

2013년 한국아동·청소년패널 제3차 콜로키움 자료집 데이터분석방법론 표본설계와 추정

일시 | 2013년 12월 17일(화) 10:30 ~ 12:30

장소 | 한국청소년정책연구원 10층 세미나실

주최 | 한국청소년정책연구원

세 부 일 정

「2013년 한국아동·청소년패널2010 데이터분석방법론 세미나」

사회: 이종원 (한국청소년정책연구원 선임연구위원)

시 간	내 용	비 고
10:30~10:40	개회 및 발표자 소개	이종원(사회자)
10:40~10:45	환영인사	이재연 (한국청소년정책연구원장)
10:45~12:00	표본설계와 추정 강의	강현철 (호서대 응용통계학과 교수)
12:00~12:20	질의응답	참석자 전원
12:20~12:30	폐회	이종원(사회자)

목 차

□ 표본설계와 추정	1
□ 한국아동·청소년패널조사2010 I 표본설계	30
1. 조사목적	30
2. 모집단 정의	30
3. 조사방법	30
4. 표본설계	31
4.1 모집단 분석	31
4.2 표본크기	35
4.3 표본할당	36
4.4 표본추출	40
4.5 표본관리	40
5. 가중치와 모수추정	41
5.1 가중치 계산	41
5.2 모수추정	44
6. 결언	45
□ 참고문헌	47
□ 부 록	
1-1. 초등학교 1학년과 4학년 시도별 집락별 분포 현황	51
1-2. 중학교 2학년 시도별 집락별 분포현황	52
2-1. 초등학교 1학년 대상 3차년도 가중치 부여방법	53
2-2. 초등학교 4학년 대상 3차년도 가중치 부여방법	58
2-3. 중학교 1학년 대상 3차년도 가중치 부여방법	62

표본설계와 추정

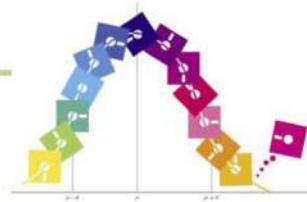
강 현 철

(호서대학교 응용통계학과 교수)

표본설계와 추정

2013. 12. 17.

강 현 철
호서대학교 응용통계학과
hychkang@hoseo.edu



목차

1. 표본설계의 주요 용어
2. 표본추출(표집) 방법
3. 오차한계
4. 모수에 대한 추정
5. 패널조사
6. 사례

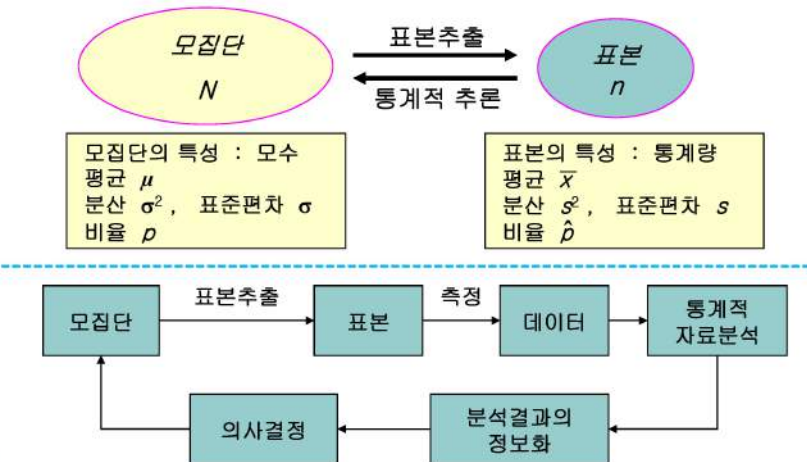
✓ 참고문헌

표본조사방법론 (성내경, 자유아카데미)

표본조사의 이해와 활용 (제6판, Scheaffer et al., 김영원 등, Thomson)

통계적 추론 (Statistical Inference)

관심 대상이 되는 전체집단(**모집단**)으로부터 모집단의 일부(**표본**)를 추출하고, 표본으로부터 관측된 내용(**통계량**)을 근거로 하여 모집단의 특성(**모수**)을 추측



- 3 -

표본추출 (Sampling; 표집)

모집단, 표본

- 모집단 (Population) : 정보를 원하는 전체 대상
- 표본 (Sample) : 관측하는 모집단의 일부
- 표집단위 (Unit; 개체) : 조사의 기본 단위 (예: 응답자, 가구, 사업체, 학교)
- 표본추출틀 (Sampling Frame) : 표본이 추출될 개체들의 목록 (예: 전화번호부)
- 전수조사 : 모집단 전체를 조사하는 방법 (예: 인구주택총조사)
- 표본조사 : 모집단의 일부를 조사하는 방법

모수, 통계량

- 모수 (Parameter) : 모집단의 특성치
모합(母合), 모평균(母平均), 모비율(母比率), ...
상수(constant)이나 표본조사에서는 그 값을 알 수 없다.
- 통계량 (Statistic; 추정량, 추정치) : 표본의 특성치
표본합(標本合), 표본평균(標本平均), 표본비율(標本比率), ...
변수(variable)로서 표본에 따라서 달라진다.

- 4 -

모집단 (Population)

- **무한모집단, 유한모집단**

- 무한모집단 (Infinite Population) : 예) 공장에서 생산된(될) 제품
- 유한모집단 (Finite Population) : 예) 유권자

- **표적모집단, 표집모집단**

- 표적모집단 (Target Population; 목표모집단)
조사에서 얻어진 추론 결과가 적용될 모집단
예) 유권자
- 표집틀모집단 (Sampling Frame Population)
표집틀이 대표하는 모집단
예) 전화번호부
전화가 없는 가구는 조사에서 제외됨
- 표집모집단 (Sampled Population; 조사모집단)
실제 표본이 추출된 모집단
예) 부적격 개체(예: 상가), 전화를 받을 수 없는 가구(예: 입원, 해외), 응답 거부, 편의상 조사에서 제외되는 지역(예: 읍/면 지역, 제주도) 등

- 5 -

표집단위, 표본오차

- **표집단위**

- 일차표집단위 (PSU; Primary Sampling Unit)
- 이차표집단위 (SSU; Secondary Sampling Unit)
예) 서울시에서 10개의 동(일차표집단위)을 추출한 후, 선택된 동에서 100개의 가구(이차표집단위)를 추출하는 경우

- **표본오차, 비표본오차**

- 표본오차 (Sampling Error; 표집오차)
모집단 전체를 조사하지 않고 표본을 추출하여 조사하기 때문에 발생하는 오류.
표본설계 방법 및 표본크기에 의하여 조절할 수 있음.
- 비표본오차 (Nonsampling Error; 비표집오차)
측정편향(예: 잘못된 조사환경 및 질문), 무응답, 선택편향(예: 조사시간) 등.
세심한 조사원칙의 수립 및 조사원 교육, 적절한 보상, 데이터 코딩 확인, 분석과정의 검토, 재조사(callback), 무응답 대체(imputation) 등이 필요함.

- 6 -

표본추출 방법

- **확률표집 (Probability Sampling)**
 - 모집단 내 각 원소들이 표본으로 선택될 확률을 알 수 있음.
 - 단순임의표집 (Simple Random Sampling)
 - 층화표집 (Stratified Sampling)
 - 집락표집 (Cluster Sampling)
- **비확률표집 (Nonprobability Sampling)**
 - 할당표집 (Quota Sampling)
 - 모집단을 대표할 수 있는 분류기준에 따라서 이를 반영하는 표본을 추출.
 - 예) 1948년 미국의 대통령 선거, 민주당 Truman vs. 공화당 Dewey
조사 결과 : 44.5% vs. 49.5% → 실제 결과 : 49.5% vs. 45.1%
 - 예) 1992년 영국의 총선거, 보수당 vs. 노동당
조사 결과 : 38.2% vs. 40.0% → 실제 결과 : 42.8% vs. 35.2%
 - 의도적표집 (Purposive Sampling, Judgmental Sampling; 편의표집)
 - 모집단을 잘 대표한다고 생각되는 개체들을 주관적으로 선택.

- 7 -

(참조) 통계청 자체통계품질진단 매뉴얼

- **표본추출방법 가이드라인**
 - 표본추출방법의 결정은 조사여건과 모집단의 분포 등을 고려하여 결정하여야 하나, 가능한 확률추출방법을 이용하도록 한다.
 - (1) 비확률추출, (2) 질사법추출, (3) 시간대별추출,
 - (4) 단순임의추출, (5) 계통추출, (6) 집락추출,
 - (7) 층화추출, (8) 확률비례추출, (9) 기타 방법
- **표본추출방법 가이드라인**
 - 적용 가능한 다양한 표본추출법을 고려하여 가장 효율적인 표본추출방법을 선택해야 한다.
 - 표본추출방법은 조사의 용이성, 조사비용, 응답부담 등을 감안하여 결정한다.
 - 기존의 유사한 조사에서 적용한 표본추출방법을 참고하여, 이보다 개선된 표본추출방법을 적용하도록 한다.
 - 조사에 적용한 표본추출방법을 상세히 기술해야 한다.
 - 부득이하게 비확률추출방법을 적용했다면, 그 이유를 상세히 밝히도록 한다.

- 8 -

(1) 단순임의표집

• 단순임의표집 (Simple Random Sampling)

- 가장 간단한 확률추출법으로 다른 여러 확률추출법의 기초가 되는 추출방법이다.
- 크기 N 인 모집단에서 크기 n 인 표본을 추출하고자 할 때, n 개의 개체 각각이 동일한 추출기회를 갖도록 표본추출이 시행되는 경우를 말한다.
- 예) A 제품에 대한 소비자 만족도 조사를 한다고 할 때, 소비자 각각에 대해 일련번호를 부여 한 후 난수(random number)를 발생하여 500명의 소비자를 추출하는 경우.

• 계통표집 (Systematic Sampling)

- 모집단의 크기가 크고 표본추출단위들을 숫자화하는 것이 불편하거나 가능하지 않을 때 사용되는 확률추출법이다.
- 모집단 명부가 임의로 나열되어 있다고 가정하고 명부를 몇 개의 추출간격으로 구분하여 추출간격에서 계통적으로 표본을 추출한다.
- 예) 50,000명의 모집단에서 500명을 조사한다고 할 때, $100(=50,000/500)$ 이하의 수에서 하나를 무작위로 추출하고 해당간격 100을 계속 더해 가면서 표본을 추출한다(최초 무작위 추출번호가 5라면 5, 105, 205, 305, ..., 49905를 추출한다).

- 9 -

(2) 층화표집 (Stratified Sampling)

- 모집단의 추출단위를 층(strata)이라고 부르는 서로 중복되지 않는 몇 개의 그룹으로 나누어서 이 들 각 층에서 확률표본을 얻는 것을 말한다.
- 일반적으로 지역, 거주형태, 성, 연령대, 소득수준, 직업 등이 층을 구분하는 보조변수(auxiliary variable)로 이용되며, 층화를 시행하면 추정값의 정확도가 향상된다.
- 일반적으로 모든 층에 대하여 조사되어야 하며, 층 간에는 이질적이고 층 내에서는 동질적일 때 층화의 효율이 높다.
- 모집단의 특성을 잘 반영하고 단순임의표집에 비하여 더 정확한 결과를 얻을 수 있으나, 조사비용이 증가하고 조사가 더 어려워질 가능성이 있다.
- 예) 서울시민의 월평균소득 조사
층 : 고소득, 중소득, 저소득
- 예) 서울시 가구환경 조사
제1층 : 구, 25개
제2층 : 주택유형(단독/공동, 아파트, 다세대, 연립/기타), 4개
- 예) 선거 조사 ($n=20,000$)
제1층 : 성(남자, 여자), 2개
제2층 : 연령대(20대 이하, 30대, 40대, 50대, 60대, 70대 이상), 6개

- 10 -

사후층화 (Post-Stratification)

- 일단 단순임의표집을 하여 조사를 수행한 후, 조사된 결과를 기초로 층화하여 층화추정량을 구하는 방법. 즉, 단순임의표집을 실시하였으나 조사 결과가 층화표집에서 나온 것으로 간주하는 것이다.
- 각 층의 크기는 알고 있어야 한다.
- 층화표집에는 관심이 없었으나 관심 변수에 영향을 미치는 주요 변수가 모집단을 잘 대표하지 못하는 현상을 보정할 목적으로 사용되기도 한다.
- 사후층화는 과소포함(undercoverage)을 보정하는 수단으로 사용될 수 있으나, 지나친 남용은 자제해야 한다.
- 사전에 조사계획을 잘 수립해야 하며, 특정한 층에서 응답거절이나 접촉불가 등이 많이 발생하면 그 이유를 파악하고 적절한 대처방안을 마련해야 한다.
- 예) 가구 방문조사
주택유형을 층으로 구분하고 싶지만 조사 전에는 주택유형을 알 수 없는 경우
- 예) 전화 여론조사
성/연령을 층으로 구분하고 싶지만 조사 전에는 응답자의 성과 연령을 알 수 없는 경우

- 11 -

(참조) 통계청 자체통계품질진단 매뉴얼

- **통계단위의 분류오류 존재 여부**
 - 통계 조사대상(단위)은 산업, 직업, 연령, 지역 등의 특성에 따라 분류기준에 의해서 분류를 하게 된다. 이 때 분류를 정확히 해야 통계작성목적에 부합되는 정확한 통계를 생산할 수 있다. 그러나 조사대상에 분류오류가 발생하면 부정확한 통계가 생산되며, 비교성과 일관성 등에 문제가 생기게 된다. 따라서 가능한 분류오류가 발생하지 않도록 해야 한다. 그러므로 조사단계에서 조사대상(단위)에 대한 적절한 분류가 이루어졌는지를 점검해야 하며, 만일 분류오류가 발생했다면 조사 이전 단계에서 반드시 수정하도록 해야 한다.
- **층화 가이드라인**
 - 모집단을 층화할 경우에는 층화의 필요성을 먼저 검토해야 한다.
 - 층화변수는 변수의 속성이 지속적으로 유지되는 변수를 선택해야 한다.
 - 층의 수는 여러 개의 층을 먼저 고려한 다음 이들을 결합해가는 방법으로 층수를 결정한다.
 - 층별 표본크기는 되도록 다양한 측면에서 검토하여 보다 효율적인 표본 배분 방법을 적용 한다.

- 12 -

(3) 집락표집 (Cluster Sampling)

- 서로 인접한 기본단위들을 묶어 집락을 구성하여 먼저 집락을 추출하고 추출된 집락 내의 일부 또는 전체를 조사하는 추출법을 말한다.
- 일반적으로 집락들이 서로 동질적일 때 집락표집의 효율이 높다.
- 집락추출을 이용하면 조사 지역 또는 대상이 제한되므로 조사비용과 시간의 소모를 줄일 수 있으나, 일반적으로 단순임의표집 혹은 층화표집에 비하여 정확도는 떨어진다.
- 예) 서울에 있는 가구들의 연간 가구당 평균 통신비지출액을 추정하고자 하는 경우 서울시 22개 구에서 일부 구를 추출하고, 구내에서 일부 동을 추출하며, 동 내에서 일부 통을 추출하고, 최종적으로 통 내에서 가구를 무작위로 추출하는 경우를 말한다. 이 때 구, 동, 통은 각각 집락의 1차, 2차, 3차 추출단위가 된다.
- **층화집락표집 (Stratified Cluster Sampling)**
 - 층화표집과 집락표집을 혼용하는 방법
 - 예) 서울시 가구환경 조사
층 : 구(25개), 동(460개)
집락 : 통/반

- 13 -

크기비례확률표집 (PPS)

- **Sampling with Probabilities Proportional to Size**
 - 각 집락이 추출될 확률이 집락의 크기에 비례하여 결정되도록 하는 방법
- 예) 12개의 집락에서 3개의 집락을 추출
 - **누적크기법 (Cumulative-size Method)**
 - (1) 1~642 사이의 난수를 발생. $R=111$.
집락 3을 추출. 계속 반복.
 - **기각적 방법 (Rejective Method)**
 - (1) 1~12 사이의 난수를 발생. $R1=5$.
 - (2) 1~132 사이의 난수를 발생.
 $R2=31$ 이면 집락 5를 추출.
 $R2=74$ 이면 집락 5를 추출하지 않음. 계속 반복.
 - **계통법 (Systematic Method)**
 - (1) 1~214(=642/3) 사이의 난수를 발생. $R=129$.
집락 4를 추출.
 - (2) $129+214=343$. 집락 8을 추출.
 - (3) $343+214=557$. 집락 10을 추출.

