# 12강 2단집락추출법

정보통계학과 이기재교수

# 학/습/목/차

# 2단집락추출법

우선 1단계에서 모집단의 집락 중에서 표본집락을 추출하고, 2단계에서 추출된 표본집락 내에서 일부 조사단위를 추출하여 조사하는 방법

- 1차 추출단위(PSU: Primary Sampling Unit)
  - ▶ 1단계에서 추출되는 집락
- 2차 추출단위(SSU: Secondary Sampling Unit)
  - ▶ 2단계에서 추출되는 조사단위

# 2단집락추출법의 사례(1)

사례 1 : 어느 시에서 실시한 가구조사

- 해당 시는 $A$개의 동으로 구성되고, $i$번째 동은 $B_i$개의 통으로 구성
- $A$개 동에서 $a$개의 동을 표본으로 추출하고, 각 표본동에서 $b_i$개의 통을 표본으로 추출하여 조사함
  ➔ 2단집락추출법
- 1차추출단위(PSU) : 1단계에서 표본으로 추출되는 동
  2차추출단위(SSU) : 각 표본 동에서 표본으로 추출되는 통

# 2단집락추출법의 사례(2)

사례 2 : 초등학생을 대상으로 B형 간염에 대한 면역 비율 추정 목적

- 조사대상 : 초등학교에 다니고 있는 전체 학생
- 단순임의추출법 적용 : 추출틀 마련 곤란, 조사비용 과다 소용

  ➔ 2단집락추출법 또는 다단계추출법 적용
- 초등학교를 1차추출단위(PSU)로 하고,

  학급을 2차추출단위(SSU)로 하는 2단집락추출법을 적용
- 표본 초등학교에서 일부 학급을 추출하여 학급 내 모든 초등학생을 조사할 수 있고, 그 일부를 뽑아서 조사할 수도 있음

# 2단집락추출법의 장점

| | |
|---|---|
| 1 | 다른 표본추출법에 비해 조사가 편리하고, 조사비용이 적게 듬 |
| 2 | 비교적 쉽게 자체가중설계를 할 수 있음 |
| 3 | 우리나라의 행정조직과 같이 모집단이 여러 단계의 계층적인 조직으로 구성되어 있는 경우에 효과적임 |

〈**참고**〉 대규모 사회조사 또는 보건조사(health survey)에서는 각 추출단계에서 적절하게 표본추출법과 추출률을 정함으로써 표본 내 모든 조사단위가 같은 추출확률을 갖도록 할 수 있는데, 이러한 표본추출법을 등확률추출방법(EPSEM:Equal Probability Sampling Method) 또는 자체가중설계(self-weighting design)이라고 한다.

# 사용할 기호

- 모집단 $A$ 개의 PSU 중에서 $a$ 개의 표본 PSU를 단순임의추출하고, $i$ 번째 표본 PSU 내의 전체 $B_i$ 개 SSU 중에서 $b_i$ 개를 단순임의추출
- $A$ : 모집단내 1차추출단위(PSU)의 총수
- $B_i$ : $i$ 번째 1차추출단위(PSU)내 2차추출단위(SSU)의 총수
- $N$ : 모집단 내 SSU의 총수 ( $N = \sum_{i=1}^{A} B_i$ )
- $y_{ij}$ : $i$ 번째 PSU 내 $j$ 번째의 SSU의 조사값
- $n$ : 모집단에서 표본으로 추출할 SSU의 총수 ( $n = \sum_{i=1}^{a} b_i$ )

# 학/습/목/차

1. 2단집락추출법의 개념과 장점
2. 모수 추정법
3. 자체가중표본의 개념
4. 엑셀을 활용한 실습

# 모총계 추정 (1)

- 모총계의 수학적 표현
  - ▶ $\tau = \sum_{i=1}^{A} \sum_{j=1}^{B_i} y_{ij}$

- 모총계 추정량
  - ▶ $\hat{\tau} = A\left(\frac{1}{a}\sum_{i=1}^{a}\hat{\tau}_i\right) = A\left\{\frac{1}{a}\sum_{i=1}^{a}\frac{B_i}{b_i}\sum_{j=1}^{b_i} y_{ij}\right\} = \frac{A}{a}\sum_{i=1}^{a} B_i \bar{y}_i$
  - ▶ 단, $\bar{y}_i = \sum_{j=1}^{b_i} y_{ij}/b_i$

