12강. 2단집락추출법

◈ 담당교수 : 이기재 교수

■ 주요용어

용어	해설
2단집락추출법	모집단의 집락들 중에서 표본집락을 추출하고, 추출된 표본집락 내의 일부 조사단위들을 추출하여 조사하는 방법이다.
1차추출단위(PSU: Primary Sampling Unit)	2단집락추출법 또는 다단집락추출법에서 1단계에서 추출되는 집락을 말한다.
2차추출단위(SSU: Secondary Sampling Unit)	2단집락추출법 또는 다단집락추출법에서 2단계에서 추출되는 추출단위을 말한다.
자체가중설계 (self-weighting design)	각 단계에서 적절하게 표본추출법과 추출률을 정함으로써 표본 내의 모든 조사단위들이 전체적으로 같은 추출확률을 갖도록 하는 표본설계를 말한다. 이러한 표본추출법을 등확률추출방법(EPSEM:Equal Probability Sampling Method)이라고도 한다.
집락간 변동 (variation between cluster)	집락추출법이나 2단집락추출법에서 집락을 표본추출함에 따라 생기는 변동을 말한다. 집락추출법에서 추정량의 분산은 표본집락 내의 모든 기본단위들을 조사하기 때문에 집락간 변동(variation between cluster)에 의해서만 나타난다.
집락내 변동 (variation within cluster)	단집락추출법에서 집락 내의 모든 기본단위를 조사하지 않고 일부를 조사함으로서 발생하는 변동을 말한다.

■ 실습하기

- 교재 224쪽 2단집락추출법에 의한 모수 추정
  * 모총계 추정, 표준오차, 신뢰구간 작성
  * 모평균 추정, 표준오차, 신뢰구간 작성

■ 연습문제

1. 전국의 성인을 대상으로 하는 건강조사와 같은 대규모 표본조사에서 다단계추출법을 사용하는 가장 중요한 이유는 무엇인가?

① 추정량의 표본오차를 줄이기 위해서
② 각 지역별 추정을 위해서
③ 조사의 비표본오차를 측정하기 위해서
④ 조사비용을 절감하고, 조사가 편리하기 때문에

정답 : ④
해설 : 전국의 성인을 대상으로 건강조사를 실시할 때 가장 먼저 고려해야 할 사항은 조사비용과 조사의 편리성이다. 다단계추출법을 사용하면 같은 크기의 단순임의추출법에 비해서 추정량의 표본오차는 다소 증가하지만 조사비용을 많이 절감할 수 있고, 조사가 편리하다.

2. 다음의 설명과 가장 관련이 깊은 것은?

사람을 대상으로 하는 대규모 조사에서는 표본 내의 모든 조사단위들이 전체적으로 같은 추출확률을 갖도록 한다. 이렇게 표본의 모든 조사단위의 추출확률이 동일하면 추정이 대단히 간편하다.

① 설계효과(DEFF)
② 비확률추출법
③ 할당추출법
④ 자체가중설계(self-weighting design)

정답 : ④
해설 : 표본으로 추출된 모든 조사단위의 추출확률이 동일한 표본설계를 자체가중설계(self-weighting design)라고 한다.

3. 위의 표본추출 과정에 대한 다음의 설명 중 <u>옳지 않은</u> 것은?

※(3~4) 다음과 같은 방법을 표본을 추출하여 조사하고자 한다. 물음에 답하시오. 서울시에 소재한 근로자 100인 이상을 고용하고 있는 사업체에 종사하는 근로자를 대상으로 직무만족도를 파악하고자 한다. 이를 위해서 먼저 산업분류와 사업체 규모에 따라서 사업체를 구분하고, 각 층에서 10개씩의 사업체를 랜덤하게 추출하여 각 표본 사업체에서 50명의 근로자를 단순임의추출법으로 추출하여 조사하였다.

① 기본단위는 사업체이다.
② 추출틀은 서울시에 소재한 근로자 100인 이상을 고용하고 있는 사업체 리스트이다.
③ 이 경우의 층화변수는 산업분류와 사업체 규모 구분이다.
④ 사용된 표본추출법은 층화2단집락추출법이다.

정답 : ①
해설 : 근로자를 대상으로 직무만족도를 조사하는 경우이다. 따라서 이 조사의 기본단위는 근로자이다.

4. 위의 표본추출 과정에서 1차추출단위와 2차추출단위는 각각 무엇인가?

※(3~4) 다음과 같은 방법을 표본을 추출하여 조사하고자 한다. 물음에 답하시오.
서울시에 소재한 근로자 100인 이상을 고용하고 있는 사업체에 종사하는 근로자를 대상으로 직무만족도를 파악하고자 한다. 이를 위해서 먼저 산업분류와 사업체 규모에 따라서 사업체를 구분하고, 각 층에서 10개씩의 사업체를 랜덤하게 추출하여 각 표본 사업체에서 50명의 근로자를 단순임의추출법으로 추출하여 조사하였다.

① 1차추출단위 : 사업체, 2차추출단위 : 근로자
② 1차추출단위 : 근로자, 2차추출단위 : 사업체
③ 1차추출단위 : 서울시, 2차추출단위 : 근로자
④ 1차추출단위 : 서울시, 2차추출단위 : 사업체

정답 : ①
해설 : 표본추출과정을 보면 각 층에서 먼저 표본사업체를 추출하고, 추출된 사업체에서 50명의 근로자를 단순임의추출법으로 추출하여 조사하고 있다. 따라서 1차추출단위는 사업체이고, 2차추출단위는 근로자이다.