집락	집락크기	누적
1	36	36
2	18	54
3	64	118
4	42	160
5	38	198
6	55	253
7	69	322
8	132	454
9	53	507
10	82	589
11	25	614
12	28	642
합계	642	

- 14 -

사례) 표본설계 및 배분

2006년 12월 31일 기준 주민등록통계(통계청)

- 16개 시도 만 13세 이상, 만 69세 이하 인구수에 비례하도록 지역별 표본크기 배정 후, 지역 내에서 성, 연령 비율에 맞추어 표본크기 재할당

할당표 작성 예

지역	성별	계	13~19세	20대	30대	40대	50대	60대
서울	남자	324	37	68	77	66	47	29
	여자	324	33	69	74	68	50	30
	계	648	70	137	151	134	97	59
부산	남자	116	14	24	23	25	19	11
	여자	114	12	22	22	26	20	12
	계	230	26	46	45	51	39	23
대구	남자	77	11	15	17	17	11	6
	여자	76	9	14	17	17	12	7
	계	153	20	29	34	34	23	13
인천	남자	80	17	14	16	17	9	7
	여자	79	16	13	16	16	8	10
	계	163	21	31	38	39	21	13
광주	남자	42	6	9	10	9	5	3
	여자	44	6	9	10	9	6	4
	계	86	12	18	20	18	11	7

- 15 -

... 사례) 실사지점 결정

- PPS (Probability Proportionate to Size) 방식 이용.
- 실사지점은 시 지역의 경우 '통'까지, 군 지역의 경우 '리'또는 '자연부락'을 추출.
- 한 실사지점에서 10표본 조사하는 것을 원칙으로 함.
- 실사지점의 추출은 '표본추출용 소프트웨어'를 이용하여 실시.
- 표본추출용 Data-Base에는 시 지역의 경우 동별 인구수가, 군 지역의 경우 읍면별 인구수가 다음과 같이 저장되어 있음.

번호	시도	시군구	읍면동	통수	인구수
★ 17	서울시	종로구	창신2동	33	18,669
18	서울시	종로구	창신3동	14	9,398
19	서울시	종로구	송인1동	16	11,683
20	서울시	종로구	송인2동	25	11,962
21	서울시	중구	태평로1가동	9	1,193
22	서울시	중구	소공동	8	2,665
23	서울시	중구	남대문로5가동	10	3,048
24	서울시	중구	회현동	18	7,716

- 16 -

... 사례) 조사대상 가구 선정 및 대체

- ❏ 실사 지점을 배당 받은 면접원은 실사 슈퍼바이저로부터 그 지점의 지도를 받게 되는데, 그 지도에는 면접을 해야 할 지역과 방문대상 가구가 표시되어 있음.
- ❏ 원표본 대상가구 조사 실패 시 대체원칙에 따라 조사 가구를 대체하여 조사를 진행함.



- 17 -

... 사례) 응답자 선정

- ❏ 선정된 조사대상 가구를 방문하여 조사 실시.
- ❏ 면접원은 실사본부에서 사전에 배부한 할당표 기준에 맞는 응답자를 선정하여 조사 실시.
- ❏ 사전 배부 받은 할당표 기준에 맞는 적격 응답자가 2명 이상일 경우 생일이 빠른 사람을 우선적으로 조사함.
- ❏ 면접을 해야 할 응답자가 집에 없을 경우 일단 그 응답자의 성별, 연령, 직업을 질문지에 적어서 나옴.
- ❏ 다시 만날 시간, 장소를 약속하여 재 면접을 시도하되, 응답자의 거절 등으로 인하여 면접이 어렵다고 판단되면, 실사지점의 '인근'에서 최초로 선정된 응답자와 성별, 연령대, 직업이 같은 응답자를 찾아 면접을 진행.

- 18 -

통계적 추론

- 통계적 추정 (Statistical Estimation)

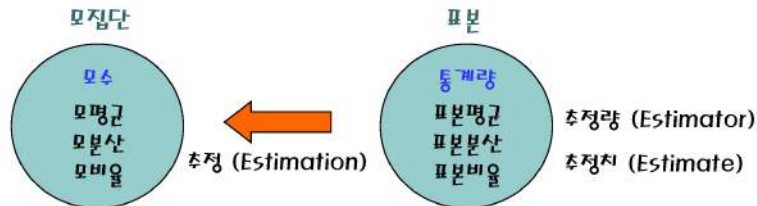
표본의 정보를 통해 실제로 관찰되지 않은 부분을 포함한 모집단 전체에 대한 추측

- 점추정 (Point Estimation)

하나의 값으로 모수를 추정하는 것

- 구간추정 (Interval Estimation)

모수가 속하리라고 생각되는 적절한 구간을 설정하여 추정하는 것



- 19 -

신뢰문(confidence statement)의 예

예) 정부의 이번 정책에 대해 표본조사를 실시한 결과 찬성한다는 비율이 74%이었다. ... 이번 조사의 신뢰수준은 95%이고 오차한계는 $\pm 4\%$ 이다.

⇒ 점 추정치: 74%, 신뢰구간: (70%, 78%)

예) 이번 선거에 대해 표본조사를 실시한 결과 A 후보의 지지율은 51%이고 B 후보의 지지율은 49%이었다. ...

이번 조사의 오차한계는 3%로서(신뢰수준 95%), 두 후보의 지지율은 오차범위 내에 존재한다.

⇒ A 후보: (48%, 54%), B 후보: (46%, 52%)

예) 금번 고객만족지수에 대한 조사에서 A사의 평균은 91점이고 B사의 평균은 83점이다(신뢰수준은 95%, 오차한계는 $\pm 3.5\%$).

⇒ A 사: (87.5, 95.4), B 사: (79.5, 86.5)

- 20 -

공직선거법 제108조 (<http://www.nec.go.kr>)

- ③ 누구든지 공표 또는 보도를 목적으로 선거에 관한 여론조사를 하는 경우에는 피조사자에게 여론조사기관·단체의 명칭, 주소 또는 전화번호와 조사자의 신분을 밝혀야 하고, 당해 조사대상의 전체층을 대표할 수 있도록 피조사자를 선정하여야 하며, 다음 각호의 1에 해당하는 행위를 하여서는 아니된다. <신설 1997.11.14>
1. 특정 정당 또는 후보자에게 편향되도록 하는 어휘나 문장을 사용하여 질문하는 행위
 2. 피조사자에게 응답을 강요하거나 조사자의 의도에 따라 응답을 유도하는 방법으로 질문하거나, 피조사자의 의사를 왜곡하는 행위
 3. 오락 기타 사행성을 조장할 수 있는 방법으로 조사하는 행위
 4. 피조사자의 성명이나 성명을 유추할 수 있는 내용을 공개하는 행위
- ④ 누구든지 선거에 관한 여론조사의 결과를 공표 또는 보도하는 때에는 조사의뢰자와 조사기관·단체명, 피조사자의 선정방법, 표본의 크기, 조사지역·일시·방법, 표본오차율, 응답율, 질문내용등을 함께 공표 또는 보도하여야 하며, 선거에 관한 여론조사를 실시한 기관·단체는 조사설계서·피조사자선정·표본추출·질문지작성·결과분석등 조사의 신뢰성과 객관성의 입증에 필요한 자료와 수집된 설문지 및 결과분석자료등 당해 여론조사와 관련있는 자료일체를 당해 선거의 선거일후 6월까지 보관하여야 한다. <신설 1997.11.14>

✓ 제256조 제2항 : 다음 각호의 어느 하나에 해당하는 자는 2년이하의 징역 또는 400만원이하의 벌금에 처한다. <개정 2005.8.4>

- 21 -

오차한계와 신뢰수준

- 오차한계 (margin of error)
모집단의 참값과 표본에서 산출되는 통계량과의 차이에 관한 일정한 경계값.
- 신뢰수준 (level of confidence)
가능한 모든 표본들 중에서 그 한계를 만족시키는 표본들의 비율.

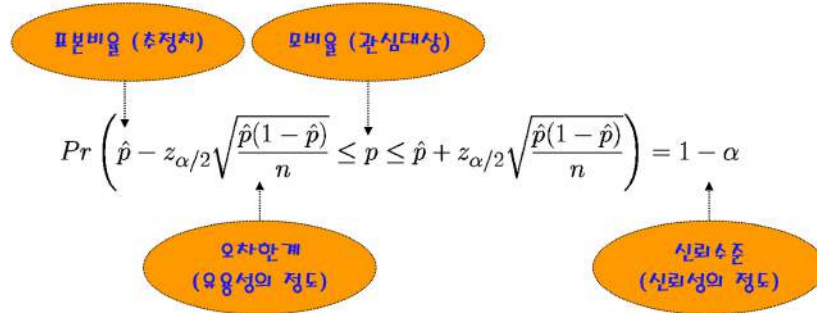
$$P(|\text{모수}-\text{추정량}| \leq \text{오차한계}) \geq \text{신뢰수준}$$

- 오차한계와 신뢰수준의 관계
 - 신뢰수준을 크게 하면 오차한계는 커진다.
 - ✓ 1,055의 임의표본에서 신뢰수준 95%일 때 오차한계는 $\pm 3.1\%$ 이고, 신뢰수준 99%일 때는 $\pm 4.0\%$ 가 된다.
 - 동일한 신뢰수준에서 오차한계를 줄이는 방법
 - ✓ 표본크기를 크게 하거나, 보다 정밀한 표본추출방법을 사용한다.

- 22 -

단순임의표집에서 모비율에 대한 추론

모비율에 대한 $(1-\alpha)\%$ 신뢰구간 (Confidence Interval)



✓ 참조: 간편공식 (모비율에 대한 95% 신뢰구간)

$$z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \simeq 2 \sqrt{\frac{0.5(1-0.5)}{n}} = \frac{1}{\sqrt{n}}$$

신뢰수준	$Z_{\alpha/2}$
90%	1.65
95%	1.96
99%	2.58

- 23 -

(1) 단순임의표집

• 모집단에 대한 표기

- 모집단 크기 (population size) : N
- 모합 (population total) : $\tau = \sum_{i=1}^N y_i$
- 모평균 (population mean) : $\mu = \frac{\tau}{N} = \sum_{i=1}^N y_i / N$
- 모비율 (population proportion) : $p = \frac{\tau}{N}$ ($= \mu$, 단 y_i 들은 이분속성 데이터)
- 모분산 (population variance) : $\sigma^2 = \sum_{i=1}^N (y_i - \mu)^2 / N (= p(1-p))$
- 모표준편차 (population standard deviation) : $\sigma = \sqrt{\sigma^2}$

- 24 -

추정 (표본에 대한 표기)

- 표본크기 (sample size) : n
- 표본합 (sample total) : $t = \sum_{i=1}^n y_i$
- 표본평균 (sample mean) : $\hat{\mu} = \bar{y} = \frac{t}{n} = \sum_{i=1}^n y_i / n$
- 표본비율 (sample proportion) : $\hat{p} = \frac{t}{n}$ ($= \bar{y}$, 단 y_i 들은 이분속성 데이터)
- 표본분산 (sample variance) : $\hat{\sigma}^2 = s^2 = \sum_{i=1}^n \frac{(y_i - \bar{y})^2}{n-1} \left(= \frac{n\hat{p}(1-\hat{p})}{n-1} \right)$
- 표본표준편차 (sample standard deviation) : $\hat{\sigma} = s = \sqrt{s^2}$

- 25 -

오차한계 (Error Bound; 표본오차)

• 모평균의 추정에 대한 오차한계

$$\bullet z_{\alpha/2} \times \widehat{se}(\bar{y}) = z_{\alpha/2} \times \sqrt{\frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

• 모합의 추정에 대한 오차한계

$$\bullet z_{\alpha/2} \times \widehat{se}(\hat{\tau}) = z_{\alpha/2} \times \sqrt{N^2 \frac{s^2}{n} \left(\frac{N-n}{N} \right)}$$

• 모비율의 추정에 대한 오차한계

$$\bullet z_{\alpha/2} \times \widehat{se}(\hat{p}) = z_{\alpha/2} \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1} \left(\frac{N-n}{N} \right)}$$

(참조) 표준오차 (se; Standard Error)

- 어떤 추정량의 표준편차 (즉, 추정량의 분산에 제곱근을 취한 것)

- 26 -

표본조사 이전의 오차한계

- 적정한 표본크기는 조사에서 어느 정도까지의 표집오차를 용인할 수 있는가?(즉, 오차한계를 어느 정도까지 유지하겠는가?) 그리고 비용과 시간의 문제 등을 고려하여 결정한다.
- 상대정밀도(relative precision: 상대오차한계) = 오차한계/모수
- 모평균의 추정**
 - 오차한계는 표준편차 s 와 표본크기에 의존한다.
 - 표준편차는 예비조사에 의한 시험표본(pilot sample) 또는 비슷한 선행 연구로부터 구할 수 있다.
 - 정보가 전혀 없는 경우 데이터들의 범위(range)를 기초로 표준편차를 추측하기도 한다. $s \approx \text{range}/4$
- 모비율의 추정**
 - 오차한계는 비율 p 와 표본크기에 의존한다.
 - 비율은 예비조사에 의한 시험표본(pilot sample) 또는 비슷한 선행 연구로부터 구할 수 있다.
 - 정보가 전혀 없는 경우 $p=1/2$ 을 사용하기도 한다.

- 27 -

사례) 오차한계와 표본크기

- 모비율의 추정**
 - 어느 보호시설의 청소년 $N=1,500$ 명을 대상으로 흡연률을 조사하려고 한다. $n=500$ 명을 조사할 경우 오차한계는 어느 정도인가? (95% 신뢰수준 하에서)
 - $$e = 1.96 \times \sqrt{\frac{0.5(1-0.5)}{500-1} \left(\frac{1500-500}{1500} \right)} = 3.58\%P$$
- 모평균의 추정**
 - 어느 보호시설의 청소년 $N=1,500$ 명을 대상으로 평균수면시간을 조사하려고 한다. 선행 연구에 의하면 청소년의 수면시간에 대한 표준편차는 2시간 정도라고 한다. $n=500$ 명을 조사할 경우 오차한계는 어느 정도인가? (95% 신뢰수준 하에서)
 - $$e = 1.96 \times \sqrt{\frac{2^2}{500} \left(\frac{1500-500}{500} \right)} = 0.143\text{시간}$$
 - 평균수면시간이 9시간 정도인 경우 : 상대정밀도 = $0.143/9 = 1.59\%P$

- 28 -

(2) 층화표집

● 모집단에 대한 표기

- y_{hj} : 층 h 의 j 번째 단위 ($h = 1, \dots, H$; $j = 1, \dots, N_h$)
- N_h : 층 h 의 크기
- $W_h = N_h/N$: 층 h 의 층비중(stratum weight)
- $\tau_h = \sum_{j=1}^{N_h} y_{hj}$: 층 h 의 모합
- $\mu_h = \frac{\tau_h}{N_h} = \sum_{j=1}^{N_h} y_{hj}/N_h$: 층 h 의 모평균
- $p = \frac{\tau_h}{N_h}$: 층 h 의 모비율 (단, y_{hj} 들은 이분속성 데이터)
- $\sigma_h^2 = \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2 / N_h$: 층 h 의 모분산

- 29 -

... 모집단에 대한 표기

- $\tau = \sum_{h=1}^H \tau_h$: 모합
- $\mu = \frac{\tau}{N} = \sum_{h=1}^H \tau_h / N = \sum_{h=1}^H \left(\frac{N_h}{N} \right) \mu_h = \sum_{h=1}^H W_h \mu_h$: 모평균
- $p = \sum_{h=1}^H W_h p_h$: 모비율 (단, y_{hj} 들은 이분속성 데이터)
- $\sigma^2 = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 / N$: 모분산

