[Korea Research Center 특강]

표본설계와 모수추정

2017. 8. 17 (목)

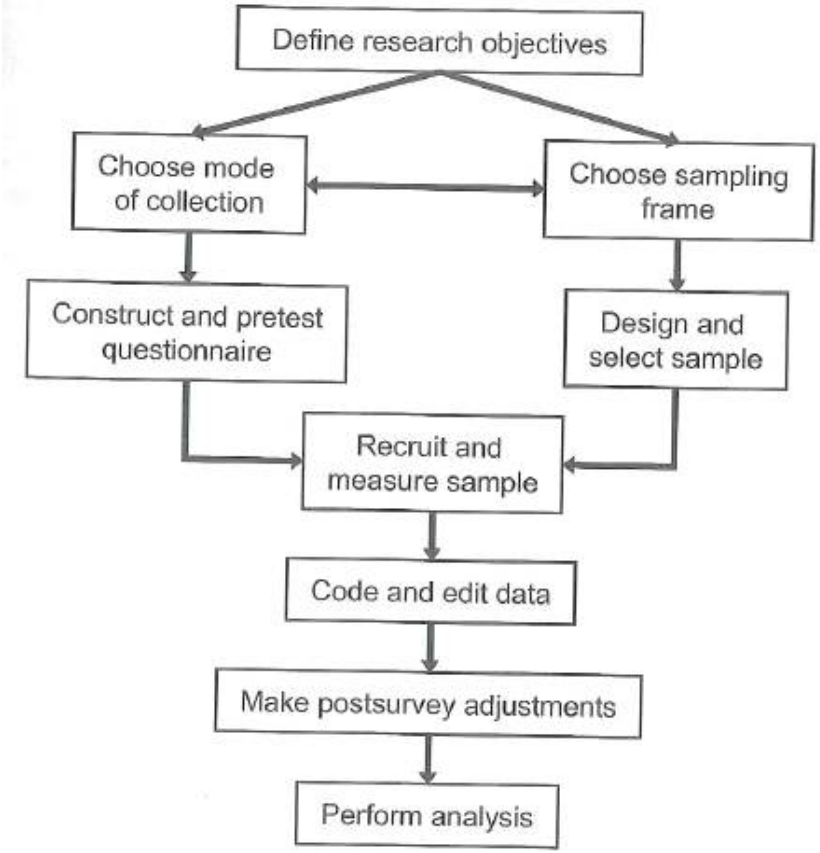
김 규 성

(서울시립대학교 통계학과, kskim@uos.ac.kr)

차례

1. 표본설계 개요
2. 표본추출틀
3. 확률표집 방법
4. 층화 및 표본배분
5. 무응답 처리
6. 사후층화
7. 가중치와 추정
참고문헌

1. 표본설계 개요

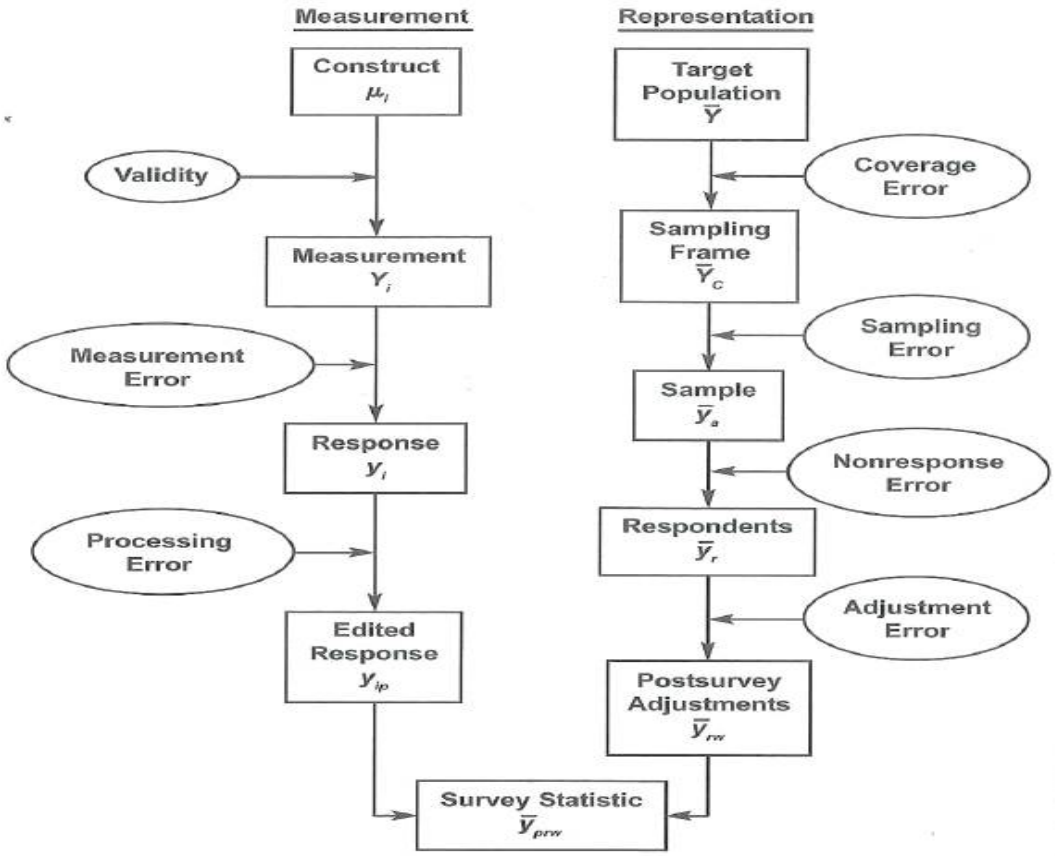

Define research objectives
Choose mode of collection
Choose sampling frame
Construct and pretest questionnaire
Design and select sample
Recruit and measure sample
Code and edit data
Make postsurvey adjustments
Perform analysis

