 부동산업 및 임대업	15,958	14,046	18,239	8,867	10,211	17,609	84,930
M. 사업 서비스업	89,020	104,561	95,524	45,102	51,039	66,528	451,774
O. 교육 서비스업	6,641	8,394	1,766	394	731	0	17,926
P. 보건 및 사회 복지사업	365	674	529	0	0	0	1,568
Q. 오락, 문화 및 운동관련업	7,876	17,128	11,242	2,452	2,925	10,989	52,612
R. 기타 , 개인 서비스업	26,239	23,208	6,791	1,709	1,506	3695	63,148
전 체	759,726	924,252	780,343	281,210	354,333	1,146,217	4,246,081

(자료 : 통계청, 「기업체 실태현황」, 2003년 12월 말 기준)

다. 새로운 표본설계의 모집단

원칙적으로 기업체노동비용조사의 조사대상은 농업, 수렵업, 임업 및 어업 등을 제외한 한국표준산업분류 상의 전 산업(단, 국가 또는 지방행정기관, 군·경찰 및 국·

공립교육기관은 제외)에 대하여 상용근로자 10명 이상을 고용하고 있는 기업체이다. 기업체노동비용조사에서 조사대상을 상용근로자 10명 이상의 기업체로 제한한 것은 소규모 기업체의 경우 임금 관련 세부 기록이 불충분하거나, 무응답 우려가 크기 때문에 조사결과에 편향을 야기할 수 있기 때문이다.

기업체는 동일 자금에 의하여 소유되고 통제되는 제도적 단위 또는 경영단위로서 수입, 지출 및 자금관리에 관한 손익계산서 및 대차대조표와 기타 기록을 유지, 관리하는 단위이다. 이러한 기업체는 하나 이상의 사업체로 구성될 수 있다는 점에서 사업체와 구분된다. 자금의 조달 및 운용, 광고활동 등은 기업체 단위로 이루어지기 때문에 각 산업의 자금 원천 및 용도, 생산자금에 관한 자료 등을 파악하고 상호 비교하는 데 필요한 재무 관련 통계작성에 유용한 통계단위이다. 그러나 하나의 기업체는 통상 다른 산업 활동 분야에 종사하는 여러 개의 사업체로 되어 있으므로 기업체를 산업분류의 적용단위로 활용하기에 곤란한 경우도 있다(김민경 등, 2004).

따라서 원칙적으로 기업체노동비용조사의 조사대상은 개인사업체와 회사법인의 기업체라고 할 수 있다. 그러나 개인사업체는 상용근로자 30인 미만을 고용하고 있는 소규모 사업체에 집중되어 있고, 수입, 지출 및 자금관리에 관한 손익계산서 및 대차대조표와 기타 기록의 관리가 소홀하여 현실적으로 조사하는 데 어려움이 많다. 이와 같은 이유로 새로운 표본설계에서 기업체노동비용조사의 조사대상은 상용근로자 10인 이상을 고용하고 있는 회사법인(주식, 유한, 합자, 합명회사)으로 한다.

<표 Ⅱ-1-14>는 2003년 12월 말 기준의 사업체기초통계조사 결과를 분석한 것이다. 상용근로자 10~29명을 고용하고 있는 전체 사업체 중에서 개인사업체의 비율은 25.3%인 것으로 나타났고, 회사법인은 61.3%로 나타났다. 상용근로자 30명 이상을 고용하고 있는 사업체 중에서 개인사업체의 비율은 9.4%인 것으로 나타났고, 회사법인은 75.3%로 나타났다.

한편, 산업대분류별 사업체 조직형태 현황을 살펴보면 다른 산업대분류에 비해서 'O. 교육서비스업'과 'P. 보건 및 사회복지사업'의 경우에는 회사법인은 거의 없고, 회사법인 이외의 법인(학교법인, 의료법인, 종교법인, 특수법인 등)과 개인사업체가 거의 대부분을 차지하고 있다. 이러한 이유로 새로운 표본설계에서 기업체노동비용조사의 조사대상은 농업, 수렵업, 임업 및 어업 등과 교육서비스업, 보건 및 사회복

지사업을 제외한 한국표준산업분류 상의 전 산업(단, 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관은 제외)에 대하여 상용근로자 10인 이상을 고용하고 있는 회사법인(주식, 유한, 합자, 합명회사)으로 하였다.

<표 Ⅱ-1-14> 산업대분류별 사업체 조직형태별 사업체 수 현황

(단위 : 개)

구분	상용근로자 10~29명 사업체				상용근로자 30명 이상 사업체			
	개인 사업체	회사법인	회사법인 이외	합계	개인 사업체	회사법인	회사법인 이외	합 계
C	51	275	0	326	12	83	2	97
	15.6%	84.4%	0.0%	100.0%	12.4%	85.6%	2.1%	100.0%
D	13077	20531	195	33803	2044	14843	132	17019
	38.7%	60.7%	0.6%	100.0%	12.0%	87.2%	0.8%	100.0%
E	85	66	79	230	1	100	267	368
	37.0%	28.7%	34.3%	100.0%	0.3%	27.2%	72.6%	100.0%
F	619	8444	181	9244	101	1455	42	1598
	6.7%	91.3%	2.0%	100.0%	6.3%	91.1%	2.6%	100.0%
G	1596	10911	341	12848	63	2469	91	2623
	12.4%	84.9%	2.7%	100.0%	2.4%	94.1%	3.5%	100.0%
H	2176	1014	93	3283	149	563	39	751
	66.3%	30.9%	2.8%	100.0%	19.8%	75.0%	5.2%	100.0%
I	490	3006	475	3971	66	2791	137	2994
	12.3%	75.7%	12.0%	100.0%	2.2%	93.2%	4.6%	100.0%
J	98	491	30	619	2	499	7	508
	15.8%	79.3%	4.8%	100.0%	0.4%	98.2%	1.4%	100.0%
K	18	5903	3650	9571	2	790	817	1609
	0.2%	61.7%	38.1%	100.0%	0.1%	49.1%	50.8%	100.0%
L	218	2274	2810	5302	21	566	590	1177
	4.1%	42.9%	53.0%	100.0%	1.8%	48.1%	50.1%	100.0%
M	1039	6629	590	8258	266	3277	339	3882
	12.6%	80.3%	7.1%	100.0%	6.9%	84.4%	8.7%	100.0%
O	1980	715	1739	4434	196	318	1857	2371
	44.7%	16.1%	39.2%	100.0%	8.3%	13.4%	78.3%	100.0%
P	2565	20	1766	4351	559	6	1001	1566
	59.0%	0.5%	40.6%	100.0%	35.7%	0.4%	63.9%	100.0%
Q	230	623	224	1077	26	440	132	598
	21.4%	57.8%	20.8%	100.0%	4.3%	73.6%	22.1%	100.0%
R	1695	1883	1476	5054	115	677	293	1085
	33.5%	37.3%	29.2%	100.0%	10.6%	62.4%	27.0%	100.0%
합 계	25937	62916	13825	102678	3623	28925	5852	38400
	25.3%	61.3%	13.5%	100.0%	9.4%	75.3%	15.2%	100.0%

(자료 : 통계청, 「사업체기초통계조사 결과」, 2003년 12월 말 기준)

따라서 기업체노동비용조사의 새로운 표본설계에서 조사모집단은 <표 Ⅱ-1-10>에 제시된 전체 71,790개 회사법인 중에서 'O. 교육서비스업'(569개소)과 'P. 보건 및 사회복지사업'(39개소)의 회사법인 608개소를 제외한 71,182개 기업체이다.

(6) 새로운 표본설계

가. 새로운 표본설계의 기본원칙

① 현재 기업체노동비용조사 결과에 대한 통계는 크게 산업중분류별 통계와 산업대분류·기업체 규모별 노동비용 통계로 구분하여 작성되고 있다. 산업중분류별로 전체 규모와 상용근로자 30인 이상 규모로 구분하여 작성하고 있고, 산업대분류별 통계는 기업체 규모(상용근로자 10-299인, 상용근로자 300인 이상, 1규모(10-29인), 2규모(30-99인), 3규모(100-299인), 4규모(300-499인), 5규모(500-999인), 6규모(1000인 이상))에 따라서 작성되고 있다. 따라서 산업대분류, 기업체 규모별 구분을 부차모집단으로 간주하여 목표오차 관리를 하는 것이 합리적이고, 산업대분류 내의 기업체 수와 상용근로자 수 등을 고려해서 목표오차에 차등을 두고, 산업중분류별 오차 관리를 함께 고려하도록 한다.

② 새로운 표본설계에서 목표오차 관리를 위한 기준변수는 노동비용 총액으로 한다. 기업체노동비용조사에서 중요한 조사항목 중에서 추정량의 상대표준오차 관점에서 보면 평균 간접노동비용에 대한 상대표준오차가 가장 크게 나타나고 있지만, 이 조사에서 가장 중요한 조사항목은 노동비용 총액이기 때문에 새로운 표본설계의 목표오차 관리를 위한 기준변수로 노동비용 총액을 이용하는 것이 합리적이다.

③ 새로운 표본설계는 산업중분류, 기업체 규모, 지역구분(서울, 지방) 등을 고려하여 층화한다. 다만, 1,000인 이상을 고용하고 있는 대규모 기업체들을 전수층으로 설정하는 방안을 검토한다.

나. 층화, 표본크기 및 배정

① 층화

이 조사의 대상은 농업, 수렵업, 임업, 어업 등을 제외한 한국표준산업분류 상의 전 산업(단, 교육서비스업, 보건 및 사회복지사업, 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관은 제외)에 대하여 상용근로자 10인 이상을 고용하고 있는 회사법인으로 한다.

상용근로자 10인 이상 전체 기업체를 고용 규모별로 6개의 군으로 계층화하여 이를 다시 산업중분류와 지역 구분에 따라 층화한다. 기업체의 규모는 1규모(10-29인), 2규모(30-99인), 3규모(100-299인), 4규모(300-499인), 5규모(500-999인), 6규모(1000인 이상) 등으로 구분한다. 지역 구분은 서울과 지방으로 구분한다. 다만, 1,000인 이상을 고용하고 있는 대규모 기업체들을 전수층으로 설정하는 방안을 검토한다.

② 표본크기

본 연구에서는 2003년도와 2004년도 기업체노동비용조사 데이터의 표본오차를 분석하여 표본의 크기를 결정하기 위한 기초자료로 이용하였다. 새로운 표본설계의 목표오차는 특수 산업을 제외하고 산업대분류-규모별로 노동비용총액의 대한 허용오차를 5% 이내로 하였다. 층별 표본크기는 목표정도를 유지하면서 지방사무소의 조사능력을 감안하여 결정하였다. 본 연구에서는 다음에 제시되는 결과와 같이 각 통계작성 단위별 목표오차 크기에 따라서 네 가지 방안을 고려하였다.

[방안 1 : 3,874개 기업체]

구분	전 규모	10~29명	30~99명	100~299명	300~499명	500~999명	1000명 이상
D.제조업	2%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%
F. 건설업, G. 도소매업, I. 운수업, M. 사업서비스, R. 기타, 개인 서비스업	3%	4%	4%	4%	4%	4%	4%
산업대분류(나머지)	4%	5%	5%	5%	5%	5%	5%
산업중분류	5%	-	-	-	-	-	-

[방안 2 : 4,040개 기업체]

구분	전 규모	10~29명	30~99명	100~299명	300~499명	500~999명	1000명 이상
D.제조업	2%	2.5%	2.5%	2.5%	2.5%	2.5%	전수
F. 건설업, G. 도소매업, I. 운수업, M. 사업서비스, R. 기타, 개인 서비스업	3%	4%	4%	4%	4%	4%	전수
산업대분류(나머지)	4%	5%	5%	5%	5%	5%	전수
산업중분류	5%	-	-	-	-	-	-

[방안 3 : 3,347개 기업체]

구분	전 규모	10~29명	30~99명	100~299명	300~499명	500~999명	1000명 이상
D.제조업	2%	2.5%	2.5%	2.5%	2.5%	2.5%	2.5%
산업대분류(나머지)	3%	5%	5%	5%	5%	5%	5%
산업중분류	5%	-	-	-	-	-	-

[방안 4 : 3,536개 기업체]

구분	전 규모	10~29명	30~99명	100~299명	300~499명	500~999명	1000명 이상
D.제조업	2%	2.5%	2.5%	2.5%	2.5%	2.5%	전수
산업대분류(나머지)	3%	5%	5%	5%	5%	5%	전수
산업중분류	5%	-	-	-	-	-	-

최종 표본기업체 수는 노동부와 협의과정을 거쳐서 현장 조사능력을 고려하여 방안 4인 3,536개 기업체로 하였다. 산업중분류 내의 각 규모에 대한 최소 표본수를 5로 하였고, 만약 산업중분류 내의 기업체 수가 5개 미만인 경우에는 전수조사하여 추정의 정도(精度)를 높였다. 다음 <표 Ⅱ-1-15>는 새로운 표본설계에 대한 표본기업체들의 산업대분류 및 기업체 규모별 현황을 정리한 것이다.

표본배정 방법은 추정의 정도를 높이는 여러 가지 절충배정 방안을 검토하여 결정되었다. 본 연구에서 산업대분류·기업체 규모 구분 내의 산업중분류·규모별 표본크기는 멱배정(power allocation)의 일종인 제곱근 비례배정법에 따라서 결정되었다. 이렇게 절충배정법을 사용한 것은 산업대분류·기업체 규모 구분에서 통계작성뿐만 아니라 산업중분류별 통계작성도 요구되기 때문에 산업중분류별 노동비용 통계의 안정적인 생산을 위한 조치이다.

<표 II-1-15> 방안 4에 대한 산업대분류별 기업체 규모별 표본 기업체 수 현황

(단위: 개)

규모 / 산업대분류	10~29명	30~99명	100~299명	200~499명	500~999명	1,000명이상	합 계
C. 광업	23	21	9	3	0	1	57
D. 제조업	229	213	297	111	117	178	1,145
E. 전기, 가스 및 수도사업	19	17	11	5	2	4	58
F. 건설업	32	32	26	19	14	10	133
G. 도매 및 소매업	62	54	43	23	19	25	226
H. 숙박 및 음식점업	48	44	29	15	13	12	161
I. 운수업	123	60	34	30	21	10	278
J. 통신업	42	31	15	8	5	8	109
K. 금융 및 보험업	51	90	35	12	15	56	259
L. 부동산업 및 임대업	67	115	33	16	12	11	254
M. 사업 서비스업	67	165	100	54	37	37	460
Q. 오락, 문화 및 운동관련산업	34	105	40	7	4	5	195
R. 기타 , 개인 서비스업	40	127	26	4	2	2	201
전체	837	1,074	698	307	261	359	3,536

③ 표본배정 및 표본추출

새로운 표본설계에서는 산업중분류, 기업체 규모, 지역 구분 등을 층화 기준으로 사용하였다. 각 층에서 표본추출은 각 규모에 배정된 표본크기만큼을 계통추출법으로 추출하였다. 이를 위해서 각 산업중분류별 기업체 규모별 기업체 리스트를 작성하여 이를 행정구역에 따라 정렬한 후 계통추출법을 적용하였다. 또한 표본교체가 필요한 경우를 대비해서 예비표본을 추출하여 표본 리스트와 함께 제공하였다.

(7) 추정

가. 가중치 산정 방법

새로운 표본설계는 산업중분류, 기업체 규모, 지역 구분(서울, 지방)을 층화변수로 이용하였다. 일반적으로 기업체노동비용조사와 같은 복합표본조사(complex sample survey)의 가중치는 ㉠ 설계가중치, ㉡ 무응답에 대한 조정, ㉢ 사후층화에 대한 조정 등의 세 가지 요인을 통합하여 산정된다.

기업체노동비용조사의 가중치는 산업중분류 내의 6개 기업체 규모 및 2개 지역 구

분에서 상용근로자 수와 응답 기업의 조사 근로자 수의 복원배율로 계산한다. 이는 설계가중치에 무응답에 대한 조정과 사후 층화에 대한 조정을 층 내에서 근로자수를 이용한 비조정(Ratio adjustment)을 사용한 개별 비추정(Separate Ratio Estimation)의 전형적인 추정 방법이다. 이러한 개별 비추정법은 층 내에서 각 기업체의 총 노동비용의 기대값과 분산이 상용근로자 수에 비례하는 초모집단 모형에서 그 기대분산을 최소화하는 추정법으로 많이 사용된다.

일반적으로 추정단계에서 가중치를 이용하면 모집단에 대한 특성치인 모수에 대한 비편향추정량(unbiased estimator)을 얻을 수 있다. 만약 통계분석 과정에서 가중치를 무시하고 분석한 추정치는 심각한 편향(bias)이 발생할 수 있다. 표본의 크기가 큰 대규모 조사에서 문제가 되는 것은 추정량의 편향이기 때문에 추정과정에서 반드시 가중치를 이용해야 한다.

나. 각 분류별 각종 평균치에 대한 추정

기업체 노동비용조사에서 중요한 추정모수는 근로자 일인당 월평균 노동비용이다. 근로자 일인당 월평균 노동비용의 모수는 다음과 같이 나타낼 수 있다.

$$\Theta = \frac{\sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i} \quad : \text{근로자 일인당 월평균 노동 비용}$$

여기서, x_i 는 기업체 내의 월급여 지급인원(12개월 합계)이고, y_i는 각종 급여의 12개월 합계를 나타낸다.

근로자 일인당 월평균 노동비용의 모수는 다음과 같은 비추정량을 이용하여 추정할 수 있다.

$$\hat{\Theta} = \frac{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi} y_{hi}}{\sum_{h=1}^{L}\sum_{i=1}^{n_h} w_{hi} x_{hi}} \tag{1}$$

여기서, n_h는 층 h의 표본 기업체 수를 나타내고, w_{hi}는 각 표본 기업체에 부여된 가중치이고, y_{hi}는 각 표본 기업체에서 얻은 변수값 (직접노동비용, 간접노동비용,

노동비용 총액 등)이며, x_{hi}는 표본기업체의 1년간 월급여 지급인원이다.

