

학생 대상 조사의 표본설계와 조사데이터 분석

2012년 4월 5일

//////
이기재 [한국방송통신대학교, 정보통계학과]

kjlee@knou.ac.kr
//////

차 례

I. 복합표본설계 개요

- 표본설계, 복합표본조사의 구성요소, 사례 등

II. 학생 대상 조사 표본설계 사례

- 아동·청소년인권실태조사, 알레르기 질환조사 등

III. 설계효과의 개념과 활용

- 설계효과, 집락내상관계수, 사례 등

IV. 조사데이터 분석

- 기술통계분석, 회귀분석, 카이제곱 검증 등

V. 결론

1. 표본설계란?

- 모집단을 효과적으로 대표해서 적은 비용으로 정확한 조사가 가능하게 하는 표본추출에 대한 전반적인 계획
- 표본설계의 구성
 - 조사단위, 모집단 정의
 - 목표오차, 표본크기 결정
 - 층화변수, 층화 기준, 표본배분법 등
 - 모수 추정방법, 추정의 정확도 계산
 - 통계의 공표 수준 검토

2. 표본설계 비교

- 전국 고등학생을 대상으로 청소년 인권실태 조사 (n=2,000명)
- 어떻게 표본 학생을 선정할 것인가?

■ 표본설계 1

- 전국 고등학생 개인 명부에서 2,000명을 랜덤추출

■ 표본설계 4 (현실)

- 전국 고등학교를 계열과 지역에 따라 6개 층으로 구분
 - ➔ 전문계와 기타 지역에서 과표집
 - ➔ 각 표본학교에서 2개 학급 추출
 - ➔ 각 표본학급에서 20명씩 조사

- 같은 표본크기이지만 조사의 정확도는 어떨까?
- 조사의 편리성과 비용은?
- 분석 방법은 어떻게 달라질까?

3. 복합표본설계

- 단순임의표집이외의 다른 방법으로 추출되는 모든 경우
- 실제 조사는 비용, 조사관리, 표본추출틀 마련 등의 이유로 단순임의표집을 적용하기 곤란함
- 대규모 표본조사는 대부분 랜덤화, 층화, 집락화, 불균등추출 등 이용

4. 전국 단위의 청소년 대상 조사의 특징

- 복합표본설계에 의해서 표본학교 및 학급 선정

- * 층화, 다단계 추출, 불균등 추출확률 등

- 다수의 조사항목 대상 : 설문지 이용

- 다양한 추정 대상

- * 전국, 지역간 비교, 인구사회학적 비교 등

5. 복합표본설계 구성요소

- 층화(stratification)
- 집락화(clustering)
- 불균등확률추출
- 가중치 사용

(1) 층화(stratification)

- 모집단에 대한 대표성을 높이기 위해서 사용
 - 예) 지역, 성별, 연령, 학력 등이 층화 변수로 사용됨
 - 예) 지역, 학교구분(중학교, 일반계, 전문계), 남녀공학 여부 등
- 일반적으로 추정의 정확도를 높여줌
- 층화를 무시하고 분석하면 추정 편향이 발생하고, 추정량의 분산도 커지게 됨

(2) 집락화(clustering)

- 모집단에 대한 최신 추출틀이 없는 경우 표본추출을 위한 현실적 방법
- 조사원의 업무 부담과 조사비용을 고려한 방법
- 지역적으로 인접한 가구나 사람들이 표본으로 추출됨
 - ➔ 지역적 인접성 및 구성원이 유사할 수 있어 응답결과에서도 유사성 발생 가능성이 있음(intracluster correlation coefficient: 집락내상관계수)
 - ➔ 학교나 학급을 추출단위로 접근하는 청소년 대상 조사에서 고려해야 함

(3) 불균등확률추출

- 주로 집단간의 비교를 목적으로 Oversampling하는 경우에 발생
→ 지역간 비교(대도시/중소도시/군지역), 학교 유형별 비교(일반계/전문계)
- 표본추출 상황에서 불가피하게 발생하는 경우
예) 전화조사에서 전화보유 대수
- 추출확률을 무시하고 분석하면 추정 편향 발생
⇒ 가중치 사용

(4) 가중값

- 조사단위가 전체 모집단 중 얼마만큼 대표하는가를 나타냄
 - 과표집 (oversampling)에 따른 추출률 보정
→ 추정의 편향 (어긋남, bias)를 줄여줌
 - 무응답으로 인한 응답확률의 차이로 발생하는 편향을 줄여줌
 - 추출틀의 부정확, noncoverage 등으로 인한 편향을 줄여줌
- ※ 조사데이터에 가중값이 포함되어 있어야 분석자(analyst)가 다양한 분석을 할 수 있음.

(5) 가중값 사용 효과

- 불균등추출확률, 무응답, 추출틀의 부정확 등으로 발생하는 추정의 편향(bias)을 줄여줌
- 가중값의 변동이 커짐 $\Rightarrow$ 추정량의 정도가 떨어짐
- $F = 1 + CV(w_i)^2$: variance inflation factor
 $CV(w_i)$: 가중값(w_i)의 변동계수

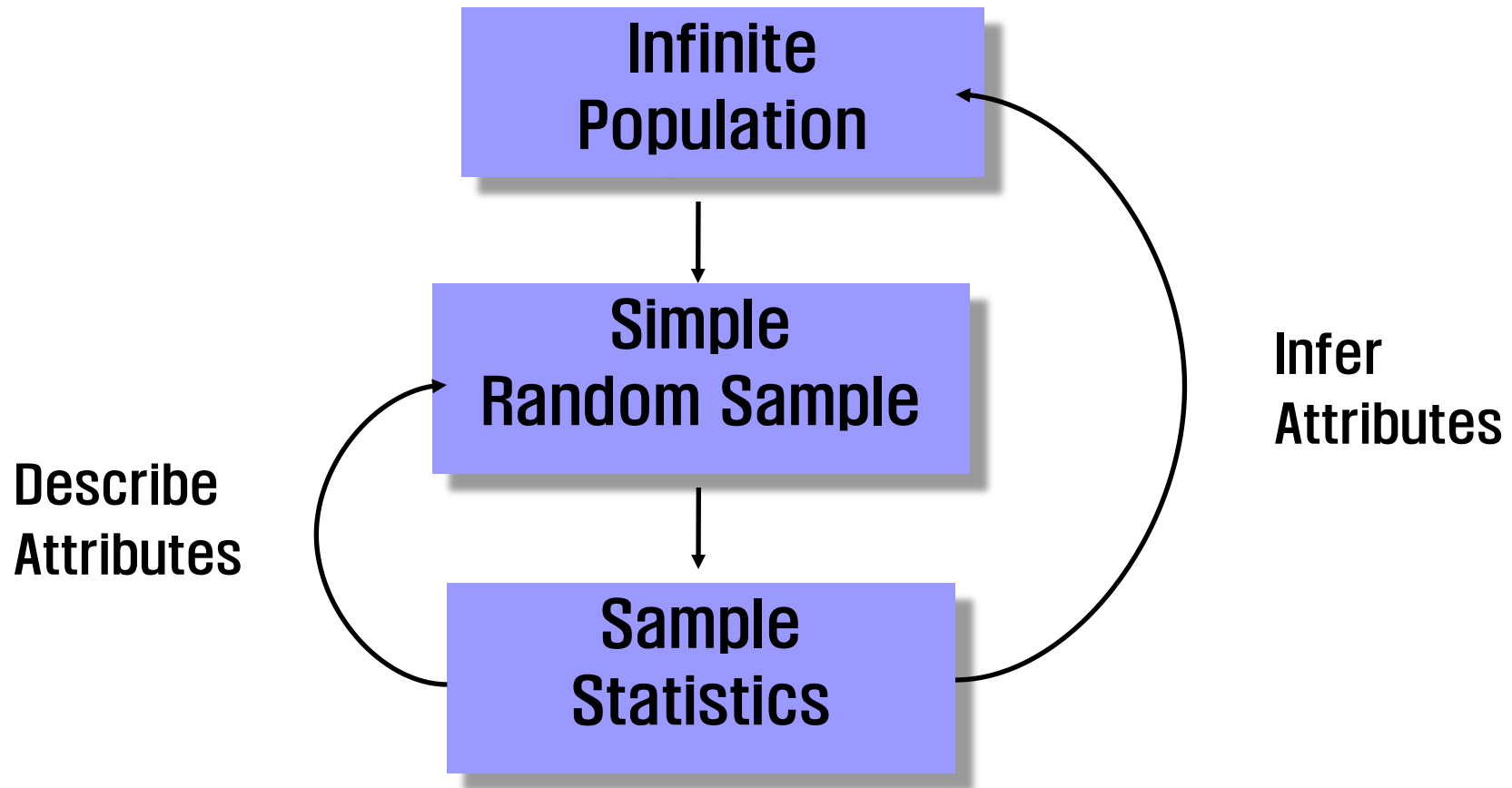
[6] 가중치 현황 : 아동·청소년인권실태조사 (2009)