조사오차(Survey Error)

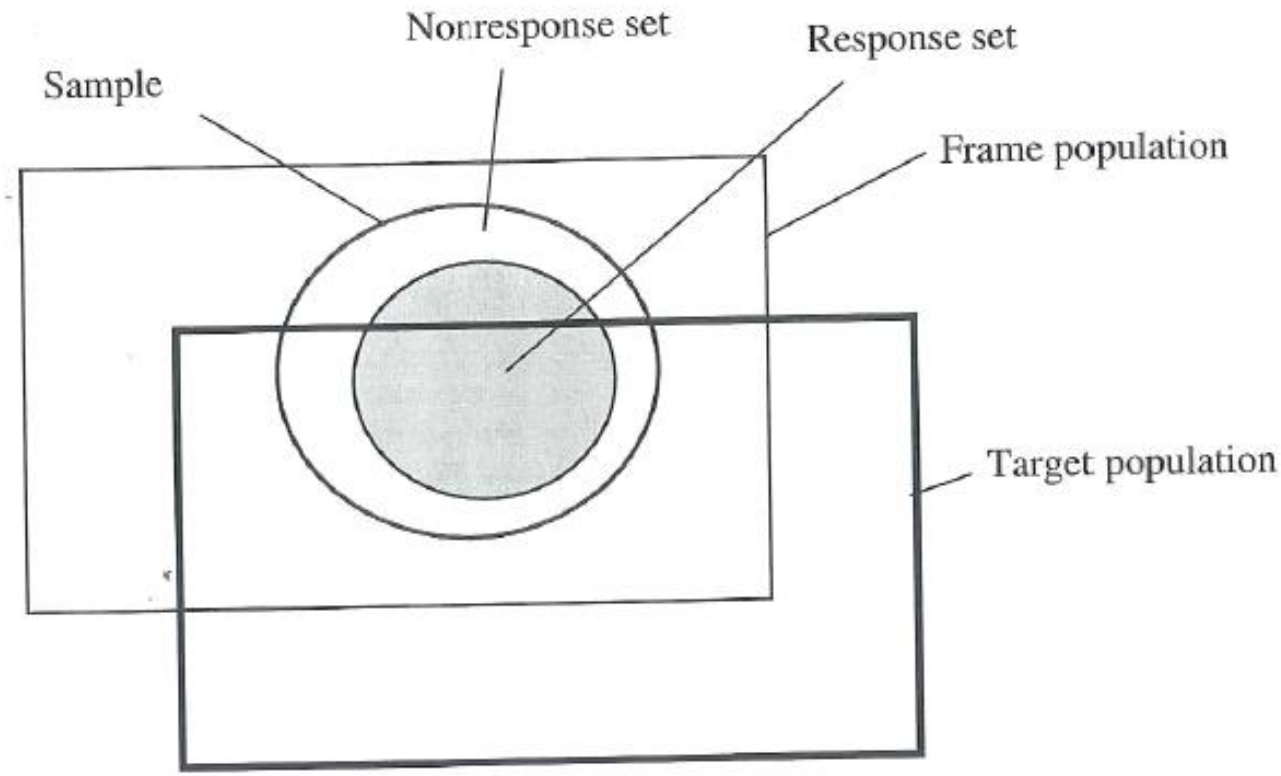
- **조사오차**(survey error) : 연구에서 참값이라고 기대되는 값(**모수**, θ, parameter)과 조사결과로 얻어진 값(**추정값**, $\hat{\theta}$)의 차이

- **비표집오차**(non-sampling error)
  - **포함범위 오차**(coverage error)
  - **무응답오차**(nonresponse error)
  - **측정오차**(measurement error)
  - **조사과정 오차**(processing error)
- **표집오차**(sampling error)

조사오차


Measurement
Representation
Construct
μ_i
Target Population
$\bar{Y}$
Validity
Coverage Error
Measurement
Y_i
Sampling Frame
$\bar{Y}_C$
Measurement Error
Sampling Error
Sample
$\bar{y}_a$
Response
y_i
Nonresponse Error
Respondents
$\bar{y}_r$
Processing Error
Adjustment Error
Edited Response
y_{ip}
Postsurvey Adjustments
$\bar{y}_{rw}$
Survey Statistic
$\bar{y}_{prw}$

미관측오차


Nonresponse set
Response set
Sample
Frame population
Target population

조사설계와 표본설계 (혹은 표집설계)

- 모집단 : 목표모집단, 조사모집단
- **표본추출틀 확보**
- **표본선정** (Sampling)
  - **층화, 표본배분**
  - **표본크기 결정**
- 조사도구(질문지) 설계
- 조사방법 결정 및 실사
- **모수 추정** (Parameter Estimation)
  - **가중치**(weight) **산정**
- 데이터 분석 및 보고서 작성

표본선정

- **Sampling** is a means of selecting a subset of units from a population for the purpose of collecting information for those units to draw inferences about the population as a whole. (SMP, 897쪽)

- 표본의 종류
  - **확률표본** (Probability Sample, SMP, 91쪽)
  - **비확률표본** (Non-probability Sample, SMP, 87쪽)

Non-Probability Sampling

- **Non-probability sampling** uses a **subjective method** of selecting units from a population.
- It provides a **fast, easy** and **inexpensive** way of selecting a sample.