- 30 -

추정 (표본에 대한 표기)

- y_{hj} : 층 h 에서 표집된 j 번째 단위 ($h = 1, \dots, H$; $j = 1, \dots, n_h$)
- n_h : 층 h 의 크기
- $f_h = n_h/N_h$: 층 h 의 표집률(sampling fraction)
- $\bar{y}_h = \sum_{j=1}^{n_h} y_{hj}/n_h$: 층 h 의 표본평균
- $\hat{\tau}_h = N_h \bar{y}_h$: 층 h 의 표본합
- $s_h^2 = \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 / (n_h - 1)$: 층 h 의 표본분산
- 모함에 대한 추정
 - $\hat{\tau} = \sum_{h=1}^H \hat{\tau}_h = \sum_{h=1}^H N_h \bar{y}_h$: 표본합
 - $e = z_{\alpha/2} \times \sqrt{\sum_{h=1}^H N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}}$: 오차한계

- 31 -

... 추정

- 모평균에 대한 추정
 - $\hat{\mu} = \bar{y} = \frac{\hat{\tau}}{N} = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^H W_h \bar{y}_h$: 표본평균
 - $e = z_{\alpha/2} \times \sqrt{\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}}$: 오차한계
- 모비율에 대한 추정
 - $\hat{p} = \sum_{h=1}^H \frac{N_h}{N} \hat{p}_h$: 표본비율
 - $e = z_{\alpha/2} \times \sqrt{\sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}}$: 오차한계

- 32 -

표본배분 (표본배정)

- 층화표집에서는 표본크기를 결정할 때 두 가지 면을 동시에 고려해야 한다.
- 즉, 총 표본크기 n 을 결정해야 하고, 각 층의 표본크기 n_h 들을 결정해야 한다.
- **비례배분 (Proportional Allocation)**
 - 각 층의 모집단 크기에 비례하여 표본크기를 결정.
- **변형 비례배분**
 - 각 층에 일정 수준 이상의 표본크기가 배정되도록 하면서 비례배분.
- **제곱근(또는 로그) 비례배분**
 - 각 층의 모집단 크기의 제곱근에 비례하여 표본크기를 결정.
- **절사층화배분**
 - 어떤 층에 대해서는 전수조사를 실시하고 나머지 층에는 표본조사를 실시.
 - 표본조사 층에는 비례배분이나 네이만 배분 등을 사용.
- **네이만 배분**
 - 각 층의 모집단 크기와 표준편차(변이)를 고려하여 표본크기를 결정.
- $$n_h = \left(\frac{N_h s_h}{\sum_{h=1}^H N_h s_h} \right) \times n$$

- 33 -

사례) 표본배분

	모집단	네이만배분			비례배분		전년도
		주거환경	교육환경	월소득	단순	제곱근	
종로구	61,926	399	386	337	341	530	516
중구	42,059	229	232	236	232	437	458
용산구	88,958	538	530	513	490	634	629
성동구	115,173	656	662	662	634	722	737
광진구	147,604	819	838	835	813	817	808
동대문구	140,814	746	747	664	775	798	793
종랑구	163,145	881	909	866	898	859	835
성북구	164,021	936	910	919	903	861	869
강북구	129,641	744	683	722	714	766	755
도봉구	128,082	698	695	742	705	761	770
노원구	203,486	1,034	1,088	1,081	1,121	959	936
...	...	...	...	...	...	...	...
은평구	166,883	728	813	684	919	869	873
서대문구	125,286	748	770	720	690	753	763
강동구	149,223	791	740	811	822	822	847
합계	3,631,787	20000	20,000	20,000	20,000	20,000	20000

- 34 -

사례) 표본배분

모집단 크기

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	291	391	119	30	831
1512	1,226	566	35	7	1,834
1513	428	292	30	1	751
1514	31	60	12	7	110
1520	19	47	16	37	119
...	...	...	...	...	...

강제 할당

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	10	10	10	10	40
1512	10	10	10	7	37
1513	10	10	10	1	31
1514	10	10	10	7	37
1520	10	10	10	10	40
...	...	...	...	...	...

모집단 크기 - 강제 할당

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	281	381	109	20	
1512	1,216	556	25	0	
1513	418	282	20	0	
1514	21	50	2	0	
1520	9	37	6	27	
...	...	...	...	...	...

- 35 -

... 사례) 표본배분

식료품 구입액 표준편차

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	623.34	849.83	2,363.01	9,764.93	3,191.32
1512	111.86	560.81	1,040.32	5,346.65	1,008.75
1513	602.59	1,992.46	4,279.57	전수조사	2,489.71
1514	97.46	1,321.71	13,135.25	전수조사	67,889.88
1520	20.26	2,309.11	4,198.05	36,990.03	85,460.94
...	...	...	...	...	...

모집단 크기 *표준편차

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	175,160	323,783	257,568	195,299	
1512	136,027	311,813	26,008	0	
1513	251,881	561,872	85,591	전수조사	
1514	2,047	66,086	26,271	전수조사	
1520	182	85,437	25,188	998,731	
...	...	...	...	...	...

표본조사 총 표본배분

업종 코드	10억 미만	10~100억	100~400억	400억 이상	전체
1511	8	15	12	9	44
1512	6	14	1	0	22
1513	12	26	4	0	41
1514	0	3	1	0	4
1520	0	4	1	46	51
...	...	...	...	...	...

- 3b -

사례) 표본조사 이전의 오차한계

모집단크기

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업	9,122	5,424	3,059	516	90	33
도소매	19,753	17,695	1,870	157	21	10
기타	22,423	18,751	2,844	557	177	94
Total	51,298	41,870	7,773	1,230	288	137

표본크기

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업	500	100	204	74	90	33
도소매	500	331	120	18	21	10
기타	500	133	69	27	177	94
Total	1,500	564	393	119	288	137

표준편차

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업	22.493	19.598	22.507	20.834	31.722	28.233
도소매	18.247	14.070	31.722	22.942	0.000	0.000
기타	10.585	11.546	10.426	0.000	0.000	19.245
Total	19.148	15.530	22.140	18.335	24.454	23.330

(Nh/N)^2

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업		0.011	0.004	0.000	0.000	0.000
도소매		0.119	0.001	0.000	0.000	0.000
기타		0.134	0.003	0.000	0.000	0.000
Total						

(Nh-nh)/Nh

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업		0.982	0.933	0.857	0.000	0.000
도소매		0.981	0.936	0.885	0.000	0.000
기타		0.993	0.976	0.952	0.000	0.000
Total						

sh^2/nh

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업		3.841	2.483	5.865	11.181	24.155
도소매		0.598	8.386	29.240	0.000	0.000
기타		1.002	1.575	0.000	0.000	3.940
Total						

(Nh/N)^2*((Nh-nh)/Nh)*(sh^2/nh)

	전체	5인미만	5-19인	20-49인	50-99인	100이상
제조업		0.042	0.008	0.001	0.000	0.000
도소매		0.070	0.010	0.000	0.000	0.000
기타		0.133	0.005	0.000	0.000	0.000
SUM	0.269					

95% 신뢰수준 하에서의 오차한계

Total	1.017
-------	-------

- 37 -

(3) 총화집락표집

Definition and Notation

For a stratified clustered sample design, together with the sampling weights, the sample can be represented by an $n \times (P+1)$ matrix

$$(\mathbf{w}, \mathbf{Y}) = (w_{hij}, y_{hij}) \\ = (w_{hij}, y_{hij}^{(1)}, y_{hij}^{(2)}, \dots, y_{hij}^{(P)})$$

where

- $h = 1, 2, \dots, H$ is the stratum number, with a total of H strata
- $i = 1, 2, \dots, n_h$ is the cluster number within stratum h , with a total of n_h clusters
- $j = 1, 2, \dots, m_{hi}$ is the unit number within cluster i of stratum h , with a total of m_{hi} units
- $p = 1, 2, \dots, P$ is the analysis variable number, with a total of P variables
- $n = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ is the total number of observations in the sample
- w_{hij} denotes the sampling weight for observation j in cluster i of stratum h
- $\mathbf{y}_{hij} = (y_{hij}^{(1)}, y_{hij}^{(2)}, \dots, y_{hij}^{(P)})$ are the observed values of the analysis variables for observation j in cluster i of stratum h , including both the values of numerical variables and the values of indicator variables for levels of categorical variables.

For a categorical variable C , let l denote the number of levels of C , and denote the level values as $c_1, c_2, \dots, c_l$. Then there are l indicator variables associated with these levels. That is, for level $C=c_k$ ($k = 1, 2, \dots, l$), a $y_{hij}^{(q)}$ ($q \in \{1, 2, \dots, P\}$) contains the values of the indicator variable for the category $C=c_k$ with the value of observation j in cluster i of stratum h .

$$y_{hij}^{(q)} = I_{\{C=c_k\}}(h, i, j) = \begin{cases} 1 & \text{if } C_{hij} = c_k \\ 0 & \text{otherwise} \end{cases}$$

Therefore, the total number of analysis variables, P , is the total number of numerical variables plus the total number of levels of all categorical variables.

- 38 -

추정

Mean

When you specify the keyword MEAN, the procedure computes the estimate of the mean (mean per element) from the survey data. Also, the procedure computes the mean by default if you do not specify any [statistic-keywords](#) in the PROC SURVEYMEANS statement.

PROC SURVEYMEANS computes the estimate of the mean as

$$\hat{Y} = \left(\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \right) / w_{...}$$

where

$$w_{...} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$$

is the sum of the weights over all observations in the sample.

Proportions

If you specify the keyword MEAN for a categorical variable, PROC SURVEYMEANS estimates the proportion, or relative frequency, for each level of the categorical variable. If you do not specify any [statistic-keywords](#) in the PROC SURVEYMEANS statement, the procedure estimates the proportions for levels of the categorical variables, together with their standard errors and confidence limits.

The procedure estimates the proportion in level c_q for variable C as

$$\hat{p} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}^{(q)}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

- 39 -

... 추정

Variance and Standard Error of the Mean

When you specify the keyword STDERR, the procedure computes the standard error of the mean. Also, the procedure computes the standard error by default if you specify the keyword MEAN, or if you do not specify any [statistic-keywords](#) in the PROC SURVEYMEANS statement. The keyword VAR requests the variance of the mean.

PROC SURVEYMEANS uses the Taylor series expansion theory to estimate the variance of the mean $\hat{Y}$. The procedure computes the estimated variance as

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

where if $n_h > 1$,

$$\begin{aligned} \hat{V}_h(\hat{Y}) &= \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2 \\ e_{hi\cdot} &= \left(\sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \hat{Y}) \right) / w_{...} \\ \bar{e}_{h\cdot\cdot} &= \left(\sum_{i=1}^{n_h} e_{hi\cdot} \right) / n_h \end{aligned}$$

and if $n_h = 1$,

$$\hat{V}_h(\hat{Y}) = \begin{cases} \text{missing} & \text{if } n_{h'} = 1 \text{ for } h' = 1, 2, \dots, H \\ 0 & \text{if } n_{h'} > 1 \text{ for some } 1 < h' < H \end{cases}$$

The standard error of the mean is the square root of the estimated variance.

$$\text{StdErr}(\hat{Y}) = \sqrt{\hat{V}(\hat{Y})}$$

- 40 -

사례) 표본조사 이전의 오차한계

- 집락효과를 무시하는 경우(단순임의추출을 가정하는 경우), 최대 표본크기를 구하도록 $p=0.5$ 로 할 때, 모비율에 대한 오차한계를 계산하면(95% 신뢰수준 하에서)

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

- 아동·청소년: $N=49,010$, $n=3,594 \Rightarrow e=\pm 1.57\%P$

- 보호자: $N=17,659$, $n=1,180 \Rightarrow e=\pm 2.76\%P$

- 대략 $M=10$ (집락크기), $\rho=0.04$ (급내상관계수; 통계청 사회통계조사의 경우)로 간주하여, $deff = 1 + (M-1)\rho = 1.36$ 을 단순임의추출에 대한 집락추출의 설계효과로 하고 최대 표본크기를 구하도록 $p=0.5$ 로 할 때, 모비율에 대한 오차한계를 계산하면(95% 신뢰수준 하에서)

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

- 아동·청소년: $N=49,010$, $n=3,594 \Rightarrow e=\pm 1.84\%P$

- 보호자: $N=17,659$, $n=1,180 \Rightarrow e=\pm 3.21\%P$

- 41 -

사례) 표준오차와 신뢰구간

Variable	Label	N	Mean	Std Error of Mean	95% CL for Mean	
Q1_1	나는 현재 여가생활에 만족한다	2140	2.179	0.039	2.099	2.258
Q1_2	지금까지 살아온 삶에 대해 만족한다	2140	2.347	0.031	2.282	2.411
Q1_3	나는 지금 행복하다	2140	2.141	0.032	2.075	2.207
Q1_4	나는 현재 몸과 마음이 건강하다	2140	2.017	0.037	1.941	2.094
Q1_5	나의 미래는 밝다	2140	2.195	0.035	2.123	2.266
Q2_1	일과 여가생활 중요도_일	2031	7.858	0.076	7.702	8.015
Q2_2	일과 여가생활 중요도_여가생활	2078	7.639	0.068	7.499	7.778
Q4_1	인생에서 중요한 정도_가족	2140	1.148	0.014	1.118	1.178
Q4_2	인생에서 중요한 정도_동성친구	2140	1.395	0.019	1.356	1.433
Q4_3	인생에서 중요한 정도_이성친구	2140	1.900	0.038	1.821	1.978
Q4_4	인생에서 중요한 정도_이웃	2140	2.320	0.032	2.254	2.386
Q4_5	인생에서 중요한 정도_여가	2140	1.540	0.018	1.504	1.577
Q4_6	인생에서 중요한 정도_건강	2140	1.168	0.014	1.141	1.196
Q4_7	인생에서 중요한 정도_일	2140	1.469	0.022	1.425	1.514
Q4_8	인생에서 중요한 정도_종교	2140	2.857	0.038	2.779	2.936
Q4_9	인생에서 중요한 정도_학력	2140	1.841	0.032	1.775	1.906
Q4_10	인생에서 중요한 정도_권력	2140	2.091	0.036	2.017	2.166
Q4_11	인생에서 중요한 정도_돈	2140	1.533	0.022	1.487	1.578
Q7_1	나는 가족과의 관계가 친밀하다	2140	1.549	0.025	1.497	1.600
Q7_2	나의 부모님은 나를 잘 이해하신다	2140	2.038	0.036	1.964	2.113
Q7_3	나는 가족과의 생활에 만족한다	2140	1.863	0.028	1.805	1.921

- 42 -

사후층화를 위한 가중치 처리

- 설계 가중치 산출 → 무응답 가중치 조정 → 사후층화
- 셀 가중법(cell weighting)
 - 모집단의 결합분포와 표본의 결합분포를 비교하여 그 비를 가중치로 이용하는 방법으로, 모집단의 다차원 결합 분포가 알려진 경우에 유용하게 이용될 수 있다. 이 방법은 모집단과 표본의 인구사회학 특성의 다차원 결합분포 내 각 셀의 비율을 조정하는 방법을 이용한다. 모집단의 다차원 결합분포가 알려진 경우가 드물며, 표본의 다차원 결합분포의 칸이 비어 있는 경우 이용이 어렵다는 단점이 있다.
- 림 가중법(rim weighting), 레이킹 방법(raking method)
 - 반복비례가중법은 모집단의 인구사회학 특성의 주변분포를 알고 있는 경우 사용 수 있는 방법으로, 각 특성 변수의 주변분포를 이용하여 반복으로 칸 내의 모집단수를 구하는 방법이다. 이 방법은 모집단의 주변분포만 알면 사용이 가능하므로 고려할 특성 변수가 많은 경우에도 활용이 가능하다. 사용 방법의 유용성 때문에 다양한 분야에 걸쳐 활발하게 이용되고 있다.