가중치 w_{hi}을 산출하는 방법으로는 크게 두 가지를 생각해 볼 수 있다. 첫 번째는 h층 내의 모집단 전체 기업수를 그 층의 표본 기업수로 나누어 주는 방법이고, 두 번째는 h층 내의 모집단 전체 기업 상용근로자 수를 그 층의 표본 기업체의 상용근로자 수로 나누어 준 것으로 사용한다. 첫 번째 방법은 일종의 Horvitz-Thompson 추정의 형태이고 두 번째 방법은 일종의 일반화 회귀 추정량(Generalized Regression Estimator ; GREG 추정량)의 특수한 경우인 개별 비추정량(separate ratio estimator)의 형태이다. 본 연구에서는 두 번째 방법으로 가중치를 사용하는 것을 추천하는데 그 이유로는 이 두 번째 방법에서 제시된 가중치가 추정량의 변동이 적어 보다 안정적인 통계를 낼 수가 있게 되기 때문이다. 실제로 한 기업체에서 지불한 총 노동 비용은 그 기업체의 상용근로자 수에 비례할 것이므로 상용근로자 수의 비를 사용한 가중치가 더 효율적인 추정을 만들어 낼 것이다. 또한 이러한 가중치를 사용하면 층 내 가중치의 합계가 해당 층의 모집단 총 근로자 수가 나온다는 장점이 있다.

이렇게 얻어진 가중치는 무응답이 있는 경우에도 그대로 적용할 수 있다. 예를 들어 각 층에서 n_h개의 표본 기업체 중에 r_h개의 기업체만이 응답한다고 하면 (1)의 추정량은 다음과 같이 계산이 된다.

$$\hat{\Theta}^* = \frac{\sum_{h=1}^{H}\sum_{i=1}^{r_h} w_{hi}^* y_{hi}}{\sum_{h=1}^{H}\sum_{i=1}^{r_h} w_{hi}^* x_{hi}} \qquad (2)$$

이때 w^*_{hi}는 응답 기업체에게 최종적으로 부여되는 가중치로써 h층 내의 모집단 전체 기업 상용근로자 수를 그 층의 응답 기업 상용근로자 수로 나누어 준 것으로 사용한다.

다. 분산 추정

앞서 제시한 월 평균 노동비용 총액 등의 비추정량에 대한 분산 추정값은 다음과

같이 계산된다.

- $\widehat{V}(\hat{\Theta}) = \sum_{h=1}^{H} \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} e_{hi}^2$
- $e_{hi} = w_{hi}(y_{hi} - \hat{\Theta}_h x_{hi}) / w_{..}$
- $\hat{\Theta}_h = \frac{\sum_{i=1}^{n_h} y_{hi}}{\sum_{i=1}^{n_h} x_{hi}}$
- $w_{..} = \sum_{h=1}^{H} \sum_{i=1}^{n_h} w_{hi}$
- $f_h = n_h / N_h$

또한 표준오차 추정량과 상대표준오차 추정량은 각각 다음과 같다.

- $\widehat{SE}(\hat{\Theta}) = \sqrt{\widehat{V}(\hat{\Theta})}$: 표준오차 추정량
- $\widehat{RSE}(\hat{\Theta}) = \frac{\widehat{SE}(\hat{\Theta})}{\hat{\Theta}} \times 100$: 상대표준오차 추정량

무응답이 있어서 (2)의 추정량을 사용한 경우에는 분산추정량은 다음과 같다.

$$\widehat{V}(\hat{\Theta}^*) = \sum_{h=1}^{H} \frac{r_h(1-f_h^*)}{r_h-1} \sum_{i=1}^{r_h} (e_{hi}^*)^2$$

이 때, $e_{hi}^* = w_{hi}^*(y_{hi} - \hat{\Theta}_h x_{hi}) / w_{..}$이고 $f_h^* = r_h / N_h$이다.

(8) 모집단 및 표본의 관리

2005년부터 사용될 『기업체 노동비용조사』의 새로운 표본설계에서는 모집단과 표본의 규모가 확대되고 이에 따라 조사내용도 많아져서 현행조사보다 많은 노력이

필요할 것이다.

기업체 노동 비용 조사의 조사모집단으로 사용되는 2003년 기준 통계청 기업체 모집단 자료 중에서 농업, 수렵업, 임업 및 어업 부분을 제외한 한국표준산업분류 상의 전 산업(단, 교육서비스업, 보건 및 사회복지사업, 국가 또는 지방행정기관, 군·경찰 및 국·공립교육기관은 제외)에서 상용근로자 10인 이상을 고용하고 있는 회사법인은 모두 71,182개소로 나타났다. 이중에서 상용근로자 10-29인 기업체가 46,535개소로 전체 모집단의 65.4%를 차지하는 것으로 나타났다. 이렇게 소규모 기업체가 차지하는 비중이 매우 높은데, 소규모 기업체는 특성상 창업, 폐업, 휴업, 그리고 전업이 경기변동에 민감하므로 정확한 『기업체 노동비용조사』를 위해서는 모집단과 표본의 변화에 따른 관리가 매우 중요하다.

가. 모집단 관리

기업체노동비용조사의 새로운 표본설계에서 고려되는 모집단은 표본설계 당시 시점에서의 모집단을 대상으로 하지만 실제 조사는 매년 정기적으로 조사하기 때문에 조사시점의 모집단이 표본 설계 당시 시점에서의 모집단과는 다른 분포를 가질 것이다. 또한 표본설계에 사용되는 모집단 프레임은 2003년 자료를 바탕으로 만들어진 것이므로 실제 조사시점과 모집단 시점 간에 2년 이상의 시차가 존재하게 된다. 그러므로 모집단 기업체가 조사시점에 다른 산업 또는 다른 규모의 기업체로 바뀌는 경우도 있고 심지어는 휴업이나 폐업으로 기업체가 존재하지 않는 경우도 상당수 예상할 수 있다. 마찬가지로 모집단에는 없으나 그 후 기업체가 창업되어 실제 조사에 누락되는 경우도 발생한다. 이와 같이 변화하는 모집단의 특성이 조사시점에 충분히 반영되어야 한다.

『기업체노동비용조사』에서 모집단 관리라 함은 조사대상 모집단 기업체들의 변동을 표본이 정확히 반영할 수 있도록 하는데 발생되는 여러 가지 문제점들을 지칭한다. 이론적으로는 추출된 표본 기업체는 모집단 기업체들의 변동을 반영하나 실제적으로는 다음과 같은 이유로 차이가 있을 것이다.

(가) 모집단 프레임이 실제 모집단을 정확히 반영하지 못하는 경우(모집단 프레임

이 정확하지 못하여 모집단의 실제 기업체를 다 수록하지 못하는 경우)

(나) 모집단 프레임이 작성된 시점과 조사 당시의 시점에 차이가 있어 추출된 표본이 모집단의 변화를 반영하지 못하는 경우

(다) 표본설계 당시에는 표본이 모집단을 반영하였으나 특정 산업이 번창, 또는 쇠퇴하여 해당 층의 모집단 기업체 수가 변하는 경우

(가)의 경우는 일종의 포괄범위오차(coverage error)의 전형적인 경우로 현행의 단일 추출틀의 문제를 보완해 줄 수 있는 또 다른 추출틀의 사용을 검토해 볼 필요가 있다. 예를 들어 list frame 외에 area frame 등을 보완하여 사용한다든지 또는 다른 출처(source, 소스)로부터 프레임을 구한다든지(예를 들어 사업자 등록 리스트나 국세청 자료 등을 사용하는 것 등) 하여 추출틀의 완벽성을 기하려는 노력이 더욱 필요하다.

(나)의 경우 또한 포괄범위오차(coverage error)로 볼 수 있는데 그 원인은 (가)의 경우와는 조금 다르다. 이 경우는 모집단 자체의 변동 때문으로 인한 시차가 원인이 될 것이다. 특히 소규모 기업체는 규모가 큰 기업체에 비해 근로자 수가 적고 자본규모가 영세하여 경기 변동에 민감하고 폐업, 휴업, 전업 및 창업 등이 빈번하게 발생하여 모집단 변동이 심하다. 따라서 변화하는 모집단에 대한 관리가 요구된다. 새로운 표본설계에 사용되는 최신의 모집단이 2003년도 말 기준의 자료가 되므로 실제 조사시점과 모집단 시점 간에 2년 이상의 시차가 존재하게 된다. 그러므로 모집단 기업체가 조사시점에 다른 산업 또는 다른 규모의 기업체로 바뀌는 경우도 있고 심지어는 휴업이나 폐업으로 기업체가 존재하지 않는 경우도 상당수 예상할 수 있다. 마찬가지로 모집단에는 없으나 그 후 기업체가 창업되어 실제 조사에 누락되는 경우도 발생한다. 이와 같이 변화하는 모집단의 특성이 조사시점에 충분히 반영되어야 한다. 이를 위해서는 매년 새로이 창업하는 기업체를 모집단에 반영하여 표본기업체를 추가로 선정하고, 폐업, 휴업 또는 전업한 표본기업체를 다른 기업체로 교체하는 등의 방법을 사용할 수 있다.

(다)의 경우는 (나)의 경우와 원인은 같으나 그 처방이 다르다. 기업체노동비용조사는 각 규모별 업종별 평균 노동 비용에 관심이 있을 뿐 모집단 기업체 수의 분포에

는 크게 관심이 없다. 따라서 이러한 경우 원칙적으로는 모집단 기업체 수의 변화가 층별 평균 노동 비용 추정 결과에 크게 영향을 미치지는 않는다. 다만 해당 층에 새로 진입하는 기업체가 표본에 선정되지 않음으로써 생기는 편향이 존재할 가능성 때문에 (나)에서 설명한 표본 추가가 필요할 것이다. 또한 층별이 아닌 전체 단위의 추정에서 가중치가 달라져야 할 것이다. 이러한 가중치의 변화는 표본 자체에서는 얻어낼 수가 없고 다른 행정 자료나 통계청 자료로부터 얻어내어 계산해주어야 할 것이다.

나. 표본 관리

새로운 『기업체 노동비용조사』는 모집단에서 추출된 3,536 개의 표본기업체를 대상으로 매년 실시될 예정이다. 표본기업체를 대상으로 매년 조사를 하다 보면 표본기업체에 변동이 생긴다. 이는 모집단 기업체가 변동함으로써 발생하게 되므로 모집단의 변동에 따라 표본 기업체의 관리가 함께 요구된다. 앞에서 언급한 바와 같이 규모가 작은 기업체가 모집단에서 큰 비중을 차지하고 있고, 이들 소규모 기업체는 경기변동에 민감하게 반응하여 변동이 심하게 되므로 이들을 고려하지 않고는 정확한 조사결과를 기대하기 어렵다. 따라서 매월 조사에서 표본기업체의 변동 상황을 파악하여 폐업, 휴업, 전업 등이 발생한 기업체를 업종별, 규모별, 그리고 지역별로 분류하여 체계적으로 관리한다면 향후 표본관리에 유용하게 활용할 수 있게 된다. 또한 새로이 창업하는 기업체에 대해서는 별도로 모집단을 설정해서 이들로부터 추가 표본을 얻게 한다. 이와 같이 모집단의 변화를 파악해서 이를 바탕으로 표본의 크기를 조절해 주어야 한다.

새로운 표본설계 시마다 제시한 표본 기업체들에 대해서 직접 실사를 실시해서 기업체의 조사 불응 비율, 표본 기업체의 소재 파악 불능 비율, 폐업 및 전업 기업체 비율 등에 대한 자료를 확보할 수 있도록 표본관리를 철저하게 시행해야 한다.

① 창업에 따른 표본관리

새롭게 창업된 기업체들을 조사에 반영하는 것은 상당한 노력이 필요한 작업이다. 모집단 관리에서 다룬 것처럼 기업체노동비용조사에서는 모집단 기업체 분포의 변화

를 중요시하는 조사가 아니고 상용근로자 10인 미만의 기업체를 모집단에 포함시키지 않고 있으므로 표본 추출한 후 첫 2-3년간은 창업에 따른 표본 관리를 매 조사시마다 표본 추가하는 작업에 집중하기 보다는 기존 표본이 얼마나 대표성을 가지는가에 대해 더 주의를 기울어야 할 것이다. 예를 들어 업종별로 비교해 보았을 때 노동 비용의 변동이 심한 업종에 대해서만 표본 추가를 실시한다든지 하는 작업은 제한된 조사 업무량을 고려했을 때 현실적인 대안이 될 것이다. 물론 조사가 시작되고 4-5년이 지나면 새로운 프레임을 사용한 새 표본 추출을 고려해야 할 것이다.

② 폐업에 따른 표본관리

우선 모집단 자료에 대해서 산업 대분류 및 기업체 규모별로 폐업하는 기업체의 비율을 정확히 파악할 수 있어야 한다. 이 자료는 기업체의 창업과 폐업에 따른 표본대체의 문제를 연구할 때 중요한 기초 자료가 된다. 새로운 표본설계에 의해서 조사가 시행되는 첫 해에는 표본으로 추출된 기업체들을 방문해서 해당 기업체가 폐업했다면 이를 조사표에 기입하고, 해당 지역 같은 층의 예비표본 중에서 랜덤하게 추출해서 표본 기업체로 선정한다. 만약 표본 기업체 중에서 폐업 비율이 20%를 넘는다면 표본 대체를 하기 전에 모집단과 표본 기업체 자료의 정확성에 대한 검토가 필요하다. 이후 연도부터는 폐업한 기업체에 대해서는 조사원으로 하여금 그 현황을 보고하게 하고, 다른 업체로 대체하지 않고 표본에서 제외한다. 이 경우에도 기업체 중 폐업 비율이 지나치게 높다라면 그 원인에 대한 충분한 검토가 필요하다. 모집단 자료가 정비되어 새롭게 창립된 기업체의 현황을 알 수 있다면 이들 기업체 중에서 표본설계 당시의 추출률에 따라서 표본 기업체를 추가로 선정한다.

③ 기업체 규모변동에 따른 표본관리

표본으로 추출된 기업체에서 규모변동이 일어나는 경우에는 현재 기업체 내의 상용근로자 수에 따라서 기업체 규모를 구분해서 이용하는 것을 원칙으로 한다. 기업체의 규모 변동은 대부분의 기업체 조사(Establishment survey)에서 종종 발생하는 현상으로 흔히 층간 이동(stratum jump)라고도 불린다. 다만 대규모 기업체의 경우 표본 기업체 수가 작기 때문에 특정 표본 기업체의 규모 변동으로 통계의 시계열 유

지가 어려운 경우에는 규모 변동 전의 기업체 규모로 통계를 작성하는 방안을 검토해야 한다.

2. 가구표본설계의 사례

1) 최저생계비 계측조사(보건복지부)

(1) 조사목적

저소득층 가구의 가계수지 및 생활실태를 정확히 파악함으로써 건강하고 문화적인 삶을 영위할 수 있는 최저생계비를 추정하기 위한 기초 자료를 생산함에 그 목적이 있다.

(2) 조사범위 및 대상

① 조사지역 : 전국

② 조사대상 : 가구조사를 기본으로 하되 복지서비스대상에서는 가구원조사도 병행한다.

(3) 조사의 내용

① 최저생계비를 가구원(1인~7인) 규모별로 구분하여 계측하고, 이를 위한 표준가구설정 및 합리적인 가구균등화지수를 산출한다.

② 최저생계비를 가구의 주거점유형태(자가, 전세, 월세 등)와 가구원의 인구경제학적 특성(성, 연령, 장애여부, 질병여부, 학생 등)별로 나누어 계측한다.

③ 최저생계비계측결과를 토대로 차상위 계층의 규모를 파악하고 이의 특성과 생

활실태 및 복지요구를 파악한다.

(4) 표본설계

가. 모집단 분석

현행 표본조사를 위한 주요변수에 대한 이용가능한 모집단 정보가 없기 때문에 이와 유사한 조사로부터 주요변수에 대한 자료를 이용하는 것이 바람직하다. 그러나 2005년도 인구주택 총조사 자료분석 결과를 조사 설계 당시 이용할 수 없고, 2004년도 계측조사에서는 3개의 조사구를 결합하여 조사가 이루어진 관계로 현행 조사 설계와 유사한 모집단 자료를 이용하기 어려운 점이 있었다. 따라서 가구소득과 지출에 어느 정도 관계가 있는 것으로 파악된 주택유형 및 조사구당 가구수 등을 기준으로 모집단 분석을 수행하였다.

<표 II-2-1> 시도별 모집단 규모(90% 조사구)

(단위 : 개)

시 도	조사구수	가구수	시 도	조사구수	가구수
계	224,523	14,298,415			
서 울	49,440	2,978,901	강 원	7,871	468,565
부 산	18,179	1,067,740	충 북	7,499	454,683
대 구	12,128	733,127	충 남	9,891	593,884
인 천	12,814	740,721	전 북	9,296	557,962
광 주	67,92	414,081	전 남	9,887	599,687
대 전	72,68	430,979	경 북	14,025	844,956
울 산	5,053	305,186	경 남	15,932	950,406
경 기	49,804	2,996,259	제 주	2,704	161,279

나. 조사구명부 작성

표본설계 당시 2005년도 인구주택총조사 90% 조사구에 대해 이용 가능한 자료로서는 앞에서 언급한대로 조사구별 가구 수, 조사구 형태, 주택형태 뿐이었기 때문에, 이를 바탕으로 517개 조사의 기초 자료를 집계하여 지역별, 조사구 유형별, 읍면동

별, 주택형태별로 분류하여 분포를 파악하였다.

<표 II-2-2> 지역별 표본조사구 분포

(단위: 개)

시 도	계	일반	아파트	계	구	시	군	동	읍	면
서울	110	68	42	25	25	-	-	110	-	-
부산	43	29	14	15	14	-	1	41	2	-
대구	30	23	7	8	7	-	1	27	3	-
인천	30	22	8	8	7	-	1	27	1	2
광주	16	6	10	5	5	-	-	16	-	-
대전	17	8	9	5	5	-	-	17	-	-
울산	14	9	5	5	4	-	1	10	3	1
경기	91	59	32	25	17	16	2	72	4	15
강원	18	9	9	11	-	5	6	10	5	3
충북	16	11	5	8	-	2	6	8	3	5
충남	21	14	7	10	-	3	7	4	6	11
전북	21	16	5	8	2	2	4	13	3	5
전남	22	16	6	11	-	5	6	7	9	6
경북	30	23	7	17	-	7	10	11	4	15
경남	33	19	14	15	-	8	7	15	6	12
제주	5	3	2	3	-	2	1	4	1	-
전체	517	335	182	179	86	59	53	392	50	75

다. 추출틀

본 조사는 추출틀은 2005년 인구주택총조사 25만 여개의 조사구중 90%조사구인 22만3천여 개의 조사구를 추출틀로 하며, 이때 사용한 자료는 조사구 특성(일반조사구, 아파트조사구), 주택유형(단독, 아파트, 연립 및 다세대), 조사구당 가구 수 등으로 구성된 리스트를 사용하였다.