- **However**, in order to make inferences about the population from the sample, the data analyst must assume that the sample is **representative** of the population. (표본의 대표성)

Types of Non-probability Sampling

- Haphazard sampling
- Volunteer sampling
- Judgement sampling
- **Quota** sampling
- Modified probability sampling

Probability Sampling

- Probability sampling involves the selection of units from a population based on the **principle of randomization or chance**.
- **Two main criteria** for **probability sampling**:

  (i) the units be **randomly selected**

  (ii) all units in the survey population have a **non-zero inclusion probability** in the sample and that these probabilities **can be calculated**.
- Probability sampling is more **complex, time consuming** and usually more **costly** than non-probability sampling.

모수와 추정량

- A **parameter (모수)** is a population characteristic that the client or data user is interested in estimating.
- An **estimator (추정량)** is a **formula** by which an estimate of the parameter is **calculated from the sample** and an **estimate (추정값)** is **the value of the estimator** using the data from the **realized sample**.

예제 : 모평균과 표본평균

- 모집단: 미국 어느 지역의 270개 블록 데이터 (1950년 총조사)
- x_i : 블록내 거주자 수, y_i : 블록내 세입자 수

단순 통계량							
변수	N	평균	표준편차	합	최소값	최대값	상관계수
x	270	25.137	24.999	6787	1	149	0.9699
y	270	16.885	20.728	4559	0	131	

표본 추출 (n=20)

- 크기 20인 단순확률표본 추출

i	난수	x_i	y_i	i	난수	x_i	y_i
1	689	5	3	11	701	29	17
2	537	9	5	12	566	31	14
3	545	18	5	13	680	5	0
4	420	68	52	14	735	2	0
5	436	32	21	15	528	4	2
6	385	48	34	16	541	102	54
7	575	11	3	17	564	20	11
8	727	1	0	18	380	15	11
9	753	1	0	19	730	1	0
10	451	4	0	20	376	29	23

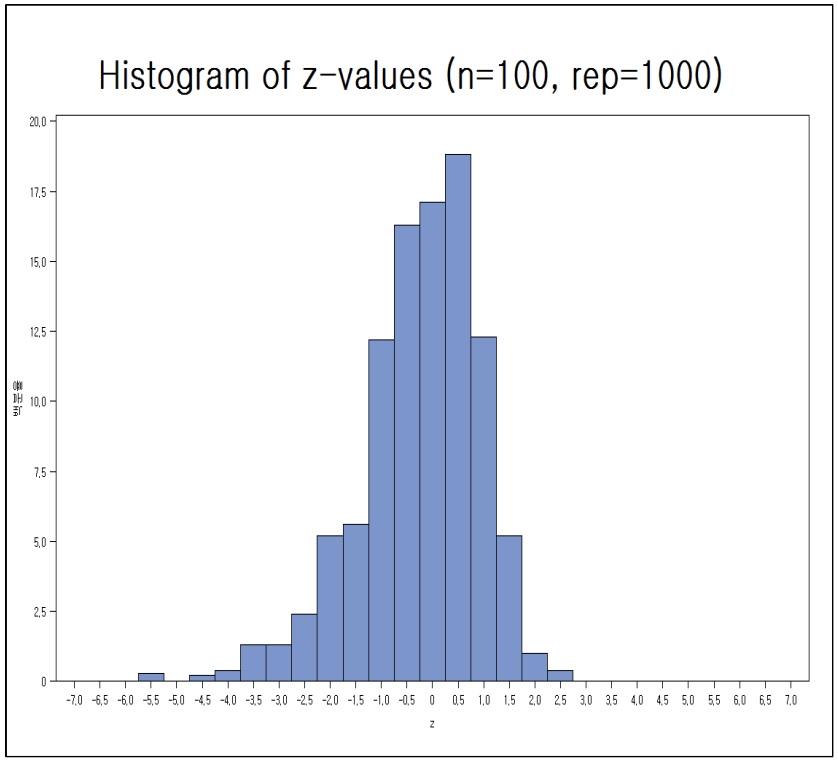
모평균 추정

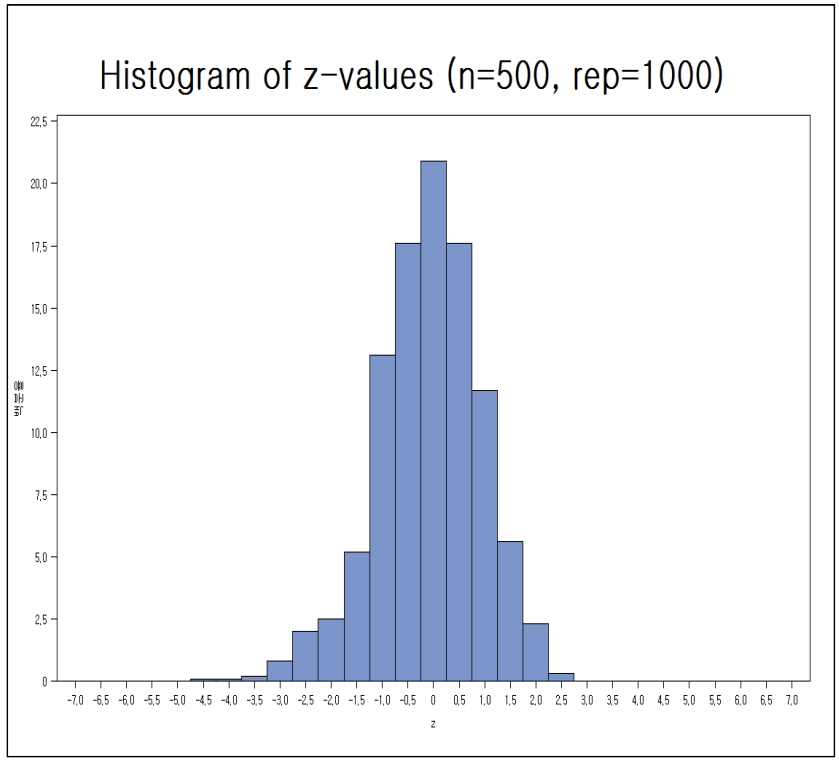
- $\bar{x}_U$ 추정 : $\bar{x}_s = \frac{1}{n}\sum_{i=1}^{n} x_i =$ 435/20=**21.75**
- $\bar{y}_U$ 추정 : $\bar{y}_s = \frac{1}{n}\sum_{i=1}^{n} y_i =$ 255/20=**12.75**