- 43 -

《예》 Rim 가중치 (Raking 가중치)

- ① 입주유형별 주변비율 조정 : 이전 단계에서 얻어진 표본가중치에 대하여 다음과 같이 가중치를 조정한다.

$$W_{j(qb)} = W_j \times \frac{N_{qb}}{T_{qb}}$$
 여기서 T_{qb} 와 N_{qb} 는 각각 (q 번째 구, b 번째 입주유형)에 대한 W_j 의 합계와 모집단크기를 나타낸다.
- ② 세대원수별 주변비율 조정 : 이전 단계에서 얻어진 표본가중치에 대하여 다음과 같이 가중치를 조정한다.

$$W_{j(qc)} = W_j \times \frac{N_{qc}}{T_{qc}}$$
 여기서 T_{qc} 와 N_{qc} 는 각각 (q 번째 구, c 번째 세대원수 구분)에 대한 W_j 의 합계와 모집단크기를 나타낸다.
- ③ 주택유형별 주변비율 조정 : 이전 단계에서 얻어진 표본가중치에 대하여 다음과 같이 가중치를 조정한다.

$$W_{j(qa)} = W_j \times \frac{N_{qa}}{T_{qa}}$$
 여기서 T_{qa} 와 N_{qa} 는 각각 (q 번째 구, a 번째 주택유형)에 대한 W_j 의 합계와 모집단크기를 나타낸다.
- ④ 가중치에 더 이상의 변화가 없을 때까지 위의 ①~③ 과정을 반복한다.

- 44 -

패널자료 형태

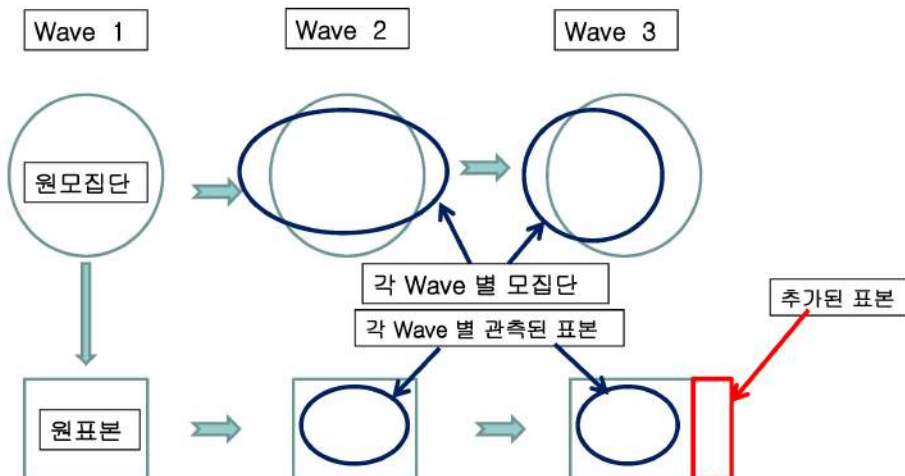
고정변수 (Wave independent Variables)				Wave 1			Wave p	
ID	나이	성별		Y_{11}		Y_{1k}	Y_{p1}	Y_{pk}
1	26	남(1)		1		500	2	 500
2	52	남(1)		2		missing	3	 200
3	44	여(0)		2		200	missing	 missing
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
n	33	남(1)		4		300	2	 300

항목무응답

개체무응답

- 45 -

패널자료 형태



- 46 -

패널자료 분석

Wave 1

Wave 2

Wave 3

종단면 분석(Longitudinal Analysis)

- 관심변수 혹은 관심모수의 시간에 따른 변화 대한 고찰.
- Mixed model, Bayesian Model .
- 초기 표본 가중치의 사용여부 결정
- Panel Attrition으로 인한 표본크기 감소 및 이를 보정하기 위한 적절한 처리가 요구됨.
 - Attrition의 성격에 따라
 - 가중치 보정, Imputation, 표본추가

- 47 -

패널자료 분석

Wave 1

Wave 2

Wave 3

횡단면 분석(Cross-Sectional Analysis)

- 각 wave별 관심변수의 모집단 특성에 대한 추론.
- Design-Based or Model Based
- 초기 표본 가중치의 각 wave별 모집단 정보에 따른 가중치 보정이 요구
- Panel Attrition으로 인한 표본크기 감소 및 이를 보정하기 위한 적절한 처리가 요구됨.
 - Attrition의 성격에 따라
 - 가중치 보정, Imputation, 표본추가

- 48 -

《예》 아동청소년패널 1차년도 가중치

(1) 표본추출확률에 대한 가중치(기본가중치)

한국아동·청소년패널조사 2010의 표본추출은 일종의 층화2단집락추출법이라 할 수 있으며 시도와 도시규모를 층으로 간주할 수 있다. 따라서 각 층별로 표본을 할당하였으므로 각 층에서 1차추출단위인 학교는 학생수를 기준으로 확률비례추출법으로 선정하고 추출된 표본학교에서는 학급을 랜덤하게 선정한 후에 학생을 조사하므로 추출된 표본학교에서 학생을 랜덤하게 선정하여 조사하는 2단집락추출법으로 생각할 수 있으므로 표본추출률을 아래 식으로 나타낼 수 있다.

$$f_{hij} = \frac{B_{hi}}{\sum_{i=1}^{N_h} B_{hi}} \cdot \frac{n_{hi}}{B_{hi}} = \frac{n_{hi}}{\sum_{i=1}^{N_h} B_{hi}}$$

여기서 N_h 는 h 층에 속한 학교의 수이고, B_{hi} 는 h 층에서 i 번째 학교의 학생수를 나타내며 n_{hi} 는 추출된 표본학교에서 조사한 학생수를 나타낸다.

표본추출률에 대한 가중치는 추출률의 역수이므로 아래와 같은 식으로 표현할 수 있다.

$$W_{1hij} = \frac{\sum_{i=1}^{N_h} B_{hi}}{n_{hi}}$$

- 49 -

... 《예》 아동청소년패널 1차년도 가중치

(2) 표본의 무응답 조정을 위한 가중치

표본으로 선정된 조사단위들 중에서 일부는 응답거부로 조사를 성공하지 못하거나 조사된 내용이 부실하여 무응답처리한 조사단위에 대한 조정이 무응답 조정의 가중치이며 아래 식으로 계산할 수 있다. $W_{2hij} = \frac{n_{hi}}{r_{hi}}$ 이며 r_{hi} 와 n_{hi} 는 각각 완전 응답한 표본크기와 추출된 표본크기이다.

(3) 사후층화 가중치(poststratification weight)

표본설계 시에는 층화변수로 고려하지 않았으나 표본조사 완료 후에 연구변수에 영향을 미치는 변수들중에서 정확한 외부보조정보를 확보하게 되면 외부보조정보 데이터의 구조와 표본조사 데이터의 구조를 분석하여 층으로 구분하고 표본조사의 각 셀에 가중치의 합계를 계산한 결과와 외부보조정보의 층별 합계와 비교하여 가중치를 계산한다.

$$BF_{hij} = \frac{PS_{hij}}{\sum W_{1hij} \cdot W_{2hij}}$$

여기서 PS_{hij} 는 외부보조정보에 의한 셀의 빈도수이고 분모의 내용은 실제가중치와 무응답 조정위한 가중치를 곱한 후에 합계를 계산결과이다.

최종가중치는 세 가지 가중치를 곱하여 아래와 같은 식으로 산출한다.

$$W_{fij} = W_{1hij} \cdot W_{2hij} \cdot BF_{hij}$$

- 50 -

... <예> 아동청소년패널 1차년도 가중치

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,180	52.1	1,083	47.9	2,263	100	315,756	52.3	287,410	47.7	603,164	100
서울	135	51.7	126	48.3	261	11.5	53,541	52.1	49,281	47.9	102,822	17.0
부산	64	48.9	67	51.1	131	5.8	19,081	52.9	16,980	47.1	36,061	6.0
대구	75	54.3	63	45.7	138	6.1	16,361	53.2	14,377	46.8	30,738	5.1
인천	54	50.9	52	49.1	106	4.7	17,238	51.9	15,948	48.1	33,186	5.5
광주	57	50.4	56	49.6	113	5.0	11,280	52.3	10,275	47.7	21,555	3.6
대전	60	53.6	52	46.4	112	4.9	10,248	51.7	9,555	48.3	19,803	3.3
울산	49	50.0	49	50.0	98	4.3	7,913	52.7	7,090	47.3	15,003	2.5
경기	185	51.5	174	48.5	359	15.9	80,811	52.1	74,346	47.9	155,157	25.7
강원	57	53.3	50	46.7	107	4.7	9,625	52.8	8,589	47.2	18,214	3.0
충북	67	48.6	71	51.4	138	6.1	10,023	52.5	9,082	47.5	19,105	3.2
충남	64	52.9	57	47.1	121	5.3	13,355	52.3	12,159	47.7	25,514	4.2
전북	70	53.8	60	46.2	130	5.7	12,060	51.8	11,217	48.2	23,277	3.9
전남	47	50.0	47	50.0	94	4.2	11,904	52.3	10,841	47.7	22,744	3.8
경북	70	58.8	49	41.2	119	5.3	16,066	53.1	14,193	46.9	30,258	5.0
경남	78	54.2	66	45.8	144	6.4	22,077	53.1	19,509	46.9	41,586	6.9
제주	48	52.2	44	47.8	92	4.1	4,173	51.3	3,968	48.7	8,141	1.3

- 51 -

<예> 아동청소년패널 2차년도 횡단면 가중치

(1) 기본 가중치

기본 가중치는 표본탈락률을 보정하기 위하여 1차년도 대비 응답률(2,280/2,351=97.0%)의 역수를 1차년도 횡단면 가중치에 곱하여 계산되었다.

$$\text{기본 가중치} = \text{1차년도 가중치} \times \frac{1}{\text{1차년도 대비 응답률}}$$

(2) 최종 가중치 (사후층화 가중치)

최종 가중치는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2011년의 중학생 2학년 기준).

$$\text{최종 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

- 52 -

... <예> 아동청소년패널 2차년도 종단면 가중치

(1) 기본 가중치

2차년도 종단면 기본 가중치는 1차년도 종단면 가중치에 응답확률 추정값의 역수를 곱하여 계산되었다.

$$\text{기본 가중치} = \text{1차년도 가중치} \times \frac{1}{\text{응답확률 추정값}}$$

<표 1-3> 로지스틱 회귀분석을 이용한 변수선택 결과

변수	자유도	Wald Chi-Square	p-value
P6A1W1_r	15	24.8300	0.0523
Q5A2W1_r	12	22.2378	0.0349
P4O1W1_r	4	16.5380	0.0024
Q13D2W1_r	4	6.8078	0.1464
Q13D3W1_r	3	9.7425	0.0209
Q19A15W1_r	2	7.0577	0.0293
Q20A03W1_r	3	11.9174	0.0077
Q22A1W1_r	3	6.1916	0.1026

(2) 최종 가중치 (사후층화 가중치)

최종 가중치(사후층화 가중치)는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2011년의 중학생 2학년 기준).

$$\text{사후층화 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

- 53 -

Q & A

Question & Answer

감사합니다.

- 54 -

한국아동·청소년패널조사2010 I 표본설계

1. 조사목적

급변하는 사회경제적 환경에 놓인 우리 사회가 아동과 청소년을 미래사회의 동력으로 양성하는데 필요한 제반여건을 조성하고 아동·청소년 관련정책을 수립하는데 필요한 실효성있는 기초자료를 제공하는 데 조사목적이 있다.

패널을 초등학교 1학년과 4학년 그리고 중학교 1학년의 3개 코호트로 구성하여 우리나라 아동·청소년의 발달 및 성장과정을 종단 축으로 하고, 각 시기의 변화양상과 변화의 실태를 횡단 축으로 하여 아동·청소년시기를 시간의 흐름속에서 다각적이고 역동적인 시각에서 패널조사를 수행하여 아동·청소년 관련정책 수립에 필요한 기초정보를 제공하고자 한다.

2. 모집단 정의

2010년 현재 초등학교 1학년과 4학년 및 중학교 1학년에 재학중인 모든 학생과 학부모를 조사대상으로 정한다. 따라서 조사모집단은 2009년 교육연감에 수록된 초등학교와 중학교의 초등 1학년과 초등 4학년 및 중학교 1학년 학생들과 학부모들로 구성한다.

3. 조사방법

각 코호트인 초등학교 1학년과 4학년 및 중학교 1학년을 확률추출방법으로 선정하고 조사에 참여한 학생들을 통해서 학부모를 접촉한 후에 가구방문조사로 진행한다. 특히 초등학교 1학년은

설문의 이해능력을 고려하여 간단한 내용을 조사하며 조사원의 설명을 통해서 타계식 조사방법을 적용하지만 초등학교 4학년과 중학교 1학년은 가능한 자기기입방법으로 조사를 하고 필요할 경우에는 조사원이 응답의 기입을 보조하는 방법으로 조사를 진행한다. 학교를 통해서 설문조사에 참여한 학생에게 가정통신문을 통해서 본 패넬조사에 대한 의의와 내용을 설명한 후에 개별적으로 방문조사를 실시한다. 초등학교 1학년의 경우에는 개인발달과 환경발달 등에 관한 내용은 학생보다는 학부모를 통해서 수집한다.

4. 표본설계

4.1 모집단 분석

2009년 교육연감에는 학교에 관한 사항만 있으므로 모집단분석은 학생에 관한 내용에 대해서만 언급하겠다. 초등학교 1학년과 4학년은 학교별로 교육환경이나 특성이 유사할 수 있고 또한 학부모에 대한 정보도 좀더 다양한 지역에서 수집하기 위해서 표본학교는 초등학교 1학년과 4학년은 독립적으로 추출할 것이다.

모집단의 분석은 학생들의 개인발달과 환경발달에 영향을 미치는 요소중에서 모집단자료에서 분석할 수 있는 것이 지역과 도시규모이므로 이를 중심으로 모집단을 분석하겠다.

먼저 16개 시도별 초등학교의 1학년과 4학년의 학교 수와 학급 수 및 학생 수를 <표 1>에 정리하였다.

표 1 초등학교 1학년 시도별 분포현황

시도	초등학교 1학년			초등학교 4학년		
	학교 수	학급 수	학생 수	학교 수	학급 수	학생 수
강원	386	590	14,298	396	735	18,057
경기	1,141	3,852	120,729	1,144	4,702	150,891
경남	508	1,196	31,798	523	1,505	41,342
경북	540	948	23,196	545	1,203	30,549
광주	146	529	15,815	147	684	20,820
대구	214	793	23,028	215	1,032	30,945
대전	139	520	15,516	139	659	19,437
부산	299	1,039	27,317	299	1,308	36,040
서울	584	2,800	81,776	584	3,429	102,001
울산	118	373	11,064	119	497	15,179
인천	232	916	25,661	233	1,171	33,179
전남	496	808	17,403	517	998	22,940
전북	413	752	17,992	419	912	23,098
제주	115	235	6,292	115	276	7,825
충남	442	796	19,517	446	948	24,424
충북	273	560	14,866	276	701	18,838
합계	6,046	16,707	466,268	6,117	20,760	595,565

초등학교 1학년이 있는 학교 수는 6,046개교이고 학급수는 16,707개 학급이며 학생수는 466,268명이지만 4학년이 있는 초등학교는 6,117개교이고 학급수는 20,760개이며 학생수는 595,565명이다. 위의 <표 1>에서 보는바와 같이 학교수는 71개교가 1학년에는 줄어들었으며 학급수도 4053개 학급이 감소되었고 학생수도 129,297명이 1학년이 적은 것으로 나타났다.

시도별로는 서울시, 대전시와 부산시가 학교 수에서 감소하지 않았으나 학급수와 학생 수에서는 모두 줄어들었다.

다음에는 중학교 1학년이 있는 중학교의 시도별 현황을 <표 2>에 정리하였다.