라. 표본추출방법

2005년도 인구주택 총 조사구 27만여 개의 조사구중 90%조사구인 24만 3천여 개 조사구중 예비조사구를 포함하여 517개 조사구를 지역별 조사구 규모에 따라 층화추출하였으며, 층화의 주요기준변수로는 지역(16), 조사구형태(2), 주택유형(3) 등을

사용하여 총 96개 층으로 나누어 각 층별로 확률 비례 추출하였다.

각 조사구별로 평균 60가구 중에서 조사원의 업무 할당을 고려하여 일률적으로 51가구를 조사하도록 하였으며, 조사 불능 가구 또는 조사대상 제외가구가 발생할 경우 가구명부의 순서에서 바로 다음 가구를 조사하도록 하였다. 조사구의 크기가 51가구 이하가 되는 조사구는 조사구내의 모든 가구를 조사토록 하였으며, 조사 불능이나 제외가구의 발생으로 51가구 이하가 되는 조사구에서는 조사가능 가구만을 표본으로 산정하도록 하였으며, 일부 조사구에 대해서는 예비조사구의 대체 가구로 대체조사 하였다.

<표 II-2-3> 지역별 표본 가구 수

(단위 : 개)

	계			동 부			읍면 부		
	합계	단독	아파트	합계	단독	아파트	합계	단독	아파트
전 국	30,000	18,624	10,824	24,008	13,616	9,839	5,991	5,008	985
서 울	6,568	4,601	1,967	6,568	4,601	1967	0	0	0
부 산	2,457	1,337	1,120	2,457	1,337	1120	0	0	0
대 구	1,587	978	609	1,587	978	609	0	0	0
인 천	1,614	931	683	1,614	931	683	0	0	0
광 주	842	405	437	843	405	437	0	0	0
대 전	851	464	387	851	464	387	0	0	0
울 산	625	343	282	626	343	282	0	0	0
경 기	4,925	2,634	2,291	4,288	1,790	1,969	1,166	844	322
강 원	1,052	717	335	576	298	263	490	419	72
충 북	980	643	337	529	280	268	431	363	69
충 남	1,235	885	350	387	184	195	856	701	155
전 북	1,312	884	428	783	382	402	528	502	26
전 남	1,320	915	405	760	413	379	528	502	26
경 북	1,853	1,348	505	814	452	350	1,051	896	155
경 남	1,944	1,298	646	1,113	625	486	833	673	160
제 주	283	241	42	212	133	42	108	108	0

(주: 소수점처리로 인하여 전체가구 수는 29,448 가구임)

마. 표본규모

표본의 규모는 예비조사구를 포함하여 총 517개 조사구의 약 30,000가구를 표본으로 결정하였다. 이때, 2004년 계측조사와는 다르게 2007년 계측조사에서는 표본으로

선정된 조사구는 인근 조사구를 결합하는 방식이 아닌 1개 조사구씩을 표본으로 추출하였다.

과거의 조사 자료를 근거로 주요 연구변수에 대한 상대오차를 이용하여 새로운 표본규모를 결정하는 방법은 다음과 같은 방법으로 간단히 구할 수 있다.

$$n' = n\left(\frac{CV_1}{CV_2}\right)^2 \quad (1)$$

여기서 n'은 새로운 표본규모이며, n은 기존의 표본규모, CV_1은 기존의 조사로부터 구한 상대표준오차, CV_2는 새로운 표본에 대한 목표 상대 표준오차이다.

<표 II-2-4> 시도별 목표정도

(단위 : 개,%)

시 도	2002년 가계조사			현행 표본설계			
	조사구수	가구수	목표정도	조사구수	목표정도	가구수	목표정도
전 국	697	5,089	1.62	500	1.91	29,448	0.67
서 울	127	995	2.87	110	3.08	6,568	1.12
부 산	70	518	3.45	41	4.51	2,457	1.58
대 구	49	357	5.62	27	7.57	1,587	2.67
인 천	50	377	3.09	27	4.20	1,614	1.49
광 주	52	340	5.07	14	9.77	842	3.22
대 전	50	357	6.17	14	11.66	851	4.00
울 산	23	190	4.11	10	6.23	625	2.27
경 기	50	384	5.73	91	4.25	4,925	1.60
강 원	27	191	7.85	18	9.61	1,052	3.34
충 북	27	193	5.81	16	7.55	980	2.58
충 남	30	184	6.96	21	8.32	1,235	2.69
전 북	29	191	6.31	21	7.42	1,312	2.41
전 남	29	209	5.75	22	6.60	1,320	2.29
경 북	32	230	4.67	30	4.82	1,853	1.65
경 남	34	253	5.57	33	5.65	1,944	2.01
제 주	18	120	5.48	5	10.40	283	3.57

바. 추정방법

① 전국단위의 추정

소득, 지출 등의 주요변수에 대한 전국 단위의 추정치를 계산하기 위한 공식은 다

음과 같이 표본추출과정을 고려하여 계산된다.

$$\bar{y} = \frac{\sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij}} \tag{2}$$

여기서 h는 전국 6개 층으로서 지역별 조사구 유형을 나타낸다. $h=1,2,\cdots,6$. 또는 전국 16개 광역시 및 시도와 동부 및 읍면부를 나타낸다. i는 표본 조사구를 나타내는 첨자로서 $i=1,2,\cdots,n_h$이다. 그리고 j는 표본 조사구 내의 가구를 나타내는 첨자로서 $j=1,\cdots,m_{hi}$이다. w_{hij}는 h지역의 i번째 표본조사구내의 j번째 가구에 부여된 가중치이다. 또한 $W_{hi}=\sum_{j}^{m_{hi}} w_{hij}$으로 h층의 i번째 조사구내의 가구들의 가중치 합을 나타낸다.

표본평균 $\bar{y}$의 분산추정치는 다음과 같이 정의할 수 있다.

$$\widehat{V}(\bar{y}) = \frac{\sum_{h}^{L} \frac{n_h}{n_h-1}(1-f_h)\sum_{i=1}^{n_h}\left[W_{hi}(\bar{y}_{hi}-\bar{y}) - \frac{1}{n_h}\sum_{s=1}^{n_h} W_{hs}(\bar{y}_{hs}-\bar{y})\right]^2}{(\sum_{h}^{L}\sum_{i}^{n_h} W_{hi})^2} \tag{3}$$

여기서 $\bar{y}_{hi} = \frac{\sum_{j}^{m_{hi}} w_{hij} y_{hij}}{\sum_{j}^{m_{hi}} w_{hij}}$ 는 h층의 i번째 조사구내의 가구들의 소득, 또는 지출 등의 주요변수들의 평균이다.

② 지역별(또는 층별) 추정- 지역별, 가구규모별, 가구유형별 소득

만일 층별 또는 부차모집단에 대한 추정치를 얻고자 한다면, 다음과 같은 추정 공식을 사용할 수 있다. 즉, 관심대상이 되는 h번째 부차집단의 평균의 추정치는

$$\bar{y}_h = \frac{\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij} y_{hij}}{\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij}} \tag{4}$$

으로서 만일 h층의 i번째 표본조사구의 j번째 가구가 어떤 특성을 가지면, $y_{hij}=1$, 그 외에는 0이라고 할 때, 모비율의 추정치로 사용할 수도 있다.

층별 표본평균에 대한 분산추정량은 다음과 같이 정의된다.

$$\hat{V}(\bar{y}_h) = \frac{\frac{n_h}{n_h-1}(1-f_h)\sum_{i=1}^{n_h}\left[W_{hi}(\bar{y}_{hi}-\bar{y}_h) - \frac{1}{n_h}\sum_{s=1}^{n_h} W_{hs}(\bar{y}_{hs}-\bar{y}_h)\right]^2}{(\sum_{i}^{n_h} W_{hi})^2} \tag{5}$$

③ 가구유형별/점유형태별 추정량

특정 층에 속한 비율을 추정하고자 할 때에는 식(3)을 변형한 다음과 같은 추정산식을 이용하면 된다. 이에 해당되는 추정대상은 주로 노인가구, 장애인가구 등의 비율 등이 될 수 있다.

$$\bar{y}_G = \frac{\sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij} y_{hij} I[hij \in G]}{\sum_{h}^{L}\sum_{i}^{n_h}\sum_{j}^{m_{hi}} w_{hij} I[hij \in G]} \tag{6}$$

여기서 G는 1인 가구, 2인 가구, 노인가구, 장애인가구 등이 될 수 있으며, 또한 $I[hij \in G]$는 h층의 i번째 조사구내의 j번째 가구가 어떤 특성을 가지면 1 그렇지 않으면 0을 갖는 지시함수이다.

식(6)에 대한 분산추정량은 다음과 같다.

$$\hat{V}(\bar{y}_G) = \frac{\sum_{h}^{L}\frac{n_h}{n_h-1}(1-f_h)\sum_{i=1}^{n_h}\left[W_{hiG}(\bar{y}_{hiG}-\bar{y}_G) - \frac{1}{n_h}\sum_{s=1}^{n_h} W_{hsG}(\bar{y}_{hsG}-\bar{y}_G)\right]^2}{(\sum_{h}^{L}\sum_{i}^{n_h} W_{hiG})^2} \tag{7}$$

여기서 가중치의 합은 $W_{hiG} = \sum_{j}^{m_{hi}} w_{hij} I[hij \in G]$ 이고, 그룹별 평균은 다음과 같다.

$$\bar{y}_{hiG} = \frac{\sum_{j}^{m_{hi}} w_{hij} y_{hij} I[hij \in G]}{\sum_{j}^{m_{hi}} w_{hij} I[hij \in G]} \tag{8}$$

만일 조사무응답이 존재하는 경우에는 무응답을 조정한 가중치를 $w^{*}_{hij} = w_{hij} \times r_{hij}$라 하면 이 값을 위의 추정량의 w_{hij}대신 대입하여 무응답 조정이 된 추정량을 구할 수 있다. 여기서 r_{hij}는 h층의 i번째 조사구의 j번째 가구의 응답률이다. 이때, 개별 가구의 응답률보다는 각 층내의 조사구별 응답률을 이용하여 무응답에 대한 가중치 조정이 가능하다.

$$r_{hi} = \frac{h\text{층의 } i\text{번째 조사구에서 응답한 총 가구 수}}{h\text{층의 } i\text{번째 조사구내의 총 가구 수}} \tag{9}$$

2) 생명보험 성향 조사

(1) 조사목적

생명보험협회에서는 3년 주기로 생명보험 수요자의 성향을 조사하여 생명보험시장의 변동상황에 비치는 제 요인을 분석하고 과거의 자료와 비교하여 생명보험에 관한 지식, 태도, 행동, 보험료 등의 변화에 대한 검토 및 분석을 수행한다. 분석결과를 이용하여 향후 수요가 예상되는 보험상품을 개발하고 보험에 대한 일반 국민의 인식을 제고하는데 필요한 기초 자료로 사용된다.

(2) 표본설계

가. 모집단의 정의

전국의 모든 가구가 조사대상이 되어야 하지만, 조사업무의 지도와 조사비용의 제한 등으로 인하여 전국의 50개 도시에 있는 모든 가구를 표본추출대상으로 한다. 즉, 조사모집단은 『1985년도 한국행정편람』에 기재되어 있는 50개 도시의 모든 가구인 5,634,701개 가구이다.

나. 표본크기의 결정

다항목 조사의 경우 표본크기를 결정하는 방법으로는 크게 2가지 방법을 고려할 수 있다. 첫 번째로 조사목적상 가장 중요한 항목에 대해서만 허용오차를 정하고, 허용오차를 만족하는 최소표본크기를 계산하는 방법이다. 두 번째 방법은 모든 조사항목에 대해 허용오차를 정하고 각각의 항목별로 허용오차를 만족하는 최소 표본크기를 계산한 다음, 표본의 크기가 가장 큰 값을 최종 표본크기로 결정하는 방법이다.

본 표본조사설계에서는 앞에서 언급한 방법 중 전자의 방법을 적용하였으나, 생명보험 가입여부를 묻는 항목과 가입금액을 묻는 항목에 대해서는 표본크기를 계산한 후 무응답 비율을 고려하여 표본의 크기를 다음과 같이 계산하였다.

추정하고자 하는 모집단 비율을 p라 하고, 표본으로부터 얻어지는 표본비율을 $\hat{p}$라 하면 추정량 $\hat{p}$의 분산은

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

이며 모비율 p의 $(1-\alpha)100\%$ 신뢰구간은 다음과 같다.

$$\hat{p} \pm z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n}}$$

여기서 $z_{\alpha/2}\sqrt{\left(1 - \frac{n}{N}\right)\frac{p(1-p)}{n}}$는 모비율 p의 추정에 대한 오차의 한계라 한다.

이와 같이 모비율 p의 추정문제에서 최대 허용오차가 e로 정해질 때 표본크기를 계산하는 공식은 다음과 같다.

$$n = \frac{p(1-p)}{\frac{e^2}{(z_{\alpha/2})^2} + \frac{p(1-p)}{N}}$$

따라서 95% 신뢰수준에서 허용오차의 한계를 $e = 0.022$로 하는 표본크기 n은 아래와 같이 계산할 수 있다. 이때 구한 표본크기 n은 모집단 비율을 $p = 0.3$으로 가정하여 계산한 결과이다. 모집단 비율 p는 1982년도에 실시한 생명보험성향조사의 모집단 비율이 0.725로 추정되어 그간의 증가분을 고려하여 0.3으로 가정한 것이다.

$$n = \frac{0.3(0.7)}{\frac{0.022^2}{1.96^2} + \frac{0.3(0.7)}{5634701}} = 1667$$

결국 1667가구를 단순임의 추출하여 모비율 p를 표본비율로 추정하게 되면 $|\hat{p} - p| \leq 0.022$이 성립하여 95%신뢰수준에서 2.2%이상 틀리지 않게 될 것이다.

다음으로 모집단 평균을 추정하는 것이 관심의 대상이라면 관심변수에 대한 모집단 평균 μ의 추정문제에서 허용오차를 e라 하면 표본크기 n을 계산하는 공식은 다음과 같다.

$$n = \frac{S^2}{\frac{e^2}{(z_{\alpha/2})^2} + \frac{S^2}{N}} = \frac{z_{\alpha/2}^2 \left(\frac{S}{\mu}\right)^2}{\frac{e^2}{\mu^2} + \frac{z_{\alpha/2}^2}{N}\left(\frac{S}{\mu}\right)^2}$$

여기서 e/μ는 모집단 평균에 대한 상대허용오차로서 보통 0.05라 하면 충분하며 S/μ는 모집단 변동계수로서 일반적으로 1보다 작은 값을 가진다. 따라서 이 값을 최대 1로 고려하여 계산하면 다음과 같은 값을 얻을 수 있다.

$$n = \frac{1.96^2(1)^2}{0.05^2 + \frac{1.95^2}{5634701}(1)^2} = 1536$$

따라서 표본크기를 1536으로 결정할 경우 모집단 평균 μ를 추정하는데 $\bar{y}$를 사용하면 $0.95 \leq \bar{x}/\mu \leq 1.05$에 있게 되며 이와 같은 신뢰를 가질 확률이 0.95가 된다.

결과적으로 모집단 평균이나 비율을 추정하고자 한때 무응답을 고려하여 표본크기는 약 1800개 가구를 취하면 충분할 것으로 본다.

이와 함께, 전국도시가구를 단순임의 추출하지 않고 층화추출 한다면 더욱 정확하

게 추정할 수 있는 장점이 있기 때문에 추정오차는 95% 신뢰도로서 모비율과 모평균의 허용오차는 2.2%와 5%이내로 예상된다.

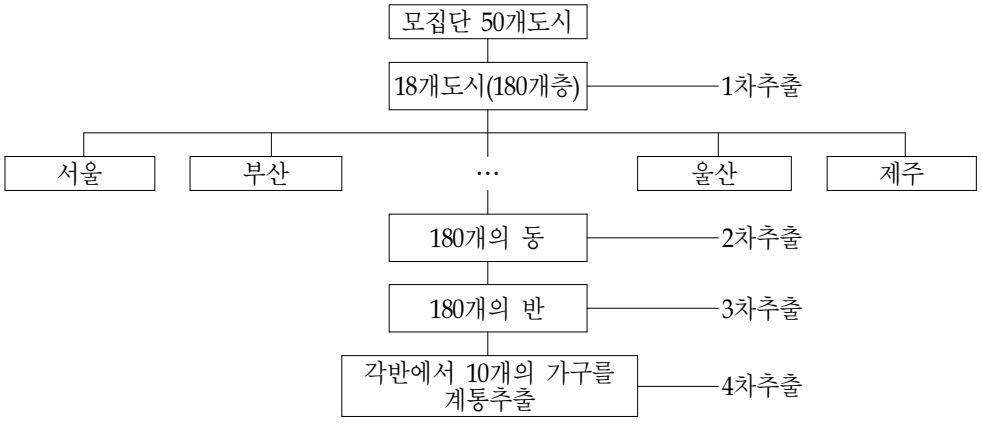
다. 표본추출

본 표본추출은 층화4단계 확률비례추출법을 사용하였다. 1차 추출단위는 도시로서 모집단인 전국 50개 도시 중에서 18개 도시를 도(都)별로 층화한 다음 추출하였다. 2차추출 단위는 동(洞)으로서 1차추출단위에서 뽑힌 18개 도시 중에서 성질이 유사한 동끼리 묶어서 180개의 층을 만들고 각 층에서 하나의 동을 추출하였다. 180개의 층은 지리적으로 입접하고 가구수의 증가비율이 유사한 동끼리 묶어서 층을 만들고 각 층은 비슷한 크기의 가구수를 갖도록 고려하였다. 각 층에서 하나의 동을 추출할 때에는 가구수에 비례하여 확률비례추출을 하였다. 3차추출단위는 반(班)으로서 추출된 각 동에서 하나의 반을 단순임의 추출하였고, 4차추출단위는 가구(家口)로서 추출된 각 반에서 10개의 가구를 계통추출하였다.