구분	초등학교 (n=2,889)	중학교 (n=3,117)	일반고 (n=2,174)	전문고 (n=1,310)
최대값	1.17	1.25	1.63	1.85
Q3	1.05	1.09	1.08	1.25
중위수	0.96	1.07	1.06	0.86
Q1	0.92	0.96	1.03	0.59
최소값	0.83	0.60	0.37	0.34
평균	1.00	1.00	1.00	1.00
CV(%)	9.71	14.35	29.82	46.37

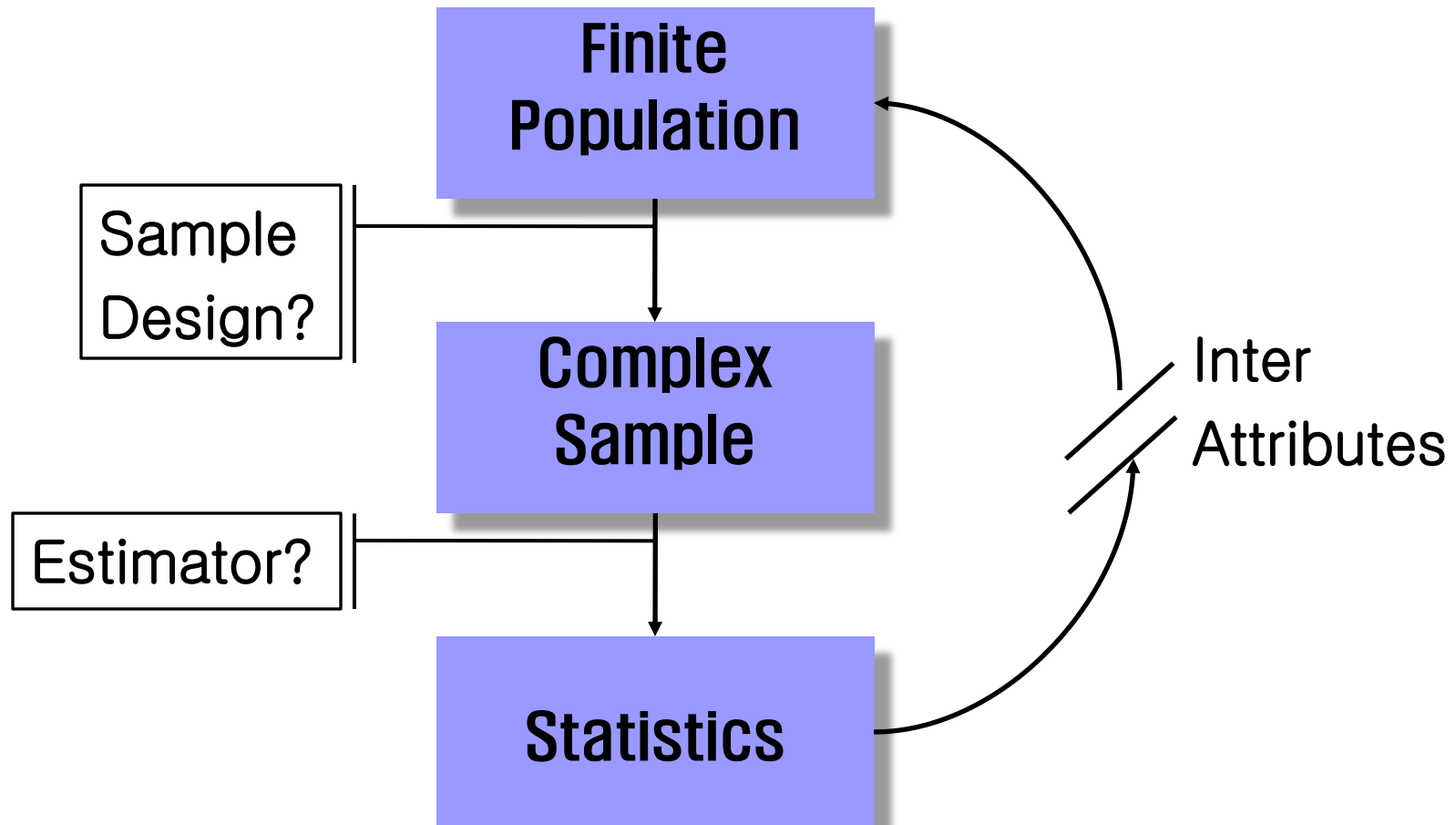
6. 복합표본조사 데이터의 특징

- 유한모집단 대상
- 복합표본설계에 의한 추출
 - 층화, 집락추출 : 표본 단위들의 비독립성
 - 불균등확률추출
- 가중치 부여
 - 불균등확률선택 보정 방법
 - 무응답 및 사후층화 보정 방법

7. 단순임의표본에서 추론



8. 복합표본조사에서 추론



차 례

I. 복합표본설계 개요

- 표본설계, 복합표본조사의 구성요소, 사례 등

II. 학생 대상 조사 표본설계 사례

- 아동·청소년인권실태조사, 알레르기 질환조사 등

III. 설계효과의 개념과 활용

- 설계효과, 집락내상관계수, 사례 등

IV. 조사데이터 분석

- 기술통계분석, 회귀분석, 카이제곱 검증 등

V. 결론

1. 아동·청소년인권실태조사 [1]

■ 조사개요

- 우리나라 아동·청소년의 인권실태 파악을 목적으로 함
- 초등학교 4학년부터 고등학교 3학년까지의 재학생 대상 자기기입식조사

■ 모집단 및 조사대상

- 모집단 : 전국 초등학교 4-6학년, 중학교 1-3학년, 고등학교 1-3학년
- 표본 추출틀 : 전국 초등학교, 중학교 및 고등학교 리스트
- 1차 추출단위 : 학교, 2차 추출단위 : 학급, 3차 추출단위: 학생

1. 아동·청소년인권실태조사 [2]

■ 총화

- 지역구분: 광역시도(16개)
- 학교구분: 초등학교, 중학교, 일반계고등학교, 전문계 고등학교(4개)
- 학년구분: 학년(3개)

■ 표본크기

- 조사비용과 시간을 고려하여 8,500명으로 결정
- 단순임의추출법을 가정하는 경우 95% 신뢰수준에서 $\pm 1.1\%P$ 예상
➔ 총화집락추출법을 적용하는 경우에는 약간 크게 될 것으로 예상

1. 아동·청소년인권실태조사 [3]

■ 표본배분

- 지역구분(16) × 학교급(4) = 64개 층에 대해서 학생 수 비례배정
- 표본학교에서는 한 학년, 한 학급을 선정하여 표본 학급 내 전체 학생 조사
- 전체 표본학교 수: 300개 학교

■ 표본추출

- 각 층에서 학교별 학생 수 기준의 확률비례추출법에 따라 표본학교 선정
 - ➔ 층 내에서 자체가중설계(self-weighting design) 만족

1. 아동·청소년인권실태조사 [4]

■ 가중치 산정

- 층화집락확률비례추출법에 따라 조사대상자 선정

$$\rightarrow \text{추출확률} : f_{hij} = n_h \frac{B_{hi}}{\sum_{i=1}^{N_h} B_{hi}} \times \frac{m_{hi}}{B_{hi}}$$

- 각 세부 층의 성별 학생 수에 맞추어 사후층화 보정

■ 추정량

- 가중치를 이용한 가중표본평균 추정량을 이용하여 모평균과 모비율 추정

$$\hat{p} = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}} = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{w_{...}}$$

1. 아동·청소년인권실태조사 [5]