표 2 중학교 1학년 시도별 분포현황

시도	남여공학				여학교			남학교		
	학교수	학급수	남학생	여학생	학교수	학급수	여학생	학교수	학급수	남학생
강원	107	267	3,920	3,693	30	168	5,738	27	189	6,490
경기	476	3,746	76,013	68,744	33	244	9,086	35	264	10,142
경남	183	808	14,548	12,087	43	266	9,314	41	302	10,494
경북	200	576	9,068	7,858	53	224	7,573	44	273	9,088
광주	65	495	9,516	9,611	9	49	1,965	7	71	2,856
대구	100	829	16,239	15,033	13	65	2,349	9	104	4,008
대전	64	505	9,778	8,793	12	55	1,850	10	65	2,289
부산	95	775	14,168	12,462	42	258	8,751	33	299	10,247
서울	273	2,679	50,498	44,658	49	345	11,743	46	372	12,894
울산	50	419	8,140	6,746	5	46	1,582	5	51	1,722
인천	66	485	9,887	9,446	31	230	9,141	27	259	10,528
전남	213	546	8,169	7,454	28	146	4,828	25	159	5,236
전북	165	635	10,935	9,120	20	103	3,415	19	90	2,851
제주	28	124	2,303	2,075	7	52	2,062	7	55	2,235
충남	139	479	7,747	6,779	29	169	5,800	25	177	6,137
충북	96	421	7,294	6,394	17	107	3,710	16	117	4,050
합계	2,320	13,789	258,223	230,953	421	2,527	88,907	376	2,847	101,267

중학교 2학년이 있는 중학교수는 3,117개교이며 이중에서 남여공학은 2,320개교이고 여학교는 421개교, 남학교는 376개교이다. 학생수는 전체는 679,350명이며 이중에서 남학생은 359,490명이고 여학생은 319,860명으로 남학생이 15,630명이 더 많다.

특히 중학교에서 남여공학과 남자중학교 및 여자중학교간의 발달환경과 사회환경의 차이가 있으므로 표본할당에서 이를 고려해야할 것이며 각 시도내에서도 도시지역과 군지역에서 생활환경과 경제적인 여건이 다르기 때문에 표본학교 추출에서 이들을 고려하기 위하여 중소도시와 군지역으로 층을 구분하였다. 16개 시도별 층의 구분기준은 우선 시지역과 군지역으로 구분한 후에 이들의 지역적 근접성을 기준으로 몇 개의 집락으로 그룹핑하였다. 서울시는 강남과 강북으로 구분하고 경기도는 대도시와 중소도시로 구분하였으며 다른 광역시는 하나의 층으로 하였고 군지역은 중소도시와 군지역으로 그룹핑하였으며 232개 시군구의 그룹핑결과는 다음 <표 3>에 정리하였다.

표 3

시군구의 집락화

시/도	특성별	그룹수	그룹명	포함되는 시/군/구
서울	대도시	2	서울대2-1	성동구, 광진구, 중랑구, 성북구, 강북구, 도봉구, 노원구, 종로구, 중구, 용산구, 은평구, 서대문구, 마포구, 동대문구
			서울대2-2	서초구, 강남구, 송파구, 강동구, 양천구, 강서구, 구로구, 금천구, 영등포구, 동작구, 관악구
부산	대도시	1	부산대1-1	동구, 부산진구, 북구, 해운대구, 사상구, 강서구, 중구, 서구, 영도구, 동래구, 남구, 사하구, 금정구, 연제구, 수영구, 기장군
대구	대도시	1	대구대1-1	수성구, 달서구, 서구, 중구, 남구, 북구, 동구, 달성군
인천	대도시	1	인천대1-1	부평구, 계양구, 남동구, 서구, 중구, 동구, 남구, 연수구, 강화군, 옹진군
광주	대도시	1	광주대1-1	북구, 동구, 광산구, 서구, 남구
대전	대도시	1	대전대1-1	서구, 중구, 유성구, 대덕구, 동구
울산	대도시	1	울산대1-1	남구, 동구, 중구, 북구, 울주군
경기	대도시	1	경기대1-1	수원시, 성남시, 고양시, 부천시
	중소도시	3	경기중3-1	파주시, 포천시, 동두천시, 양주시, 연천군, 가평군, 의정부시, 남양주시, 구리시, 양평군
			경기중3-2	안산시, 군포시, 의왕시, 안양시, 시흥시, 김포시, 과천시, 광명시, 하남시
			경기중3-3	안성시, 오산시, 평택시, 화성시, 여주군, 광주시, 용인시, 이천시
강원	중소도시	1	강원중1-1	춘천시, 원주시, 강릉시, 속초시, 동해시, 삼척시, 태백시
	군지역	1	강원군1-1	고성군, 인제군, 양구군, 화천군, 철원군, 양양군, 홍천군, 횡성군, 평창군, 정선군, 영월군
충북	중소도시	1	충북중1-1	청주시, 충주시, 제천시
	군지역	1	충북군1-1	단양군, 음성군, 괴산군, 진천군, 보은군, 옥천군, 영동군, 증평군, 청원군
충남	중소도시	1	충남중3-1	아산시, 천안시, 연기군, 공주시, 논산시, 계룡시, 서산시, 보령시,
	군지역	1	충남군1-1	서천군, 부여군, 청양군, 예산군, 당진군, 금산군, 홍성군, 연기군, 태안군
전북	중소도시	1	전북중2-1	군산시, 익산시, 전주시, 남원시, 정읍시, 김제시
	군지역	1	전북군1-1	부안군, 고창군, 진안군, 장수군, 임실군, 순창군, 무주군, 완주군
전남	중소도시	1	전남중1-1	순천시, 광양시, 여수시, 나주시, 목포시
	군지역	1	전남군1-1	신안군, 영광군, 장성군, 담양군, 화순군, 곡성군, 구례군, 함평군, 무안군, 영암군, 해남군, 진도군, 완도군, 강진군, 장흥군, 보성군, 고흥군
경북	중소도시	1	경북중1-1	문경시, 상주시, 김천시, 구미시, 안동시, 영주시, 영천시, 경산시, 포항시, 경주시
	군지역	1	경북군1-1	예천군, 의성군, 군위군, 칠곡군, 성주군, 고령군, 청도군, 울진군, 영양군, 영덕군, 청송군, 울릉군, 봉화군
경남	중소도시	1	경남중1-1	밀양시, 김해시, 양산시, 통영시, 거제시, 진주시, 사천시, 창원시, 마산시, 진해시
	군지역	1	경남군1-1	합천군, 창녕군, 의령군, 함안군, 고성군, 남해군, 하동군, 산청군, 함양군, 거창군
제주	중소도시	1	제주중1-1	제주시, 북제주군, 서귀포시, 남제주군

4.2 표본크기

표본크기는 표본추출법과 모집단의 특성 및 층화방법 등의 영향을 받지만 여기서는 통계생산단 위수준에서 고려하면, 이론적으로 통계분석단위별로 통계치의 목표오차를 어떻게 통제할 것인가에 따라서 아래 식으로 계산할 수 있다.

$$n = \frac{N(1.96 \cdot S)^2}{Nd^2 + (1.96 \cdot S)^2} = \frac{n_0}{1 + \frac{n_0}{N}}, \quad \text{단 } n_0 = \frac{(1.96 \cdot S)^2}{d^2}$$

여기서 N은 모집단의 크기이고 S은 관심변수의 표준편차이며 1.96은 신뢰수준 95%에서 신뢰계수이다.

다음은 조사내용이 비율인 경우에 모비율 추정에 대한 표본크기는 아래 식으로 계산할 수 있다.

$$n = \frac{\frac{2^2 \cdot PQ}{d^2}}{1 + \frac{1}{N} \left(\frac{2^2 \cdot PQ}{d^2} - 1 \right)}$$

여기서 N은 모집단의 크기이고 P는 모비율의 값이며 Q=1-P이며 2는 신뢰수준 95%에서 신뢰계수이다.

위 식에 주어진 분모의 오른쪽 항에서 N이 충분히 큰 경우에는 1/N이 무시할 정도로 작기 때문에 아래와 같이 근사식으로 계산할 수 있다.

$$n = \frac{\frac{2^2 \cdot PQ}{d^2}}{1 + \frac{1}{N} \left(\frac{2^2 \cdot PQ}{d^2} \right)}$$

실제로 표본추출이 학교단위로 이루어지고 선정된 학교에서 표본학급을 선정하여 해당학급의 모든 학생을 조사한다면 일종의 집락추출법이므로 허용오차를 충분하기 위해서는 위에서 산출한 값보다 표본크기는 더 크게 정하는 것이 좋겠지만 조사에 따른 소요비용과 조사기간 등을 고려하여 표본크기를 결정하는 방법이 현실적으로 적용되기도 한다.

한국 아동·청소년패널조사2010에서는 조사비용과 WAVE I 차와 WAVE II 차에서 패널유지율을 참조하여 각 코호트별 WAVE I 에서 표본크기를 2,200명으로 정하였다. 이의 근거는 WAVE II 이후에 유지되는 표본크기를 각 코호트별로 2,000명을 유지한다는 목표에서 결정하였다.

그러나 1차출단위인 학교를 기준으로 할 경우에는 표본학교당 평균적으로 약 80%명의 유효응답자가 조사될 것으로 가정하여 각 코호트별 표본학교수를 결정하였다. 즉, 각 코호트별 학급당 평균 학생수가 초등학교 1학년 28명, 초등학교 4학년 29명, 중학교 1학년 35명이며, 이것들의 80%는 초등학교 1학년 22.4명, 초등학교 4학년 23.2명, 중학교 1학년 28.0명이고, 이를 기준으로 조사될 학교수를 계산하면 초등학교 1학년 98학교, 초등학교 4학년 95학교, 중학교 1학년 79개교가 된다.

4.3 표본할당

각 시도별로 통계를 생산하지는 않지만 시도별 특성이 뚜렷하고 조사된 데이터를 이용하여 권역별로 특성분석이나 통계를 생산할 수 있도록 최소 표본을 우선 할당한 후에 학생수를 기준으로 비례배분법으로 할당표본크기를 계산하였다. 각 코호트별로 16개시도에 할당된 표본크기는 아래 <표 4>에 정리하였다.

표 4 **각 코호트별 시도별 표본할당 결과**

시도	초등 1학년			초등 4학년			중학교 1학년		
	학교수	학생수	표본수	학교수	학생수	표본수	학교수	학생수	표본수
강원	386	14,298	4	396	18,057	4	164	19,841	4
경기	1,141	120,729	15	1,144	150,891	15	544	163,985	10
경남	508	31,798	6	523	41,342	6	267	46,443	5
경북	540	23,196	6	545	30,549	5	297	33,587	5
광주	146	15,815	5	147	20,820	5	81	23,948	4
대구	214	23,028	5	215	30,945	5	122	37,629	4
대전	139	15,516	5	139	19,437	5	86	22,710	4
부산	299	27,317	6	299	36,040	6	170	45,628	5
서울	584	81,776	12	584	102,001	11	368	119,793	8
울산	118	11,064	4	119	15,179	4	60	18,190	4
인천	232	25,661	6	233	33,179	5	124	39,002	5
전남	496	17,403	5	517	22,940	5	266	25,687	4
전북	413	17,992	5	419	23,098	5	204	26,321	4
제주	115	6,292	4	115	7,825	4	42	8,675	4
충남	442	19,517	5	446	24,424	5	193	26,463	4
충북	273	14,866	5	276	18,838	5	129	21,448	4
합계	6,046	466,268	98	6,117	595,565	95	3,117	679,350	78

각 코호트별로 16개 시도별로 최소할당된 표본학교는 4개교이고, 초등학교1학년과 4학년의 서울시에 할당된 학교수는 각각 12, 11개교이며 중학교의 경우는 8개교이고, 경기도에는 초등학교 1학년과 4학년 각각 15개교가 할당되었으며 중학교는 10개교가 할당되었다.

<표 4>에 주어진 3개 코호트에서 각 시도별로 할당된 결과를 <표 4>에 주어진 층별로 배분은 학급수를 기준으로 비례배분하였으며 특히 중학교 1학년의 코호트에 대해서는 남여공학과 여학교 및 남자학교로 구분하여 할당한 후에 표본학교를 추출하여 남여간의 표본학생들의 분포를 균형있고 모집단의 분포와도 유사하도록 하였다. 중학교 1학년 코호트의 시도별 남여학교 별 도시규모별 학교수와 표본배분 결과를 <표 5>와 <표 6>에 정리하였다.

표 5

초등학교 1학년, 4학년의 시도별 도시규모별 학제별 표본할당결과

시도	규모	총구분	초등학교 1학년			초등학교 4학년		
			모집단		할당된 표본학교 수	모집단		할당된 표본학교 수
			학교 수	학급 수		학교 수	학급 수	
서울	대도시	서울대1	300	1,415	6	300	1,723	6
		서울대2	284	1,385	6	284	1,706	5
부산	대도시	부산대1	299	1,039	6	299	1,308	6
대구	대도시	대구대1	214	793	5	215	1,032	5
인천	대도시	인천대1	232	916	6	233	1,171	5
광주	대도시	광주대1	146	529	5	147	684	5
대전	대도시	대전대1	139	520	5	139	659	5
울산	대도시	울산대1	118	373	4	119	497	4
경기	대도시	경기대1	288	1,138	4	287	1,451	4
	중소도시	경기중1	273	739	3	273	897	3
		경기중2	229	939	4	230	1,175	4
		경기중3	351	1,036	4	354	1,179	4
강원	중소도시	강원중1	186	389	3	188	477	3
	군지역	강원군1	200	201	1	208	258	1
충북	중소도시	충북중1	123	341	3	124	436	3
	군지역	충북군1	150	219	2	152	265	2
충남	중소도시	충남중1	261	566	4	264	674	4
	군지역	충남군1	181	230	1	182	274	1
전북	중소도시	전북중1	276	598	4	281	728	4
	군지역	전북군1	137	154	1	138	184	1
전남	중소도시	전남중1	191	437	3	196	539	3
	군지역	전남군1	305	371	2	321	459	2
경북	중소도시	경북중1	363	749	5	364	955	4
	군지역	경북군1	177	199	1	181	248	1
경남	중소도시	경남중1	343	992	5	355	1,261	5
	군지역	경남군1	165	204	1	168	244	1
제주	중소도시	제주중1	115	235	4	115	276	4
합 계			6,046	16,707	98	6,117	20,760	95

표 6

중학교 1학년의 시도별 도시규모별 학제별 표본할당결과

시도	규모	총구분	모집단 학교 수			할당된 표본학교 수			
			남여공학	여학교	남학교	남여공학	여학교	남학교	합계
서울	대도시	서울대1	127	29	31	3	0	1	4
		서울대2	146	20	15	3	1	0	4
부산	대도시	부산대1	95	42	33	3	1	1	5
대구	대도시	대구대1	100	13	9	4	0	0	4
인천	대도시	인천대1	66	31	27	3	1	1	5
광주	대도시	광주대1	65	9	7	4	0	0	4
대전	대도시	대전대1	64	12	10	3	0	1	4
울산	대도시	울산대1	50	5	5	4	0	0	4
경기	대도시	경기대1	130	10	14	2	0	1	3
	중소도시	경기중1	110	7	7	2	0	0	2
		경기중2	110	4	5	3	0	0	3
		경기중3	126	12	9	1	1	0	2
강원	중소도시	강원중1	44	9	18	1	0	0	1
	군지역	강원군1	63	21	9	1	1	1	3
충북	중소도시	충북중1	42	10	10	1	0	0	1
	군지역	충북군1	54	7	6	2	0	1	3
충남	중소도시	충남중1	77	18	15	1	0	0	1
	군지역	충남군1	62	11	10	2	1	0	3
전북	중소도시	전북중1	98	13	14	1	0	0	1
	군지역	전북군1	67	7	5	2	0	1	3
전남	중소도시	전남중1	68	14	13	1	0	0	1
	군지역	전남군1	145	14	12	2	1	0	3
경북	중소도시	경북중1	116	38	31	1	0	0	1
	군지역	경북군1	84	15	13	2	1	1	4
경남	중소도시	경남중1	114	32	30	1	0	0	1
	군지역	경남군1	69	11	11	2	1	1	4
제주	중소도시	제주중1	28	7	7	2	1	1	4
합 계			2,320	421	376	57	10	11	78

<표 5>와 <표 6>에 각 층별로 할당된 표본학교 수와 모집단의 학교수와의 구성비에서 불일치한 측면을 볼 수 있는데 이는 학급수를 기준으로 비례배분하였기 때문이다. 중소도시는 학교수는 작지만 학급수가 많고 군지역은 학교의 규모가 작기 때문에 학교 수에 비해서 학급수가 작기 때문에 나타난 현상이다.