SAS 예제

```
proc surveyselect data=pop method=srs n=20 seed=1234
     out=sample;
proc print; run;
proc surveymeans data=sample total=270;
  var x y;
run;
```

표집분포


Histogram of z-values (n=100, rep=1000)
빈도
z


Histogram of z-values (n=500, rep=1000)
빈도
z

추정의 통계적 효율성

- $\bar{x}_U$ 추정 :

$$\bar{x} \pm t_{0.025,n-1}\sqrt{v(\bar{x})} = 21.75 \pm 2.09 \times 5.58 = (10.09,\ 33.41)$$

- $\bar{y}_U$ 추정 :

$$\bar{y}_s \pm t_{0.025,n-1}\sqrt{v(\bar{y})} = 12.75 \pm 2.09 \times 3.59 = (5.24,\ 20.25)$$

2. 표본추출틀

가구조사 예제

- ○ 목표모집단 : 전국 만 19세 이상의 일반국민
- ○ 조사모집단 : 전국 만 19세~69세 일반국민
- ○ 표본추출틀 : 2015년 등록센서스 조사구 리스트
- ● 표본추출틀 제약사항 (현실적인 측면)
  - 섬, 기숙사, 특수사회시설, 관광호텔 및 외국인 조사구 제외
  - 조사구내 집단가구(기숙사 등), 외국인 가구 제외

- ○ 고려사항 : **표본추출틀 구축**, **등록센서스** 및 **과소포함** 문제

예제 : 2015년 기준 환경산업통계

환경산업의 개념

- **환경산업** : 기후◦대기, 물, 환경복원◦복구, 환경안전◦보건, 자원순환, 지속가능 환경◦자원, 환경지식◦서비스 분야 등 환경보전 및 관리를 위한 환경시설 및 측정기기 등을 설계, 제작, 설치하거나 환경기술 등에 관한 서비스를 제공하는 산업활동
- **사업체** : 일정한 물리적 장소에서 단일 또는 주된 경제활동을 독립적으로 수행하는 단위
- **통계작성 범위** : **환경산업특수분류** 342개 업종(세세분류) 57,311개 업체

환경산업통계 (계속)

○ **표본조사 부문** : 약 5,700개 사업체

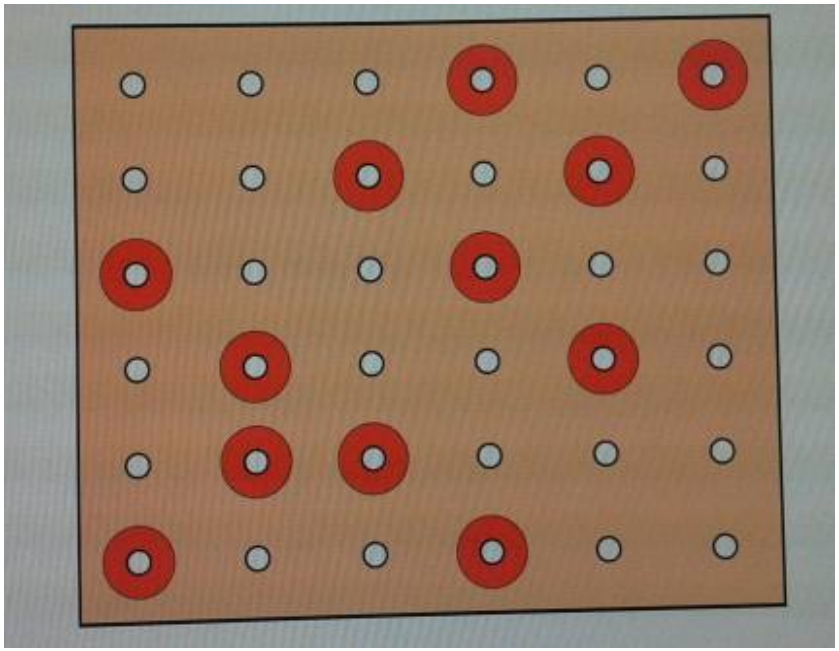
- 방문조사, 인터넷조사

○ **적격률**을 계산하여 모집단 크기 추정 (p.7)