가중치 현황 : 아동·청소년인권실태조사

구분	초등학교 (n=2,889)	중학교 (n=3,117)	일반고 (n=2,174)	전문고 (n=1,310)
최대값	1.17	1.25	1.63	1.85
Q3	1.05	1.09	1.08	1.25
중위수	0.96	1.07	1.06	0.86
Q1	0.92	0.96	1.03	0.59
최소값	0.83	0.60	0.37	0.34
평균	1.00	1.00	1.00	1.00
CV(%)	9.71	14.35	29.82	46.37

2. 알레르기질환 유병률 조사 [1]

- 표본설계의 기본원칙
 - 전국 단위 모집단 대표성 확보
 - 추정의 정확도를 높일 수 있는 층화 및 표본배분법 고안
 - 복합표본설계(complex sample survey)를 반영한 추정 방법 고안

- 모집단 및 조사대상
 - 모집단 : 전국 초등학교 1학년(6-7세) 및 중학교 1학년(13-14세)
 - 표본 추출틀 : 전국 초등학교 및 중학교 리스트
 - 1차 추출단위 : 학교, 2차 추출단위 : 학급, 3차 추출단위: 학생

2. 알레르기질환 유병률 조사 [2]

■ 총화 기준

- 권역 구분(서울, 경기+강원, 충청권, 호남+제주, 영남권)
- 도시 규모(서울, 광역시, 중소도시, 군 지역)
- 시군구별 인구 및 인구밀도, 아파트 가구 비율을 고려하여 세부 총화함

■ 표본배정 방안

- 각 권역별 표본배정 : 권역별 학생 수의 제곱근 비례배정 원칙
 - * 서울과 지방의 유병률 비교를 목적으로 일부 층의 표본 학교 수 조정
- 각 권역 내 세부층별 표본배분 : 학생 수 비례배정 원칙

2. 알레르기질환 유병률 조사 [3]

- 추정 모수 : 모비율
 - 주요 알레르기질환에 대한 유병률 추정을 목적으로 함
- 모비율 추정에 대한 95% 신뢰수준의 최대 허용오차 한계
 - 초등 1학년생 : 전국 추정은 약 1.9%p
 - 중등 1학년생 : 전국 추정은 약 2.2%p
- 2000년 조사결과를 분석하여 얻은 설계효과를 기초로 산출
 - 설계효과 : $DEFF(\hat{\theta}) = \frac{V_D(\hat{\theta})}{V_{SRS}(\hat{\theta})}$
 - 초등학교 : 설계효과의 평균은 1.34, 중위수는 1.31
 - 중학교 : 설계효과의 평균은 1.98, 중위수는 2.41

2. 알레르기질환 유병률 조사 [4] : 초등학교 1학년

권역	세부층	학급		학생		표본배분 (최종)
		학급	비율	학생	비율	
서울	북동	840	5.0%	24,501	5.3%	3
	북서	575	3.4%	16,558	3.6%	2
	남동	541	3.2%	15,861	3.4%	2
	남서	844	5.1%	24,856	5.3%	3
경기·인천	인천광역시	886	5.3%	25,129	5.4%	2
	대규모	2,582	15.5%	83,791	18.0%	5
	중·소규모	1,300	7.8%	37,470	8.1%	3
강원	강원	590	3.5%	14,298	3.1%	2
충청	대전광역시	520	3.1%	15,516	3.3%	2
	중규모	567	3.4%	16,882	3.6%	2
	소규모	789	4.7%	17,501	3.8%	2
호남	광주광역시	529	3.2%	15,815	3.4%	2
	중규모	749	4.5%	20,331	4.4%	2
	소규모	811	4.9%	15,064	3.2%	2
영남	부산, 대구, 울산	1,996	11.9%	56,104	12.0%	4
	중규모	1,535	9.2%	43,762	9.4%	3
	소규모	818	4.9%	16,537	3.5%	2
제주	제주	235	1.4%	6,292	1.3%	2
전국		16,707	100.0%	466,268	100.0%	45

2. 알레르기질환 유병률 조사 [5] : 중학교 1학년

권역	세부층	학급		학생		표본배분 (최종)
		학급	비율	학생	비율	
서울	북동	991	5.2%	33,388	5.1%	2
	북서	657	3.4%	21,693	3.3%	2
	남동	736	3.8%	25,853	3.9%	2
	남서	977	5.1%	33,572	5.1%	2
경기·인천	인천광역시	942	4.9%	36,377	5.5%	2
	대규모	3,098	16.2%	117,191	17.9%	4
	중·소규모	1,252	6.5%	43,510	6.6%	2
강원	강원	632	3.3%	19,397	3.0%	2
충청	대전광역시	633	3.3%	22,003	3.4%	2
	중규모	696	3.6%	23,582	3.6%	2
	소규모	796	4.2%	23,474	3.6%	2
호남	광주광역시	618	3.2%	23,087	3.5%	2
	중규모	874	4.6%	29,680	4.5%	2
	소규모	769	4.0%	20,542	3.1%	2
영남	부산, 대구, 울산	2,617	13.7%	90,744	13.8%	4
	중규모	1,788	9.3%	60,329	9.2%	2
	소규모	858	4.5%	23,408	3.6%	2
제주	제주	230	1.2%	8,294	1.3%	2
전국		19,164	100.0%	656,124	100.0%	40

2. 알레르기질환 유병률 조사 [6]

■ 표본 학교 추출

- 표본학교 수 : 초등학교 45개교, 중학교 각 40개교 추출
 - * 초등 1학년: 4,000명, 중등 1학년 : 4,000명
- 표본추출방법 : 층화 2단확률비례추출법 적용
 - * 각 층에서 1학년 학급 수에 비례하는 확률비례추출법 적용
 - * 조사 학생의 추출확률을 유사하도록 하여 추정의 정확도 향상

■ 표본학급 선정

- 각 표본학교에서 3개 표본학급을 단순임의추출함
- 해당 표본학교의 학급 수가 3개 미만인 경우는 전수추출함
- 표본 학생 : 표본 학급 내 전체 학생 조사

2. 알레르기질환 유병률 조사 [7]

- 가중치 산정
 - 설계가중치 산정, 무응답 보정, 사후층화 보정 실시
 - ⇒ 추출확률의 차이로 인한 추정량의 편향(bias) 제거 목적
- 가중치 산정 단계
 - 설계가중치
 - * 표본설계 단계에서 발생하는 조사단위 간 추출률의 차이 보정
 - 무응답 보정
 - * 무응답으로 인한 과소추정 보정
 - 사후층화 보정
 - * 최종 가중치를 최신 모집단의 현황 분포와 일치하도록 조정
 - * 최신 모집단 정보로 2010년 4월 1일 기준의 학교 현황 이용

2. 알레르기질환 유병률 조사 [8]

가중치 현황 : 알레르기질환 유병률 조사

구분	초등 1 (n=4,003)	중등 1 (n=4,112)	전체 (n=8,115)
최대값	5.46	2.09	5.46
Q3	1.26	1.42	1.32
중위수	0.88	0.90	0.88
Q1	0.71	0.65	0.66
최소값	0.36	0.17	0.17
평균	1.00	1.00	1.00
CV(%)	51.85	47.08	49.48

2. 알레르기질환 유병률 조사 [9]

■ 추정

- 추정대상 : 천식 및 알레르기질환의 유병률과 발생률

- 추정 방법 : 가중치를 이용한 가중평균 및 비율 이용

$$* \quad \hat{p} = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}} = \frac{\sum_{h=1}^L \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij}}{w_{...}}$$

- 추정의 정확도 계산

* 복합표본설계를 반영한 표준오차(standard error) 계산

* 조사데이터 분석을 위해 SAS와 SUDAAN 이용

3. 청소년건강행태 온라인조사 (1)

- 우리나라 청소년 건강위험행태의 현황과 관련 요인 파악
 - 청소년 건강행태 보건지표의 지속적이고 체계적 생산
 - 청소년기 만성질환 예방 정책 개발 및 평가를 지원할 수 있는 기초 자료 제공
 - 관련 분야 기초 연구자료 제공
-
- 조사내용 : 흡연, 음주, 비만 및 체중조절, 신체활동, 식습관 등 13개 영역
 - 조사방법 : 익명성 자기기입식 온라인조사, 컴퓨터실 일괄조사