5. 조사비용을 고려하지 않을 때 조사의 정확도가 가장 높을 것으로 예상되는 방안은?

※(5-6) 전국의 고등학교 3학년 학생을 대상으로 표본크기 2000명인 조사를 진행하고자 한다. 이를 위해서 다음과 같은 표본설계 방안을 고려하고 있다. 다음 물음에 답하시오.
- 방안 1 : 전국에서 2000명을 단순임의추출하는 경우
- 방안 2 : 100개 학교를 추출하고 각 표본학교에서 20명씩 조사하는 경우
- 방안 3 : 50개 학교를 추출하고 각 표본학교에서 40명씩 조사하는 경우

① 방안 1
② 방안 2
③ 방안 3
④ 방안 1, 2, 3은 조사의 정확도에 차이가 없다.

정답 : ①
해설 : 대개 집락내상관계수가 0보다 크기 때문에 같은 표본크기의 경우에는 집락추

출법에 비해서 단순임의추출법을 적용할 때 추정의 정확도가 높다.

6. 다음은 위에 제시된 방안 2의 표본추출방법에 대한 설명이다. 옳지 않은 것은?

※(5-6) 전국의 고등학교 3학년 학생을 대상으로 표본크기 2000명인 조사를 진행하고자 한다. 이를 위해서 다음과 같은 표본설계 방안을 고려하고 있다. 다음 물음에 답하시오.
- 방안 1 : 전국에서 2000명을 단순임의추출하는 경우
- 방안 2 : 100개 학교를 추출하고 각 표본학교에서 20명씩 조사하는 경우
- 방안 3 : 50개 학교를 추출하고 각 표본학교에서 40명씩 조사하는 경우

① 2단집락추출법이 적용되었다.
② 1차추출단위는 학교이고, 2차추출단위는 학생이다.
③ 방안 3에 비해서 조사비용과 시간을 줄일 수 있을 것이다.
④ 집락내상관계수는 양수의 값을 나타낼 것이다.

정답 : ③
해설 : 일반적으로 2단집락추출법을 적용할 경우에는 조사비용과 시간에 가장 영향을 미치는 것은 1차추출단위의 수이다.

■ 정리하기

- 2단집락추출법의 개념- 모집단의 집락들 중에서 표본집락을 추출하고, 추출된 표본집락 내의 일부 단위들을 추출하여 조사하는 방법- 이 방법은 집락추출법과 비교할 때 추정의 정확도가 높아지고, 표본 집락 내의 기본단위에 대한 추출틀만 마련되면 적용할 수 있어널리 사용되고 있다.

- 2단집락추출법에서 모총계의 추정량과 분산

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

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

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

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

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

■ 모평균(N)의 추정량과 분산

① 모집단의 기본단위 총수(N)를 알고 있는 경우

- $\hat{\mu} = \frac{\hat{\tau}}{N} = \frac{1}{N} \times \left(\frac{A}{a} \sum_{i=1}^{a} B_i \bar{y}_i \right)$

- $V(\hat{\mu}) = V(\hat{\tau}/N) = V(\hat{\tau})/N^2$

$$= \frac{1}{N^2} \left[A^2 \left(1 - \frac{a}{A} \right) S_{ba}^2 + \frac{A}{a} \sum_{i=1}^{A} B_i^2 \left(1 - \frac{b_i}{B_i} \right) \frac{S_{wi}^2}{b_i} \right]$$

단, $S_b^2 = \frac{1}{A-1} \sum_{i=1}^{A} (\tau_i - \mu_{PSU})^2$, $S_{wi}^2 = \frac{1}{B_i - 1} \sum_{j=1}^{B_i} (y_{ij} - \mu_i)^2$

② 기본단위 총수 N을 모르는 경우

- $\hat{\mu}_r = \frac{\sum_{i=1}^{a} B_i \bar{y}_i}{\sum_{i=1}^{a} B_i}$

- $V(\hat{\mu}) = \left(1 - \frac{a}{A} \right) \frac{1}{\overline{B}^2} \frac{S_r^2}{a} + \frac{1}{aA\overline{B}^2} \sum_{i=1}^{A} B_i^2 \left(1 - \frac{b_i}{B_i} \right) \frac{S_{wi}^2}{b_i}$

단, $\overline{B} = \sum_{i=1}^{A} B_i / A$, $S_r^2 = \frac{1}{A-1} \sum_{i=1}^{A} B_i^2 (\mu_i - \mu)^2$, $S_{wi}^2 = \frac{1}{B_i - 1} \sum_{j=1}^{B_i} (y_{ij} - \mu_i)^2$

■ 모비율의 추정 : 기본단위 총수 N이 알려지지 않은 경우

- 모비율 p의 추정량 : $\hat{p} = \frac{\sum_{i=1}^{a} B_i \hat{p}_i}{\sum_{i=1}^{a} B_i}$ 단, $\hat{p}_i = \sum_{j=1}^{b_i} y_{ij} / b_i$

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

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

■ 자체가중표본의 개념

- 표본을 구성하는 각 조사단위의 추출확률이 같은 표본을 자체가중표본(self-weighting sample)이라고 함
- 2단집락추출법에서 자체가중표본을 얻는 방법으로는 PSU를 등확률로 추출하

는 방법과 확률비례추출법을 이용하는 방법이 있음

- 자체가중표본을 이용하면 표본으로 추출된 모든 조사단위가 같은 가중치를 갖기 때문에 추정이 간편하고 효율적임

■ 참고문헌

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