이와 같이 층화다단 추출법을 사용하게 된 이유는 다음과 같다.

- 모집단 전체에 대한 조사결과 뿐만 아니라 도별 지역별로 나누어 결과를 산출할 필요가 있다.
- 층내의 추출단위를 동질적으로 구성하여 조사의 정밀도를 높일 필요가 있다.
- 면접조사가 편리하고, 사후관리가 용이하다.

앞에서 설명한 추출과정을 요약하면 다음과 같다.



① 1차추출

모집단은 전국 50개 도시의 5,637,701개 가구이며, 이중 18개 도시를 추출하였다. 50개 도시가 13개의 도지역 단위별 행정구역으로 나뉘고 각 행정구역에서 도청소재지는 추출확률이 1로 뽑히도록 하고, 나머지 5개 도시는 도시의 수가 많은 도 순서대로 배정하였다. 따라서 경기도, 전라남도, 경상남도, 경상북도, 강원도는 도청소재지 이외에 추가로 하나의 도시가 더 뽑히도록 하였다. 추가로 뽑을 때에는 가구수에 비례하여 확률비례추출을 하였다. 이와 같은 방법으로 추출된 8개 도시는 다음과 같다.

<표 II-2-5> 도시의 추출 및 도별 가구수 배정

층	도시	가구수	가구수합계	층별가구수비율	표본가구수
서울	서울	2,116,334	2,116,334	37.56%	680
부산	부산	765,236	765,236	13.58%	240
대구	대구	459,223	459,223	8.15%	150
인천	인천	281,350	281,350	4.99%	90
경기도	수원 성남 의정부 안양 부천 광명 안산 동두천	85,832 97,088 34,429 70,734 82,972 47,378 15,631 15,843	449,907	7.98%	140
강원도	춘천 원주 강릉 동해 태백 속초	36,343 32,842 27,725 22,392 25,182 16,287	160,771	2.85%	50
충북	청주 충주 제천	67,082 25,265 20,147	1,124,949	2.00%	40
충남	대전 천안	150,909 28,839	180,848	3.21%	60
전북	전주 군산 이리 정읍 남원	84,674 38,283 38,159 14,859 10,873	186,848	3.32%	60
전남	광주 목포 여수 순천 금성	175,596 49,025 36,449 23,391 10,680	296,141	5.26%	90
경북	포항 경주 김천 안동 구미 영주	58,052 30,093 18,364 27,403 25,489 19,604	189,667	3.37%	60
경남	마산 울산 진주 창원 진해 충무 삼천포 김해	96,837 114,502 46,835 34,313 29,061 18,921 13,932 17,047	371,448	6.59%	120
제주	제주 서귀포	45,930 18,504	64,434	1.14%	20
합계		5,634,701	5,634,701	100.00%	1800

<표 II-2-6>에서 층별 가구수에 비례하여 표본가구수를 배정할 때 층별 가구수 비례에 1800을 곱한 다음 각 도시에서 20가구 이상이 뽑히도록 조정하고 표본가구수의 끝자리가 0이 되도록 하였다. <표 II-2-7>에서 표본도시별 표본가구수를 배정할 때는 층별 가구수비율에 따라 배정된 표본가구수를 도별로 추출된 도시들의 가구수에 비

례하여 배정하였다. 이때에도 각 도시에서 20가구이상으로 하고 가구수는 10의 배수로 배정하였다.

<표 II-2-6> 표본도시와 도시별 가구수 배정

(단위: 개)

층	표본가구수	총가구수	표본도시	표본도시별표본가구수
서 울	680	2,116,334	서 울	670
부 산	240	76,5236	부 산	240
대 구	150	459,223	대 구	150
인 천	90	28,1350	인 천	90
경기도	140	182,920	수 원	70
			성 남	70
강원도	50	69,185	춘 천	30
			원 주	20
충 북	40	67,082	청 주	40
충 남	60	150,909	대 전	60
전 북	60	84,674	전 주	60
전 남	90	199,987	광 주	70
			순 천	20
경 북	60	55,582	구 미	30
			경 주	30
경 남	120	211,339	마 산	60
			울 산	60
제 주	20	45,930	제 주	20
계	1800	4,689,751	18개도시	180

<표II-2-7> 표본도시별 층의 수

(단위: 개)

표본도시	층수	표본도시	층수
서 울	68	전 주	6
부 산	24	광 주	7
대 구	15	순 천	2
인 천	9	구 미	3
수 원	7	경 주	3
성 남	7	마 산	6
춘 천	3	울 산	6
원 주	2	제 주	2
청 주	4	계	180
대 전	6		

② 2차추출

<표 Ⅱ-2-7>에서 표본가구1800개를 18개 표본도시에 배정한 후 동을 추출하기 위해 18개 표본도시의 모든 동을 합쳐서 180개 층으로 층화하였다. 층을 구성할 때 지리적으로 인접하고, 교통, 문화적 입지조건이 비슷하고, 생활수준이 균일하다고 생각되는 동들을 하나의 층으로 묶는 것을 원칙으로 하였다. 그리고 한 층에서 10가구를 추출할 수 있도록 하기위해 <표 Ⅱ-2-8>의 표본도시별 표본가구수를 10으로 나누어 나오는 값을 그 도시의 층수로 결정하였다. 따라서 표본도시별 층수는 <표 Ⅱ-2-7>과 같다. 서울의 경우 68개의 층을 만드는 방법은 각 구(區)에서 가구수의 구성비율에 따라 층의 수를 배정하고 배정된 층의 수를 기준으로 각 구에서 동을 나누었다. 각 구에서 동을 나눌 때에는 총 구가수의 증가비율의 순서대로 나열한 후 각 층에서 총 가구수가 비슷한 동들을 나누었다. 부산, 대구, 인천, 대전은 이러한 방법으로 층을 구별하고, 기타도시는 구가 없으므로 구별로 층화하지 않았다. <표 Ⅱ-2-8>은 각 도시별로 구를 층화한 결과이다.

③ 3차추출

2차표본으로 추출된 180개의 표본동에서 하나의 반을 선정하는 방법은 먼저 표본동에 대해 "한국행정구역편람(85)"를 찾아 해당 통수를 구한다.

<표 Ⅱ-2-8> 서울지역의 표본동-통-반 리스트 일부

구	동	통-반
종로구	창신1동	20-53
종로구	혜화동	7-63
중 구	신당2동	25-35
중 구	황학동	18-30
용산구	한강로1가동	4-58
용산구	원효로1가동	11-21
용산구	후암동	20-53
성동구	하왕십리2동	2-46
성동구	군자동	10-72
성동구	송정동	23-17
성동구	성수2가2동	37-10
성동구	성수2가3동	19-94

여기서 난수표를 이용하여 표본통을 선정한다. 마지막으로 각 통에 몇 개의 반이 있는지 모르기 때문에 난수표에서 두 자리 숫자를 임의로 택하여 표본반을 추출한다. 참고로 서울지역의 경우 표본으로 추출된 표본통 번호와 표본반 리스트 중 일부가 <표 Ⅱ-2-8>과 같다.

④ 4차추출

A. 실제반 선정 방법

표본동-통-반 리스트로부터 표본반을 추출하는 방법은 다음과 같다. 표본동-통 에서 조사원이 몇 개의 반이 있는지를 확인한 후 만일 x개의 반이 있는데 표본반의 숫자가 y이면 y를 x로 나누고 나머지가 실제 표본반이 된다. 나누어떨어지면 x가 실제 표본반이 된다. 예를 들어 x와 y가 각각 30과 53이라 하면 53을 30으로 나눈 나머지가 23이므로 23반을 표본반으로 추출하게 된다.

B. 가구선정 방법

표본동-통-반에서 실제 가구수를 확인한 후 실제 가구수에서 계통 추출법으로 10개의 가구를 추출한다. 예를 들어 실제가구의 수가 56가구라 하자. 그러면 56/10 =5.6에서 5를 취한다. 이보다 작은 수를 난수표로부터 하나를 택하여 만일 난수 값이 3이라면, 표본가구는 다음번호에 해당되는 가구가 된다.

3, 8, 13, 18, 23, 28, 33, 38, 43, 48, 53, 58, 2, 7, 12, 17

표본가구는 3-48까지 10가구이지만, 예비표본가구로 5가구를 추가로 선정한다.

<표 II-2-9> 대도시의 구별 층의 구분

지 역	구	가구수	기구구성비율	가구구성비층수	실제층수
서울	종로구	64,564	3.06%	2.08	2
	중구	53,674	2.54%	1.74	2
	용산구	78,373	3.70%	2.52	3
	성동구	167,907	7.93%	5.39	5
	동대문구	208,728	9.86%	6.70	7
	성북구	134,530	6.36%	4.32	4
	도봉구	183,727	8.68%	5.90	6
	은평구	93,361	4.41%	3.00	3
	서대문구	78,373	3.70%	2.52	3
	마포구	102,098	4.82%	3.28	3
	강서구	149,067	7.04%	4.79	5
	구로구	144,702	6.84%	4.65	5
	영등포구	102,324	4.83%	3.28	3
	동작구	91,207	4.31%	2.93	3
	관악구	121,011	5.72%	3.89	4
	강남구	154,285	7.29%	4.96	5
	강동구	171,206	8.09%	5.50	5
	계	2,116,334	100.00%	68	68
부산	중구	23,568	3.08%	2.739	1
	서구	104,890	13.70%	3.288	3
	동구	46,702	6.10%	1.464	1
	영도구	47,201	6.17%	1.482	1
	부산진구	117,517	15.36%	3.686	4
	동래구	173,406	22.66%	5.438	5
	남구	115,767	15.13%	3.631	4
	북구	88,624	11.58%	2.779	3
	해운대구	47,561	6.22%	1.493	2
	계	765,236	100.00%	24	24
대구	중 구	49,072	10.96%	1.61	2
	동 구	73,245	15.95%	2.39	2
	서 구	121,409	26.44%	3.97	4
	남 구	78,779	17.15%	2.57	3
	북 구	74,837	16.29%	2.44	2
	수 성 구	61,881	13.48%	2.02	2
	계	459,223	100.00%	15	15
인천	중 구	21,205	7.54%	0.68	1
	동 구	35,401	12.58%	1.13	1
	남 구	129,067	45.87%	4.13	4
	북 구	95,677	34.01%	3.06	3
	계	281,350	100.00%	9	9

<표 II-2-9> 대도시의 구별 층의 구분

지 역	구	가구수	기구구성비율	가구구성비층수	실제층수
대전	동 구	59,472	39.41%	2.36	2
	중 구	91,473	60.59%	3.64	4
	계	150,909	100.00%	6	6
광주	동 구	46,410	26.28%	1.84	2
	서 구	74,192	42.01%	2.94	3
	북 구	55,994	31.71%	2.22	2
	계	176,596	100.00%	7	7

Ⅲ. 표본조사자료 분석을 위한 통계 프로그램

1. SAS 프로그램

SAS 프로그램에서 표본조사 자료의 분석을 위해 한 표본추출과 관련하여 SurveySelect 프로시저를 제공하고 있으며, 모수 추정과 관련하여 SurveyMeans를 제공하고, 회귀추정과 비추정에 관련하여 SurveyReg 프로시저를 제공한다.

(1) 표본추출(SurveySelect)

SurveySelect 프로시저는 확률에 근거하여 표본을 선택하기 위한 다양한 방법을 제공한다.

단순임의추출, 계통추출, 층화추출, 집락추출, 불균등확률추출을 포함하는 복합 다단계 추출설계에 따라 표본을 선택할 수 있다.

SurveySelect 프로시저를 사용하기 위해서는 먼저 표본을 추출하기 위한 표본추출틀(sampling frame) 또는 추출단위의 리스트(list)가 데이터셑으로 존재해야 한다.

이와 더불어 추출하고자 하는 표본 수 또는 추출률 등 표본추출과 연관된 모수가 지정되어야 한다. 만일 다단계로 표본을 추출하고자 한다면, 각 추출단계에 따라 프레임과 현 단계에서의 추출과 관련된 모수가 지정되어야 한다.

□ SurveySelect 문장의 일반 형식

```
PROC SurveySelect options ;
STRATA variables ;
CONTROL variables ;
SIZE variable ;
ID variables ;
RUN ;
```

□ Options 사항들

① 데이터셀의 입력과 출력의 지정

입력 : DATA= SAS 데이터셀

출력 : OUT= OUTPUT데이터셀
OUTSORT= OUTPUT데이터셀

② 표본추출방법의 지정

추출방법 : METHOD= 추출방법 또는 M= 추출방법

③ 표본크기 및 추출률의 지정

표본크기 : SAMPSIZE= 크기 및 데이터셀 , N=크기 또는 데이터셀;
추출률 : SAMPRATE= 비율 또는 데이터셀; 추출률의 지정
NMAX= n ; 최대 층화표본크기
NMIN= n ; 최소 층화표본크기

④ 반복회수의 지정 - 분산추정에 사용

REP=반복회수 r ; 표본의 반복 추출회수의 지정

⑤ 크기척도의 조정

MINSIZE= 데이터셀 또는 최대값; 층의 최대크기에 의해 크기척도의 조정
MAXSIZE= 데이터셀 또는 최소값; 층의 최소크기에 의해 크기척도의 조정

⑥ 크기척도의 지정(PPS에서 유용)

CERTSIZE= 데이터셀 또는 비율;

⑦ 정렬 방법의 지정

```
SORT=NEST | SERP ;통제변수에 따른 정렬(기본은 SERP)
```

⑧ 난수의 시드 지정

```
SEED=값; 초기 시드값의 배정
```

⑨ OUT= 내용의 통제

```
JTPROBS ; 출력에 결합확률을 포함시킴.
OUTHITS ; 동일한 단위가 표본에 1회이상 포함된 경우표시
OUTSIZE ; 출력에 추가적인 설계와 추출틀의 모수를 포함
STATS ; 출력에 추출확률과 추출가중치를 표시
```

가. 단순임의비복원추출

```
PROC SurveySelect DATA=데이터셑이름
 METHOD=SRS | URS N=표본수  OUT=SrSSAMP ;
RUN ;
```

▷ 비복원의 경우 "M=SRS"를 지정하고 복원인 경우 "M=URS"를 지정한다.

▷ 단순임의추출의 경우 추출확률이 n/N으로 표본을 추출한다.

▷ 데이터셑은 추출틀이나 리스트가 저장된 SAS 데이터셑을 의미한다.

▷ "OUT=" 은 추출된 단위를 나타내는 출력 데이터셑을 지정한다.

▷ 데이터셑에 존재하는 일부 변수만을 포함하고자 할 때에는 ID 명령으로 지정한다.

▷ 복원 추출의 경우 하나의 단위가 표본으로 선택된 회수를 "the number of hits"로 나타낸다.

나. 층화임의추출

```
PROC SurveySelect DATA=데이터셑이름
 METHOD=SRS N=표본수 SEED=시드번호 OUT=StSAMP ;
 STRATA 층화변수 ;
 CONTROL 변수 ;
 SIZE 변수 ;
 ID 변수 ;
RUN ;
```

▷ 층화추출을 위해서는 반드시 데이터셑에 층화변수가 존재해야 한다.

▷ 이러한 층화는 또한 각 층별로 서로 겹치지 않도록 나뉘어져야 한다.

▷ 표본추출은 각 층별로 추출확률이 n_h/N_h로서 독립적으로 단순임의추출한다.

▷ 층화변수를 기준으로 표본을 추출하기 전에 층화변수를 기준으로 정렬이 되어야 한다.

▷ 표본수는 각 층별 표본수 n_h를 나타낸다.

▷ 만일 층별로 서로 다른 표본수를 지정하려면 층화 확률비례추출로 가능하다.

▷ 층화계통추출의 경우 “M=SYS"를 지정하고, 층내에서 단위들의 ordering을 위해 “CONTROL 명령" 에 지정한다.

□ 정렬을 위한 프로그램

```
PROC SORT DATA=데이터셑이름;
 BY  층화변수 ;
RUN ;
```

다. 계통추출

```
PROC SurveySelect DATA=데이터셑이름
 METHOD=SYS N=표본수  SEED=시드번호 OUT=SySSAMP ;
RUN ;
```

▷ 층화계통추출의 경우 층화변수를 지정하고, "CONTROL 명령" 에서 층내에서 순서화될 수 있는 변수를 지정해야 한다.

라. 단순집락추출

```
PROC SurveySelect DATA=데이터셑이름
 METHOD=SRS N=표본수 SEED=시드번호 OUT=StSAMP ;
 STRATA 층화변수 ;
 CONTROL 변수 ;
 CLUSTER 변수;
 SIZE 변수 ;
 ID 변수 ;
RUN ;
```

▷ PSU를 단순임의비복원추출한다.

▷ 집락추출은 PSU를 단순임의추출하는 것이므로 집락을 지정하는 변수를 추출틀에서 정의해야 한다.

▷ 표본수는 표본 PSU의 크기를 나타낸다.

▷ 층화 집락 추출의 경우 층화 변수와 함께 층내에서 집락을 나타내는 변수로서 CONTROL 변수를 사용한다.

▷ 2단계집락추출의 경우 PSU와 SSU 변수를 기준으로 표본추출을 수행함.

마. 확률비례추출

```
PROC SurveySelect DATA=데이터셑이름
 METHOD=PPS N=표본수 SEED=시드번호 OUT=StSAMP ;
 STRATA 층화변수 ;
 CONTROL 변수 ;
 SIZE 변수 ;
 ID 변수 ;
RUN ;
```

▷ 관심변수와 상관이 있는 보조변수가 크기변수로 "SIZE 변수"에서 지정한다.

▷ 추출틀안에는 크기척도를 지정하는 보조변수가 존재해야 한다.
▷ 층화확률비례 추출의 경우 "M=PPS"로 지정하고, "CONTROL 변수"를 지정하면 된다.

(2) 모수추정(SurveyMeans)

조사된 표본 데이터로부터 모집단의 평균이나 총합을 추정하며, 추정분산과 신뢰구간을 구한다.