3. 청소년건강행태 온라인조사 [2]

- 모집단

중·고등학교 재학 중인 학생

- 표본

중학교 1학년 - 고등학교 3학년(약 8만 명)

- 표본학교 : 800개교 (중학교 400개교, 고등학교 400개교)

- 표본학급 : 학교당 학년별 1개 학급 선정, 총 3개 학급 대상

3. 청소년건강행태 온라인조사 (3)

- 층화

- 1단계 : 대도시, 중소도시, 군지역으로 층화 분류
- 2단계 : 시군 집락화 (45개 층)
(지리적 접근성, 학교 수 및 인구수 유사성, 경제활동 참가율, 농업/광업서비스업 종사율을 고려하여 집락 구성)

- 표본배분

- 중학교 : 시도별, 집락별 학교 수, 성비에 따라 학교 할당
- 고등학교 : 시도별, 집락별 학교 수, 학교유형, 성비에 따라 학교 할당

- 추출단위

- PSU : 학교, SSU : 학급

3. 청소년건강행태 온라인조사 (4)

가중치 현황 : 청소년건강행태 온라인 조사

구분	중학교 (n=38,943)	일반계 (n=26,808)	전문계 (n=9,487)	전체 (n=75,238)
최대값	262.66	193.46	374.43	374.43
Q3	66.86	74.44	62.45	68.90
중위수	43.91	44.00	39.67	43.34
Q1	26.83	25.14	27.44	26.45
최소값	2.52	1.61	2.79	1.61
평균	52.13	52.81	51.10	52.24
CV(%)	69.98	67.55	75.17	69.77

차 례

I. 복합표본설계 개요

- 표본설계, 복합표본조사의 구성요소, 사례 등

II. 학생 대상 조사 표본설계 사례

- 아동·청소년인권실태조사, 알레르기 질환조사 등

III. 설계효과의 개념과 활용

- 설계효과, 집락내상관계수, 사례 등

IV. 조사데이터 분석

- 기술통계분석, 회귀분석, 카이제곱 검증 등

V. 결론

1. 표본설계 비교

- 전국 고등학생을 대상으로 청소년 인권실태 조사 (n=2,000명)
- 어떻게 표본 학생을 선정할 것인가?

- 표본설계 1
 - 전국 고등학생 개인 명부에서 2,000명을 랜덤추출

- 표본설계 4 (현실)
 - 전국 고등학교를 계열과 지역에 따라 6개 층으로 구분
 - ➔ 전문계와 기타 지역에서 과표집
 - ➔ 각 표본학교에서 2개 학급 추출
 - ➔ 각 표본학급에서 20명씩 조사

- 같은 표본크기이지만 조사의 정확도는 어떨까?
- 조사의 편리성과 비용은?
- 분석 방법은 어떻게 달라질까?

2. 설계효과 (Design effect: deff) [1]

- 적용된 표본설계가 추정량의 분산에 미치는 전반적인 영향 분석 목적
 - 총화 : 일반적으로 추정의 정확도를 높임
 - 집락추출 : 같은 크기의 단순임의표본에 비해서 정확도가 떨어짐
- 추정량 $\hat{\theta}$ 에 대한 설계효과

$$D^2(\hat{\theta}) = \frac{\text{Variance of } \hat{\theta} \text{ with the complex design}}{\text{Variance of } \hat{\theta} \text{ with a SRS of the same size}} = \frac{V_C(\hat{\theta})}{V_{SRS}(\hat{\theta})}$$

$$\bullet \text{ 예 : } D^2(\bar{y}) = \frac{V_C(\bar{y})}{V_{SRS}(\bar{y})} \quad D^2(\hat{p}) = \frac{V_C(\hat{p})}{V_{SRS}(\hat{p})}$$

- 설계효과는 적용된 복합표본설계와 추정대상에 따라 다른 값을 가짐

2. 설계효과 (design effect : deff) [2]

- 설계효과가 1보다 크게 나타나는 경우
 - 단순임의표본으로 간주해서 분석하면 추정량의 분산이 실제보다 작게 계산됨
- 설계효과는 분석과정에서 표본설계를 무시함으로써 발생하는 추정량 분산의 과소 추정의 정도를 나타냄
- 설계효과는 주로 표본설계 단계에서 활용됨

3. 설계효과에 영향을 미치는 요소

- 층화: proportionate and disproportionate stratification
- 집락화 : clustering
- 불균등 확률 추출
- 가중치 보정 : 무응답 보정, 사후층화 보정 등

4. 설계효과에 영향을 미치는 요소 : 층화

- 학생 대상 조사에서 전국 추정만을 목적으로 한다면 층화비례배분법은 추정의 효율을 높여줌

$$D^2(\bar{y}) = \frac{V_C(\bar{y})}{V_{SRS}(\bar{y})} \leq 1$$

- 非비례배분법의 적용
 - * 지역 또는 인구사회학적 특성별 정확도 높은 추정이 목적인 경우
 - * 층에 따라 단위당 조사비용에 차이가 큰 경우
 - * 비례배분법에 비해서 전국 추정의 정확도는 떨어질 수 있음
 - ➔ 추출률의 차이가 커지면 전국 추정의 정확도가 떨어지게 됨

5. 설계효과에 영향을 미치는 요소 : 집락화(1)

- 학생 대상 조사는 다단계 추출법을 이용함

예 : 학교(PSU) --> 학급(SSU) --> 전체 학생

- 단순집락추출법 적용 사례 : 학급 내 학생 수가 동일한 경우

* b : 각 표본 학급에서의 학생 수

$$* V(\overline{y_{cl}}) = \frac{S^2}{n} [1 + (b-1)\rho] \rightarrow D^2(\overline{y_{cl}}) = 1 + (b-1)\rho$$

* ρ : 집락내상관계수(intra-class correlation coefficient)

“집락 내에서 조사변수의 유사성 정도를 나타냄. 대개 같은 집락에 속해 있으면 서로 비슷하기 때문에 집락내상관계수는 양의 값을 갖게 됨”

5. 설계효과에 영향을 미치는 요소 : 집락화[2]

- 집락내상관계수 ρ 는 조사변수에 따라 달라짐
 - * 인구사회학적 특성(성, 연령, 종교 등) : 대개 0근처의 값
 - * 사회조사 조사항목 : 0.03-0.05 정도의 값
 - * 조사구와 관련성이 깊은 변수(주거형태, 대중교통 수단 접근성 등)
- 집락내상관계수 ρ 는 집락 구성에 따라 달라지고, 이론적으로 음수도 가능
- 집락의 크기가 커짐에 따라 다양성이 증대됨
 - ➔ 집락내상관계수 ρ 는 0근처의 값을 갖게 됨
- 표본설계 단계에서 주요 조사변수에 대한 집락내상관계수 ρ 의 추정이 중요

5. 설계효과에 영향을 미치는 요소 : 집락화(3)

- 실제 학급대상 조사의 표본설계

* 조사구별 학생 수는 유사함

* 총화다단계비례추출법 적용

- 근사적으로 $D^2(\overline{y_{cl}}) = 1 + (b - 1)\rho$ 을 이용하여 설계효과 설명 가능

- 추출확률(가중치)에 차이가 적고, b_α 가 비슷한 경우임

$$D^2(\overline{y_{cl}}) = 1 + (\bar{b} - 1)\bar{\rho}$$

여기서 $\bar{b} = \sum_{\alpha} b_{\alpha} / a$, $\bar{\rho}$: 집락내 유사성의 정도를 나타내는 측도

6. 설계효과에 영향을 미치는 요소: 불균등 가중치 (1)

- 불균등 가중치 발생 이유

- * 불균등 추출확률에 대한 보정

- * 무응답 보정

- * 모집단 분포 정보와 일관성이 있도록 보정

➔ 학생 대상 조사는 정도의 차이는 있지만 가중치에 차이가 발생함

6. 설계효과에 영향을 미치는 요소: 불균등 가중치 (2)

- 가중치와 조사변수가 서로 관련성이 없는 경우

$$d^2(\overline{y_w}) = \frac{n \sum_j w_j^2}{\left(\sum_j w_j \right)^2} = 1 + cv^2(w_j)$$

