10강 집락추출법(1)

정보통계학과 이기재교수

학/습/목/차

1. 집락추출법의 개념과 장단점
2. 모평균과 모총계 추정방법
3. 설계효과의 개념
4. 엑셀을 활용한 실습

집락추출법(cluster sampling)

먼저 서로 인접한 기본단위들로 구성된 집락을 추출하고,
추출된 집락 내의 일부 또는 전체를 조사하는 방법

- 집락의 예
  - 가구, 반, 통, 동, 인구주택총조사구, 사업체, 학교나 학급 등
  - 컴퓨터 회로판, 사과나무 등

- 집락 내 기본단위들은 같은 외부환경을 공유하여 서로 비슷한 특성을 지님

집락추출법의 사례

예제 6-1

▶ A시의 가구당 월평균 소득이나 소비지출액 추정을 위한 조사

◈ 단순임의추출법 적용

- A시의 전체 가구에 대한 완전한 명부를 구할 수 없음
  (주민등록부를 추출틀로 사용할 수 있지만,
  주민등록상의 주소지와 실제 거주지가 다를 수 있음)
- 표본가구가 흩어져 있어 실사 과정에서 시간과 비용이 많이 소요됨

◈ 집락추출법 적용

- A시 전체를 지역적인 경계가 명확하도록 블록(집락)으로 구분
- 구성된 전체 블록 중에서 일부를 표본블록으로 추출하여 조사
  ➡ 표본추출틀의 마련이 손쉽고, 실사 과정이 편리함

집락추출법의 장단점

장점

- ▶ 추출틀 마련이 간편함
  - ▪ 표본으로 추출된 집락 내의 조사단위에 대한 명부만 필요함
- ▶ 조사비용과 노력을 줄일 수 있음
  - ▪ 뽑힌 표본이 서로 인접하여 조사가 편리함

단점

- ▶ 같은 표본크기의 다른 표본추출법에 비해서 추정의 정확도가 떨어짐

학/습/목/차

1. 집락추출법의 개념과 장단점
2. 모평균과 모총계 추정방법
3. 설계효과의 개념
4. 엑셀을 활용한 실습

집락추출법 : 집락의 크기가 동일한 경우

- A 개의 집락, 각 집락은 B 개의 기본단위로 구성된 모집단을 대상으로 함
  - 모집단 내의 전체 조사단위 수 : $N = A \times B$
  - A 개의 집락 중에서 a 개 집락을 단순임의추출법으로 추출
  - 추출된 집락 내의 모든 기본단위 B개를 조사

