

## 10강. 집락추출법(1)

◆ 담당교수 : 이기재 교수

### ■ 주요용어

용어	해설
집락추출법(cluster sampling)	서로 인접한 기본단위들로 구성된 집락을 만들어, 먼저 집락을 추출하고 추출된 집락 내 일부 또는 전체를 추출하여 조사하는 방법이다.
집락내상관계수(intra-cluster correlation coefficient)	집락 내의 단위들이 동질적인가 아니면 이질적인가를 나타내는 측도를 말한다.
설계효과(design effect : DEFF)	어떤 표본추출방법 D에서 모수 $\theta$ 에 대한 추정량으로 $\hat{\theta}_D$ 를 사용할 때 설계효과(DEFF)는 다음과 같이 정의된다. $DEFF(\hat{\theta}) = \frac{V_D(\hat{\theta})}{V_{SRS}(\hat{\theta})}$ 여기서, $V_{SRS}(\hat{\theta})$ 는 같은 표본크기의 단순임의표본에서 구한 추정량의 분산이다.

### ■ 실습하기

- 교재 187쪽 예제 6-2
  - \* 집락추출법에서 모평균 추정
- 교재 193쪽 예제 6-3
  - \* 설계효과 및 집락내상관계수 추정

### ■ 연습문제

1. 다음은 집락추출법에 대한 설명이다. 올바른 것끼리 짝지어진 것은?

가. 최신 정보를 포함한 추출틀 작성이 곤란한 경우에 사용할 수 있다.
나. 조사가 편리하고 조사비용을 절감할 수 있다.
다. 같은 표본 크기의 다른 추출방법에 비해서 추정의 정확도가 높다.

- ① 가, 나
- ② 가, 다
- ③ 나, 다
- ④ 가, 나, 다

정답 : ①

해설 : 집락추출법은 표본추출이 편리하고, 조사비용을 절감할 수 있지만, 같은 크기의 단순임의추출법에 비해서 추정의 효율이 떨어진다.

2. 다음은 설계효과(design effect: DEFF)에 대한 설명이다. 옳지 않은 것은?

① 추정량의 분산 측면에서 특정 표본추출법과 단순임의추출법을 비교하기 위하여 구한다.

② 어떤 표본추출방법 D의 설계효과는  $DEFF(\hat{\theta}) = \frac{V_D(\hat{\theta})}{V_{SS}(\hat{\theta})}$  으로 정의된다.

③ 집락추출법(집락의 크기 동일)의 설계효과는  $DEFF(\bar{y}_{cl}) = [1 + (B-1)\rho]$  이다.

④ 일반적으로 집락추출법을 적용하면 설계효과는 1보다 작게 나타난다.

정답 : ④

해설 : 어떤 표본추출법 D의 설계효과는  $DEFF(\hat{\theta}) = \frac{V_D(\hat{\theta})}{V_{SS}(\hat{\theta})}$  으로 정의된다.

따라서 설계효과가 1보다 크게 나타나면 특정한 표본설계가 같은 표본크기의 단순임의추출법에 비해서 효율이 떨어짐을 나타낸다. 보통 집락추출법의 설계효과(DEFF)는 1보다 크다.

3. 다음은 층화임의추출법에서 사용되는 층과 집락추출법에서 사용되는 집락을 비교하여 설명한 것이다. 다음 설명 중에서 옳지 않은 것은?

① 층화임의추출법에서는 모든 층에서 표본을 추출한다.

② 집락추출법은 집락 내의 조사단위들이 동질적이도록 구성될 때 추정의 효율이 높다.

③ 층과 집락은 모집단을 분할하여 구성한다는 점에서 유사하다.

④ 집락추출법에서는 전체 집락 중에서 일부를 추출하여 조사한다.

정답 : ②

해설 : 집락추출법은 집락 내에서는 이질적이고, 집락 간에는 동질적인 경우에 추정의 효율이 높다.