여기서 w_j : 각 조사단위에 대한 가중치

- 사후층화 보정에 사용된 변수나 이와 관련성이 높은 변수에 대해서 구한 가중치로 인한 설계효과 추정값은 잘못된 값이 될 수 있음

7. 복합표본설계에서 설계효과

$$- d^2(\bar{y}) = d_w^2(\bar{y}) \times d_{cl}^2(\bar{y}) = (1 + cv^2(w_j)) \times (1 + (\bar{b} - 1)\bar{\rho})$$

“복합표본설계에 의한 설계효과는 가중치의 효과와 집락화 효과의 곱으로 표현됨”

$$\rightarrow \hat{\bar{\rho}} = \frac{[d^2(\bar{y}) / (1 + L)] - 1}{\bar{b} - 1}$$

- 표본설계 측면에서 보면 설계효과보다 집락내상관계수 $\bar{\rho}$ 의 추정이 중요함
- PSU가 같고 표본 SSU 수도 같은 경우라면 $\bar{\rho}$ 추정값은 표본설계에 필수적인 정보임

〈참고〉 위의 설계효과는 근사적인 복합표본설계 효과를 살펴보기 위한 측도

8. 설계효과의 활용

- 과거 : 표본설계와 조사데이터 분석 목적으로 활용
 - * 추정량의 분산 계산이 어렵고 비용이 많이 들었음
 - ➔ 일부 중요 변수에 대해서 분석 후 나머지 변수는 설계효과 이용 근사 계산
- 오늘날 : 주로 표본설계 목적으로 활용
 - * 조사데이터 분석용 통계소프트웨어 개발
 - ➔ 추정량의 분산 계산이 비교적 쉽고 비용이 적게 소요됨
 - * 설계효과를 통해서 집락화나 불균등 가중치 효과를 파악하여 표본설계에 반영
 - * 집락내상관계수는 다단계추출법을 적용할 때 중요한 정보로 사용됨

9. 설계효과 사례 : 알레르기질환 유병률 조사 [1]

항목(결과)	초등학생			중학생		
	유병률	설계효과	질락내상관계수	유병률	설계효과	집락내상관계수
현재까지 천식증상 여부(sx ever ¹⁾)	19.1%	1.33	0.0005	15.4%	1.04	-0.0015
12개월 이내 천식증상 여부(12m sx ²⁾)	10.3%	1.26	-0.0001	8.2%	1.05	-0.0014
현재까지 천식 진단 여부(dx ever ³⁾)	10.5%	3.88	0.0234	7.5%	1.31	0.0007
12개월 이내 천식 치료 여부(12m tr ⁴⁾)	4.1%	1.65	0.0034	1.8%	1.13	-0.0007
운동 후 쌉쌉소리 여부(12m exer ⁵⁾)	3.0%	1.75	0.0043	7.3%	1.36	0.0011
마른 기침 (12m dry ⁶⁾)	18.9%	1.86	0.0053	13.3%	2.05	0.0067

9. 설계효과 사례 : 알레르기질환 유병률 조사 [2]

항목(결과)	초등학생			중학생		
	유병률	설계효과	질락내상관계수	유병률	설계효과	질락내상관계수
현재까지 알레르기 비염 증상여부 (sx ever ¹⁾)	47.4%	2.06	0.0071	47.5%	1.31	0.0007
최근 12개월 이내 알레르기 비염 증상여부 1 (12m sx type I ²⁾)	44.3%	2.04	0.0069	41.8%	1.69	0.0038
최근 12개월 이내 알레르기 비염 증상여부 2 (12m sx type II ³⁾)	18.4%	2.93	0.0149	18.6%	1.53	0.0025
알레르기 비염 진단 여부 (dx ever ⁴⁾)	37.7%	6.98	0.0512	29.9%	1.63	0.0033
최근 12개월 이내 알레르기 비염 치료 여부 (12m tr ⁵⁾)	29.4%	5.89	0.0414	19.8%	2.07	0.0068

9. 설계효과 사례 : 알레르기질환 유병률 조사 [3]

항목(결과)	초등학생			중학생		
	유병률	설계효과	집락내상관계수	유병률	설계효과	집락내상관계수
현재까지 아토피 피부염 증상 여부 (sx ever ¹⁾)	26.90%	1.85	0.0052	19.90%	1.7	0.0038
최근 12개월 이내 아토피 피부염 증상 여부 (12m sx ²⁾)	20.50%	1.65	0.0034	12.60%	1.87	0.0052
아토피 피부염 진단여부 (dx ever ³⁾)	35.50%	3.40	0.0191	24.00%	1.68	0.0037
최근 12개월 이내 아토피 피부염 치료 (12m tr ⁴⁾)	15.70%	1.65	0.0034	8.60%	1.53	0.0025

10. 설계효과 사례 : 청소년건강행태 온라인조사 [1]

1. 설계효과: 현재 흡연을

구분	남자	여자	전체
중 1	2.03	1.55	1.88
중 2	2.24	1.99	2.34
중 3	2.41	1.86	2.72
고 1	3.56	3.70	4.16
고 2	3.22	3.54	4.66
고 3	3.44	2.89	3.78

10. 설계효과 사례 : 청소년건강행태 온라인조사 [2]

2. 설계효과: 현재 음주율

구분	남자	여자	전체
중 1	1.70	1.37	1.50
중 2	1.40	1.29	1.40
중 3	1.29	1.56	1.71
고 1	1.78	3.33	2.45
고 2	2.32	2.55	2.56
고 3	2.36	2.70	2.55

10. 설계효과 사례 : 청소년건강행태 온라인조사 [3]

- 중학생에 비해서 고등학생의 설계효과가 크게 나타남
 - ⇒ 고등학생의 학교 내 유사성이 높다는 의미
 - ⇒ 같은 수준의 통계적 정확도를 유지하려면 고등학생의 경우에 중학생 표본에 비해서 약 1.5 - 2배의 표본이 필요할 것임
- 흡연여부가 음주에 비해서 학교 내 유사성이 높게 나타남
- 중학교 표본을 상대적으로 적게 배정하고 고등학교 표본크기를 늘리는 방안
모색

〈참고〉 SAS 프로그램

```
proc surveymeans data=sss rate=cdc.rate08;  
  weight we08;  
  strata clus gubun; /* clus : 집락, gubun : 학교구분[중, 일반, 전문] */  
  cluster code; /* code : 표본학교 고유번호 */  
  var al_p; /* al_p : 현재 음주 여부 */  
  domain x194 x194*x170 ; /* x194 : 학년, x170 : 성별 */  
run ;
```

차 례

I. 복합표본설계 개요

- 표본설계, 복합표본조사의 구성요소, 사례 등

II. 학생 대상 조사 표본설계 사례

- 아동·청소년인권실태조사, 알레르기 질환조사 등

III. 설계효과의 개념과 활용

- 설계효과, 집락내상관계수, 사례 등

IV. 조사데이터 분석

- 기술통계분석, 회귀분석, 카이제곱 검증 등

V. 결론

IV. 조사 데이터 분석

1. 조사데이터분석을 위한 준비

2. 기술통계분석

3. 회귀분석

4. 카이제곱 검증

1. 분석에 앞서

- 유의한 분석 결과를 얻을 수 있는가에 대한 검토
- 분석하는 각 부분집단별로 충분한 수의 표본이 있는가에 대한 검토
- 결측값과 극단값 등에 대한 검토
- 충분한 수의 1차추출단위(PSU)가 있는가 검토

2. 분석에 필요한 정보들

- 가중치

추정치 계산과 분산 계산에 필수적인 정보

- 층, PSU, 추출단위 등에 표본설계 관련 변수 정보

⇒ 각 층에서는 적어도 두 개의 PSU 필요

3. 조사데이터 분석용 소프트웨어

- ① SPSS Complex samples (SPSS 13.0부터 지원됨)
- ② SAS
- ③ SUDAAN
- ④ STATA

4. 조사데이터 분석용 소프트웨어 : SPSS

- Sampling from a Complex Design
- [Preparing a Complex Sample for Analysis](#)
- Complex Samples Plan
- Complex Samples Frequencies
- [Complex Samples Descriptives](#)
- [Complex Samples Crosstabs](#)
- Complex Samples Ratios
- Complex Samples General Linear Model
- Complex Samples Logistic Regression