    ➔ 조사되는 기본단위 수 : $n = a \times B$

    추출률 : $f = a \times B / A \times B = a / A$

집락추출법 : 집락의 크기가 동일한 경우

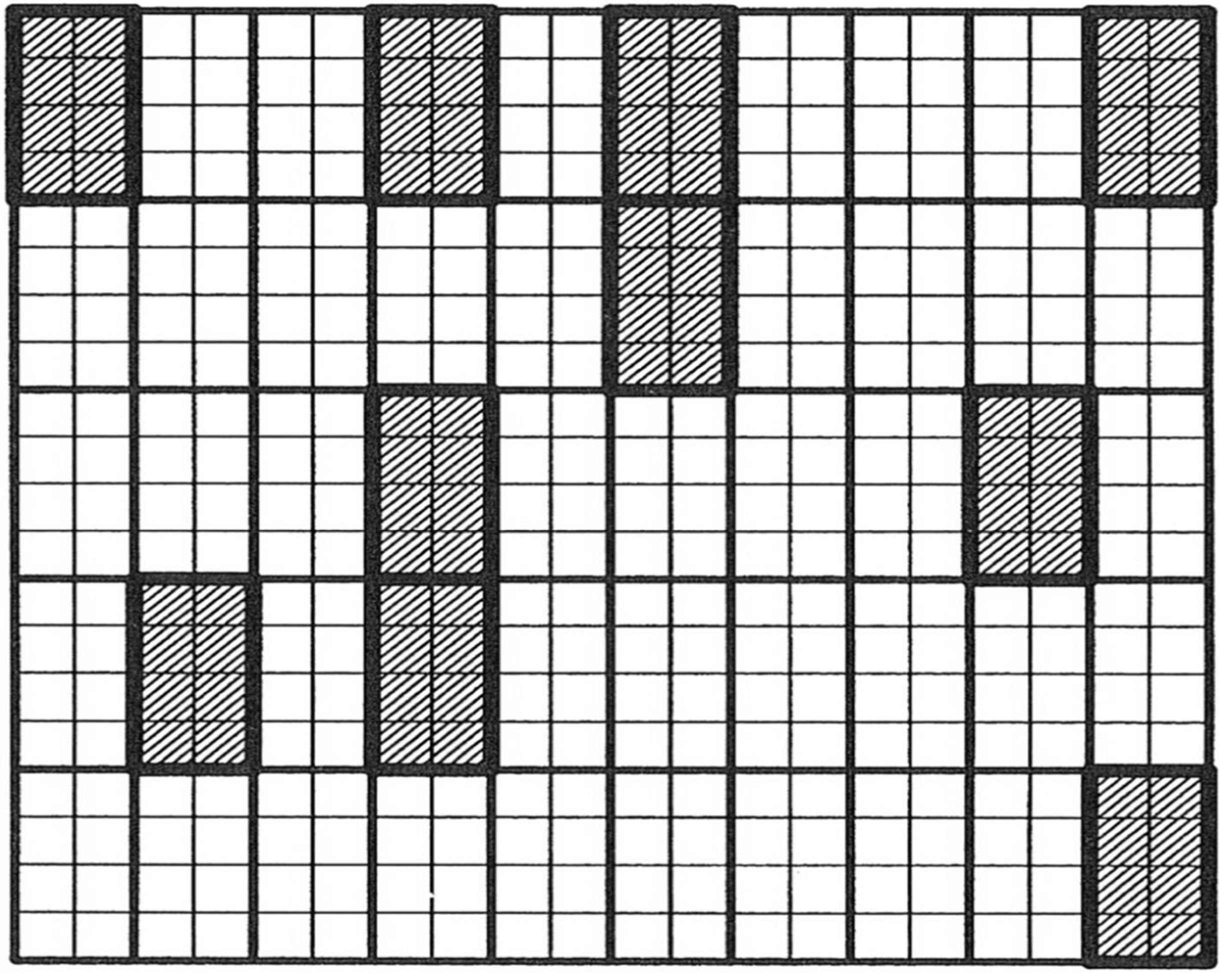

Figure 12.1. Cluster sample.

추정 : 집락의 크기 동일(1)

- 사용 기호

$N = A \times B$	모집단에서 기본단위의 총수
y_{ij}	i 번째 집락 내의 j 번째 기본단위의 조사값
$y_i = \sum_{j=1}^{B} y_{ij}$	i 번째 집락 내의 조사값의 합계
$\mu = \sum_{j=1}^{A} y_i / A = \sum_{i=1}^{A} \sum_{j=1}^{B} y_{ij} / N$	모평균
$\tau = \sum_{j=1}^{A} y_i = \sum_{i=1}^{A} \sum_{j=1}^{B} y_{ij}$	모총계

추정 : 집락의 크기 동일(2)

- 모평균 μ 에 대한 추정
  - $\overline{y}_{cl} = \sum_{i=1}^{a} \overline{y}_i / a$
  - 단, $\overline{y}_i = \sum_{j=1}^{B} y_{ij} / B$: 표본 집락 내의 조사값 평균
  - $\hat{V}(\overline{y}_{cl}) = \left(1 - \frac{a}{A}\right) \frac{s_b^2}{a}$, 여기서 $s_b^2 = \frac{\sum_{i=1}^{a} (\overline{y}_i - \overline{y}_{cl})^2}{a-1}$

※ 표본 집락 내의 모든 기본단위의 조사값 평균을 집락에 대한 조사값으로 간주하여 계산하기 때문에 단순임의추출법의 분산 추정과 유사함

추정 : 집락의 크기 동일(3)

- 모총계 τ 에 대한 추정
  - 모평균 추정량 $\bar{y}_{cl}$ 에 모집단의 총수(N)를 곱하여 구함
  - $\hat{\tau}_{cl} = N \times \bar{y}_{cl}$
  - $V(\hat{\tau}_{cl}) = N^2 V(\bar{y}_{cl})$
- 모비율 p 에 대한 추정
  - 모평균 추정의 특수한 경우임

모평균 추정 사례

- 예제 6-2
  - 어느 지역 의료보험조합에서 의료보험에 가입한 40,000가구를 대상으로 자기 집을 소유한 가구 비율 을 추정하고자 함

- 10가구씩을 묶어서 구성한 전체 A=4,000개 집락에서 a=40 집락 추출

집락	집락크기(B)	y_i	$\bar{y}_i$
1	10	10	1.0
2	10	8	0.8
3	10	6	0.6
4	10	5	0.5
5	10	9	0.9
6	10	8	0.8
7	10	8	0.8
8	10	5	0.5
9	10	9	0.9
10	10	9	0.9

집락	집락크기(B)	y_i	$\bar{y}_i$
11	10	9	0.9
12	10	10	1.0
13	10	4	0.4
14	10	3	0.3
15	10	1	0.1
16	10	2	0.2
17	10	3	0.3
18	10	4	0.4
19	10	0	0.0
20	10	6	0.6

집락내상관계수(intra-cluster correlation) (1)