4. 다음은 실제 조사에서 집락추출법을 널리 사용되는 이유를 기술한 것이다. 옳지 않은 것은?

- ① 조사비용과 시간을 줄일 수 있다.
- ② 같은 크기의 단순임의표본나 층화임의표본에 비해서 추정량의 표준오차를 줄일 수 있다.
- ③ 모집단에 대한 완전한 추출률이 마련되지 않은 경우에도 사용할 수 있다.
- ④ 집락 내의 가구들이 지리적으로 가까이 위치하여 조사가 편리하다.

정답 : ②

해설 : 집락추출법은 표본추출이 편리하고, 조사비용을 절감할 수 있지만 같은 크기의 단순임의추출법이나 층화임의추출법에 비해서 추정의 효율이 떨어진다.

5. 어떤 조사에 사용된 표본추출법에 대해서 설계효과(design effect: DEFF)를 계산해 보니 2.35로 나타났다. 그 의미를 적절하게 설명한 것은?

- ① 이 조사에서 추정량의 분산은 2.35이다.
- ② 이 조사에서 추정량의 표준오차는 2.35이다.
- ③ 같은 크기의 단순임의표본을 이용할 때와 비교하면 추정량의 분산이 2.35배 커진다.
- ④ 같은 크기의 단순임의표본을 이용할 때와 비교하면 추정량의 분산이 2.35배 작아진다.

정답 : ③

해설 : 어떤 표본추출방법 D의 설계효과는 으로 정의된다. 따라서 같은 크기의 단순임의표본인 경우와 비교하여 추정량의 분산이 2.35배 커진다는 의미이다.

## ■ 정리하기

- 집락추출법(cluster sampling)은 기본단위들로 구성된 집락을 만들어, 먼저 집락을 추출하고 추출된 집락 내의 일부 또는 전체를조사하는 방법을 말한다.
- 집락추출법은 같은 표본크기의 다른 표본추출법에 비해서 추정의 정확도는 떨어지지만 조사가 편리하고, 비용을 절감할 수 있다.
- 모평균 추정 : 집락의 크기가 동일한 경우

$$\text{- 추정량 : } \bar{y}_{\alpha} = \sum_{i=1}^a \bar{y}_i / a$$

여기서,  $\bar{y}_i = \sum_{j=1}^B y_{ij} / B$  : 표본 집락 내의 조사값 평균

$$\text{- } V(\bar{y}_{\alpha}) = \left(1 - \frac{a}{A}\right) \frac{S_y^2}{a}, \quad S_y^2 = \frac{\sum_{i=1}^A (\bar{y}_i - \mu)^2}{A-1}$$

$$\text{- } \hat{V}(\bar{y}_{\alpha}) = \left(1 - \frac{a}{A}\right) \frac{s_y^2}{a}, \quad s_y^2 = \frac{\sum_{i=1}^a (\bar{y}_i - \bar{y}_{\alpha})^2}{a-1}$$

4. 집락내 상관계수는  $-\frac{1}{B-1} \leq \rho \leq 1$ 의 값을 갖는다. 모든 변동이 집락 내에서 이루어지고, 집락간 분산이 0인 경우에  $\rho = -\frac{1}{B-1}$ 로 최소값을 갖는다.

반대로 집락내의 분산이 0인 경우는  $\rho = 1$ 로 최대값을 갖는다. 각 집락 내 단위들이 랜덤하게 배치되어 있다면  $\rho = 0$ 이다.

5. 집락의 크기가 동일한 경우에 설계효과(DEFf)는 다음과 같다.

$$DEFf(\bar{y}) = \frac{V_{\text{독}}(\bar{y}_{\alpha})}{V_{\text{독}}(\bar{y})} = [1 + (B-1)\rho]$$

## ■ 참고문헌

- 이계오, 박진우, 이기재, 표본조사론, 한국방송통대학교출판부, 2013. 제1장
- 통계청 홈페이지 : <http://www.nso.go.kr>