5. 조사데이터 분석용 소프트웨어 : SAS

- SURVEYSELECT : 추출틀에서 표본 추출
- SURVEYMEANS : 각종 기술 통계량들 산출
- SURVEYREG : 회귀모형 적합, 유의성 검증 등
- SURVEYFREQ : 교차표 작성, 카이제곱 검증 등
- SURVEYLOGISTIC : 로지스틱회귀모형 적합

IV. 조사 데이터 분석

1. 조사데이터분석을 위한 준비

2. 기술통계분석

3. 회귀분석

4. 카이제곱 검증

1. 평균 추정의 편향(bias)

- 2009년 아동·청소년인권실태조사 결과 분석
- Y=“스트레스 관련 다섯 문항의 점수 합계” (연속형 변수로 간주하여 평균 추정)
“학업 문제, 가정 불화, 친구 관계, 경제적 어려움, 외모 및 신체 조건, 미래 불안”
- 상대편향 = $100 \times (\text{비가중 평균} - \text{가중평균}) / \text{가중평균}$

구분	비가중 평균	가중평균	상대편향
초등학교	11.90	11.90	0.00%
중학교	13.97	13.99	-0.14%
일반계	14.72	14.69	0.20%
전문계	14.77	14.81	-0.27%

2. 설계효과

구분	표본크기	가중치 반영	가중치+총화	가중치+집락	복합설계반영
중학교	3,084	1.02	1.02	1.97	2.06
일반계	2,166	1.10	1.10	1.65	1.83
전문계	1,302	1.20	1.20	3.82	4.91
전체	6,552	1.08	1.07	2.71	2.65

- 전문계 학생 조사에서 설계효과가 높게 나타남.
- 가중치 효과나 총화효과는 크지 않음.
- 설계효과의 대부분은 집락효과임.

3. 급내상관계수

구분	표본크기	표본학교	평균 응답자수	가중치 효과	급내상관계수
중학교	3,084	85	36.3	1.02	0.0291
일반계	2,166	56	38.7	1.10	0.0178
전문계	1,302	41	31.8	1.20	0.1006
전체	6,552	182	36.0	1.08	0.0417

$$d^2(\bar{y}) = d_w^2(\bar{y}) \times d_{cl}^2(\bar{y}) = (1 + cv^2(w_j)) \times (1 + (\bar{b} - 1)\bar{\rho})$$

4. 일반적인 분석법 사용의 문제

- 단순랜덤표집으로 간주하여 조사 데이터를 분석하면
 - 추정 편향 (bias, 어긋남) 발생
 - 추정량 분산의 과소 추정 (집락표본에서 심각)
 - ⇒ 신뢰구간의 폭이 실제보다 작아짐
 - ⇒ 가설검정에서 실제로는 통계적으로 有意하지 않지만, 有意하다는 결론 도달 (제2종오류의 가능성)