- 모분산의 분해
  - $$\sigma^2 = \frac{\sum_{i=1}^{A}\sum_{j=1}^{B}(y_{ij}-\mu)^2}{N} = \frac{\sum_{i=1}^{A}\sum_{j=1}^{B}(y_{ij}-\bar{y}_i+\bar{y}_i-\mu)^2}{N}$$
    $$= \frac{\sum_{i=1}^{A}\sum_{j=1}^{B}(y_{ij}-\bar{y}_i)^2}{A\times B} + \frac{\sum_{i=1}^{A}\sum_{j=1}^{B}(\bar{y}_i-\mu)^2}{A\times B}$$
    $$= \frac{\sum_{i=1}^{A}\sum_{j=1}^{B}(y_{ij}-\bar{y}_i)^2}{A\times B} + \frac{\sum_{i=1}^{A}(\bar{y}_i-\mu)^2}{A}$$
    $$= \sigma_w^2 + \sigma_b^2 = (\text{집락내 분산}) + (\text{집락간 분산})$$
  - 집락 내 기본단위들이 이질적으로 구성되어 집락내분산이 큰 경우에 집락추출법이 효과적임

집락내상관계수(intra-cluster correlation) (2)

집락 내의 단위들이 동질적인가 아니면 이질적인가를 나타내는 척도

- $$\rho = \frac{E(y_{ij}-\mu)(y_{ij'}-\mu)}{E(y_{ij}-\mu)^2} = \frac{\sigma_b^2 - \sigma_w^2/(B-1)}{\sigma^2}$$
$$= \frac{B}{B-1} \times \frac{\sigma_b^2}{\sigma^2} - \frac{1}{B-1}$$

$$-\frac{1}{B-1} \leqq \rho \leqq 1$$

〈참고〉 $$V(\overline{y}_{cl}) = \left(1-\frac{a}{A}\right)\frac{S_b^2}{a} = \left(1-\frac{a}{A}\right)\frac{A\,\sigma_b^2}{a(A-1)}$$

집락내상관계수(intra-cluster correlation) (3)

- $\sigma_b^2 = 0$ 인 경우

  ➔ $\rho = -\dfrac{1}{B-1}$ 로 최소값

- $\sigma_w^2 = 0$ 인 경우

  ➔ $\rho = 1$ 로 최소값

- 각 집락 내의 단위들이 완전히 랜덤하게 배치된 경우

  ➔ $\rho = 0$

※ 집락 내에서는 동질적이고, 집락 간에 이질적인 모집단에
집락추출법을 적용하면 추정의 효율이 크게 떨어진다.

학/습/목/차

1. 집락추출법의 개념과 장단점
2. 모평균과 모총계 추정방법
3. 설계효과의 개념
4. 엑셀을 활용한 실습

설계효과(design effect) (1)

어떤 표본추출법과 같은 크기의 단순임의추출법을 추정의 정확도 측면에서 비교

- $DEFF(\hat{\theta}) = \dfrac{V_D(\hat{\theta})}{V_{SRS}(\hat{\theta})}$

  $V_{SRS}(\hat{\theta})$: 같은 표본크기의 단순임의표본에서 구한 추정량의 분산

- $DEFF(\hat{\theta}) > 1$: 특정한 표본설계 D가 단순임의추출법에 비해서 비효율적

  $DEFF(\hat{\theta}) < 1$: 특정한 표본설계 D가 단순임의추출법에 비해서 비효율적

설계효과(design effect) (2)

층화추출법 : 설계효과(DEFF)가 1보다 작게 나타남
집락추출법 : 대개 설계효과(DEFF)가 1보다 크게 나타남

- 집락추출법의 설계효과(집락의 크기 동일)

$$DEFF = \frac{V_{집락}(\bar{y}_{cl})}{V_{SRS}(\bar{y})} = [1 + (B-1)\rho]$$

- 대부분의 경우 집락내상관계수 ρ 는 양수값임

➔ 단순임의추출법에 비해서 추정의 정확도가 떨어짐

설계효과와 집락내상관계수 추정 사례

✤ 예제 6-3

▶ A=4,000개 집락에서 a=40개 집락 추출 (B=10)

- $\bar{y}_{cl} = \hat{p} = 0.463$

$$\hat{V}(\bar{y}_{cl}) = \left(1 - \frac{a}{A}\right)\frac{s_b^2}{a} = 0.99 \times \frac{0.1045}{40} = 25.86 \times 10^{-4}$$

$$\hat{V}_{SRS}(\hat{p}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}(1-\hat{p})}{n-1} = 6.17 \times 10^{-4}$$

- $\widehat{DEFF} = \dfrac{\hat{V}_{집락}(\bar{y}_{cl})}{\hat{V}_{SRS}(\bar{y})} = 4.19$

$$\hat{\rho} = \frac{\widehat{DEFF} - 1}{B - 1} = \frac{4.19 - 1}{10 - 1} = 0.354$$

학/습/목/차

1. 집락추출법의 개념과 장단점

2. 모평균과 모총계 추정방법

3. 설계효과의 개념

4. 엑셀을 활용한 실습

〈실습하기〉에서 자세히 다룸

이 강의는

강의용 휴대폰(U-KNOU 서비스 휴대폰)으로도

다시 볼 수 있습니다.