4.4 표본추출

표본추출은 학교추출과 학급추출로 나누어 2단계 추출법을 적용한다. 먼저 학교추출에 대해서 설명한다.

초등학교 2개 코호트에서는 시도별, 도시규모별로 할당된 표본학교들을 확률비례추출법으로 선정한다. 즉, 시도별 도시규모별 층내에서 학급수를 기준으로 확률비례추출법으로 표본학교를 선정하되, 선정된 표본학교는 최소한 각 코호트별로 2개 학급 이상이고 학생수가 50명 이상이 되도록 추출한 표본학교의 학급수와 학생수를 확인하고 만일에 조건을 충족하지 않을 경우에는 바로 인접한 학교를 표본학교로 대체한다.

중학교 1학년 코호트에서 표본학교 추출은 <표 6>에 주어진 시도별 도시규모별 학제별(남여공학, 여학교와 남학교)의 할당결과를 이용하여 확률비례추출법으로 선정한다. 즉, 시도별 도시규모별 학제별 층내에서 학급수를 기준으로 확률비례추출법으로 표본학교를 선정하되, 추출된 표본학교가 2개 학급 이상이고 학생수가 50명 이상인지의 여부를 확인하여 만일에 이 조건을 충족하지 않을 경우에는 바로 인접한 학교로 표본학교를 대체하여 선정된 표본학교내에서 표본학급을 선정하는데 최소한 2개 학급 이상에서 1개 학급을 선정할 수 있도록 조정한다.

표본학급의 선정은 3개 코호트별로 추출된 표본학교의 특성이나 소재지를 확인하고 한국아동·청소년패널조사2010에 참여에 대한 동의를 받은 후에 학급수와 학급별로 학생수(중학교는 남여별 학생수 포함)에 대한 정보를 확보한 후에 랜덤으로 표본학급을 선정한다.

본 보고서에서는 1차 추출단위인 표본학교의 추출에 대해서 주로 언급하였으며 추출된 표본학교의 명부는 <별첨>에 포함하였다.

4.5 표본관리

2010년 조사를 위해서 추출된 표본학교는 조사대상자인 학생을 선정하여 접촉하기 위한

표본추출과정이며 패널조사에 참여한 학생은 매년 개인별로 접촉하여 조사해야하므로 관리해야 할 것이다. 특히 학생이 조사에 참여해야만 학부모를 조사할 수 있으므로 완전한 응답자의 데이터는 학생과 학부모가 모두 응답한 경우이기 때문에 다른 패널조사보다 표본관리가 더 중요하다. 한국아동·청소년패널조사 2010은 앞으로 7년 동안 학생과 학부모를 추정해서 조사를 수행하기 때문에 응답거절이나 표본이탈을 최소화할 수 있도록 조사원의 노력과 사명의식이 절대 필요하다. 또한 응답자들이 본 패널조사에 참여할 수 있도록 금전적인 보상과 심리적인 성취감 등을 느낄 수 있도록 관리하는 방안이 필요하다. 각 코호트별로 목표유효응답자는 학생과 학부모를 쌍으로 한 2,000명으로 가정하였으므로 만일에 조사진행이 3년이나 4년이 지난 후에 패널조사 탈락이 많아지면 중간에 표본을 보충할 수 없으므로 표본의 대표성이 훼손될 수 있다. 만일에 표본의 대표성이 훼손될 경우에는 조사된 데이터로부터 모집단 전체에 대한 추론을 할 수 없거나 추론에 대한 신뢰성과 정확성이 낮아져서 통계로써 역할을 할 수 없을 것이다.

패널조사에서 패널의 탈락률을 낮추기 위한 다양한 방법들이 있을 수 있으나 중요한 것은 패널을 구축하는 첫 해의 조사에서 조사참여자들에게 본 패널의 목적과 취지를 충분히 설명하고 패널조사에 참여하면 앞으로 7년 동안은 조사에 참여해야함을 이해시켜야한다. 그리고 패널참여자들이 조사에 응답한 내용에 대한 피드백같은 유용한 정보를 제공하는 방안도 마련하여 응답자들에게 자긍심을 느낄 수 있도록 하여 심리적인 성취감을 갖도록 하는 것도 중요하다.

5. 가중치와 모수추정

5.1 가중치 계산

한국아동·청소년2010패널조사는 조사 웨이브를 여러 번 진행하면서도 동일표본을 조사하므로 조사된 표본들의 조사년도간의 순수변동은 정확하게 추정할 수 있으나, 패널효과와 모집단의 시간에 따른 변동에 의한 특성은 추정하기가 쉽지 않다, 조사 시점마다 패널유지율이 상이하므로 설계 당시에 자체가중설계를 했을지라도 조사 웨이브가 진행되면서 불균등확률표본설계가 되기 때문에 매 웨이브마다 가중치를 계산해야한다.

따라서 매 WAVE마다 조사시점을 기준으로 모집단 특성을 분석하는데 횡단적인 가중치를 사용하여야 할 것이며 2개 이상의 WAVE들간의 변동에 관한 분석을 위해서 종단적인 가중치를 사용해야 한다. 그러나 종단적인 가중치는 매 WAVE마다 조사가 종료된 후에 가중치를 계산하게 되므로 본 연구에는 포함되지 않았으며 다음 설명은 횡단적 가중치의 계산을 중심으로 설명하겠다.

먼저 첫 회의 조사에서 가중치 계산은 일반표본조사에서 가중치 계산과 같으므로 계산절차는 다음 세 가지 단계로 설명하겠다.

(1) 표본추출확률에 대한 가중치(기본가중치)

한국아동·청소년패널조사 2010의 표본추출은 일종의 층화2단집락추출법이라 할 수 있으며 시도와 도시규모를 층으로 간주할 수 있다. 따라서 각 층별로 표본을 할당하였으므로 각 층에서 1차추출단위인 학교는 학생수를 기준으로 확률비례추출법으로 선정하고 추출된 표본학교에서는 학급을 랜덤하게 선정한 후에 학생을 조사하므로 추출된 표본학교에서 학생을 랜덤하게 선정하여 조사하는 2단집락추출법으로 생각할 수 있으므로 표본추출률을 아래 식으로 나타낼 수 있다.

$$f_{hij} = \frac{B_{hi}}{\sum_{i=1}^{N_h} B_{hi}} \cdot \frac{n_{hi}}{B_{hi}} = \frac{n_{hi}}{\sum_{i=1}^{N_h} B_{hi}}$$

여기서 N_h 는 h 층에 속한 학교의 수이고, B_{hi} 는 h 층에서 i 번째 학교의 학생수를 나타내며 n_{hi} 는 추출된 표본학교에서 조사한 학생수를 나타낸다.

표본추출률에 대한 가중치는 추출률의 역수이므로 아래와 같은 식으로 표현할 수 있다.

$$W_{1hij} = \frac{\sum_{i=1}^{N_h} B_{hi}}{n_{hi}}$$

(2) 표본의 무응답 조정을 위한 가중치

표본으로 선정된 조사단위들 중에서 일부는 응답거부로 조사를 성공하지 못하거나 조사된 내용이 부실하여 무응답처리한 조사단위에 대한 조정이 무응답 조정의 가중치이며 아래 식으로 계산할 수 있다. $W_{2hij} = \frac{n_{hij}}{r_{hij}}$ 이며 r_{hij} 와 n_{hij} 는 각각 완전 응답한 표본크기와 추출된 표본크기이다.

(3) 사후층화 가중치(poststratification weight)

표본설계 시에는 층화변수로 고려하지 않았으나 표본조사 완료 후에 연구변수에 영향을 미치는 변수들중에서 정확한 외부보조정보를 확보하게 되면 외부보조정보 데이터의 구조와 표본조사 데이터의 구조를 분석하여 층으로 구분하고 표본조사의 각 셀에 가중치의 합계를 계산한 결과와 외부보조정보의 층별 합계와 비교하여 가중치를 계산한다.

$$BF_{hij} = \frac{PS_{hij}}{\sum W_{1hij} \cdot W_{2hij}}$$

여기서 PS_{hij} 는 외부보조정보에 의한 셀의 빈도수이고 분모의 내용은 설계가중치와 무응답 조정위한 가중치를 곱한 후에 합계를 계산결과이다.

최종가중치는 세 가지 가중치를 곱하여 아래와 같은 식으로 산출한다.

$$W_{f hij} = W_{1hij} \cdot W_{2hij} \cdot BF_{hij}$$

패널조사에서 1차년도 조사에서 가중치는 위의 일반 표본조사의 가중치작성방법을 적용하였지만 2차년도 이후에는 종단면가중치와 횡단면 가중치를 별도로 계산해야한다. 종단면 가중치는 1차년도의 가중치에 무응답 보정가중치를 곱하여 계산할 수 있다. 무응답 보정가중치는 응답률의 역수로 계산하지만 무응답률을 어느 단계에서 계산할 것인가는 표본설계와 패널추적조사의 규칙과 관계가 있다. 횡단면가중치는 조사시점에서 모집단의 상황을 반영하여 계산해야하므로

일종의 사후층화보정 가중치를 종단면가중치에 곱하여 계산하는 것이 일반적인 규칙이다. 그러나 분석단위와 패널추적조사의 규칙에 따라서 개인단위 가중치와 가구단위 가중치를 별도로 계산해야하는 경우가 있을 수 있으나 여기서는 개인단위 가중치에 관한 내용만 언급하였다.

5.2 모수추정

패널조사에서 모수추정은 생산할 통계의 종류에 따라서 다양하게 주어질 수 있으나 기본적인 모수추정량은 표본가중추정형식이다. 통계생산단위를 전국단위, 시도별과 성별로 한다는 가정에서 모비율이나 모평균의 추정량을 살펴보자.

우선 사용될 기호는 아래와 같다.

N_{ijk} : i 시도, j 도시규모, k 성별의 층의 규모(조사 학생수)

n_{ijk} : i 시도, j 도시규모, k 성별의 표본크기(조사 성공한 학생수)

W_{ijkl} : i 시도, j 도시규모, k 성별 층내에서 l 번째 조사된 사람의 가중치

y_{ijkl} : i 시도, j 도시규모, k 성별 층내에서 l 번째 조사된 사람의 관찰치

i 시도의 모비율 추정량은 표본가중추정량으로 아래와 같다.

$$\hat{P}_i = \frac{\sum_{j=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl} y_{ijkl}}{\sum_{j=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl}}$$

j 도시규모의 모비율 추정량도 표본가중추정량의 형식으로 아래와 같다.

$$\hat{P}_{+j} = \frac{\sum_{i=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl} y_{ijkl}}{\sum_{i=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl}}$$

또한 전국단위의 모평균이나 모비율의 추정치는 아래와 같은 식으로부터 계산할 수 있다.

$$\widehat{P}_{++} = \frac{\sum_{i=1} \sum_{j=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl} y_{ijkl}}{\sum_{i=1} \sum_{j=1} \sum_{k=1} \sum_{l=1}^{n_{ijk}} W_{ijkl}}$$

위 식에서 사용된 가중치는 표본추출률, 응답률 및 벤치마킹보정정보를 이용하여 계산하며 추정량의 편향(bias)을 최대한 줄일 수 있도록 연구되어야 할 것이다.

6. 결론

“한국아동·청소년패널조사2010 I”의 1차년도 조사를 위한 표본학교 추출에 관한 내용을 연구하였다. “한국아동·청소년패널조사2010 I”가 3개의 코호트로 구성되었으며 각 코호트별로 모집단을 정의하였으며 아동·청소년의 발달과 성장과정에 영향을 줄 수 있는 요소들중에서 모집단의 데이터에서 분석할 수 있는 요소가 지역과 도시규모이므로 이를 층화변수로 하여 모집단을 분석하였다.

조사방법은 학생과 학부모를 하나의 세트로 구성하여 조사해야하므로 먼저 학생을 조사한 후에 학교에서 가정통신문이나 전화 등의 연락수단을 이용하여 학부모를 조사하는 방법을 적용키로 하였으며 자료수집방법은 방문면접조사를 원칙으로 하였다.

표본설계에서 표본크기는 각 코호트별로 2,200명을 1차년도 목표유효표본으로 정하였으며 1차 추출단위인 표본학교수는 표본학교당 한 학급에서 평균적으로 약 80%명의 유효응답자가 조사될 것으로 가정하여 각 코호트별 표본학교수를 결정하였다. 이를 기준으로 조사될 학교수를 계산하면 초등학교 1학년 98학교, 초등학교 4학년 95학교, 중학교 1학년 79개교가 된다. 조사모집단을 초등학교에서는 시도와 도시규모로 층화하였고 각 층에서 표본학교는 학생수를 기준으로 확률비례추출법으로 추출하였다. 그러나 중학교는 시도와 도시규모로 층화한 후에 남녀공학, 여자학교와 남자학교로 층을 다시 나눈 후에 표본을 할당하였으며 할당된 표본학교는 각 층에서 학생수를 기준으로 확률비례추출법으로 추출하였다.

추출된 표본학교에서 학생의 선정은 우선 선정된 학교로부터 패널조사의 협조를 구한 후에 조사대상 학생들을 랜덤하게 추출하는 것으로 하였다.

앞에서 언급한 표본추출법은 층화2단집락추출법으로 실제로 1차년도와 2차년도의 조사를 완료한 후에 조사된 데이터를 기준으로 표본가중치를 계산하는 방법을 설명하였으며 표본가중치를 이용한 모수추정법을 설명하였다.

‘한국아동·청소년패널조사2010’의 조사를 성공적으로 수행하기 위해서는 패널에 참여한 학생과 학부모로부터 적극적인 협조가 우선적으로 필요하므로 이들의 협조를 받을 수 있는 방안과 조사원들의 적극적인 노력과 사명의식을 유발시킬 수 있는 방안에 대한 연구가 중요함을 강조하고 싶다.

본 연구에서는 언급되지 않았으나 표본의 대표성이나 표본오차의 관리보다 중요하다고 할 수 있는 비표본오차를 통제관리할 수 있는 방안이 마련되어 앞으로 아동과 청소년 관련 정책수립에 필요한 기초정보를 제공하는데 본 표본연구가 도움이 되었으면 한다.