IV. 조사 데이터 분석

1. 조사데이터분석을 위한 준비
2. 기술통계분석
3. 회귀분석
4. 카이제곱 검증

1. 선형회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

여기서, $\beta_0, \beta_1, \dots, \beta_k$: 추정될 회귀계수들,

ε_i : 오차항

2. 회귀모형의 적합

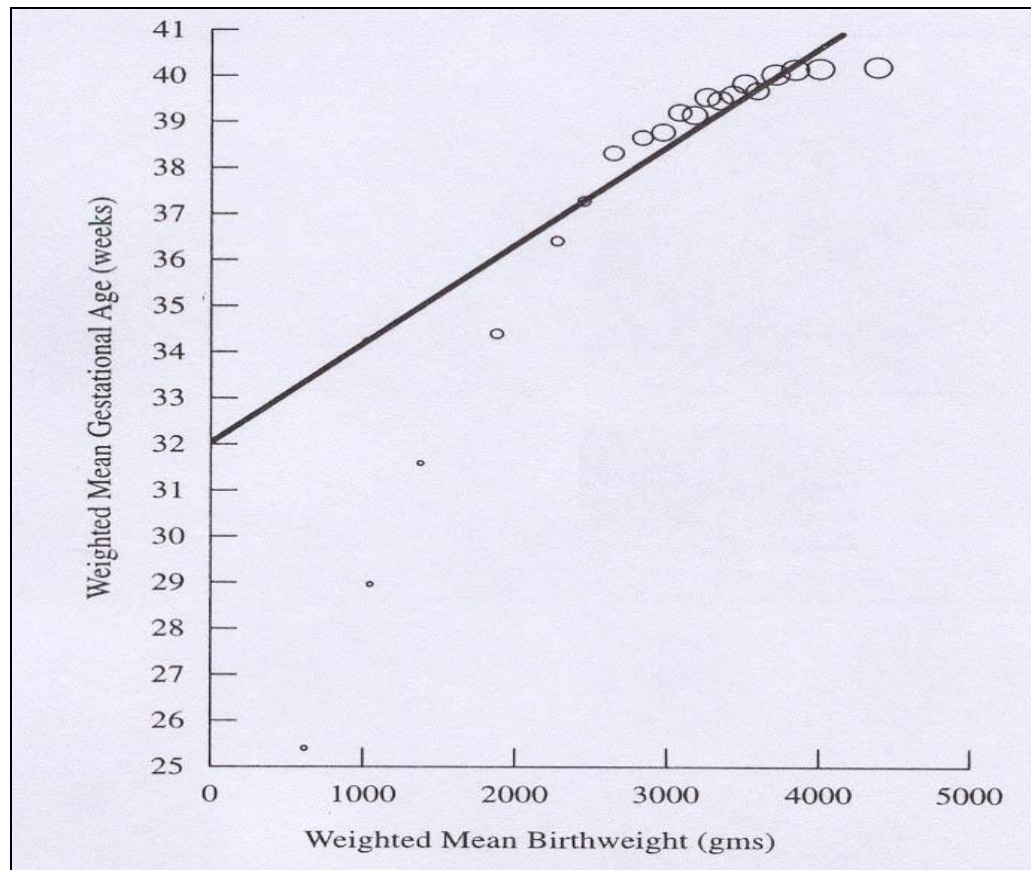
① SRS 표본인 경우

- 전통적인 통계기법 적용
- 회귀계수 추정을 위해서 최소제곱법(Ordinary Least Squares: OLS) 이용

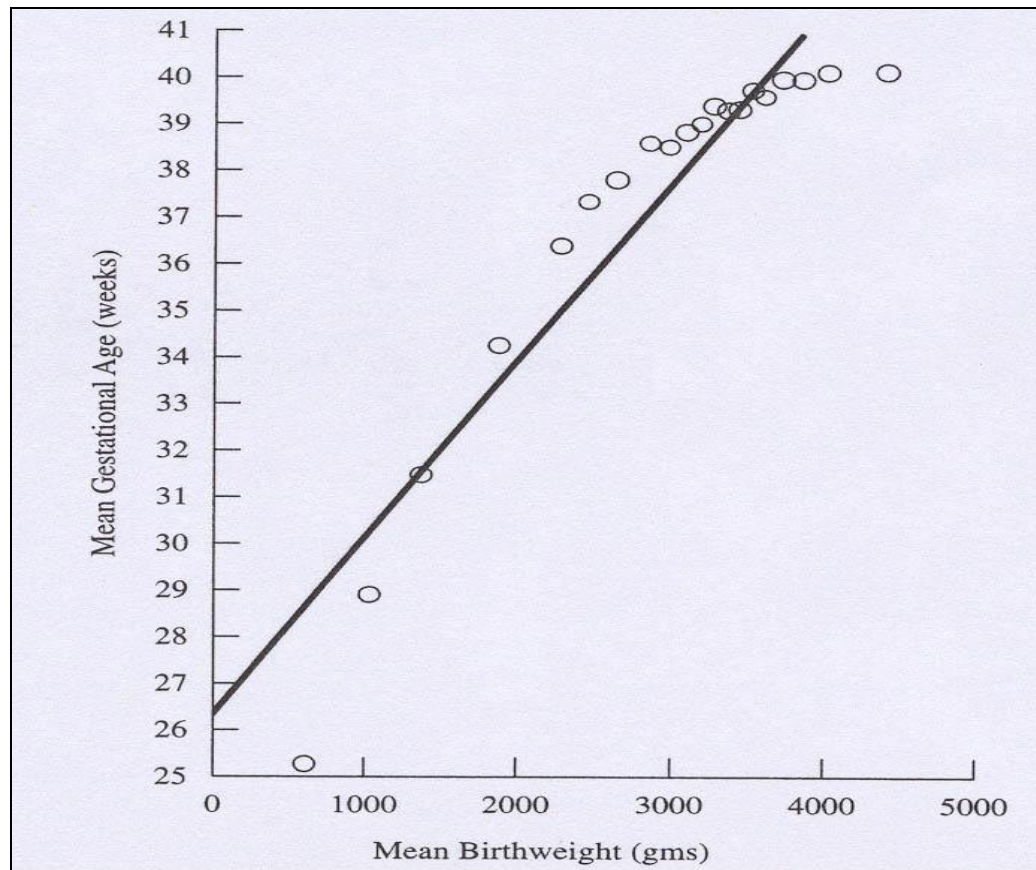
② 복합표본조사 데이터의 경우

- 회귀계수 추정 : 가중최소제곱법 이용 (조사 데이터의 가중치)
- 추정치의 분산 계산
 - ⇒ 표본설계가 SRS 가정에서 크게 벗어남
 - ⇒ 복합표본설계를 반영한 분산 추정법 사용
- 설명변수 누락과 오차항 가정 위배에 대해서 강건성(robustness) 유지

[a] 가중치를 사용한 회귀직선 적합



(b) 가중치를 사용하지 않는 회귀직선 적합



3. 회귀계수의 추정

- 회귀계수의 표현

$$B_1 = \frac{\sum_{i=1}^N x_i y_i - \frac{\left(\sum_{i=1}^N x_i\right)\left(\sum_{i=1}^N y_i\right)}{N}}{\left(\sum_{i=1}^N x_i^2\right) - \left(\sum_{i=1}^N x_i\right)^2 / N}$$

$$B_0 = \frac{\left(\sum_{i=1}^N y_i\right) - B_1 \left(\sum_{i=1}^N x_i\right)}{N}$$

■ 회귀계수의 추정 : 가중최소제곱법 이용

$$\hat{B}_1 = \frac{\sum_{i \in S} w_i x_i y_i - \frac{(\sum_{i \in S} w_i x_i)(\sum_{i \in S} w_i y_i)}{\sum_{i \in S} w_i}}{\sum_{i \in S} w_i x_i^2 - \frac{(\sum_{i \in S} w_i x_i)^2}{\sum_{i \in S} w_i}}$$

$$\hat{B}_0 = \frac{\sum_{i \in S} w_i y_i - \hat{B}_1 \sum_{i \in S} w_i x_i}{\sum_{i \in S} w_i}$$

■ 추정량의 분산추정

- 선형화 방법, 잭나이프 방법 이용

〈참고〉 분산 추정 소개

① 선형화 방법 적용 : SRS 표본인 경우

$$\hat{V}_L(\hat{B}_1) = \frac{n \sum_{i \in S} (x_i - \bar{x}_S)^2 (y_i - \hat{B}_0 - \hat{B}_1 x_i)^2}{(n-1) \left[\sum_{i \in S} (x_i - \bar{x}_S)^2 \right]^2}$$

② 전통적인 방법 적용

$$\hat{V}_M(\hat{\beta}_1) = \frac{n \sum_{i \in S} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2) \left[\sum_{i \in S} (x_i - \bar{x}_S)^2 \right]^2}$$

선형화 방법 적용 : 일반적인 경우

$$\hat{V}_L(\hat{B}_1) = \frac{\hat{V}(\sum_{i \in S} w_i q_i)}{\left[\sum_{i \in S} w_i x_i^2 - \frac{(\sum_{i \in S} w_i x_i)^2}{\sum_{i \in S} w_i} \right]^2}$$

여기서

$$q_i = (y_i - \hat{B}_0 - \hat{B}_1 x_i)(x_i - \bar{x})$$

〈참고〉 The SURVEYREG Procedure

```
PROC SURVEYREG <options>;  
  CLASS variables ;  
  STRATA variables;  
  MODEL dependent=<effects></options>;  
  WEIGHT variables;  
  ESTIMATE 'label' effect values <effect values>  
RUN;
```

〈표 A-2〉 월평균 임금총액에 대한 회귀계수 추정치
 [종속변수 : $\ln(\text{월평균 임금총액})$]

변수명	OLS 방법 회귀계수 (s.e)	가중치, 표본설계 반영 회귀계수 (s.e)	반영 meff
광업+제조업 ^a	0.129 ^c (0.0069 ^d)	0.154 ^c (0.0130 ^d)	3.31 ^e
수도, 가스, 전기+건설업 ^b	0.127 (0.0110)	0.107 (0.0198)	4.37
도소매업	0.133 (0.0073)	0.147 (0.0128)	4.73
음식 숙박업	0.138 (0.0090)	0.122 (0.0153)	4.52
운수, 창고, 통신업	0.066 (0.0076)	0.084 (0.0152)	1.63
금융·보험업	0.221 (0.0088)	0.267 (0.0150)	1.39
부동산업	0.118 (0.0077)	0.144 (0.0151)	3.13
교육 서비스업	0.010 (0.0121)	0.030 (0.0181)	3.59
건강, 사회 서비스업	0.207 (0.0125)	0.216 (0.0175)	3.54
기타 서비스업	0 ^f	0 ^f	

변수명	OLS 방법 회귀계수 (s.e)	가중치, 표본설계 반영 회귀계수 (s.e)	반영 meff
서울	0.073 (0.0039)	0.078 (0.0071)	3.32
광역시	-0.012 (0.0040)	-0.016 (0.0073)	3.34
시·군지역	0 [∗]	0 [∗]	
사업체 내의 근로자 수	0.023 (0.0016)	0.025 (0.0029)	3.45
관리자 및 입법자	0.312 (0.0111)	0.241 (0.0238)	4.09
전문가	0.169 (0.0125)	0.103 (0.0206)	3.16
기술공 및 준전문가	0.177 (0.0091)	0.122 (0.0163)	3.21
사무직원	0.143 (0.0076)	0.097 (0.0135)	2.96
판매원	0.091 (0.0083)	0.048 (0.0139)	3.52
기능 근로자	0.105 (0.0080)	0.060 (0.0138)	2.52
장치 조작용	0.130 (0.0084)	0.074 (0.0146)	2.48
단순 노무직 근로자	0 [∗]	0 [∗]	

변수명	OLS 방법 회귀계수 (s.e)	가중치, 표본설계 반영 회귀계수 (s.e)	반영 meff
중학교 졸업 이하	-0.152 (0.0072)	-0.166 (0.0128)	3.05
고등학교 졸업	-0.088 (0.0052)	-0.100 (0.0092)	3.08
전문대학교 졸업	-0.076 (0.0064)	-0.091 (0.0101)	2.58
대학 졸업 이상	0 ^f	0 ^f	
남성	0.286 (0.0039)	0.261 (0.0061)	2.49
1년 이하	-0.242 (0.0062)	-0.228 (0.0103)	2.81
1-3년	-0.166 (0.0056)	-0.154 (0.0093)	2.68
3-4년	-0.125 (0.0064)	-0.115 (0.0100)	2.35
4-5년	-0.108 (0.0067)	-0.100 (0.0103)	2.24
5-10년	-0.065 (0.0053)	-0.067 (0.0087)	2.45
근로자 연령	0.046 (0.0012)	0.043 (0.0021)	3.25
연령의 제곱	-0.001 (0.0000)	-0.000 (0.0000)	3.34
총 근로시간(월 단위)	0.001 (0.0000)	0.001 (0.0001)	5.05
절편	5.308 (0.0274)	5.364 (0.0493)	3.55
R	0.464	0.449	

IV. 조사 데이터 분석

1. 조사데이터분석을 위한 준비
2. 기술통계분석
3. 회귀분석
4. 카이제곱 검증

1. 간단한 사례

① SRS로 추출된 500쌍의 부부 조사

(i) 가정에 개인용 PC가 있는가?,

(ii) 케이블 TV에 가입했는가?

➔ 전체 1,000명을 조사하였다면 어떻게 교차표를 작성할 것인가?