□ 기본형식

```
PROC SurveyMeans  DATA=데이터셋이름 options  통계키워드 ;
 BY 변수명 ;
 CLASS 변수명;
 CLUSTER 변수명;
 DOMAIN 변수명< 변수1*변수2  변수1*변수2*변수3 ... > ;
 STRATA 변수명< / option > ;
 VAR 변수명 ;
 WEIGHT 변수명 ;
 ODS OUTPUT ODS테이블 이름 = 데이터셋 ;
RUN ;
```

□ SurveyMeans의 options 사항

① 신뢰계수의 지정 (기본값은 0.05 : 95% 신뢰구간)

```
 ALPHA=α ;
```

② 추출률의 지정(fpc)

```
 RATE = 값 또는 SAS 데이터셋 ;
 R=값 또는 데이터셋 ;
```

▷ 다단계 추출에서는 1단계 추출률을 나타낸다.
▷ 모집단의 전체 PSU 중 표본으로 추출된 PSU의 비율
▷ 층별 추출률이 같은 경우 값으로 나타내고, 각 층별 추출률이 다른 경우에는 데이터셀으로 나타낸다.

③ 모집단 크기의 지정

```
TOTAL = 값 또는 SAS 데이터셀 ;
 N=값 또는 데이터셀 ;
```

▷ 다단계 추출의 경우 PSU의 크기를 나타내며, 각 층별 크기를 데이터셀으로 지정함.
▷ 단순임의추출의 경우 모집단 크기를 나타냄.

□ SurveyMeans의 통계량 키워드

통계량	의미	통계량	의미
ALL	모든 통계량	NOBS	관찰값의 개수
CLM	모평균의 신뢰구간	RANGE	범위(=MAX - MIN)
CLSUM	모집단 총합의 신뢰구간	STD	총합의 표준편차
CV	변동계수	STDERR	표본평균의 표준오차
DF	자유도	SUM	가중합계
MAX , MIN	최대값과 최소값	SUMWGT	가중값의 합계
MEAN	표본평균, 표본비율	T	귀무가설 : μ=0 하에서의 t-값
NCLUSTER	집락의 개수	VAR	평균의 분산
NMISS	결측값의 개수	VARSUM	총합의 분산

가. 단순임의추출 하에서의 모수 추정

```
PROC SurveyMeans DATA=데이터셀이름 TOTAL=N ;
 VAR 변수명 ;
RUN ;
```

나. 층화추출 하에서의 모수 추정

```
PROC SurveyMeans DATA=데이터셋이름 TOTAL=데이터셋 ;
 STRATUM 층화변수 /LIST ;
 <CLUSTER 집락변수 ;>/*층화 집락 추출의 모수 추정시 사용*/
 VAR 변수명 ;
RUN ;
```

▷ 층화추출의 경우 층별 분산과 총 분산을 구해줌.

▷ "TOTAL=데이터셑"은 층별 크기가 지정된 데이터셑을 의미함.

다. 계통추출 하에서의 모수 추정

```
PROC SurveyMeans DATA=데이터셋이름 TOTAL=데이터셋 ;
 VAR 변수명 ;
RUN ;
```

라. 집락추출 하에서의 모수 추정 - 단순집락추출의 경우

```
PROC SurveyMeans DATA=데이터셋이름 TOTAL=데이터셋 ;
 CLUSTER 집락변수 ;
 VAR 변수명 ;
RUN ;
```

마. 비추정과 회귀 추정

표본조사 데이터에 대해 회귀분석을 수행하고, 층화, 집락, 복합 표본설계에 대한 회귀 추정량을 구한다. 비 추정의 경우 회귀추정에서 선택사항으로 "NOINT"를 성정하면 된다.

□ SurveyReg 의 기본형식

```
PROC SurveyReg DATA= 데이터셑이름  options  ;
 BY 변수명 ;
 CLASS 변수명 ;
 CLUSTER 변수명 ;
 CONTRAST 'label' effect values < ... effect values > < / options >;
 ESTIMATE 'label' effect values < ... effect values > < / options > ;
 MODEL 종속변수(관심변수) =  독립변수(보조변수)  < / options > ;
 STRATA 변수명 ;  < / options > ;
 WEIGHT 변수명 ;
RUN;
```

□ MODEL 명령문의 options

① CLPARM : 모수의 신뢰 한계
② COVB : 추정된 회귀 계수의 공분산 행렬
③ DEFF : 회귀계수 추정치의 설계효과
④ DF=value : 자유도
⑤ I | INVERSE : X'X 행렬의 역행렬 또는 일반화 역행렬
⑥ NOINT : 모형의 절편
⑦ SOLUTION : 정규방정식의 해
⑧ X |XPX : X'X 또는 X'WX(가중변수가 있을 때)

2. R 프로그램

R 프로그램은 SAS 프로그램과 달리 interactive하게 사용자의 명령에 즉각적으로 출력을 나타내주는 프로그램으로서, 복합 표본 추출설계에 유용하게 사용될 수 있는 프로그램이다. 또한 타 프로그램과는 달리 R 프로그램은 사용비용이 없는 무료 프로그램이라는 데 매우 큰 장점이 있다. 표본설계와 관련하여 R 프로그램은 표본추출보다는 추정과정에 주안점이 맞춰져 있으며, 향후 표본추출과 관련된 모듈 또한 개발될 것으로 기대된다.

(1) Svydesign( ) 함수

☞ 복합표본추출설계를 지정한다.

□ 기본형식

```
> svydesign(id, probs=NULL, strata=NULL, variables=NULL, fpc=NULL, data=NULL,
                nest=FALSE, checks.stratra=!nest, weights=NULL)
```

□ 인수들

인수	설명
id	집락을 정의하는 식이나 데이터프레임으로서 ~0 또는 ~1이면, 집락 설계가 아님을 명시함. 즉, 단순히 층화추출설계인 경우 ~0 또는 ~1 을 명시함.
prob	집락 추출확률을 정의한 식이나 데이터프레임.
strata	층화를 정의하는 식이나 벡터를 정의한 변수.
variables	조사에서 측정된 변수를 정의하는 식 또는 데이터프레임. 만일 "NULL"값이면, 데이터 연산식이 사용된다.
fpc	유한모집단 수정계수
data	변수가 속한 데이터프레임
nest	층화 집락 추출에서 1단계 층에서 뽑힌 단위들은 서로 다른 층 또는 같은 층에서 서로 다른 집락이 뽑힐 수 있기 때문에 집락이 층에 대해 nested 됨.
check.strata	"TRUE" 값이면, 집락이 층에 대해 nested 되었는지를 검사함.
weights	"prob"의 대체로서 추출가중치를 지정한 식이나 벡터.

(2) Surveysummary

☞ 복합 표본추출 설계에 따른 데이터에 대한 요약 통계량을 계산한다.
다양한 통계값을 계산하며, 표본설계에 따라 계산할 수 있다.

□ 통계량 계산 함수

```
> svymean(x, design, na.rm=FALSE, deff=FALSE)/*표본평균*/
> svrepmean(x, design, na.rm=FALSE, rho=NULL, return.replicate=NULL,
   deff=FALSE)/*반복평균*/
> svyvar(x, design, na.rm=FALSE)/*표본분산*/
> svrepvar(x, design, na.rm=FALSE,rho=NULL,return.replicate=NULL)/*반복표본분산 */

> svytotal(x, design, na.rm=FALSE, deff=FALSE)/*표본총합*/
> svreptotal(x, design, na.rm=FALSE, rho=NULL, return.replicate=NULL,
  deff=FALSE)/*반복 총합*/
> cv(object, ...) /*변동계수*/
> coef(object, ...) /*회귀계수*/
> vcov(object, ...)/*분산 공분산*/
> deff(object, quietly=FALSE, ...)/*설계효과*/
```

□ 인수들

인 수	설 명
x	식이나, 벡터 또는 행렬변수.
design	survey.design 또는 svyrep.design의 object.
na.rm	결측치를 drop시킬 것인지를 결정. na.rm=TRUE이면 모든 결측치를 가지는 case들은 drop됨.
rho	BRR 설계에서 Fay의 분산추정량에 대한 값.
return.replicate	반복평균을 return할지를 결정.
deff	설계효과를 return함.
object	survey summary 함수중 하나의 결과.
quitely	deff가 계산되지 않더라도 경고를 나타내지 않음.

(3) 추출설계에 따른 추정

가. 층화추출설계

▷ svydesign( ) 함수에서 "id=~1" 또는 "~0"을 지정하고, 층화변수를 나타내는 "strat=층화변수"를 지정하고, 데이터셑 "data=데이터셑" 과 유한모집단 수정인자인 "fpc"를 지정한다.
survey.design object를 이용하여 svymean( ) 함수로부터 모평균의 추정값을 구한다.

[추출설계 지정의 예]

```
> library(survey)
> data(api)
> dstrat<-svydesign(id=~1, strata=~stype, weights=~pw, data=apistrat, fpc=~fpc)
```

나. 집락추출설계

▷ 1단계 집락 추출과 2단계 집락 추출설계에 대한 모수 추정결과를 표현한다.
▷ PSU와 SSU의 기준 변수를 반드시 지정해야 하며, 2단계 추출설계의 경우에는 PSU와 SSU를 동시에 지정해야 한다.
▷ 만일, PSU나 SSU를 지정하지 않으면, 층화추출설계로 간주한다.

[집락추출설계 지정의 예]

```
> dclus1<-svydesign(id=~dnum, weights=~pw, data=apiclus1, fpc=~fpc)
```

☞ id=~dnum : 집락변수로 dnum을 지정한다.
weights=~pw : pw 변수가 가중치를 지정한다.
data=apiclus1 : 데이터프레임을 지정한다.

fcp=~fpc : 유한모집단 수정인자를 지정한다.

[2단계 추출설계 지정의 예]

```
> dclus2<-svydesign(id=~dnum+snum, weights=~pw, data=apiclus2)
```

☞ id=~dnum+snum : 1단계 집락변수로 dnum을 2단계집락변수로 snum을 지정함.
weights=~pw : pw 변수가 가중치를 지정함.
data=apiclus2 : 데이터프레임을 지정함.

다. 층화집락추출설계

▷ svydesign( ) 함수에서 "id"로 PSU를 지정하고, 층화변수를 나타내는 "strat=층화변수"를 지정하면 층화집락추출설계가 된다.
▷ survey.design object를 이용하여 svymean( ) 함수로부터 모평균의 추정값을 구한다.

[층화집락추출설계 지정의 예]

```
> data(fpc)
> dfpc<-svydesign(id=~psuid, strat=~stratid, weight=~weight, data=fpc, nest=TRUE)
> dsub<-subset(dfpc,x>4)
```

☞ 동일한 psuid를 갖는 단위가 서로 다른 층에 존재하기 때문에 "nest=TRUE"를 지정한다.

(4) 분산추정

▷ 전형적인 추출설계의해 정의된 object에 대해 반복가중값을 사용하기 위해서는 추

출설계를 지정한 object를 as.svrepdesign( )으로 변환시켜야 한다.

▷ 이러한 변환 작업은 잭나이프(Jackknife)분산, 균형반복(BRR)분산, Fayd의 분산 추정을 위해서는 반드시 필요하다.

① as.svrepdesign( ) 함수

□ 기본형식

```
> as.svrepdesign(design, type=c("auto","JK1","JKn", "BRR", "Fay"))
```

□ 인수들

인 수	설 명
design	survey.design 또는 svyrep.design의 object.
type	반복가중치의 형태를 지정 "auto"는 층화에서 JKn을 사용하고, 층화하지 않은 경우에는 JK1을 사용함.

(5) 사후 가중값 조정

가. 사후층화(poststratification)

Poststratification은 추출가중값과 반복가중값을 조정하기위해 일련의 사후층화변수의 결합분포와 기지의 모집단의 결합분포를 서로 일치시키는 작업을 의미한다. 만일 결합분포를 알 수 없는 경우 주변분포를 이용할 수 있으며, 이때에는 raking작업을 수행하게 된다.

□ 기본형식

```
> postStratify(design, strata, population, partial=FALSE)
```

□ 인수들

인 수	설 명
design	반복가중치를 가지는 조사 설계
strata	사후층화 변수의 식 또는 데이터 프레임
population	모집단 빈도수를 가지는 table, xtable, 또는 데이터프레임
partial	partial=TRUE 이면, 표본에 나타나지 않는 모집단 층을 무시한다.

나. 래킹(Raking)

Raking은 분할표상에서 기지의 모집단 주변 분포를 이용하여 표본의 주변분포를 반복적인 방법으로 일치시키는 가중치 조정 작업으로서 셀 값이 0을 갖지 않는 한 수렴하는 것으로 알려져 있다. 특정한 조건을 만족할 때 까지 반복적으로 분할표 상의 각 셀 가중치를 조정해가는 방법이다.

□ 기본형식

```
> rake(design, sample.margins, population.margins, control=list(maxit=10, epsilon=1,
        verbose=FALSE), compress=NULL )
```

□ 인수들

인 수	설 명
design	반복가중치를 가지는 조사 설계
sample.margins	표본의 주변분포를 나타내는 식 또는 데이터 프레임. postStratifiy와 같은 형태를 취한다.
population.margins	모집단의 주변분포를 나타내는 table의 list
control	maxit : 최대반복회수 epsilon : 수렴조건 (< 1)
compress	반복설계이면, 새로운 반복가중치 행렬로 압축을 시도함.

(6) 무응답 가중치 조정

조사 무응답 단위에 대해 무응답가중치 조정을 수행해준다. 또한, 작은 규모나 큰 가중치를 가지는 층들을 결합하여 무응답가중치를 간단히 구해준다. 층화추출설계에서 불완전한 응답이 발생했을 때 무응답을 반영하기 위해 추출가중치를 적절하게 재조정해준다.

□ 무응답 가중치 조정을 위한 함수들

함 수	설 명
nonresponse( )	모집단 크기, 표본크기, 표본가중치에 대한 object를 생성
sparseCells( )	결합이 필요한 셀들을 정의함.
neighbours( )	특정한 셀과 인접한 셀들을 나타냄.
joinCells( )	특정셀 들을 결합(collapse)시킴.
weights( )	결합이 끝난 셀들에 대해 무응답 가중치를 구해줌.

① nonresponse( ) 함수

무응답 가중치를 갖는 object를 생성한다.

□ 기본형식

```
> nonresponse(sample.weights, sample.counts, population)
```

□ 인수들

인 수	설 명
sample.weights	층별 추출가중치 테이블
sample.counts	층별 표본수 테이블
population	층별 모집단 크기 테이블

② sparseCells( ) 함수

무응답 가중치를 갖는 object에 대해 “sparse” 셀을 지정하는 함수이다.

□ 기본형식

```
> sparseCells(object, count=0, totalweight=Inf, nrweight=1.5)
```

□ 인수들

인 수	설 명
object	nonresponse( ) 함수에 의해 생성된 object
count	이 값보다 적은 수의 셀들.
totalweight	이 값보다 큰 무응답가중치와 평균 추출가중치 곱을 가지는 셀들.
nrweight	이 값보다 높은 무응답가중치를 갖는 셀들.

③ neighbours( )함수

특정셀에 인접한 셀들을 정의한다.

□ 기본형식

```
> neighbours(index, object)
```

□ 인수들

인 수	설 명
index	인접 셀이 될 수 있는 셀의 수
object	nonresponse( ) 함수에 의해 정의된 object

④ joinCells( ) 함수

□ 기본형식

```
> joinCells(object, a, ...)
```

□ 인수들

인 수	설 명
object	nonresponse( ) 함수에 의해 정의된 object
a, …	결합할 셀들을 지정

⑤ weights( ) 함수

nonresponse( ) 에 의해 생성된 object를 인수로 하면, 무응답 가중치가 생성되며, svydesign( ) 함수에 의해 생성된 object를 인수로 하면, sampling weight를 생성해준다.

□ 기본형식

```
> weights(object, type=c("replication", "sampling", "analysis" ), … )
```

□ 인수들

인 수	설 명
object	nonresponse object나 survey design object.
type	가중치의 유형을 지시함. "analysis"는 sampling과 replication weight를 결합함.

3. STATA 프로그램

(1) Survey data 분석의 개요

STATA 프로그램에서 survey 데이터를 분석하기 위한 이용 가능한 명령어들은 상당히 다양하게 존재한다. 그 중에서 다른 프로그램의 내용과 균형을 이루기 위해 모수의 추정, 비추정과 회귀추정, 분산추정 등으로 구분하여 다루기로 한다.

Survey 데이터를 분석하기 위한 첫 번째 단계로서는 데이터셀을 어떠한 조사 설계에 의해 얻어졌는가를 구명하는 일이다. 이를 위해 STATA에서는 "**svyset**"과 "**svydes**" 명령어를 제공한다. 다음으로 모수추정에 필요한 명령어로는 "**svy:mean**", "**svy:proportion**", "**svy:total**", "**svy:ratio**" 등을 제공한다. 이들 이외에 조사 자료의 회귀분석(회귀추정)에 필요한 명령어를 제공한다. 이때, STATA에서는 **"regression**" 명령과 "**svy :regree**" 명령은 각각 서로 다른 분석을 수행하는데, 전자는 조사 자료가 아닌 경우의 회귀분석을 위한 명령이고, 후자는 조사 자료에 대한 회귀추정을 나타낸다.

분산추정을 위해 "**svy brr**", "**svy jackknife**"를 제공하며, 사후추정을 위해서 "**svy poststrata**"을 제공하고 있다. 조사 자료에 대해 사후추정 통계량으로 설계효과, 부차모집단의 크기, 공분산, 상관계수 등을 계산해주는 명령어로 "**estat**" 명령어가 있다.

STATA 프로그램을 이용하여 조사자료에 대한 분석을 위해서는 다음과 같은 추출설계에 대한 3가지 인자를 규정할 필요가 있다.

① 추출가중치의 정의 - "pweight" 의 지정

표본을 추출할 때 표본의 추출확률의 역수로 계산되는 추출가중치를 정의한다.

② 집락 추출설계 - PSU의 지정

대부분의 대규모 조사 설계에서는 조사개체를 직접 표본으로 추출하는 경우는 거의 없다. 즉, 개체들의 집합 -집락- 을 표본으로 추출하게 된다. 이러한 집락의 1차단계가 바로 "1단계추출단위" 인 PSU가 된다. 결과적으로 조사 자료의 분석에서 PSU가 어떤 것인지를 먼저 정의할 필요가 있다.