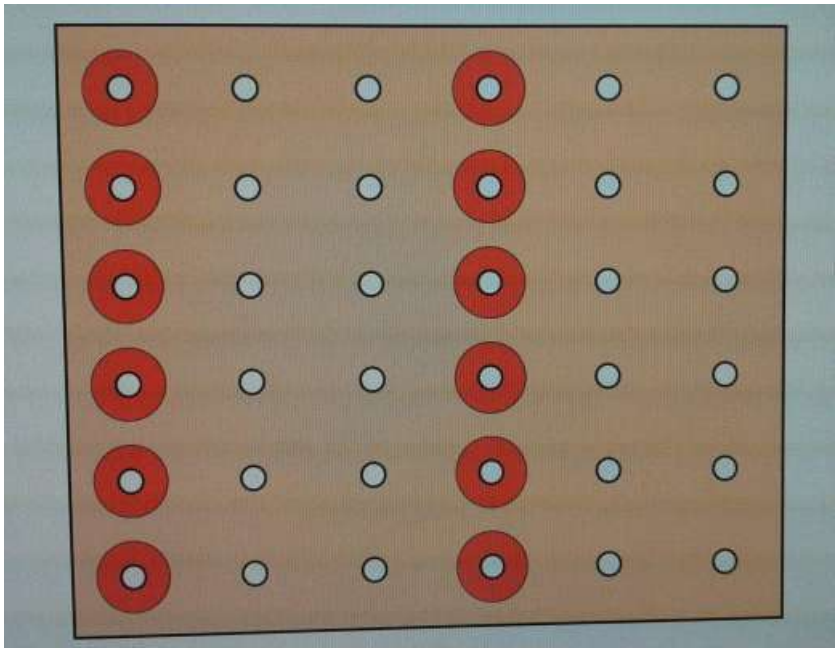
3. 확률표집 방법

- 단순확률표집
- 계통표집
- 집락표집
- 층화표집
- 다단계 표집
- 다중표집

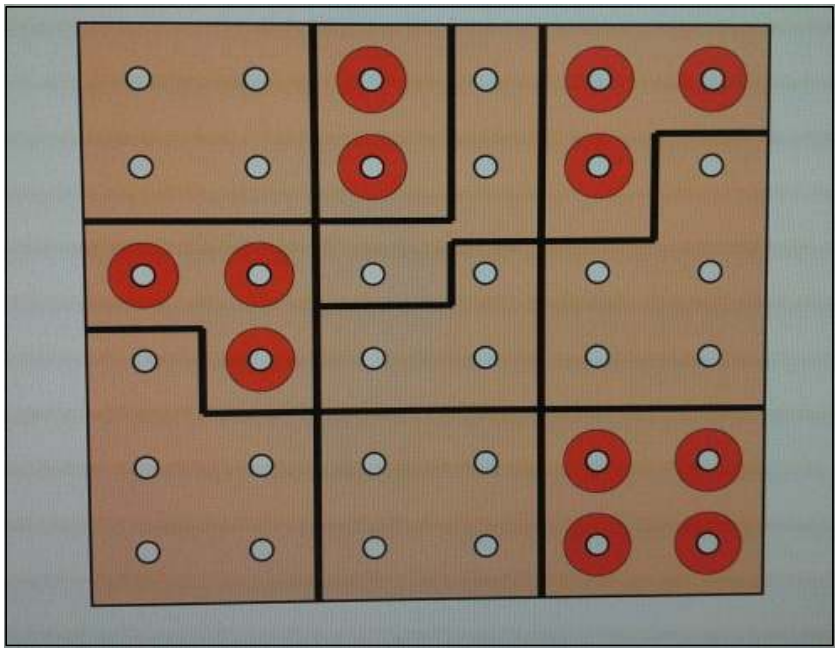
단순확률표집(n=12)



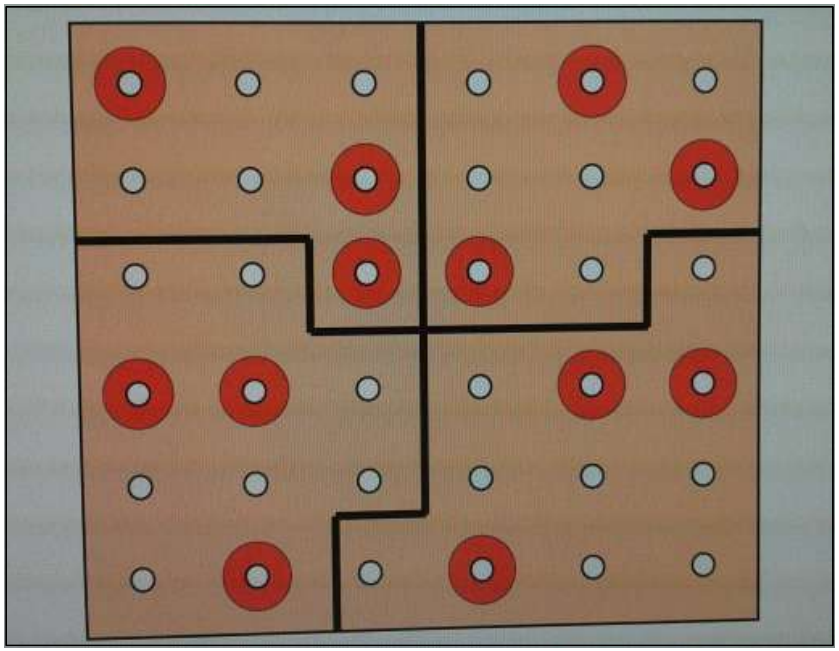
계통 표집(n=12, N=36, k=3)



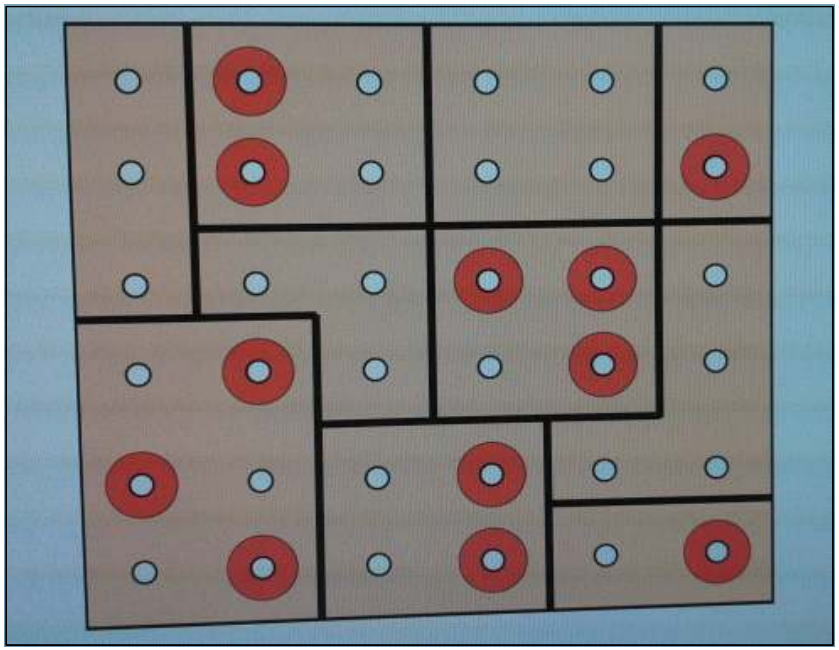
집락표집 (4 clusters are sampled)