참 고 문 헌

- 이경상 외4인(2003), “한국 청소년 패널조사(KYPS) I”, 한국청소년개발원
- 이계오 외 2인(2007), “표본조사론”, 한국방송통신대학교 출판부
- 이계오(2007), ‘고령화연구 패널조사의 표본설계’, 노동연구원
- Daniel Kasprzyk(1989), "Panel Surveys", John Wiley and Sons
- Gregory B. Markus(1979), "Analyzing Panel Data", Sage university paper
- Scott Menard(1991), "Longitudinal Research", Sage university paper
- Whitney Moore etc.(2000), "Technical Sampling Report of National Longitudinal Survey of Youth 1997(NLSY1997)", NORC. Chicago University

부 록

- 1-1. 초등학교 1학년과 4학년 시도별 집락별 분포 현황
- 1-2. 중학교 2학년 시도별 집락별 분포현황
- 2-1. 초등학교 1학년 대상 3차년도 가중치 부여방법
- 2-2. 초등학교 4학년 대상 3차년도 가중치 부여방법
- 2-3. 중학교 1학년 대상 3차년도 가중치 부여방법

부록 1-1 초등학교 1학년과 4학년 시도별 집락별 분포 현황

시도구분	집락구분	초등학교1학년			초등학교4학년		
		학교수	학급수	학생수	학교수	학급수	학생수
강원	강원군1_1	200	201	3,539	208	258	4,373
	강원중1_1	186	389	10,759	188	477	13,684
	소계	386	590	14,298	396	735	18,057
경기	경기대1_1	288	1,138	37,054	287	1,451	48,185
	경기중3_1	273	739	21,921	273	897	27,257
	경기중3_2	229	939	30,318	230	1,175	38,835
	경기중3_3	351	1,036	31,436	354	1,179	36,614
	소계	1,141	3,852	120,729	1,144	4,702	150,891
경남	경남군1_1	165	204	3,279	168	244	4,489
	경남중1_1	343	992	28,519	355	1,261	36,853
	소계	508	1,196	31,798	523	1,505	41,342
경북	경북군1_1	177	199	3,706	181	248	4,725
	경북중1_1	363	749	19,490	364	955	25,824
	소계	540	948	23,196	545	1,203	30,549
광주	광주대1_1	146	529	15,815	147	684	20,820
대구	대구대1_1	214	793	23,028	215	1,032	30,945
대전	대전대1_1	139	520	15,516	139	659	19,437
부산	부산대1_1	299	1,039	27,317	299	1,308	36,040
서울	서울대2_1	300	1,415	41,059	300	1,723	51,077
	서울대2_2	284	1,385	40,717	284	1,706	50,924
	소계	584	2,800	81,776	584	3,429	102,001
울산	울산대1_1	118	373	11,064	119	497	15,179
인천	인천대1_1	232	916	25,661	233	1,171	33,179
전남	전남군1_1	305	371	6,227	321	459	8,260
	전남중1_1	191	437	11,176	196	539	14,680
	소계	496	808	17,403	517	998	22,940
전북	전북군1_1	137	154	2,448	138	184	3,137
	전북중2_1	276	598	15,544	281	728	19,961
	소계	413	752	17,992	419	912	23,098
제주	제주중1_1	115	235	6,292	115	276	7,825
충남	충남군1_1	181	230	4,574	182	274	5,829
	충남중3_1	261	566	14,943	264	674	18,595
	소계	442	796	19,517	446	948	24,424
충북	충북군1_1	150	219	4,786	152	265	5,814
	충북중1_1	123	341	10,080	124	436	13,024
	소계	273	560	14,866	276	701	18,838
총합계		6,046	16,707	466,268	6,117	20,760	595,565

부록 1-2 중학교 2학년 시도별 집락별 분포현황

시도	집락구분	중2학년남여공학				중2학년 여자학교			중2학년 남자학교			중2학년 합계			
		학교	학급	남학생	여학생	학교	학급	여학생	학교	학급	남학생	학교	학급	남학생	여학생
강원	강원군1_1	63	103	1,257	1,101	9	30	935	9	35	1,097	81	168	2,354	2,036
	강원중1_1	44	164	2,663	2,592	21	138	4,803	18	154	5,393	83	456	8,056	7,395
	합계	107	267	3,920	3,693	30	168	5,738	27	189	6,490	164	624	10,410	9,431
경기	경기대1_1	130	1,274	26,926	23,327	10	100	3,698	14	92	3,597	154	1,466	30,523	27,025
	경기중3_1	110	654	12,883	11,775	7	46	1,673	7	47	1,736	124	747	14,619	13,448
	경기중3_2	110	1,076	22,081	19,691	4	33	1,248	5	36	1,334	119	1,145	23,415	20,939
	경기중3_3	126	742	14,123	13,951	12	65	2,467	9	89	3,475	147	896	17,598	16,418
	합계	476	3,746	76,013	68,744	33	244	9,086	35	264	10,142	544	4,254	86,155	77,830
경남	경남군1_1	69	106	1,203	995	11	40	1,242	11	49	1,521	91	195	2,724	2,237
	경남중1_1	114	702	13,345	11,092	32	226	8,072	30	253	8,973	176	1,181	22,318	19,164
	합계	183	808	14,548	12,087	43	266	9,314	41	302	10,494	267	1,376	25,042	21,401
경북	경북군1_1	84	136	1,783	1,467	15	28	791	13	35	921	112	199	2,704	2,258
	경북중1_1	116	440	7,285	6,391	38	196	6,782	31	238	8,167	185	874	15,452	13,173
	합계	200	576	9,068	7,858	53	224	7,573	44	273	9,088	297	1,073	18,156	15,431
광주	광주대1_1	65	495	9,516	9,611	9	49	1,965	7	71	2,856	81	615	12,372	11,576
대구	대구대1_1	100	829	16,239	15,033	13	65	2,349	9	104	4,008	122	998	20,247	17,382
대전	대전대1_1	64	505	9,778	8,793	12	55	1,850	10	65	2,289	86	625	12,067	10,643
부산	부산대1_1	95	775	14,168	12,462	42	258	8,751	33	299	10,247	170	1,332	24,415	21,213
서울	서울대2_1	127	1,247	23,515	20,167	29	219	7,399	31	213	7,199	187	1,679	30,714	27,566
	서울대2_2	146	1,432	26,983	24,491	20	126	4,344	15	159	5,695	181	1,717	32,678	28,835
	합계	273	2,679	50,498	44,658	49	345	11,743	46	372	12,894	368	3,396	63,392	56,401
울산	울산대1_1	50	419	8,140	6,746	5	46	1,582	5	51	1,722	60	516	9,862	8,328
인천	인천대1_1	66	485	9,887	9,446	31	230	9,141	27	259	10,528	124	974	20,415	18,587
전남	전남군1_1	145	243	3,067	2,799	14	46	1,389	12	51	1,524	171	340	4,591	4,188
	전남중1_1	68	303	5,102	4,655	14	100	3,439	13	108	3,712	95	511	8,814	8,094
	합계	213	546	8,169	7,454	28	146	4,828	25	159	5,236	266	851	13,405	12,282
전북	전북군1_1	67	102	1,084	953	7	20	629	5	23	693	79	145	1,777	1,582
	전북중2_1	98	533	9,851	8,167	13	83	2,786	14	67	2,158	125	683	12,009	10,953
	합계	165	635	10,935	9,120	20	103	3,415	19	90	2,851	204	828	13,786	12,535
제주	제주중1_1	28	124	2,303	2,075	7	52	2,062	7	55	2,235	42	231	4,538	4,137
충남	충남군1_1	62	128	1,777	1,546	11	49	1,616	10	51	1,733	83	228	3,510	3,162
	충남중3_1	77	351	5,970	5,233	18	120	4,184	15	126	4,404	110	597	10,374	9,417
	합계	139	479	7,747	6,779	29	169	5,800	25	177	6,137	193	825	13,884	12,579
충북	충북군1_1	54	145	2,141	1,860	7	29	932	6	32	1,014	67	206	3,155	2,792
	충북중1_1	42	276	5,153	4,534	10	78	2,778	10	85	3,036	62	439	8,189	7,312
	합계	96	421	7,294	6,394	17	107	3,710	16	117	4,050	129	645	11,344	10,104
총합계		2,320	13,789	258,223	230,953	421	2,527	88,907	376	2,847	101,267	3,117	19,163	359,490	319,860

부록 2-1 초등학교 1학년 대상 3차년도 가중치 부여방법

1. 횡단면 가중치 (Cross-sectional Weights)

1) 기본 가중치

기본 가중치는 표본탈락률을 보정하기 위하여 1차년도 대비 응답률(2,200/2,342=93.9%)의 역수를 1차년도 횡단면 가중치에 곱하여 계산되었다.

$$\text{기본 가중치} = \text{1차년도 가중치} \times \frac{1}{\text{1차년도 대비 응답률}}$$

2) 최종 가중치 (사후층화 가중치)

최종 가중치는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 초등학교 3학년 기준).

$$\text{최종 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

3) 가중치의 분포

3차년도 횡단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 1> 및 <표 2>에 제시되어 있다.

표 1 3차년도 횡단면 가중치의 요약 통계량

	<i>n</i>	최소값	최대값	평균	합계
전체	2,200	54	661	216	474,629
남자	1,134	54	661	218	247,027
여자	1,066	57	577	214	227,602

표 2 3차년도 횡단면 가중치의 지역별/성별 분포

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
전체	1,134	51.5	1,066	48.5	2,200	100.0	247,027	52	227,602	48	474,629	100.0
서울	127	55.2	103	44.8	230	10.5	42,031	51.7	39,245	48.3	81,276	17.1
부산	67	51.5	63	48.5	130	5.9	14,468	52.4	13,128	47.6	27,596	5.8
대구	59	50	59	50	118	5.4	12,485	52.6	11,229	47.4	23,714	5.0
인천	69	55.6	55	44.4	124	5.6	13,784	51.3	13,096	48.7	26,880	5.7
광주	61	53.5	53	46.5	114	5.2	8,800	52.7	7,884	47.3	16,684	3.5
대전	56	47.5	62	52.5	118	5.4	8,280	51.7	7,744	48.3	16,024	3.4
울산	52	51.5	49	48.5	101	4.6	6,121	53.6	5,297	46.4	11,418	2.4
경기	171	48.9	179	51.1	350	15.9	65,110	51.8	60,579	48.2	125,689	26.5
강원	44	53	39	47	83	3.8	7,256	51.9	6,726	48.1	13,983	2.9
충북	61	49.2	63	50.8	124	5.6	7,779	52.2	7,116	47.8	14,895	3.1
충남	46	41.8	64	58.2	110	5.0	10,451	52	9,648	48	20,099	4.2
전북	65	57.5	48	42.5	113	5.1	9,144	51.4	8,637	48.6	17,781	3.7
전남	62	53.4	54	46.6	116	5.3	8,755	51.9	8,101	48.1	16,856	3.6
경북	79	54.5	66	45.5	145	6.6	12,078	52.5	10,942	47.5	23,020	4.9
경남	68	51.5	64	48.5	132	6.0	17,065	53	15,140	47	32,206	6.8
제주	47	51.1	45	48.9	92	4.2	3,419	52.5	3,089	47.5	6,509	1.4

2. 종단면 가중치 (Longitudinal Weights)

1) 기본 가중치

3차년도 종단면 가중치는 1, 2, 3차년도에 모두 응답한 학생들(2,166명)만을 대상으로 계산되었다(1차년도 대비 응답률 $2,166/2,342=92.5\%$, 2차년도 대비 응답률 $2,166/2,264=95.7\%$).

3차년도 가중치 산출을 위해서 2차년도 조사를 통해 수집된 정보와 3차년도 응답여부 변수를 이용한 적절한 통계적 모형을 설정하고 이 모형에 근거한 3차년도 조사대상자의 응답확률 추정값을 계산하였다. 이를 위해 3차년도의 응답여부 변수를 응답의 경우 '1' 그리고 무응답의 경우 '0'을 갖는 이항변수로 정의하고 로지스틱 회귀모형을 이용한 응답확률 예측방법을 적용하였다.

로지스틱 회귀모형의 적합에 앞서 주요 설명변수의 응답률 예측에 대한 효율성을 알아보기 위하여 먼저 각 설명변수에 대하여 3차년도 응답자와 무응답자 간의 차이를 비교하는 것이 필요하다. 응답자 그룹과 무응답자 그룹 간의 차이를 변수별로 분석하기 위해 근사 χ^2 -검정을 시행하였다. 각 범주별 빈도수가 적어서 χ^2 -검정의 결과를 신뢰할 수 없는 경우 빈도수가 적은 범주를 묶은 후에 검정을 실시하였다.

여러 설명변수들의 동시적 예측력을 반영하며 동시에 최대우도 추정량의 유일성을 살피기 위하여 로지스틱 회귀모형의 변수선택 방법을 고려하였다. 변수선택을 위한 방법으로 각 변수들의 추가 또는 제거 후 모형의 설명력에 대한 Wald의 적합도 검정을 바탕으로 한 단계적 방법(stepwise method)을 사용하였다($\alpha=0.15$).

<표 3>에는 변수선택 과정을 통해 선택된 변수들과 로지스틱 회귀분석의 결과를 보여주고 있다. 3차년도 응답확률을 추정하기 위하여 사용된 변수들은 '교육수준-모(p2102bw2_r)', '근로일수-부-주중(p2104a1w2_r)', '근로일수-모-주중(p2104b1w2_r)', '학교숙제(분)-비등교일(p0403b2w2_r)', '친구들과 놀기(분)-비등교일(p0409b2w2_r)', 전학경험 유무(p11w2_r) 등이다.

<표 4>에는 응답확률 추정값에 대한 요약 통계량이 제시되어 있다. 응답확률의 추정값은 대략 0.500~0.999의 범위 내에 존재하며, 평균과 표준편차는 각각 0.965와 0.094 정도이다.

3차년도 종단면 기본 가중치는 3차년도 종단면 가중치에 응답확률 추정값의 역수를 곱하여 계산되었다.

$$\text{기본 가중치} = \text{2차년도 가중치} \times \frac{1}{\text{응답확률 추정값}}$$

표 3 로지스틱 회귀분석을 이용한 변수선택 결과

변수	자유도	Wald Chi-Square	p-value
p2102bw2_r	3	24.9	< .0001
p2104a1w2_r	3	171.1	< .0001
p2104b1w2_r	3	18.0	0.0004
p0403b2w2_r	8	16.5	0.0631
p0409b2w2_r	6	81.6	< .0001
p11w2_r	1	22.3	< .0001

표 4 응답확률 추정값의 요약 통계량

<i>n</i>	최소값	최대값	평균	표준편차
2,351	0.5000	0.9999	0.9647	0.0939

2) 최종 가중치 (사후층화 가중치)

최종 가중치(사후층화 가중치)는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 초등학교 3학년 기준).

$$\text{사후층화 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

3) 가중치의 분포

3차년도 종단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 5> 및 <표 6>에 제시되어 있다.

표 1 3차년도 종단면 가중치의 요약 통계량

	<i>n</i>	최소값	최대값	평균	합계
전체	2,166	53	1,088	219	474,626
남자	1,115	53	739	222	247,027
여자	1,051	54	1,088	217	227,599

표 2 3차년도 종단면 가중치의 지역별/성별 분포

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
전체	1,115	51.5	1,051	48.5	2,166	100.0	247,027	52.0	227,599	48.0	474,626	100.0
서울	126	55.5	101	44.5	227	10.5	42,031	51.7	39,245	48.3	81,276	17.1
부산	66	52.4	60	47.6	126	5.8	14,468	52.4	13,128	47.6	27,596	5.8
대구	58	49.6	59	50.4	117	5.4	12,485	52.6	11,229	47.4	23,713	5.0
인천	66	54.5	55	45.5	121	5.6	13,784	51.3	13,096	48.7	26,880	5.7
광주	57	52.3	52	47.7	109	5.0	8,800	52.7	7,884	47.3	16,684	3.5
대전	55	47.4	61	52.6	116	5.4	8,280	51.7	7,744	48.3	16,024	3.4
울산	52	51.5	49	48.5	101	4.7	6,121	53.6	5,297	46.4	11,418	2.4
경기	169	49.0	176	51.0	345	15.9	65,110	51.8	60,580	48.2	125,690	26.5
강원	44	53.7	38	46.3	82	3.8	7,256	51.9	6,726	48.1	13,982	2.9
충북	58	48.3	62	51.7	120	5.5	7,779	52.2	7,116	47.8	14,895	3.1
충남	46	42.6	62	57.4	108	5.0	10,451	52.0	9,648	48.0	20,099	4.2
전북	62	56.4	48	43.6	110	5.1	9,144	51.4	8,637	48.6	17,781	3.7
전남	62	53.4	54	46.6	116	5.4	8,756	51.9	8,101	48.1	16,857	3.6
경북	79	54.9	65	45.1	144	6.6	12,078	52.5	10,941	47.5	23,019	4.8
경남	68	51.5	64	48.5	132	6.1	17,065	53.0	15,140	47.0	32,205	6.8
제주	47	51.1	45	48.9	92	4.2	3,419	52.5	3,089	47.5	6,508	1.4

부록 2-2 초등학교 4학년 대상 3차년도 가중치 부여방법

1. 횡단면 가중치 (Cross-sectional Weights)

1) 기본 가중치

기본 가중치는 표본탈락률을 보정하기 위하여 1차년도 대비 응답률(2,219/2,378=93.3%)의 역수를 1차년도 횡단면 가중치에 곱하여 계산되었다.

$$\text{기본 가중치} = \text{1차년도 가중치} \times \frac{1}{\text{1차년도 대비 응답률}}$$

2) 사후층화 가중치

사후층화 가중치는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 초등학교 6학년 기준).

$$\text{사후층화 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

표 1 3차년도 횡단면 가중치의 요약 통계량

	<i>n</i>	최소값	최대값	평균	합계
전체	2,219	62	770	272	602,682
남자	1,167	62	750	270	315,567
여자	1,052	75	770	273	287,115

표 2 3차년도 횡단면 가중치의 지역별/성별 분포

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,166	52.6	1,052	47.4	2,218	100.0	315,423	52.3	287,115	47.7	602,538	100.0
서울	130	51.8	121	48.2	251	11.3	53,240	52.1	48,955	47.9	102,195	17.0
부산	68	49.3	70	50.7	138	6.2	19,018	52.8	16,979	47.2	35,997	6.0
대구	74	54.0	63	46.0	137	6.2	16,412	53.2	14,453	46.8	30,865	5.1
인천	51	53.7	44	46.3	95	4.3	17,300	51.9	16,009	48.1	33,309	5.5
광주	55	51.4	52	48.6	107	4.8	11,373	52.4	10,339	47.6	21,712	3.6
대전	61	55.5	49	44.5	110	5.0	10,281	51.7	9,594	48.3	19,875	3.3
울산	51	51.0	49	49.0	100	4.5	7,945	52.9	7,086	47.1	15,031	2.5
경기	184	52.0	170	48.0	354	16.0	80,626	52.1	74,232	47.9	154,858	25.7
강원	53	52.0	49	48.0	102	4.6	9,619	52.9	8,560	47.1	18,179	3.0
충청	66	47.8	72	52.2	138	6.2	10,117	52.6	9,123	47.4	19,240	3.2
전라	65	51.6	61	48.4	126	5.7	13,375	52.3	12,205	47.7	25,580	4.2
경상	72	57.1	54	42.9	126	5.7	12,067	51.7	11,251	48.3	23,318	3.9

* 기타(대안학교, 학교 안다님, 무응답) 1명은 사후층화의 과정이 생략됨.