③ 층화추출설계

집락의 정의와 마찬가지로, 집락들의 서로 다른 그룹인 층(stratum)을 정의해야 한다. 예를 들어 254개 군부(counties)를 각각 도시지역과 그 외 지역의 2개 층으로 구분할 수 있다. 이와 같이 조사자료 분석을 위해 먼저 모집단 수준에서 정의된 층을 나타내는 변수를 지정할 필요가 있다.

(2) 추출설계의 지정

① svyset 명령

데이터셋에 대한 조사설계를 지정한다.

□ **기본형식** - 1단계 추출설계

```
.svyset [psu] [weight] [, design_options  options]
```

□ **기본형식** - 다단계 추출설계

```
.svyset psu [weight] [,design_options] [ ||ssu, design_options] [options]
```

☐ design_options

선택사항	설 명
strata(변수명)	층화변수를 지정한다.
fpc(변수명)	유한모집단 수정계수를 지정한다.

☐ options

선택사항	설 명
brrweight(변수리스트)	BRR(Balanced repeated replicated) 가중치의 지정
jkrweight(변수리스트)	jackknife replication weight를 지정
vce(linearized)	테일러 선형화 분산추정치를 계산
vce(brr)	BRR(Balanced repeated replicated) 분산추정치의 계산
vce(jacknife)	jackknife replication 분산을 계산
poststrata(변수명)	사후층화 변수의 지정
postweight(변수명)	사후층의 모집단 크기의 지정

② svydes 명령

svyset에서 지정한 추출설계에 대한 자세한 설명을 제공한다.

☐ 기본형식

```
.svydes [변수리스트] [,if] [in] [,options]
```

☐ options

선택사항	설 명
stage(#)	추출단계를 지정. 기본값은 stage(1) 이다.
finalstage	최종단계의 추출단위당 정보를 제공
single	1단계추출단위를 가진 층에 대한 정보를 제공
generate(새로운 변수)	1단계추출단위를 가지는 층화변수를 생성

(3) 모평균, 모비율, 모총계의 추정

STATA 프로그램에서 조사자료에 대한 추정을 위한 명령으로 "**svy:mean**", "**svy:proportion**", "**svy:total**" 등을 제공한다. 각 명령의 형식과 선택사항들에 대해 살펴보기로 한다.

☞ 명령어의 형식

조사자료의 모평균, 모비율, 모총계을 추정하며, 그에 대한 표준오차를 각 분산추정방법에 따라 제공한다. 또한 각 부차모집단에 대한 추정값도 계산한다.

□ 기본형식 - 모평균의 추정

```
.svy, [분산추정방법] [,svy_options] :mean 변수리스트 [if] [in] [over] [,options]
```

□ 기본형식 - 모비율의 추정

```
.svy, [분산추정방법] [,svy_options] :proportion 변수리스트 [if] [in] [over] [,options]
```

□ 기본형식 - 모총계의 추정

```
.svy, [분산추정방법] [,svy_options] :total 변수리스트 [if] [in] [over] [,options]
```

□ 분산추정방법의 선택사항

표준오차의 계산법	설 명
linearized	테일러 선형화 분산추정방법을 적용
brr	BRR 분산추정방법을 적용
jackknife	잭나이프 분산추정방법을 적용

☐ svy_options

선택사항	설 명
subpop()	부차모집단의 정의
brr_options	추가적인 BRR 분산추정방법의 선택사항
jackknife_options	추가적인 잭나이프 분산추정방법의 선택사항
levels(#)	신뢰계수의 지정. 기본값은 level(95) 이다.

☐ options

선택사항	설 명
stdize(변수명)	표준화를 위한 층화변수의 지정
stdweight(변수명)	표준화 가중변수의 지정
nostdrescale	표준 가중치 변수의 rescale 을 하지 않음
over(변수리스트)	다중 부차모집단의 정의

(4) 비추정과 회귀추정

비추정과 회귀추정을 위해 "**svy:ratio**"와 "**svy:regress**" 명령어를 제공한다.

① svy:ratio 명령어

조사자료의 유한모집단 총합에 대한 비추정치와 표준오차를 제공한다. 분산추정방법의 선택사항과 svy_options, options 사항은 svy:mean 과 같다.

☐ 기본형식

```
.svy : ratio [새로운 변수이름:] 변수명1 [/] 변수명2
```

□ 완전형식

```
.svy, [분산추정방법] [,svy_options] :ratio ( [새로운 변수:] 변수명1 [/] 변수명2)
                                                    [if] [in] [,options]
```

② svy:regress 명령어

조사자료에 대해 전통적인 회귀분석을 수행한다. 이때 종속변수와 독립변수를 지정하며, 종속변수는 조사자료에서는 관심변수가 되며 독립변수는 보조변수의 역할을 한다. 이와 같은 회귀분석으로부터 얻어진 추정량은 일반화회귀추정량(generalized linear regression : GREG estimator) 가 된다.

□ 기본형식

```
.svy, [분산추정방법] [,svy_options] :regress 종속변수 [독립변수들] [if] [in] [,options]
```

(5) 사후층화

사후층화는 표본의 대표성을 유지시키기 위한 방법으로 추출가중치를 조정하는 방법 중의 하나이다. 즉 사후층화가중치의 합은 각 사후층내의 모집단 크기를 모두 더한 값과 같다. 따라서 모집단에 대한 표본의 무응답이나 과소대표성에 기인한 편향을 감소시킬 수 있다. svyset 명령에서 poststrata() 와 postweight() 옵션을 지정하여 사후층화를 수행할 수 있다.

□ 기본형식

```
.svyset, poststrat(변수명) postweight(변수명) fcp(변수명)
```

(6) 분산추정

복합표본설계에서의 분산추정을 위해 "**svy brr**", "**svy jackknife**" 명령어를 제공한다.

① svy brr 명령어

조사 자료에 대한 균형반복 분산추정치를 계산한다. 이때 가중치를 "brrweight "에 의해 계산된 값을 사용한다. 각각의 층에 2개의 PSU가 뽑히도록 한 표본설계에 대해 분산을 추정하는 방법이다. 이러한 설계에서는 선형화 분산보다 더 좋은 분산추정치를 구할 수 있다.

□ 기본형식

.[svy] brr exp_list [,svy_options brr_options] : 명령어

□ svy_options

선택사항	설 명
subpop()	부차모집단의 정의
levels(#)	신뢰계수의 지정. 기본값은 level(95) 이다.

□ brr_options

표준오차의 계산법	설 명
hadamard(행렬)	Hadamard 행렬을 정의
fay(#)	Fay의 조정방법을 적용함. fay(0)는 brr과 같음
saving(파일명)	매 반복 때마다 결과를 저장함.
mse	분산추정에 대해 MSE식을 사용함.
verbose	전체 표의 범례를 표현
nodots	반복때 표현되는 dot들의 통제
nondrop	관찰치를 drop 하지 않음
reject(수식)	부적합한 결과의 정의

② svy jackknife 명령어

조사데이터에 대한 잭나이프분산 추정량을 구한다. 이때 잭나이프 반복가중치 "jkrweight" 를 적용한다.

□ 기본형식

```
.svy jackknife exp_list [,svy_options brr_options] : 명령어
```

이때, svy_options는 svy brr과 같다.

□ jackkinfe_options

표준오차의 계산법	설 명
eclass	e(N)에 있는 관찰치의 수
rclass	r(N)에 있는 관찰치의 수
n(exp)	사용된 관찰치의 수
saving(파일명)	매 반복 때마다 결과를 저장함.
keep	pseudo-value 의 보존
mse	분산추정에 대해 MSE식을 사용함.
verbose	전체 표의 범례를 표현
nodots	반복때 표현되는 dot들의 통제
nondrop	관찰치를 drop 하지 않음
reject(수식)	부적합한 결과의 정의

Ⅳ. 적용상의 한계 및 맺음말

1. 표본설계

표본조사에서 가장 먼저 고려할 사항은 해당 표본조사를 왜 수행해야 하는가? 하는 점으로 이는 조사의 목적과 관련된 사항이다. 표본조사에 대한 기본적인 지식이나 경험이 없이 조사과정을 수행할 경우 인력과 예산 및 시간을 낭비할 뿐만 아니라 조사로부터 얻은 데이터는 무의미한 자료가 될 뿐이기 때문이다. 따라서 표본설계의 중요성은 아무리 강조해도 지나치지 않다.

또한 표본설계과정에서 고려할 사항으로는 먼저 유사한 조사설계가 이루어 졌는가를 고려하고, 다음으로 주요변수의 오차의 한계 또는 허용오차를 얼마로 생각하고 있는가? 이다. 이는 조사해야할 표본의 크기와 밀접한 관련이 있기 때문이다. 두 번째로 본 조사 이외에 다른 조사에 현재 추출할 표본이 이용되는가?, 즉, 다목적 표본인지 아니면 현재 조사만을 위한 표본인지를 고려하는 것이다. 만일 다목적 표본이라면, 현재 조사를 통해 고려해야할 변수 이외에 추후에 고려해야할 변수까지 현재의 표본설계과정에서 함께 고려하는 것이 바람직하기 때문이다. 셋째는 표본추출방법을 결정해야한다. 층화 추출방법을 가장 많이 사용하는데, 과연 층화 추출설계를 고려한다면 층화의 기준변수는 어떤 변수를 사용해야 하는지, 또는 층을 몇 개로 고려해야 하는지에 관한 고려가 있어야 할 것이다. 너무 많은 층을 고려할 경우에는 층화 추출을 이용함으로서 추정량의 효율성이 떨어지는 문제가 있고, 반대로 너무 적은 수의 층으로 고려하게 되면, 모집단의 특성을 제대로 반영하지 못하는 문제가 발생할 수 있기 때문에 층수의 결정 또한 신중하게 생각해야할 부분이다. 이와 더불어 복합 추출설계를 고려한다면, 집락변수로는 어떤 변수를 고려하는 것이 타당한지? 또는 PSU를 어떤 방법으로 추출하는 것이 현재 정의된 모집단의 분포를 적절히 반영할 수 있는지? 등 표본 추출방법에 대한 신중한 고려가 필요하다.

대체로 가구표본인 경우에는 층화 집락 추출설계가 일반적이며, 기업체 또는 사업

체 표본인 경우에는 층화 추출설계가 사용되고 있다. 이러한 측면에서 본 매뉴얼은 어떤 추출방법이 가장 최선의 표본추출방법인가를 결정하는 것이 아니라, 조사의 목적과 예산, 시간, 등 제반 여건을 고려하여 사용가능한 표본추출방법을 제시함으로서 이용자 측면에서 다양한 표본추출방법을 고려할 수 있음을 언급하고자 한다. 매뉴얼 내용에서도 지적한 바 있지만, 최적의 표본설계는 존재하지는 않는다. 다만, 조사여건과 제반 비용을 고려할 때 현실성 있고, 모집단을 적절히 대표할 수 있는 표본추출방법을 가장 최선의 표본추출방법으로 제안하고자 한다.

특별히 가구표본의 경우에는 행정구역 또는 조사구 단위의 집락을 PSU로 고려할 수 있고, 서울시, 인천, 경기도 등의 지역은 층으로 고려하는 것이 일반적이다. 즉, 층화기준을 지역으로 고려하고, 지역내 조사구를 집락으로, 해당 표본집락 내의 가구를 최종 추출단위로 고려하게 된다.

또한 기업체 또는 사업체 조사의 경우 종업원 수, 매출액, 업종 등이 층화의 기준으로 사용할 수 있으며, 해당 층에서 기업체 또는 사업체를 단순임의 또는 계통추출에 의해 표본사업체 또는 기업체를 표본으로 추출하는 방법이 일반적이다. 층화기준을 어떤 변수로, 또는 표본사업체 또는 기업체의 수는 적절한 이론에 따라 추출하면 될 것이다. 이전에 유사한 조사가 수행된 적이 있을 경우에는 표본의 크기를 결정할 때 이러한 정보를 적절히 이용할 수 있다. 한편, 표본설계과정에서 흔히 이용되는 통계량으로 변동계수(CV)가 있다. 만일 이전의 조사로부터 CV값을 이용할 수 있다면 현재의 표본설계에 이 정보를 적절히 이용할 수 있다. 주어진 허용오차 또는 목표오차 하에서 CV가 큰 변수를 사용하여 표본수를 결정하는 것이 타당하다. 변동이 큰 변수일수록 추정의 정도를 높이기 위해서는 상대적으로 많은 표본이 요구되기 때문이다. 이와 함께, 표본추출과정이 결정되면, 추정량의 형태 또한 동시에 결정해야 한다. 표본설계가 적절히 반영된 추정량은 편향이 없기 때문이다. 대부분 대규모 조사의 경우 복합 표본추출방법을 적용하고 있으며, 추정량 또한 통상적인 선형추정량(linear estimator)의 형태가 아니기 때문에 분산추정량을 구함에 있어 별도의 추정식을 고려하는 것이 바람직하다. 결과적으로 표본설계과정은 사전의 이용 가능한 정보를 충분히 활용하여 허용오차 범위 내에서 적절한 표본수를 구하고, 모집단의 분포를 적절히 대표할 수 있는 표본을 추출하기 위해 필수적인 과정이라 할 수 있다. 따

라서 표본조사를 수행하고자 한다면, 반드시 적절한 표본설계가 기본이 되어야 하며, 이러한 과정에 전문가의 의견을 반영함으로서 보다 정도 높은 추정량을 구할 수 있을 것으로 본다. 표본설계는 최선(best)의 방법을 구하는 것이지, 최적(optimal)의 방법을 구하는 것이 아님을 항상 생각해야 하며, 따라서 다양한 설계방법을 함께 고려하는 것이 바람직하다고 할 수 있다.

2. 표본추출방법

추출틀(sampling frame)로부터 어떠한 표본을 추출해야 하는가는 매우 자명하다. 즉, 모집단의 분포를 적절히 대표할 수 있는 표본을 추출하는 것이 최선의 표본추출방법이라 할 수 있다. 따라서 다양한 표본추출방법을 고려할 수 있으며, 때에 따라서는 비확률 표본추출방법을 고려해야 하는 경우도 있다. 그러므로 표본설계과정에서 항상 고려해야 하는 것이 모집단 분석이다. 모집단 분석을 통해 가장 최선의 표본추출방법을 고려할 수 있으며, 이와 더불어 이용 가능한 정보, 그리고 추정하고자 하는 변수가 어떤 것인지에 대한 사전 고려가 충분히 전제되어야 한다.

표본추출과정에서 가장 범하기 쉬운 잘못 중의 하나는 층화추출(stratified random sampling)과 할당추출(quota sampling)간의 오해이다. 층화추출방법은 확률추출방법으로서 모집단을 층화하고, 모집단의 층에서 특정한 수의 표본을 단순임의 추출하는 방법이며, 할당추출은 정해진 표본수가 될 때까지 조사를 수행하는 방법이다. 즉, 층화추출은 표본을 추출하는 과정에 임의성(randomness)이 반영된 방법인 반면, 할당추출은 추출과정 보다는 조사과정에서 정해진 표본 수만큼을 채우는 조사 또는 면접(interview)성공 수에 관심이 있는 방법이다. 조사통계학자들의 대부분은 이러한 할당추출의 문제점을 매우 심각하게 고려하지만, 실제로 대다수의 여론조사 기관에서는 이러한 측면을 고려하지 못하고 있는 듯 하다. 향후 본 매뉴얼을 조사과정에 적용할 경우 두 방법간의 오해를 어느 정도 해소하기를 기대하는 바이다.

다음으로 가구표본에서 고려할 표본추출 방법으로 층화 집락 추출방법이다. 물론 보다 복잡하게 층화2단계 집락 또는 층화 다단계집락 추출방법 등을 고려할 수 있을

것이다. 앞에서도 언급한바 있지만, 어떤 방법이 최적이다 라고는 할 수 없으며, 이론적으로도 존재하지 않는다. 하지만, 조사과정에서 최선의 표본추출방법을 고려해야 한다면 모집단 분석을 통해 모집단을 가장 잘 대표할 수 있는 표본추출방법을 고려해야 한다. 물론 이전의 조사가 수행된 적이 있다면, 현재의 표본은 어느 정도 시계열성을 유지해주는 방법도 고려하는 것이 바람직 할 것이다.

적절한 예로서 층화 추출의 경우에도 이용가능한 정보가 무엇인지에 따라 표본의 배분문제 또한 여러 가지로 고려할 수 있듯이, 결정적(deterministic)으로 어떠한 조사에는 어떠한 표본추출방법이 최적이라는 공식은 성립하지 않는다. 따라서 많은 경험과 표본조사 이론을 통해 합당한 표본추출방법을 결정해야 할 것이며, 이는 전문가적인 견해를 충분히 반영하는 것도 고려할 만하다.

3. 가중치 및 추정식

수많은 표본조사 이론을 다룬 교재에서는 정해진 표본추출방법에 따른 추정량과 분산 그리고 분산추정량을 제시하고 있다. 이론적으로 이러한 추정식 또는 계산식을 실제 표본 추출이론에 적용하기란 매우 어렵다. 왜냐하면, 대부분의 현실세계에서의 표본조사는 단순히 한 가지 추출방법만을 사용하지 않기 때문이다. 이론적으로 성립하는 추정산식은 실제 표본이론을 연구하는 바탕이 되지만, 실제로 적용하기에는 많은 한계가 있다. 따라서 최근에는 복합 표본설계의 경우 복잡한 계산식을 지양하고, 가중치를 고려한 추정산식을 적용하고 있다. 각각의 표본추출과정에서 자체가중표본이 되지 못하기 때문에 각 단계별로 가중치를 고려함으로서 추정량의 편향을 제거할 수 있기 때문이다. 이와 더불어 조사과정에서의 무응답에 대한 추정식상의 반영이 중요하다. 앞에서도 다룬바 있지만, 가중치를 고려할 때는 먼저 추출확률, 무응답 조정, 사후층화 등을 함께 고려해야 하며, 이러한 측면에서 추정산식은 가중치를 고려한 가중평균, 가중총합과 같은 추정량을 도입하는 것이 이론적으로 타당하다. 이에 따라 분산추정 또한 비선형(nonlinear)형태를 취하게 됨으로 테일러분산(Taylor variance), 또는 잭나이프 분산(Jackknife variance)과 같은 점근적인 방법으로 추정오

차를 계산하고 있다. 가장 최선의 방법은 자체가중이 되도록 설계하면, 전통적인 추정산식을 적용할 수 있으며 통계 분석 프로그램 또한 간단히 적용할 수 있다. 그러나 복합 표본추출의 경우 다양한 가중치를 동시에 고려해야 하며, 그에 따라 추정산식 또한 가중추정식이 사용되게 되며, 결과적으로 분석 프로그램 또한 그에 적합한 프로그램을 이용해야 한다. 최근 통계분석 S/W에서는 이러한 측면을 고려한 모듈이 많이 개발되어 있으므로 이를 적절히 활용하는 것이 표본추출과정을 반영한 추정치를 얻을 수 있다.