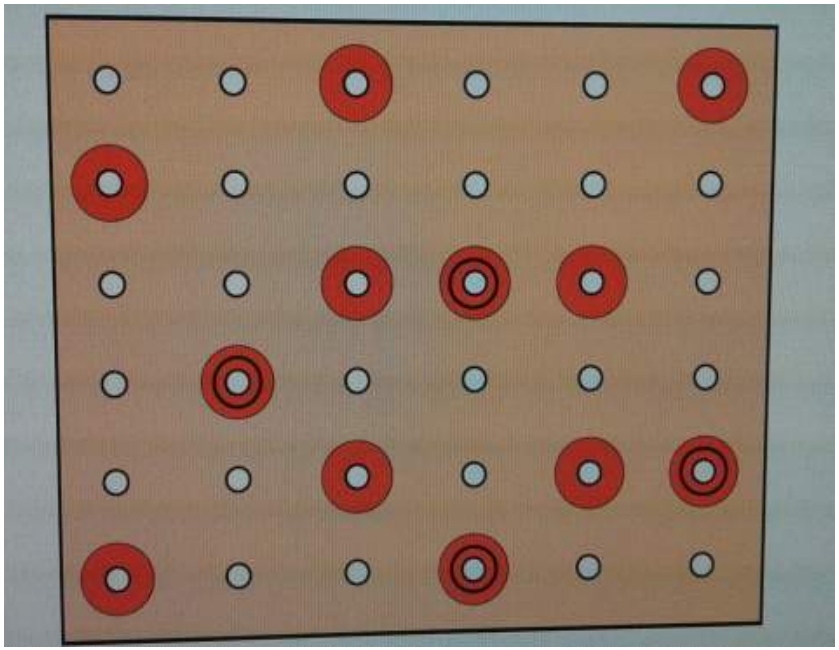
층화 표집(4 strata, 3 units selected per stratum)



다단계 표집(2 stage cluster design, 6 PSU's selected with 1 to 3 SSU's selected within each PSU)



다중표집(12 units selected at the first phase, 4 at the second)



4. 층화 및 표본배분

층화

- 모집단을 서로 겹치지 않는 여러 개의 층들(strata)로 나누고, 각 층에서 표본을 선정
- 모집단에 대한 표본의 대표성을 높이기 위하여 사용
- 일반적으로 추정의 효율을 높여줌

층화변수

- 사업체조사 : 산업분류, 사업체 규모(종업원수 등), 지역 등
- 가구조사 : 지역(시◦도), 세부지역(동부, 읍면부 등), 주택유형 등

층화평균 : 모평균 추정량

- 모집단이 H개의 층으로 층화되고, 각 층에서 n_h개 표본 추출
  - $N = N_1 + N_2 + \ldots + N_H$: 모집단크기
  - $n = n_1 + \ldots + n_h + \ldots + n_H$: 표본크기

- **층화평균**(stratified sample mean)
  - $\bar{y}_{st} = \frac{1}{N}\{N_1\bar{y}_1 + \ldots + N_L\bar{y}_H\} = \sum_{h=1}^{H} W_h\bar{y}_h$
  - $\bar{y}_h$: 층 표본평균, $W_h = \frac{N_h}{N}$: 층 가중치

표본의 배분

○ 표본배분 시 고려사항

- 사회조사 : 표본의 대표성 강조
- 경제조사 : 추정의 효율성 제고 강조

○ 표본배분의 방법

- **비례배분**
- **제곱근 배분**
- **최적배분**(**네이만** 배분)

비례 배분(Proportional Allocation)

- 각 층에서 모집단 크기에 비례하도록 표본배분

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

- **층화평균** : $\bar{y}_{st} = \sum_{h=1}^{L} W_h \bar{y}_h = \sum_{h=1}^{L} w_h \bar{y}_h = \bar{y}$
- 쉽고, 간편. 자기 가중(self weighting) 성립. 2차 분석에 유리
- 큰 층에서는 표본이 많이 뽑히고, 작은 층에서는 작게 뽑힘.
- 층별 오차관리에 어려움 발생

제곱근 비례 배분

- 각 층에서 모집단 제곱근에 비례하도록 표본배분

$$\frac{n_h}{n} \propto \frac{\sqrt{N_h}}{\sum \sqrt{N_h}}$$

- 소규모 층에서 과대, 대규모 층에서 과소 표집
- 비례배분의 단점 보완
- 모집단과 표본 구성비 불일치

최적 표본배분(Optimal Allocation)

○ 선형 비용 함수 : $Cost = c_0 + \sum_{h=1}^{H} c_h n_h$,

여기에서 c_0 : 고정비용, c_h: h층 단위비용

○ 표본배분에서 고려할 사항

- 각 층 내의 총 조사단위 수 : N_h
- 각 층 내의 관측값들 간의 변동 : S_h^2
- 각 층 내에서 관측에 드는 비용 : c_h

최적 표본배분(계속)