② 조사 결과

		Computer?		
		Yes	No	
Cable?	Yes	119	188	307
	No	88	105	193
		207	293	500

③ 독립성 가정에서의 기대 도수

		Computer?		
		Yes	No	
Cable?	Yes	127.1	179.9	307
	No	79.9	113.1	193
		207	293	500

$$\hat{m}_{ij} = n \hat{p}_{i+} \hat{p}_{+j} = n \frac{x_{i+}}{n} \frac{x_{+j}}{n}$$

④ 피어슨의 카이제곱 검증통계량

$$X^2 = \sum_{\text{모든 셀}} \frac{(\text{관측도수} - \text{기대도수})^2}{\text{기대도수}} = 2.281$$

$$G^2 = 2 \sum_{\text{모든 셀}} \text{관측도수} \ln \left(\frac{\text{관측도수}}{\text{기대도수}} \right) = 2.275$$

⑤ X^2 과 G^2 은 귀무가설 하에서 자유도 $(r-1)(c-1)$ 인 카이제곱분포 따름

$$\chi^2_{(1, 0.05)} = 3.84$$

2. 단순임의표본에 대한 독립성 검증

① 자료구조

		C				
		1	2	$\dots$	c	
R	1	p_{11}	p_{12}	$\dots$	p_{1c}	p_{1+}
	2	p_{21}	p_{22}	$\dots$	p_{2c}	p_{2+}
	$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	r	p_{r1}	p_{r2}	$\dots$	p_{rc}	p_{r+}
		p_{+1}	p_{+2}	$\dots$	p_{+c}	1

② 귀무가설과 기대도수 계산

$$- H_0: p_{ij} = p_{i+} p_{+j}$$

for $i = 1, \dots, r$ and $j = 1, \dots, c$.

$$- \hat{m}_{ij} = n \hat{p}_{i+} \hat{p}_{+j} = n \frac{x_{i+}}{n} \frac{x_{+j}}{n}$$

③ 검증통계량

$$\begin{aligned}
 X^2 &= \sum_{i=1}^r \sum_{j=1}^c \frac{(x_{ij} - \hat{m}_{ij})^2}{\hat{m}_{ij}} \\
 &= n \sum_{i=1}^r \sum_{j=1}^c \frac{(\hat{p}_{ij} - \hat{p}_{i+} \hat{p}_{+j})^2}{\hat{p}_{i+} \hat{p}_{+j}}
 \end{aligned}$$

$$\begin{aligned}
 G^2 &= 2 \sum_{i=1}^r \sum_{j=1}^c x_{ij} \ln \left(\frac{x_{ij}}{\hat{m}_{ij}} \right) \\
 &= 2n \sum_{i=1}^r \sum_{j=1}^c \hat{p}_{ij} \ln \left(\frac{\hat{p}_{ij}}{\hat{p}_{i+} \hat{p}_{+j}} \right)
 \end{aligned}$$

3. 복합표본조사에 대한 독립성 검증

① 귀무가설 하에서 X^2 과 G^2 은 자유도 $(r-1)(c-1)$ 인 χ^2 분포를 따르지 않음

⇒ p-값이 실제보다 작게 계산됨

⇒ 실제로는 통계적으로 有意하지 않은 현상을 有意한 것으로 판정할 수 있음

⇒ 급내상관계수가 양수인 집락표본에서 발생하는 문제

② 독립성 검정에 대한 귀무가설

$$- H_0 : p_{ij} = p_{i+} p_{+j}$$

$$\Leftrightarrow \text{for } i = 1, \dots, r \text{ and } j = 1, \dots, c.$$

$$H_0 : \theta_{11} = 0, \theta_{12} = 0, \dots, \theta_{r-1, c-1} = 0$$

여기서, $\theta_{ij} = p_{ij} - p_{i+} p_{+j}$ 로 정의

③ Wald 통계량 : 2×2 분할표의 경우 예시

$$- \theta = p_{11} - p_{1+}p_{+1} = p_{11}p_{22} - p_{12}p_{21}.$$

$$\hat{\theta} = \hat{p}_{11} \hat{p}_{22} - \hat{p}_{12} \hat{p}_{21} \text{로 정의}$$

$$- H_0 : p_{ij} = p_{i+} p_{+j} \Leftrightarrow H_0 : \theta = 0$$

- 표본크기가 충분히 크게 되면

$$H_0 : \theta = 0 \text{ 하에서 } \frac{\hat{\theta}}{\sqrt{\hat{V}(\hat{\theta})}} \sim N(0, 1),$$

$$X_w^2 = \frac{\hat{\theta}^2}{\hat{V}(\hat{\theta})} \sim \chi^2(1) : \text{Wald 통계량}$$

4. SAS SURVEYFREQ 절차

- 총화, 집락화, 불균등 가중치 등의 복합표본설계 요소를 반영하여 분석

- one-way frequency table : 적합도 검정(goodness-of-fit test)

“ Rao-Scott의 카이제곱에 의한 적합도 검정 “

- two-way table : 행 변수와 열 변수의 독립성 검정

“Rao-Scott 카이제곱 검정, Rao-Scott LR 검정, Wald 카이제곱 검정 등”

- 분산추정

Taylor 선형화 방법, 균형반분법(BRR), 잭나이프방(Jackknife) 등

5. SAS 예제 프로그램

```
proc surveyfreq data=sss rate=cdc.rate08;  
table x194*ci_p/ wchisq cv deff row nowt nostd  
nototal;  
strata clus gubun/list;  
cluster code;  
weight we08;  
run;
```

- list : stratum information 제시
- nowt : weighted frequencies를 제시하지 않음
- wchisq: “The Wald chi-square 방법 이용”

6. 분석 사례 : 아동 · 청소년인권실태조사 중 일부

- 전문고 학생 대상으로 “최근 일 년 동안 자살에 대해 생각해 본 적이 있습니까?” 질문에 대한 응답결과의 성별 차이 유무를 검토하고자 함
- 추정결과

구분	남자	여자	전체
전혀 없다	70.6%	60.7%	65.3%
가끔 생각해 본다	25.8%	33.9%	30.1%
자주 생각해 본다	3.6%	5.5%	4.6%

■ SAS를 이용하여 분석한 결과 : 비가중, 단순임의표본 가정(가중치 고려)

검증통계량	자유도	통계량값	p값
Chi-Square	1	14.35	0.0008
Likelihood Ratio Chi-Square	1	14.44	0.0007

■ SAS Survey 절차를 이용하여 분석한 결과 : 표본설계와 가중치 고려

검증통계량	자유도	통계량값	p값
Chi-Square	1	8.49	0.0241

차 례

I. 복합표본설계 개요

- 표본설계, 복합표본조사의 구성요소, 사례 등

II. 학생 대상 조사 표본설계 사례

- 아동·청소년인권실태조사, 알레르기 질환조사 등

III. 설계효과의 개념과 활용

- 설계효과, 집락내상관계수, 사례 등

IV. 조사데이터 분석

- 기술통계분석, 회귀분석, 카이제곱 검증 등

V. 결론

V. 결론

1. 조사 데이터는 표본설계와 가중치를 반영해서 분석해야 한다.

실제적으로 조사 데이터 분석을 위해서는 조사 데이터 분석용 패키지 이용

2. 조사 데이터 분석용 패키지를 사용하지 않고, 분석하는 경우

[1] 추정 편향 발생

[2] 표집분산의 과소 추정이 있을 수 있음

⇒ 신뢰구간의 폭이 지나치게 작아짐

⇒ 가설검증에서 2종 오류가 크게 나타남

3. 표본설계를 반영해서 분석해야 하는 경우

- [1] 공공 목적으로 사용되는 공식통계(official statistics) 작성
- [2] 확률추출법으로 추출되었고, 표본크기가 충분히 큰 경우

4. 조사 데이터를 분석할 때 요구되는 절차들 : Levy and Lemeshow(1999)

- [1] 표본설계에 대한 다음 사항을 명확히 할 것
 - 층화 구분 변수
 - 집락 구분 변수
 - 유한모집단 수정항을 위한 모집단 크기
- [2] 설계가중치 산정

-
- [3] 무응답 보정과 사후층화 보정 단계를 거쳐서 최종 가중치 산정
 - [4] 조사 데이터가 층화, 집락, 모집단 크기 등에 대한 정보가 있는지 확인
 - [5] 조사 분석용 소프트웨어로 분석

감사합니다.