4. 표본조사자료 분석 프로그램

대표적인 통계분석 프로그램으로는 SAS, SPSS, STAT, SUDAAN, R 등으로서 현재 국내외적으로 많이 사용되는 것으로 판단된다. 이들 중 복합 표본설계에 적합하게 모듈화 되어 있는 프로그램은 그리 많지 않으며, 대부분 필요에 따라 프로그램을 작성해서 사용해야 하기 때문에 해당 분야에 전문가가 아니면 표본 추출과정을 프로그램에 나타내기는 그리 쉬운 일은 아니다. 그러나 일부 통계분석 프로그램에서는 이러한 점들을 감안하여 특히 분산추정과 관련된 부분에 대해서는 어느 정도 프로그램에 대한 숙지가 있으면 충분히 사용할 수 있도록 설계되어 있다. 이러한 측면에서 본 보고서에서 다룬 몇 가지 분석 프로그램은 실제 표본설계담당 실무자 수준에서 무난하게 다룰 수 있는 내용만을 선택적으로 편집하여 나타낸 것이다. 프로그램의 많은 부분은 현실적으로 수정 보완하여 사용해야 하며, 가장 기본적인 내용만을 수록했음을 명시하고자 한다.

이와 관련하여 특별히 프로그램 부분에서 강조하는 것은 분산추정의 문제로서 가중추정량의 형태가 비선형(nonlinear)형태를 취하게 됨으로 테일러분산(Taylor variance), 또는 잭나이프 분산(Jackknife variance)과 같은 점근적인 방법으로 추정오차를 계산해야 함으로 그에 따른 통계 분석 프로그램을 적용해야 한다는 것이다.

5. 맺음말

표본조사론 또는 조사통계방법론은 통계학의 한 분야로서 최근 표본조사의 중요성이 대두됨으로 인해 다양한 학문에서 새롭게 조명되는 분야이다. 고전적인 표본추출이론으로부터 무응답 대체 방법, 가중치 조정방법등과 같은 현실적으로 적용해야할 새로운 학문분야들이 개발되고 있다.

이러한 측면에서 본 매뉴얼은 다양한 표본조사 실무자들의 이론적인 바탕 하에서 이용 가능한 지침서가 되기를 바란다. 표본추출 또는 조사방법론에 대한 초보자들이 현실적으로 표본설계와 표본 관리를 한다는 것은 조사로부터 생산된 데이터의 신뢰성을 의심하게 되는 단초를 제공할 뿐이다. 표본설계의 전문가는 아니더라도, 최소한의 표본추출관련 교육을 받은 실무자가 담당 업무에 대한 지침서 또는 업무의 평가를 위한 체크리스트가 되기를 바라며, 이러한 측면에서 본 매뉴얼은 조사통계 실무자들에게 보다 폭넓은 이론적 바탕을 제공하는 한편, 표본조사의 오용을 막는데 최소한의 기준이 되기를 바란다.

참고문헌

김영원 외 8인, (2005), 조사방법의 이해, 한국통계학회 조사통계연구회.

김영원, 류제복, 박진우, 홍기학, (2000), 표본조사의 이해와 활용, 자유아카데미.

김종호 외(2003) , 표본조사 입문 , 자유아카데미.

박홍래 (1993), 통계조사론, 영지문화사.

이계오, 박진우, 이기재, (2004), 표본조사론, 한국반송통신대학교 출판부.

장인식 (1996), 표본조사론, 다산출판사.

통계청, 한국통계조사현황(2006), 통계청.

Australian Bureau of Statistics, (1999), ABS Survey manual, Australian Bureau of Statistics.

Bennett, S.(1993), " The EPI Cluster Sampling Method: A Critical Appraisal, " Invited Paper, International Statistical Institute Session, Florence.

Cochran, W.(1977), Sampling Techniques, 3rd Eds, Wiley, New York.

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

Brick, J. M. and Kalton, G.(1996), Handling Missing Data In Survey Research," Statistical Methods in Medical Research, Vol.5, 215-238.

Chamber, R.L., Skinners, C.J.(2002), Analysis of Survey Data, John Wiley & Sons.

Census of Bureau Statistics,(2002) CPS technical report, Design and Methodology TP63RV, Census of Bureau Statistics.

Chaudhuri, A., and Stenger, H. (1992), Survey Sampling- Theory and Methods, Marcel Dekker Inc, New York.

Cochran, W. G.(1976), Sampling Techniques, 3rd Eds. John Wiley & Sons.

Faraway, J. J.(2005), Linear Models with R, Chapman & Hall.

Groves, R.M., Fowler, F. J., Couper, M.P., Lepkowski, J.M., Singer, E., and Tourangeau, R.(2004), Survey Methodology, John Wiley & Sons.

Hansen, M., Hurwitz, W. and Madow, W. (1953), Sample Survey Methods and Theory, Wiley, New York.

Kalton, G.(1983), Introduction to Survey Sampling, Sage, Beverly Hills.

Kalton, G. Heeringa, S. (2003), Leslie Kish selected paper, John Wliey & Sons.

Kish, L. (1995), Survey Sampling, John Wiley & Sons.

Lehtonen, R, Pahkinen, E. (2004), Practical Methods for Design and Analysis of Complex Surveys, 2nd Eds. John Wliey & Sons.

Lessler, J. and Kalsbeek, W.(1992), Nonsampling Error in Surveys, John Wiley & Sons, New York.

Levy, P.S. and Lemeshow, S.(1999), Sampling of Populations : Methods and Applications, 3rd Eds, John Wiley & Sons, New York.

Lohr, S. L., (1999), Sampling design and Analysis, Duxbury Press.

Raj, D.(1972), Design of Sample Surveys, McGraw-Hill, New York.

Statistics Canada. 2003. Survey Methods and Practices. 12-287-XPE.

Sarndal, C.E., Swensson, B., and Wretman, J.(2003), Model Assisted Survey Sampling, Springer, New York.

Sarndal, C.E., Lundstrorm S., (2005), Estimation in Surveys with Nonresponse, John Wliey & Sons.

Scheaffer, R. L. Mendenhall, W., and Ott, R. L. (2006), Elementary Survey Sampling, Duxbury Press.

Skinner, C.J. (1989), Analysis of Complex Surveys, John Wliey & Sons.

STATA survey Reference Manual, Release 9, StataCorp, Texas.

Rao, J.N.K. (2003), Small Area Estimation, John Wiley & Sons.

Thompson, S.K. (2002), Samping, 2nd Eds, John Wiley & Sons.

United Nations, Outline of the Draft Handbook on Designing of Household Sample Surveys, (2005), United Nations Statistics Division.

Valliant, R., Dorfman, A.H., and Royall, R. M.(2000), Finite Population Sampling and Inference - A prediction Approach, John Wiley & Sons.

Verma, V.(1991), Sampling Methods, Training Handbook, Tokyo.
Verzani, J. (205), Using R for Introductory Statistics, Chapman & Hall/CRC.
Westat, (1994). Current Best Methods Manual, Westat Inc.
Wolter, K.M.(1985), Introduction to Variance Estimation, Springer-Verlag, New York.

부 록

A1. 표본설계의 기본 개념

A1.1 표본조사의 발달 과정

19C 까지 정부나 사회단체에서 통계조사라 하면 전수조사(census)로 생각되었다. 일부 표본조사의 개념을 이용하여 부분적인 조사가 수행되었지만, 현재의 표본조사와 동일한 개념의 표본조사는 Kiaer(1890)이 제안한 대표표본(representative sample)이 처음이라고 언급하고 있다. 그 이전에는 표본조사의 합리성에 관한 확신이 없는 상태였으나, Kiaer 이후 표본의 대표성이라는 매우 합리적인 논리를 제공함으로서 근대의 표본조사의 틀을 마련하였다고 할 수 있다. Kiaer은 임의화 또는 확률화(randomization)를 "제비뽑기"로부터 가능하다고 언급하였다. 그 이후 확률화(randomization)를 통계조사에 최초로 도입한 사람이 Bowley이며, 그는 유의추출법(purposive sampling)을 동시에 제안한 사람이다. 또한 층화추출에서 비례배분(proportional allocation)의 문제를 다루기도 하였다.

다음으로 Neyman(1934)은 현재 표본조사의 기틀을 만든 사람이라 할 수 있다. Neyman은 첫째, 임의추출법이 유의 추출법에 비해 효과적이라는 사실을 이론적으로 증명하였으며, 유의추출법의 대표적인 모형이 선형 회귀모형임을 보였다. 둘째, 점추정과 구간추정의 이론을 제시하였다. Neyman의 영향으로 미국의 정부통계생산에서 표본조사방법을 집중적으로 연구하는 계기가 되었으며, 이 당시 주도적인 역할을 한 사람이 Hansen과 Hurwitz이다. 1933년 이전에는 공식적으로 미국에서 표본조사를 수행한 적이 없으며, 1940년에 비로소 정치여론조사에 표본조사를 이용하였다(Duncan 과 Shelton, 1978). Hansen과 Hurwitz는 현재의 복합표본추출 이론의 기틀을 마련한 사람들이며, 이들 이외에도 Fisher와 Cochran 등이 표본조사이론의 발전에 크게 공헌한 사람들이라 할 수 있고, Mahalanobis의 '다단계표본추출이론과 실제'는 현재의 복합표본추출이론의 근거를 마련했다고 할 수 있다.

이러한 표본설계에 근거한 추론 방법에 대해 새로운 방향을 모색한 사람은 Godambe(1955)으로서 표본조사에서 추정이론의 근거를 제시하였다. 특히 그는 유한모집단에서 평균에 대한 UMVE(Unbiased Minimum Variance Estimator)가 존재하지 않는다는 모형접근이론을 모색하였다. 이와 더불어 모형접근이론자로는 Royall(1970)을 들 수 있으며, 그는 유의추출법 이론을 다시 부활시킨 장본인이라 할 수 있다.

종합적으로 볼 때, 현재 표본조사이론에는 앞에서 언급한 바와 같이 설계접근(design-based)방법과 모형접근(model-based)방법이 있으며, 이들 두 방법을 적절히 병행하여 사용하는 것이 바람직하다고 할 수 있을 것이다. 즉, 대규모 다항목 조사에서는 설계접근 방법이 더 효율적이며, 소규모 소항목 조사이고 보조정보를 이용할 수 있는 경우에는 모형접근방법이 유효한 것으로 나타났다. 따라서 일반적인 가구조사나 기업체 조사 등에서는 설계접근 방법을, 소지역 통계 생산 등에서는 모형접근 방법이 유효한 방법이라 할 수 있다.

A1.2 표본설계의 개념

일반적으로 표본을 대상으로 하는 조사과정에서 가장 중요한 부분이 추출틀(sampling frame)로부터 단위를 추출하는 것이다. 전체적인 조사과정에서 ① 잘 설계된 조사표, ② 고도로 훈련된 조사원, ③ 유능한 조사현장 관리원, ④ 수집된 자료에 맞는 자료처리방법, ⑤ 잘 설계된 편집단계 등이 필수적으로 요구된다. 그러나 만일 조사에 필요한 표본을 연구자의 주관에 따라 인위적으로 추출한다면, 조사의 목표가 되는 모집단에 대한 추론에 대한 의미 없는 자료가 될 것이다.

따라서 표본설계는 표본조사의 성패를 좌우하는 핵심적인 사항 중 하나로서 중점적으로 고려해야 할 사항들은 다음과 같다고 할 수 있다.

첫째, 수집한 데이터에서 계산된 추정치에 대한 목표정도를 사전에 결정해야한다. 만약 통계 이용자가 요구하는 목표정도를 수치적으로 표현할 수 없는 경우에는 통계학자와 협의하여 허용된 예산한도에서 달성할 수 있고, 수량적으로 표현 가능한 목표정도를 정해야 할 것이다. 둘째, 추출틀과 추출단위 및 조사단위를 결정해야 한다. 통상적으로 추출단위와 조사단위는 다르며, 추출단위와 조사단위의 결정은 적용할

표본추출법의 결정에도 영향을 미치기 때문이다. 또한 추출단위는 조사목적과 밀접한 관계가 있으므로 조사목적에 부합하도록 선정되어야 할 것이다.

셋째, 조사비용과 목표정도를 결정해야 한다. 이 두 가지 요소가 결정되면 표본의 크기는 쉽게 계산할 수 있으며, 일반적으로 요구정도를 만족하는 범위 내에서 표본의 크기를 최소로 하면 조사비용이 최소가 될 것이고, 반대로 정해진 조사비용 내에서 오차를 최소로 하려면 정해진 비용 하에서 표본의 크기를 최대로 해야 하기 때문이다. 단순하게 생각하면 표본의 크기를 증가시켜서 조사결과에 대한 정도(precision)를 높이면 좋을 것으로 생각하기 쉽지만 이는 잘못된 인식이며, 작성된 통계를 이용하는 측면에서 5%오차를 허용했을 경우에는 그 이상의 정도를 갖는 통계는 큰 의미가 없다. 표본의 크기가 커지면 조사비용만 증가하게 되고, 조사원의 업무가 증가되어 조사된 자료의 질이 떨어질 수도 있다는 점도 간과해서는 안 된다.

이론적으로 표본의 크기를 결정하는 것과 실제로 표본조사를 실행할 때와는 차이가 있을 수 있으며, 조사원이 행정기관에 있는 공무원일 경우나 행정구역별로 표본배정을 균형있게 해야 하는 경우에는 이론적인 계산과는 약간 다를지라도 표본조사가 원만하게 정상적으로 이루어지도록 해야 한다.

표본설계가 위와 같을 사항을 고려하여 결정되면, 그에 따른 표본추출방법을 결정해야 하는데, 다음 사항들을 고려하는 것이 바람직한 표본추출방법이라 할 수 있다.

- 표본오차는 계산가능하며 최소화할 수 있는가?
- 추정량과 그에 대한 분산(추정오차)은 직접 계산 가능한가?
- 가용예산과 인원의 제약 아래에서 요구정도를 달성할 수 있는가?
- 표본조사를 실행하는 과정이 용이한가?
- 유사한 조사방법을 과거에 실행한 적은 있는가?

이러한 전반적인 표본설계의 내용은 전문적인 표본추출이론분야 이기 때문에 반드시 전문가의 자문을 통해 가능한 정도 높은 통계를 생산하는데 필요한 표본이 추출되도록 해야 할 것이다.

A1.3 표본관리의 개념

계속조사에서 시간이 지날수록 표본에도 여러 문제가 생길 수 있으며, 표본단위의 소멸, 응답불응 등이 대표적인 예가 될 수 있다. 추출틀 보완에 따라 신규표본이 발생하거나 또는 감소되는 경우가 생길 수가 있다. 이런 경우 매 조사마다 일정한 표본크기를 유지하는 것이 현실적으로 불가능할 때도 있다.

일반적인 표본 관리의 방법으로는 표본의 대체(substitution), 추가(addition), 삭제(deletion), 표본개편 등이 있다. 표본의 대체는 조사단위에 대한 조사가 더 이상 개의치 않을 때에 유사한 다른 조사단위로 대체하는 것을 의미한다. 추가나 삭제는 모집단의 변동을 표본에 반영하기 위해 고려하게 된다. 국내의 표본개편은 표본의 모집단에 대한 대표성이 떨어졌다고 판단될 때 실시하는데 인구주택총조사 주기에 따라 5년에 한번은 표본을 개편하는 것이 일반적이다.

A1.4 확률표본추출

확률표본추출(probability sampling) 또는 임의표본추출은 유의표본추출의 상반된 개념으로 인식되어 오다가 최근에는 모형-기반 접근과 설계-기반 접근의 방식이 대두되면서 서로 양립하고 있는 이론으로 개념화되고 있다. 즉, 확률추출만이 표본조사의 유일한 표본추출방법이 아니며, 경우에 따라서는 보다 효과적으로 비확률추출 또는 유의표본추출이 적용될 수 있다는 의미이다. 그러나 통상적으로 표본추출이라 함은 대개의 경우 확률표본추출을 의미하고 있으며, 이러한 확률추출방법에는 단순임의추출법, 층화추출법, 계통추출법, 집락추출법 등을 대표적인 확률추출법으로 고려할 수 있다. 이와 같은 확률추출방법을 사용함으로서의 장점은 미지의 모집단 정보를 통계적으로 추론할 수 있다는 점과, 비편향 추정량을 구할 수 있다는 점, 알려진 통계적 이론을 적용할수 있다는 점 등이다. 즉, 설계-기반 추론방법에 따라 확률표본으로부터 도출된 추정값이 이론적으로 모집단의 대표값으로 적용될 수 있음을 의미한다. 이러한 관점에서 확률추출이론은 매우 중요한 표본추출 이론으로 발전하고 있으며, 현재 대다수의 표본조사에서는 이와 같은 확률추출방법을 적용하고 있다.

A1.5 표본오차와 비표본오차의 관리

통계조사에서 발생하는 오차를 크게 표본오차(sampling error)와 비표본오차(nonsampling error)로 구분한다. 전자는 모집단을 조사하지 않고 그의 일부를 조사함으로서 발생하는 오차로서 모집단의 일부 개체를 추출 조사하여 측정 및 공표하기까지 오차가 전혀 개입되지 않는다고 가정하였을 때 모집단 값과 차이가 난다면 이는 표본오차로 인한 것이다. 표본오차 이외의 요인에 의해 발생하는 오차를 비표본오차라 한다. 비표본오차 발생요인 중 중요한 것은 무응답과 응답 중에서 고의적으로 잘못 응답하였거나 거짓응답 등으로 발생하는 측정오차(measurement error)를 들 수 있다. 통상적으로 표본조사의 단계를 다음과 같이 4개의 단계로 구분한다(Murthy, 1967).