- 최적 표본배분 :

$$n_h \propto \frac{W_h S_h}{\sqrt{c_h}}$$

- 층 크기가 크면 표본수 많이 배분 : W_h에 비례
- 층내 관측값이 이질적이면 표본수 많이 배분 : S_h에 비례
- 층내 단위 비용이 크면 표본수 적게 배분 : $\sqrt{c_h}$에 반비례

네이만 배분(Neyman Allocation)

◦ 네이만 표본 배분 :

$$n_h \propto W_h S_h$$

◦ 각 층의 단위조사 비용이 같은 경우에 최적배정임

5. 무응답 처리

- **단위 무응답** : 표본에서 응답을 받지 전혀 못한 경우
- **항목 무응답** : 표본에서 일부 문항의 응답을 받지 못한 경우

표본(k)	y_{1k}	y_{2k}	y_{3k}	...	y_{qk}	
1	100	121	143	...	122	완전 응답
2	110	?	134	...	154	항목 무응답
3	120	?	?	...	176	항목 무응답
4	?	?	?	...	?	단위 무응답

- **무응답으로 통계적 문제 : 잠재적 편향**

무응답 처리 방법

- **단위 무응답** : 가중치 보정으로 처리
- **항목 무응답** : 무응답 대체로 처리

단위무응답의 가중치 보정 예제

	연령대					합계
	15-24	25-34	35-44	45-64	65+	
표본 크기	202	220	180	195	203	1,000
응답수	124	187	162	187	203	863
표본가중치 합계(1)	30,322	33,013	27,046	29,272	30,451	150,104
응답가중치 합계(2)	18,693	28,143	24,371	28,138	30,451	(129,796)
가중치 합 비율 ((2)/(1))	0.6165	0.8525	0.9011	0.9613	1.0000	(0.865)
가중치 인자 ((1)/(2))	**1.622**	**1.173**	**1.110**	**1.040**	**1.000**	(1.1565)

수치 예제

○ $n_1 = 3,\ n_2 = 4,\ n_{1R} = 2,\ n_{2R} = 3$, $N_1 = 70,\ N_2 = 60$

h	i	w_{hi}	R_{hi}	w_{hi}^{*}
1	1	20	1	$20 \times 70/45$
	2	25	0	0
	3	25	1	$25 \times 70/45$
2	1	15	0	0
	2	15	1	$15 \times 60/40$
	3	17	1	$17 \times 60/40$
	4	13	1	$13 \times 60/40$

무응답 대체 예제

표본 가구원	나이	성별	교육연한	범죄피해	강력범죄피해
1	47	M	16	0	0
2	45	F	?	1	1
3	19	M	11	0	0
4	21	F	?	1	1
5	24	M	12	1	1
6	41	F	?	0	0
7	36	M	20	1	?
8	50	M	12	0	0
9	53	F	13	0	?
10	17	M	10	?	?

표본 가구원	나이	성별	교육연한	범죄피해	강력범죄피해
11	53	F	12	0	0
12	21	F	12	0	0
13	18	F	11	1	?
14	34	M	16	1	0
15	44	M	14	0	0
16	45	M	11	0	0
17	54	F	14	0	0
18	55	F	10	0	0
19	29	F	12	?	0
20	32	F	10	0	0

주1) 범죄피해 : 1=yes, 0=no. 주2) ? : 무응답

무응답 대체 방법

- 연역적 대체 : 논리적 대체
  - 9번 표본 : 범죄피해=0 ⇒ 강력범죄피해=0
- 평균대체 :
  - 응답자 평균과 무응답자 평균이 동일하다는 가정
  - 간편, 변동 과소 추정 위험
- 핫덱 대체
- 회귀 대체

평균대체 예제

○ 교육연한 무응답 처리 :

		나이	
		$\leqq 34$	$\geqq 35$
성별	남	가구원 번호 3, 5, 10, 14	가구원 번호 1, 7, 8, 15, 16
	여	가구원 번호 4^*, 12, 13, 19, 20 $(y_4^* = \bar{y}_R = 11.25)$	가구원 번호 2^*, 6^*, 9, 11, 17, 18 $(y_2^* = y_6^* = \bar{y}_R = 12.25)$

○ 교육연한 : 2, 4, 6번 무응답

6. 사후층화

예제

○ 평균 몸무게를 알아내기 위하여 n=100인 단순확률표본을 선정

구분	남자	여자	합계
표본수	20	80	100
표본평균	180(파운드)	110(파운드)	124(파운드)

○ 표본 평균 : $n=100,\ \bar{y}=124 \Rightarrow$ 과소 추정됨

○ 사후층화 평균: $\bar{y}_{st}=\left(\frac{N_1}{N}\right)\bar{y}_1+\left(\frac{N_2}{N}\right)\bar{y}_2=145$

○ 사후층 비율이 $N_1/N\approx 0.5$, $N_2/N\approx 0.5$ 임을 이용

사후층화 보정

- 가구수 보정 : 시,도/ 동읍면/ 주택유형(단독, 연립, 아파트)
- 가구원수 보정 : 연령대/ 성별 등

- 조사의 특성에 따라 사후층화 보정 달라짐

7. 가중치와 추정

- **Estimation (추정)** is the means by which the statistical agency obtains values for the population of interest so that it can draw conclusions about that population based on information gathered from only a sample of the population.
- The **principle** behind estimation in a probability survey is that each **sample unit represents not only itself, but also several units of the survey population**.
- It is common to call the ave. num. of units in the population that a unit in the sample represents the **design weight** of the unit.

가중치

- The **first step** in estimation is assigning a weight to each sampled unit or to each of the responding sampled units.
- The **design weight (설계 가중치)** can be thought of as **the average number of units in the survey population that each sampled unit represents** and is determined by the sample design.