    $\hat{\tau}_i = B_i \bar{y}_i$ ( $i$ 번째 표본 PSU에 대한 총계 추정값)

# 모총계 추정 (2)

- $$\hat{V}(\hat{\tau}) = A^2\left(1-\frac{a}{A}\right)\frac{s_b^2}{a}+\frac{A}{a}\sum_{i=1}^{a}B_i^2\left(1-\frac{b_i}{B_i}\right)\frac{s_{wi}^2}{b_i}$$

- 단, $$s_b^2 = \frac{1}{a-1}\sum_{i=1}^{a}(\hat{\tau}_i - \hat{\mu}_{PSU})^2$$

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

$$\hat{\tau}_i = B_i\bar{y}_i\ ,\ \hat{\mu}_{PSU} = \frac{1}{a}\sum_{i=1}^{a}\hat{\tau}_i\ ,\ \bar{y}_i = \sum_{j=1}^{b_i} y_{ij}/b_{ij}$$

# 모총계 추정 (3)

- $$\hat{V}(\hat{\tau}) = A^2\left(1-\frac{a}{A}\right)\frac{s_b^2}{a}+\frac{A}{a}\sum_{i=1}^{a}B_i^2\left(1-\frac{b_i}{B_i}\right)\frac{s_{wi}^2}{b_i}$$

- 집락추출법 : 표본집락 내의 모든 기본단위를 조사
  - 추정량의 변동

    : 집락간변동(variation between cluster)에 의함

- 2단계집락추출법
  - 추정량의 변동

    : 집락간 변동과 집락내 변동 모두에 의해서 영향

# 모총계 추정 사례 : 예제 7-2

전체 A=100개의 농업조사구로 이루진 어느 지역에서 현재 사육 총 한우 수를 추정하고자 함

- a=10개 조사구를 단순임의추출하여 전체 가구 수, 표본 가구 수, 각 표본가구에서 조사한 현재 사육 중인 한우 수를 조사함

| 조사구 | $B_i$ | $b_i$ | $y_{ij}$ | | | | | | | | | | $\bar{y}_i$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 63 | 6 | 6 | 5 | 2 | 1 | 0 | 1 | | | | | 2.50 |
| 2 | 25 | 3 | 2 | 1 | 1 | | | | | | | | 1.33 |
| 3 | 57 | 6 | 4 | 10 | 6 | 0 | 1 | 0 | | | | | 3.50 |
| 4 | 96 | 10 | 5 | 1 | 3 | 2 | 7 | 1 | 0 | 1 | 0 | 0 | 2.00 |
| 5 | 44 | 4 | 7 | 1 | 1 | 0 | | | | | | | 2.25 |
| 6 | 23 | 2 | 1 | 0 | | | | | | | | | 0.50 |
| 7 | 67 | 7 | 9 | 2 | 5 | 1 | 0 | 0 | 1 | | | | 2.57 |
| 8 | 31 | 3 | 1 | 0 | 0 | | | | | | | | 0.33 |
| 9 | 27 | 3 | 3 | 1 | 0 | | | | | | | | 1.33 |
| 10 | 66 | 7 | 7 | 10 | 1 | 5 | 0 | 0 | 1 | | | | 3.43 |

# 모평균 추정

- 모평균( $\mu$ )의 수학적 표현
  - $\mu = \tau / N = \sum_{i=1}^{A} \tau_i / N = \sum_{i=1}^{A} \sum_{j=1}^{B_i} y_{ij} / N$
- 모집단 내 기본단위 총수( $N$ ) : 알려진 경우
  - $\hat{\mu} = \frac{\hat{\tau}}{N} = \frac{1}{N} \times \left( \frac{A}{a} \sum_{i=1}^{a} B_i \bar{y}_i \right)$
- 모집단 내 기본단위 총수( $N$ ) : 미지의 경우
  - $\hat{\mu}_r = \frac{\sum_{i=1}^{a} B_i \bar{y}_i}{\sum_{i=1}^{a} B_i}$