3) 가중치의 분포

3차년도 횡단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 1> 및 <표 2>에 제시되어 있다.

2. 종단면 가중치 (Longitudinal Weights)

1) 기본 가중치

3차년도 종단면 가중치는 1, 2, 3차년도에 모두 응답한 학생들(2,170명)만을 대상으로 계산되었다(1차년도 대비 응답률 $2,170/2,378=91.3\%$, 2차년도 대비 응답률 $2,170/2,263=95.9\%$).

3차년도 가중치 산출을 위해서 2차년도 조사를 통해 수집된 정보와 3차년도 응답여부

변수를 이용한 적절한 통계적 모형을 설정하고 이 모형에 근거한 3차년도 조사대상자의 응답확률 추정값을 계산하였다. 이를 위해 3차년도의 응답여부 변수를 응답의 경우 '1' 그리고 무응답의 경우 '0'을 갖는 이항변수로 정의하고 로지스틱 회귀모형을 이용한 응답확률 예측방법을 적용하였다.

로지스틱 회귀모형의 적합에 앞서 주요 설명변수의 응답률 예측에 대한 효율성을 알아보기 위하여 먼저 각 설명변수에 대하여 3차년도 응답자와 무응답자 간의 차이를 비교하는 것이 필요하다. 응답자 그룹과 무응답자 그룹 간의 차이를 변수별로 분석하기 위해 근사 χ^2 -검정을 시행하였다. 각 범주별 빈도수가 적어서 χ^2 -검정의 결과를 신뢰할 수 없는 경우 빈도수가 적은 범주를 묶은 후에 검정을 실시하였다. 범주가 새롭게 정의된 변수들의 범주 값과 빈도표 및 χ^2 -검정의 결과는 부록에 제시되어 있다.

여러 설명변수들의 동시적 예측력을 반영하며 동시에 최대우도 추정량의 유일성을 살피기 위하여 로지스틱 회귀모형의 변수선택 방법을 고려하였다. 변수선택을 위한 방법으로 각 변수들의 추가 또는 제거 후 모형의 설명력에 대한 Wald의 적합도 검정을 바탕으로 한 단계적 방법(stepwise method)을 사용하였다($\alpha=0.15$).

초등학교 4학년 패널의 경우, 위와 같은 과정을 적용하였으나 통계적으로 유의하게 선택된 변수가 없었다. 따라서 3차년도 종단면 기본 가중치는 2차년도 종단면 가중치와 동일한 값으로 적용되었다.

$$\text{기본 가중치} = \text{2차년도 가중치}$$

2) 사후층화 가중치

사후층화 가중치는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 초등학교 6학년 기준).

$$\text{사후층화 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

3) 가중치의 분포

3차년도 종단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 3> 및 <표 4>에 제시되어 있다.

표 3 3차년도 종단면 가중치의 요약 통계량

	<i>n</i>	최소값	최대값	평균	합계
전체	2,170	60	1,012	278	602,674
남자	1,137	60	1,012	278	315,559
여자	1,033	76	920	278	287,115

표 4 3차년도 종단면 가중치의 지역별/성별 분포

	가중치 부여 이전						종단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
전체	1,136	52.4	1,033	47.6	2,169	100.0	315,423	52.3	287,115	47.7	602,538	100.0
서울	129	51.8	120	48.2	249	11.5	53,240	52.1	48,955	47.9	102,195	17.0
부산	66	49.6	67	50.4	133	6.1	19,018	52.8	16,979	47.2	35,997	6.0
대구	74	54.4	62	45.6	136	6.3	16,412	53.2	14,453	46.8	30,865	5.1
인천	49	52.7	44	47.3	93	4.3	17,300	51.9	16,009	48.1	33,309	5.5
광주	54	51.4	51	48.6	105	4.8	11,373	52.4	10,339	47.6	21,712	3.6
대전	59	55.1	48	44.9	107	4.9	10,281	51.7	9,594	48.3	19,875	3.3
울산	49	51.0	47	49.0	96	4.4	7,945	52.9	7,086	47.1	15,031	2.5
경기	170	50.9	164	49.1	334	15.4	80,626	52.1	74,232	47.9	154,858	25.7
강원	53	52.0	49	48.0	102	4.7	9,619	52.9	8,560	47.1	18,179	3.0
충청	66	48.2	71	51.8	137	6.3	10,117	52.6	9,123	47.4	19,240	3.2
전라	65	51.6	61	48.4	126	5.8	13,375	52.3	12,205	47.7	25,580	4.2
경상	69	56.1	54	43.9	123	5.7	12,067	51.7	11,251	48.3	23,318	3.9

* 기타(대안학교, 학교 안다님, 무응답) 1명은 사후충화의 과정이 생략됨.

부록 2-3 중학교 1학년 대상 3차년도 가중치 부여방법

1. 횡단면 가중치 (Cross-sectional Weights)

1) 기본 가중치

기본 가중치는 표본탈락률을 보정하기 위하여 1차년도 대비 응답률($2,259/2,351=96.1\%$)의 역수를 1차년도 횡단면 가중치에 곱하여 계산되었다.

$$\text{기본 가중치} = \text{1차년도 가중치} \times \frac{1}{\text{1차년도 대비 응답률}}$$

2) 최종 가중치 (사후층화 가중치)

최종 가중치는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 중학생 3학년 기준).

$$\text{최종 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

3) 가중치의 분포

3차년도 횡단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 1> 및 <표 2>에 제시되어 있다.

표 1 2차년도 횡단면 가중치의 요약 통계량

	<i>n</i>	최소값	최대값	평균	합계
전체	2,259	42	1,256	283	640,052
남자	1,140	42	1,256	292	332,677
여자	1,119	77	883	275	307,375

표 2 3차년도 횡단면 가중치의 지역별/성별 분포

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%	<i>N</i>	%	<i>N</i>	%	<i>N</i>	%
전체	1,140	50.5	1,119	49.5	2,259	100.0	332,676	52.0	307,375	48.0	640,052	100.0
서울	91	42.7	122	57.3	213	9.4	57,649	52.3	52,675	47.7	110,325	17.2
부산	73	56.6	56	43.4	129	5.7	21,193	52.4	19,250	47.6	40,444	6.3
대구	79	51.0	76	49.0	155	6.9	18,364	53.0	16,293	47.0	34,657	5.4
인천	42	28.0	108	72.0	150	6.6	18,652	51.4	17,609	48.6	36,261	5.7
광주	38	41.8	53	58.2	91	4.0	12,025	51.6	11,266	48.4	23,292	3.6
대전	51	49.5	52	50.5	103	4.6	11,128	52.2	10,196	47.8	21,324	3.3
울산	74	67.9	35	32.1	109	4.8	8,824	52.6	7,942	47.4	16,767	2.6
경기	152	45.5	182	54.5	334	14.8	81,464	51.5	76,730	48.5	158,194	24.7
강원	71	64.5	39	35.5	110	4.9	9,901	51.4	9,366	48.6	19,267	3.0
충북	61	49.2	63	50.8	124	5.5	10,545	51.2	10,066	48.8	20,611	3.2
충남	48	52.7	43	47.3	91	4.0	13,516	51.8	12,593	48.2	26,108	4.1
전북	67	59.8	45	40.2	112	5.0	12,914	52.1	11,857	47.9	24,771	3.9
전남	41	38.3	66	61.7	107	4.7	12,425	51.7	11,603	48.3	24,028	3.8
경북	87	55.8	69	44.2	156	6.9	16,724	52.4	15,183	47.6	31,907	5.0
경남	80	56.3	62	43.7	142	6.3	22,966	52.5	20,805	47.5	43,771	6.8
제주	85	63.9	48	36.1	133	5.9	4,386	52.7	3,941	47.3	8,327	1.3

2. 종단면 가중치 (Longitudinal Weights)

1) 기본 가중치

3차년도 종단면 가중치는 1, 2, 3차년도에 모두 응답한 학생들(2,226명)만을 대상으로 계산되었다(1차년도 대비 응답률 $2,226/2,351=94.7\%$, 2차년도 대비 응답률 $2,226/2,276=97.8\%$).

3차년도 가중치 산출을 위해서 2차년도 조사를 통해 수집된 정보와 3차년도 응답여부 변수를 이용한 적절한 통계적 모형을 설정하고 이 모형에 근거한 3차년도 조사대상자의 응답확률 추정값을 계산하였다. 이를 위해 3차년도의 응답여부 변수를 응답의 경우 '1' 그리고 무응답의 경우 '0'을 갖는 이항변수로 정의하고 로지스틱 회귀모형을 이용한 응답확률 예측방법을 적용하였다.

로지스틱 회귀모형의 적합에 앞서 주요 설명변수의 응답률 예측에 대한 효율성을 알아보기 위하여 먼저 각 설명변수에 대하여 3차년도 응답자와 무응답자 간의 차이를 비교하는 것이 필요하다. 응답자 그룹과 무응답자 그룹 간의 차이를 변수별로 분석하기 위해 근사 χ^2 -검정을 시행하였다. 각 범주별 빈도수가 적어서 χ^2 -검정의 결과를 신뢰할 수 없는 경우 빈도수가 적은 범주를 묶은 후에 검정을 실시하였다.

여러 설명변수들의 동시적 예측력을 반영하며 동시에 최대우도 추정량의 유일성을 살피기 위하여 로지스틱 회귀모형의 변수선택 방법을 고려하였다. 변수선택을 위한 방법으로 각 변수들의 추가 또는 제거 후 모형의 설명력에 대한 Wald의 적합도 검정을 바탕으로 한 단계적 방법(stepwise method)을 사용하였다($\alpha=0.15$).

<표 3>에는 변수선택 과정을 통해 선택된 변수들과 로지스틱 회귀분석의 결과를 보여주고 있다. 3차년도 응답확률을 추정하기 위하여 사용된 변수들은 '다른 사람 집단따돌림 시키기(qa2306w2_r)', '학교숙제(시간)-비등교일(q0303b1w2_r)' 등이다.

<표 4>에는 응답확률 추정값에 대한 요약 통계량이 제시되어 있다. 응답확률의 추정값은 대략 0.5000~1.0000의 범위 내에 존재하며, 평균과 표준편차는 각각 0.9842와 0.0746 정도이다.

3차년도 종단면 기본 가중치는 1차년도 종단면 가중치에 응답확률 추정값의 역수를 곱하여 계산되었다.

$$\text{기본 가중치} = \text{2차년도 가중치} \times \frac{1}{\text{응답확률 추정값}}$$

표 3 로지스틱 회귀분석을 이용한 변수선택 결과

변수	자유도	Wald Chi-Square	p-value
qa2306w2_r	7	14.7	< 0.0406
q0303b1w2_r	2	78.0	< 0.0001

표 4 **응답확률 추정값의 요약 통계량**

<i>n</i>	최소값	최대값	평균	표준편차
2,276	0.5000	1.0000	0.9842	0.0746

2) 최종 가중치 (사후층화 가중치)

최종 가중치(사후층화 가중치)는 해당 학생이 속한 지역별/성별 기본 가중치의 합계와 모집단 크기를 고려하여 다음과 같이 계산되었다(2012년의 중학생 3학년 기준).

$$\text{사후층화 가중치} = \text{기본 가중치} \times \frac{\text{지역별/성별 모집단 크기}}{\text{지역별/성별 기본 가중치의 합계}}$$

3) 가중치의 분포

3차년도 종단면 가중치의 요약 통계량 및 지역별/성별 분포가 <표 5> 및 <표 6>에 제시되어 있다.

표 5 **3차년도 종단면 가중치의 요약 통계량**

	<i>n</i>	최소값	최대값	평균	합계
전체	2,226	42	1,994	288	640,052
남자	1,128	42	1,302	295	332,677
여자	1,098	76	1,994	280	307,375

표 6 3차년도 종단면 가중치의 지역별/성별 분포

	가중치 부여 이전						횡단면 가중치 부여 이후					
	남자		여자		전체		남자		여자		전체	
	n	%	n	%	n	%	N	%	N	%	N	%
전체	1,128	50.7	1,098	49.3	2,226	100.0	332,677	52.0	307,375	48.0	640,052	100.0
서울	91	43.3	119	56.7	210	9.4	57,649	52.3	52,675	47.7	110,324	17.2
부산	72	56.7	55	43.3	127	5.7	21,193	52.4	19,250	47.6	40,443	6.3
대구	78	51.0	75	49.0	153	6.9	18,364	53.0	16,293	47.0	34,657	5.4
인천	42	29.0	103	71.0	145	6.5	18,652	51.4	17,609	48.6	36,261	5.7
광주	38	42.2	52	57.8	90	4.0	12,025	51.6	11,266	48.4	23,291	3.6
대전	50	49.0	52	51.0	102	4.6	11,128	52.2	10,196	47.8	21,324	3.3
울산	73	69.5	32	30.5	105	4.7	8,824	52.6	7,942	47.4	16,766	2.6
경기	151	45.8	179	54.2	330	14.8	81,464	51.5	76,730	48.5	158,194	24.7
강원	70	64.2	39	35.8	109	4.9	9,901	51.4	9,366	48.6	19,267	3.0
충북	60	48.8	63	51.2	123	5.5	10,545	51.2	10,066	48.8	20,612	3.2
충남	46	52.9	41	47.1	87	3.9	13,516	51.8	12,593	48.2	26,109	4.1
전북	67	59.8	45	40.2	112	5.0	12,914	52.1	11,857	47.9	24,771	3.9
전남	40	38.1	65	61.9	105	4.7	12,425	51.7	11,603	48.3	24,028	3.8
경북	85	55.6	68	44.4	153	6.9	16,724	52.4	15,183	47.6	31,907	5.0
경남	80	56.3	62	43.7	142	6.4	22,966	52.5	20,805	47.5	43,771	6.8
제주	85	63.9	48	36.1	133	6.0	4,386	52.7	3,941	47.3	8,327	1.3

MEMO

MEMO

MEMO

MEMO

MEMO

MEMO

**2013년 한국아동·청소년 패널
제3차 콜로키움 자료집 데이터분석방법론 표본설계와 추정**

인 쇄 2013년 12월 13일

발 행 2013년 12월 16일

발행처 **한국청소년정책연구원**
서울특별시 서초구 태봉로 114

발행인 이 재 연

등 록 1993. 10. 23 제 21-500호

인쇄처 문영사 전화 02)2263-5087 대표 김희자

사전 승인없이 보고서 내용의 무단전재·복제를 금함.
구독문의 : (02) 2188-8844(연구기획·대외협력팀)