- The **design weight**, w_k for unit k is the sample is the inverse of its inclusion probability, π_k : $w_k = \frac{1}{\pi_k}$

가중치 산출

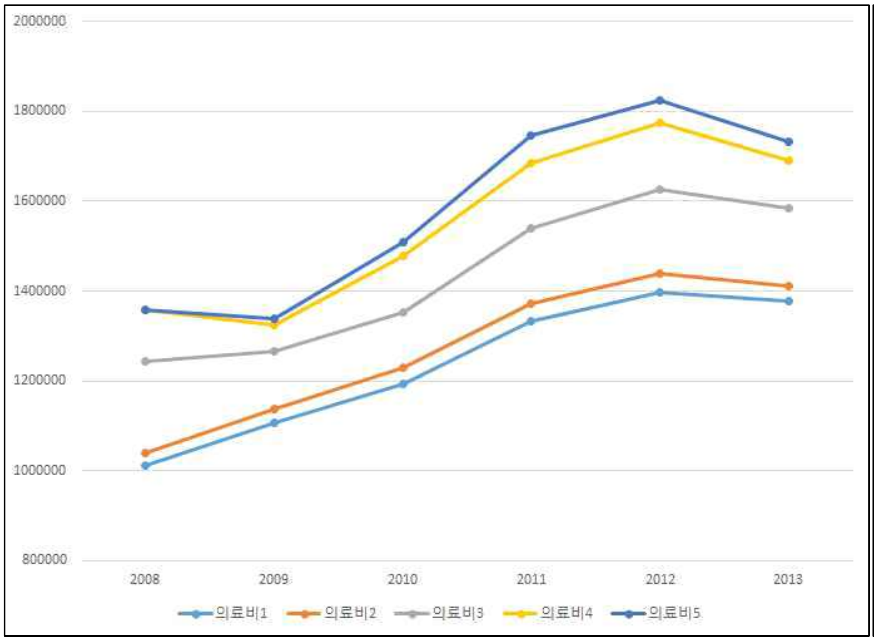
- **기본가중치** :
  - $w_k^0 = 1/\pi_k$, 표본추출확률의 역수
- **무응답 보정 가중치**
  - $F_k = 1/\hat{\phi}_k$: 무응답 보정인자
  - $w_k^1 = w_k^0 \times F_k$: 무응답 보정가중치
- **사후층화 보정가중치 : 최종가중치**
  - 사후층화 보정인자 : $\sum_{k \in r} w_k^1 G_k x_k = \sum_{k \in U} x_k$
  - $w_k = w_k^1 G_k = w_k^0 F_k G_k$: 사후층화 보정 최종가중치

추정량의 형태

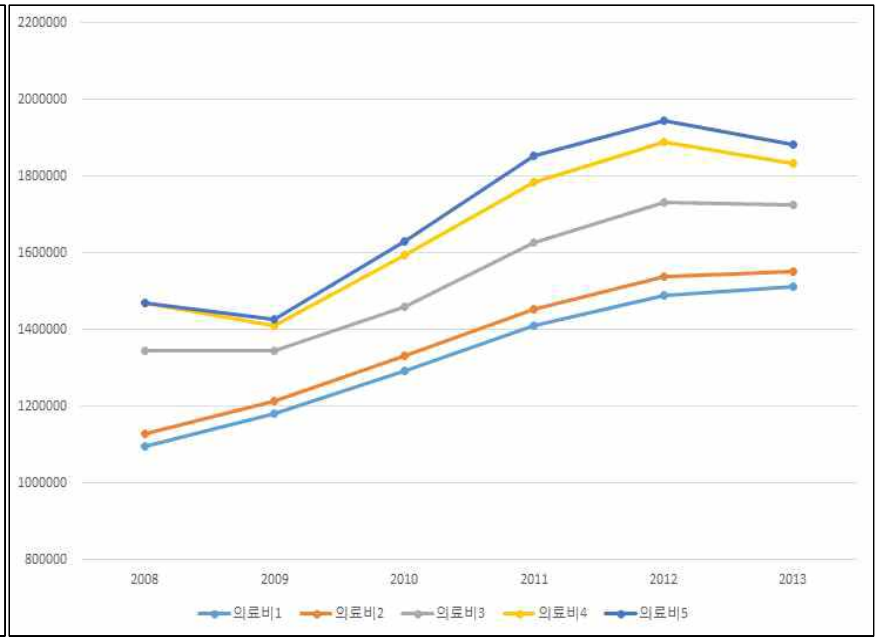
- 모집단 크기 추정량 : $\hat{N} = \sum_{i \in S_r} w_i$
- 모총계 추정량 : $\hat{t}_U = \sum_{i \in S_r} w_i y_i$
- 모평균 추정량 : $\hat{\bar{y}}_U = \dfrac{\sum_{k \in s} w_k y_k}{\sum_{k \in s} w_k}$
- 모비율 추정량 : $\hat{P} = \dfrac{\sum_{i \in S_r \cap C} w_i}{\sum_{i \in S_r} w_i} = \dfrac{\hat{N}_C}{\hat{N}}$

가중치 효과 : 의료패널 통계

○ **가구의료비1~5**



< 가중평균 >



< 비가중 평균 >

참고문헌

- 환경부, 한국환경공단 (2016). 2015년 기준 환경산업통계조사보고서.
- Leeuw, E.D., Hox, J.J. and Dillman, D.A. (2008). International handbook of survey methodology. European Association of Methodology.
- United Nations (2008). Designing household survey samples: Practical Guidelines.
- Statistics Canada (2003). Survey methods and practices.

Thank You!