# 모비율 추정

- 모집단의 기본단위 총수 $N$ 이 알려지지 않는 경우가 대부분
  - 모비율 $p$에 대한 추정량

$$\hat{p} = \frac{\sum_{i=1}^{a} B_i \hat{p}_i}{\sum_{i=1}^{a} B_i} \qquad \text{단, } \hat{p}_i = \sum_{j=1}^{b_i} y_{ij} / b_i$$

  - 추정량 $\hat{p}$에 대한 추정분산

$$\hat{V}(\hat{p}) = \left(1 - \frac{a}{A}\right) \frac{1}{\hat{\bar{B}}^2} \frac{s_r^2}{a} + \frac{1}{aA\hat{\bar{B}}^2} \sum_{i=1}^{a} B_i^2 \left(1 - \frac{b_i}{B_i}\right) \frac{\hat{p}_i \hat{q}_i}{b_i - 1}$$

단, $\hat{\bar{B}}^2 = \sum_{i=1}^{a} B_i / a$ , $s_r^2 = \frac{1}{a-1} \sum_{i=1}^{a} B_i^2 (\hat{p}_i - \hat{p})^2$ , $\hat{q}_i = 1 - \hat{p}_i$

# 학/습/목/차

1. 2단집락추출법의 개념과 장점
2. 모수 추정법
3. 자체가중표본의 개념
4. 엑셀을 활용한 실습

# 자체가중표본(self-weighting sample)

- 표본으로 추출된 모든 조사단위가 같은 가중치를 갖도록 표본설계되어 추출된 표본
  - 추정과정이 간편하고, 추정의 정확도도 높아짐
  - 등확률추출방법
    (EPSEM : Equal Probability Sampling Method)

# 자체가중표본의 예

- 단순임의표본
- 비례배정이 사용된 층화임의표본
- 계통추출법
- 집락추출법 : 단순임의추출법으로 표본 집락추출
- 2단계추출법 : 1단계와 2단계 추출에 적절한 방법을 사용할 때

# 2단계집락추출법 : 자체가중표본

- PSU를 등확률로 추출하는 경우
  - $A$개의 PSU에서 $a$개를 단순임의추출하고, 추출된 PSU에서 $b_i$개의 SSU를 단순임의추출하는 경우
  - 모총계 추정량

$$\hat{\tau} = \frac{A}{a}\sum_{i=1}^{a}\frac{B_i}{b_i}\sum_{j=1}^{b_i} y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b_i}\left(\frac{A}{a}\right)\left(\frac{B_i}{b_i}\right) y_{ij} = \sum_{i=1}^{a}\sum_{j=1}^{b_i} w_{ij} y_{ij}$$

$$w_{ij} = \left(\frac{A}{a}\right)\times\left(\frac{B_i}{b_i}\right) = \left(\frac{1}{f_1}\right)\times\left(\frac{1}{f_{2i}}\right)$$

  - 자체가중표본이 되려면 $b_i \propto B_i$ 를 만족해야 함

# 2단계집락추출법 : 자체가중표본

- PSU를 확률비례로 추출하는 경우
  - 각 PSU의 크기 $B_i$에 비례하여 $a$개의 표본 PSU를 확률비례추출하고, 표본으로 추출된 PSU에서 각각 $b$개의 SSU를 단순임의추출
  - $f_1 = a \bullet \dfrac{B_i}{\sum_{\alpha=1}^{A} B_\alpha}$ , $f_{2i} = \dfrac{b}{B_i}$

$$f = f_1 \times f_{2i} = a \bullet \frac{B_i}{\sum_{\alpha=1}^{A} B_\alpha} \times \frac{b}{B_i} = \frac{a \times b}{\sum_{\alpha=1}^{A} B_\alpha} = \frac{n}{N}$$

  - 각 표본 PSU에서 조사 업무량이 일정함

# 학/습/목/차

1. 2단집락추출법의 개념과 장점
2. 모수 추정법
3. 자체가중표본의 개념
4. 엑셀을 활용한 실습

〈실습하기〉에서 자세히 다룸

이 강의는

강의용 휴대폰(U-KNOU 서비스 휴대폰)으로도
다시 볼 수 있습니다.