1단계) 조사 개념 설정단계
2단계) 표본추출단계
3단계) 자료수집단계
4단계) 자료처리 및 추론단계

이때 1단계와 2단계에서 발생하는 오차를 설정오차(specification error), 3단계 에서 발생하는 오차를 확인오차(ascertainment error), 4단계에서 발생하는 오차를 처리오차(processing error)라 한다.

A1.6 무응답처리 방법

조사과정에서 모든 조사관계자들이 아무리 노력한다고 해도 무응답 사례는 생기게 마련이다. 이러한 무응답은 전체 조사의 일정 및 조사의 질에 영향을 미친다. 따라서 사전에 미리 무응답에 대한 대책을 세워두는 것은 조사읜 품질을 일정 수준 이상으로 유지하고 관리하기 위해 매우 필요하다.

무응답은 크게 단위무응답(unit nonresponse)과 항목무응답(item nonresponse)으로

구분된다. 단위무응답이란 응답자가 조사 자체에 불응한 경우에 생기는 것이고 항목무응답은 전체 조사항목 중 일부 조사항목에 대해 응답을 않은 경우이다. 무응답은 표본의 크기를 원래 목표한 것보다 작아지게 함으로 조사의 효율에 영향을 미치고, 무응답이 어떤 경향성을 띠게 되는 경우 추정값의 편향을 초래할 수 있다.

가능한 한 무응답이 발생하지 않도록 하는 것이 일차적인 관심이어야 하지만 부득불 무응답이 발생하였을 경우 이에 대해 적절히 대처하는 것 또한 중요하다.

<표 A1-1-1> 무응답 유형과 처리방법

무응답 유형	처리방법
단위무응답 (Unit Nonresponse)	가중셀 조정(Weighting adjustment)
	래킹비 조정(raking ratio adjustment)
	보정방법(Calibration)
항목무응답 (Item Nonresponse)	대체(Imputation)

A1.7 대체(Imputation)

데이터편집과 대체(imputation)를 혼동하여 대체라는 용어로 사용하는 경우가 있는데 점점 두 용어를 구분하는 추세이다. 데이터편집은 오류를 점검한 후 논리에 따라 기계적으로 수정하는 과정까지를 일컫는 것에 반해 대체는 무응답을 그럴싸한 값으로 대체시키는 작업이다. 데이터편집에서의 수정을 할 경우 선택의 여지없이 값이 결정되지만 대체에서는 사람에 따라 동일한 무응답을 다른 값으로 채울 수 있는 것이다.

대체를 하고서는 그 값을 마치 관찰된 값인 것처럼 간주할 경우 추정에 심각한 문제가 생길 수 있다는 사실이 널리 알려지고 있다. 따라서 특정 데이터를 대체할 경우 이 값이 대체된 값인지 관찰된 값인지를 명확히 밝혀 두어야 한다. 대체를 할 경우 추정과정에서 이론적으로 고려해야 할 요소가 많으므로 유의해야 한다.

일반적으로 모든 조사자료를 수집한 후 중앙에서 무응답 대체를 하는 것이 바람직

하다. 만일 일선 조사현상에서 각각 대체를 하도록 한다면 방법이나 기준이 달라질 우려가 많고 또 대체층을 만드는 면에서도 효과적이지 못하다.

다양한 대체방법들을 고려할 수 있으므로 이를 위해 간략하게 각 방법들의 특징을 살펴보면 다음과 같다.

□ 평균값 대체(mean imputation)
전체 표본을 몇 개의 대체층으로 분류한 다음 각 층에서 응답자의 평균값을 그 층에 속한 모든 결측값에 대체하는 방법이다. 이 방법은 매우 사용하기 편리하고, 평균이나 총합과 같은 단변량 모수에 대한 점추정량의 편향을 감소시키는데 계산상 또는 비용측면에서 상당히 효과적이다. 그러나 각 층에서 응답자들의 평균 한 개의 값으로 대체됨으로 실제 모집단의 경험적 분포를 상당히 왜곡시킬 수 있다.

□ 핫 덱 대체(hot deck imputation)
대체층 내에서 대체값을 확률추출에 의해 임의로 선택하여 결측값을 대체하는 방법이다. 이 방법은 평균값 대체방법이 모집단 분포를 왜곡시킬 수 있다는 문제점을 완화시킬 수 있다.

□ 콜드 덱 대체(cold deck imputation)
이전의 조사 자료나 역사적 자료와 같은 다른 정보로부터 얻은 값으로 결측값을 대체하는 방법이다. 대체되는 자료가 현재의 자료가 아니라는 점에서 콜드 덱이라고 부른다. 핫 덱과 마찬가지로 선택편향을 줄여주는 효과가 있다.

□ 최근방 대체(nearest neighborhood imputation)

Brooks와 Baliar(1978)에 의해 미국의 경상인구조사(CPS) 에서 발생한 결측자료를 대체하기 위한 방법으로 고안되었으며, 전체표본을 대체층으로 나눈 뒤 각 층에서 응답자료를 순서대로 정리하여 결측값이 있는 경우 그 결측값 바로 이전의 응답을 결측값 대신 대체하는 방법이다. 응답그룹에서 무응답된 자료의 특성과 가정 유사한 자료를 구하여 대체하며 이때 최초의 자료가 결측일 때에는 최초 응답자료를 결측값 대신 대체한다.

□ 회귀대체(regression imputation)

무응답이 있는 항목 y에 응답이 있는 y의 보조변수 $x_1, x_2, \cdots, x_k$를 회귀모형에 적합시키는 방법으로 i번째 결측값에 대해 대체값을 다음과 같이 구한다.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k x_{ki} + e_i$$

여기서 $\hat{\beta}$는 OLS 방법으로 추정한 회귀계수이다.

□ 비대체(ratio imputation)

조사변수 y와 높은 상관관계가 있는 보조변수 x가 있을 때, 무응답 항목 i에 비추정치를 대체하는 방법으로 회귀대체와 유사한 방법으로 무응답 항목을 대체한다. 즉 i번째 무응답 항목에 대한 비대체값은 다음과 같다.

$$y_i^* = \frac{\bar{y}_r}{\bar{x}_r} x_i$$

□ 다중 대체(multiple imputation)
앞에서 언급한 대체방법과는 다르게 각 결측항목에 대한 대체값으로 2개 이상의 가능한 벡터로 대체하는 방법으로서 하나의 결측값에 대한 대체값으로 가능한 여러 가지 경우에 따라 생성된 대체값을 대상으로 결측값으로 구하는 방법이다.

A2. 국가통계의 표본조사 현황 및 문제점

A2.1 현황파악

2006.6월 현재 통계청 승인 통계의 현황을 파악한 결과가 다음과 같다. 총 515개 통계 중에서 280여개의 조사통계가 생산되고 있으며, 정부기관에서 생산되는 조사통계는 180여종에 이르며, 통계청 이외의 여타 기관에서 작성되는 조사통계의 수는 239여종에 달하고 있다.

<표 A2-1-1> 2006년 6월 현재 국가 승인통계 현황

(단위 : 종)

부문	작성 기관수	계	승인통계				
			통계종류		작성방법		
			지정통계	일반통계	조사통계	보고통계	가공통계
계	139	515	91	424	280	176	59
정부기관	63	367	75	292	180	142	45
중앙행정기관	31	262	59	203	134	110	18
(통계청)	(1)	(52)	(36)	(16)	(41)	(1)	(10)
지방자치단체	32	105	16	89	46	32	27
지정기관	76	148	16	132	100	34	14

(통계청 홈페이지: www.nso.go.kr)

표본조사는 국내뿐만 아니라 국외 특히 미국과 캐나다가 대표적인 조사통계관련 이론과 실제 업무 매뉴얼을 작성하여 수행하고 있는 국가이다.

<표 A2-1-2> 분야별 생산 통계현황

(단위 : 종, %)

부문	작성 계수		작성방법별		
	총	구성비	조사통계	보고통계	가공통계
계	515(52)	100.0(100.0)	280	176	59
인구	28(8)	5.4(15.4)	4	19	5
고용・임금	34(1)	6.6(1.9)	31	3	-
물가・가계소비	16(5)	3.1(9.6)	16	-	-
보건・사회・복지	83(4)	16.1(7.7)	44	37	2
환경	22(0)	4.3(0.0)	10	11	1
농림・수산	46(8)	8.9(15.4)	27	18	1
광공업・에너지	31(5)	6.0(9.6)	19	8	4
건설・주택・토지	27(4)	5.2(7.7)	14	10	3
교통・정보통신	38(1)	7.4(1.9)	16	22	-
도소매・서비스	13(7)	2.5(13.5)	13	-	-
경기・기업경영	70(4)	13.6(7.7)	54	4	12
국민계정・지역계정	11(2)	2.1(3.8)	1	1	9
재정・금융	19(0)	3.7(0.0)	4	15	-
무역・외환・국제수지	10(0)	1.9(0.0)	4	4	2
교육・문화・과학	43(0)	8.3(0.0)	19	21	3
기타	24(3)	4.7(5.8)	4	3	17

(통계청 홈페이지: www.nso.go.kr)

따라서 대내외적으로 신뢰성과 표준화를 기준으로 할 때 이들 국가의 표본조사관련 매뉴얼을 검토 분석하여 현재 국내의 조사통계현황에 맞도록 벤치마킹뿐만 아니라 자체적인 개발노력이 필요하다고 할 수 있다. 이러한 관점에서 국내의 대표적인 조사통계 중 제반 수행절차를 준수한 조사통계를 발굴하여 소개하며, 미국과 캐나다의 조사통계 중 대표적인 내용을 소개함으로서 통계작성 기관에서 조사통계작성절차 중 참고자료로 활용하게 하고자 한다.

A2.2 문제점

통상적으로 각종 조사통계에서는 조사목적에 따라 조사대상 모집단과 표본에 대한 정의를 서로 다르게 표현하고 있다. 따라서 이와 같이 조사대상의 다양성에 따라 크게 가구 부문과 기업체 부문의 조사로 구분하여 표본추출 및 관리 매뉴얼을 작성하는 것이 바람직하다. 업종관련 통계의 경우 대다수의 조사통계는 “기업체” 또는 “사업체” 대상 조사이며, 그 외에는 대부분 “가구” 대상 조사로 구분할 수 있다.

즉, 현재 통계청 승인 통계 중 조사통계부분을 파악해보면 가구 부문에서는 조사

구는 40-60개의 가구들의 묶음(cluster)으로 나타나고, 사업체 조사의 경우 일정 규모(종업원 수, 또는 매출액 기준)에 따라 사업체를 층화하여 모집단 층별로 해당 표본 사업체를 표본으로 추출하고 있기 때문이다. 따라서 가구부문에서는 추출단위가 집락(조사구)이며, 조사단위는 각각의 PSU 또는 SSU내의 개별 가구가 되지만, 사업체 조사의 경우 추출단위와 조사단위가 모두 해당 표본 사업체가 됨으로 동일한 틀 속에서 가구부문과 사업체 또는 기업체 부문을 같이 다루게 되면 조사단위와 추출단위에 대해 혼동하기 쉽기 때문에 이를 각각 분리하여 정의하는 것이 바람직하다. 참고로 한국통계조사현황집(통계청, 2004년 발간)의 조사통계현황을 분석한 결과 다음과 같은 문제점을 파악할 수 있었다.

첫째, 표본수 산정식, 조사대상, 표본추출단위 등의 정의가 모호하거나 기관별로 용어의 통일성이 결여된 것으로 나타났다.

둘째, 모수추정식이 대부분 표본조사 교과서에서 소개되는 단순추정량 식으로 표현되어 가중치 적용이 거의 되어 있지 않은 것으로 파악되었다.

셋째, 통계작성 기관별로 표본추출방식에 대한 용어가 서로 다르게 표현되고 있다. 예를 들어, 유선무작위계통추출이란 용어는 통계학에서 층화임의 계통 추출로 통일해야할 것이다.

넷째, 추정을 위한 통계조사와 단순집계만을 위한 통계조사를 구분하여 매뉴얼 작성 작업이 필요한 것으로 파악되었다. 왜냐하면 단순집계에는 추정식이나, 추정오차와 같은 내용이 불필요하기 때문이다.

다섯째, 기업체 부문의 경우 사업체와 기업체 용어의 혼용으로 이용자의 혼란이 가중되고 있다. 따라서 기업체 부분의 경우 해당 통계작성 기준을 명확히 제시할 필요가 있는 것으로 파악되었다.

국내 승인통계 중 가계소비조사와 미국, 캐나다의 조사통계와의 비교를 통해 현재 국내 조사통계작성의 기준항목들이 어느 정도 갖추어 졌는가를 평가할 수 있을 것이다. 이러한 비교로부터 조사목적, 모집단의 정의, 추출프레임의 정의, 표본추출방법, 조사범위 및 대상 등에 대해서는 3개국 모두 적절한 수준으로 표현하고 있음을 알 수 있다.

<표 A2-2-1> 미국, 캐나다, 한국의 표본조사사례의 비교

구성요소	국가		
	미국	캐나다	대한민국
조사명칭	Current Population Survey(CPS)	Survey of Family Expenditures(Famex)	가계소비조사
주관부서	미국 통계국(Census Bureau)과 노동통계국(Bureau of Labor force Statistics : BLS)	Statistics Canada	통계청 사회통계과
조사기관	미국 통계국(Census Bureau)	Statistics Canada	통계청
조사목적	·표본가구의 취업 또는 실업 상태를 직접적으로 측정하기 위함. . 미국인들의 노동력과 인구학적 특성을 파악 ·개인의 소득, 노동력, 교육수준 등에 관한 월별 자료 생산	·캐나다에 거주하는 가구의 사회-경제학적인 삶의 조건 을 측정. ·소비자물가지수의 편제에 필요한 가중치 산정 자료.	·도시가구의 수입과 지출을 조사하 여 가구의 생활실태와 변동사항을 명확히 파악. ·국민소비수준변화의 측정 및 분석 을 위한 자료생산. ·소비자물가지수 편제에 필요한 가 중치 산정자료. ·각종 사회정책입안기초자료의 생산. ·국민소득추계기초자료의 생산.
조사시작년도	1942년도	1996년도	1951년도(한국은행) - 2003년도 명칭, 조사대상, 표본가 구수 변경
목표모집단	16세 이상의 미국인	캐나다 거주 가구 및 15세이상 비수용자	전국에 거주하는 가구
추출틀	10년 주기의 센서스데이터로부터 얻은 주소명부(list) Unit 프레임, Area프레임,Group quater 프레임, 신축건물프레임.	LFS 추출틀	2000년 인구주택 총 조사의 10%표본조사구중 섬지역 및 시설단위조사구를 제외한 24,998개 아파트 및 보통조사구 명부
표본추출방법	층화다단계 추출방법	층화다단계 추출방법	다단계 층화 집락 확률비례추출
표본규모	792개 표본지역으로부터 56,000가구를 표본으로 추출	LFS의 54,000가구 (약 100,000명)	7,500가구
면접방법	면접원에 의한 타계식 면접	면접원에 의한 자발적인	자계식 가계부기입 방법
보고단위	표본가구 내 개인(16세 이상)	표본가구 내 모든 개인	읍면동지역 2인 이상 비 농가구
조사주기	4-8-4 순환주기	2년	매월
컴퓨터 사용여부	CAPI 와 CATI	CAPI와 CATI	PAPI
관찰 수준	개인 및 가구	개인 및 가구	가구
웹사이트	http;//www.bls.census.gov/**cps**	http://www.statcan.ca	http://www.nso.go.kr

이를 보다 구체적으로 살펴보기 위해 대표적인 국가로서 미국과 캐나다의 표본 추출 및 관리 방법에 대한 검토와 국내 표본추출방법을 비교해보면 다음과 같은 특징을 지니고 있다.

통계청에서 사용하고 있는 PSU는 약 40-60개로 묶여진 조사구로 정의되며, 2005년도 인구주택총조사 결과 전국 약 27만개 조사구를 확정하였고, 이 중 10% 표본조사구 약 2만7천개 조사구를 통계청에서 별도로 관리하고 있다. 이때, 각 PSU의 크기는 다르며, 지역별(대도시 및 중소도시 등) PSU의 규모 또한 서로 상이하다. 이와 같이 정의된 조사구는 통상적으로 가구단위 조사에서 많이 활용되고 있는 실정이다.

<표 A2-2-2> 통계청, 미국, 캐나다의 표본추출방법의 비교

구분	한국(통계청)	미국/캐나다
PSU	조사구	County이용
USU	집락전체를 추출 - 경제인구: 20가구 가계조사: 10가구	개별가구를 계통추출
최종 조사가구수	10가구	4가구
가구의 조사응답 부담	경제인구과 가계조사를 함께 조사함으로 중복 부담	경제인구와 가계조사가 다르기 때문에 부담경감
예비표본	조사구내 5가구씩 순환하여 예비 표본을 선정함.	동일 집락내 예비표본을 미리 선정함.

미국 및 캐나다는 county나 block 등을 PSU로 사용하고 있는 것으로 파악되었다. 즉, 미국에서는 3,141개의 county를 이용하여 별도의 PSU(약 3,000여개)를 정의하고 있으며, PSU 설정기준은 다음과 같다.

① 1개의 PSU는 1개의 county 또는 적어도 2개 이상의 county로 구성되며,

② 1개의 PSU의 면적이 3,000평방마일 이하 이거나,

③ 인구수가 7,500명 이하인 경우

이때 PSU의 그룹은 대도시지역의 PSU-self representing PSU와 나머지 PSU는 층(strata)으로 묶어 사용하고 있다. 이에 비해 캐나다에서는 다음과 같은 기준으로 표본을 추출하고 있는 것으로 파악되었다.

① 산업별 취업구조, 연령별 인구구조, 가구원수별 가구, 총소득, 교육수준, 모국어 등의 층화 변수를 활용하여 모집단을 PSU(약 7,000여개)로 층화하여 이들 중 일부의 PSU를 표본으로 추출한다.

② 인구센서스 결과를 이용한 주소명부(address register)를 구축하여 표본추출업무에 부가적으로 활용하고 있다.