제 6장: 집락표집

6.1 집락표집의 개요

- 집락표집은 계통표집과 마찬가지로 조사의 편리성을 고려한 조사 방법

- 예를 들어, 서울시에서 각 가구별로 보유한 컴퓨터의 총 수를 추정하고자 하는 경우 서울시에 존재하는 모든 가구의 목록이 필요하나 이런 목록을 얻기가 어렵고 모든 가구의 목록을 얻는다 하더라고 단순임의추출을 시행하는 경우 조사 개체들이 서울시 전체에 산포되게 되어 조사비용이 커지게 된다. 이 경우 서울시에 존재하는 모든 동의 목록을 마련하고 이들 중 50개의 동을 임의로 선택하게 되고 선택된 동에서 5개의 통씩 임의로 선택하고 또 선택된 통 내의 모든 가구들을 조사하는 방식을 선택하게 된다.

- 이런 표집방법을 이단집락표집(two-stage cluster sampling)이라 하고
  일차표집단위 (primary sampling unit, psu)는 동 들이
  이차표집단위 (secondary sampling unit, ssu)는 통 들이 된다.

- 집락표집을 이용하여 모수를 추정하는 경우 동일 크기의 단순임의표집에 비하여 추정의 정밀도가 낮아지는 경향이 있을 수 있다. 이는 집락표집을 시행할 때 집락내의 구성원들의 성향이 비슷함으로서 나타나는 현상이다. 집락표본에서 최대 정밀도를 얻기 위하여는 집락내의 원소는 서로 이질적이고 집락간은 서로 동질적이어야 한다.

- 간단한 예로 크기가 $N(=Km)$인 모집단에서 $n=2m$개의 표본을 얻는 경우 크기가 m인 집락을 2개 뽑았다고 하면

  $$\widehat{\mu}_{\mathrm{cl}} = \frac{1}{2m}\sum_{i=1}^{2}\sum_{j=1}^{m} Y_{ij} = \frac{1}{2}(\overline{Y}_1 + \overline{Y}_2)$$

  이다. 이의 분산은 집락간 독립을 가정하면

  $$\mathrm{var}\big(\widehat{\mu}_{\mathrm{cl}}\big) = \frac{1}{4}\big(\mathrm{var}(\overline{Y}_1) + \mathrm{var}(\overline{Y}_2)\big)$$

이고

$$\mathrm{var}(\overline{Y}_1) = \mathrm{var}(\overline{Y}_2) = \frac{1}{m^2}\Big\{\sum_{j=1}^{m} \mathrm{var}(Y_{1j}) + \sum_{j1 \neq j2} \mathrm{cov}(Y_{1,j1}, Y_{2,j2})\Big\}.$$

- 만일 집락내가 동질적이라면 $\mathrm{var}(\overline{Y}_1) \geq \frac{\sigma^2}{m}$ 이 되어 $\mathrm{var}\big(\widehat{\mu}_{\mathrm{cl}}\big)$ 이 $\mathrm{var}\big(\widehat{\mu}_{\mathrm{sys}}\big)$ 보다 커지게 된다.

- 집락표집에서는 표집단위 (sampling unit)이 집락이므로 집락표집의 표본크기는 선택된 집락의 개수를 의미하고 조사한 개체들의 수가 아니다.

6.2 집락표집에 대한 표기

y_{ij}: i번째 psu의 j번빼 ssu에 해당하는 관측값

모수	표본통계량
N=모집단에서 psu들의 총수, 집락 수	n=표본으로 선택된 psu들의 총수, 표집된 집락수
M_i= i번째 집락의 크기: i번째 psu에 속한 ssu들의 총수	m_i :i번째 psu에서 추출하는 표본의 크기
$K=\sum_{i=1}^{N} M_i$: 모집단에서 ssu들의 총수	$k=\sum_{i=1}^{n} m_i$:총 표본크기
$\overline{M}=K/N=\sum_{i=1}^{N} M_i/N$: psu당 평균 크기	$\overline{m}=\sum_{i=1}^{n} m_i/n$: 추출된 psu당 평균 표본 크기
$\tau_i=\sum_{j=1}^{M_i} y_{ij}$: i번째 psu의 모합	$t_i=\sum_{j=1}^{m_i} y_{ij}$: i번째 psu의 표본합 $\hat{\tau}_i=\frac{M_i}{m_i}t_i$: i번째 psu의 모합에 대한 추정량
$\tau=\sum_{i=1}^{N}\tau_i=\sum_{i=1}^{N}\sum_{j=1}^{M_i} y_{ij}$: 모합	$\hat{\tau}=\frac{N}{n}\sum_{i=1}^{n}\hat{\tau}_i$: 모합에 대한 추정량
$\overline{\tau}=\tau/N$: psu당 평균합	$\overline{t}=\hat{\tau}/N$:psu당 평균합에 대한 추정량
$\sigma_b^2=\frac{1}{N-1}\sum_{i=1}^{N}(\tau_i-\overline{\tau})^2$: psu 합들에 대한 모분산	$s_b^2=\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\tau}_i-\overline{t})^2$
$\mu=\tau/K$: (ssu당)모평균	$\hat{\mu}=\hat{\tau}/K$
$\mu_i=\tau_i/M_i$: i번째 psu의 모평균	$\hat{\mu}_i=\hat{\tau}_i/M_i$
$\sigma^2=\frac{1}{K-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij}-\mu)^2$: (ssu당)모분산	$s^2=\frac{1}{k-1}\sum_{i=1}^{n}\sum_{j=1}^{m_i}(Y_{ij}-\hat{\mu})^2$
$\sigma_i^2=\frac{1}{M_i-1}\sum_{j=1}^{M_i}(y_{ij}-\mu_i)^2$: i번째 psu의 모분산	$s_i^2=\frac{1}{m_i-1}\sum_{j=1}^{m_i}(Y_{ij}-\hat{\mu}_i)^2$

6.3 일단집락표집

일단집락표집은 선택된 psu에 속한 ssu들을 전부 조사한다. 일단집락표집의 예로는 고등학생들에 대한 설문조사의 경우 psu는 학급이고 ssu는 학생, 폐렴으로 입원한 환자에 대한 조사의 경우 병원을 psu로 폐렴환자를 ssu로 조사한다.

6.3.1 동일한 집락의 크기

- 편의상 $M_1 = M_2 = \cdots = M_N = M$을 가정하고 이 때 전체 개체의 수 $K = N \cdot M$이다.
- 관측값은 $\{\tau_1, \tau_2, \cdots, \tau_n\}$이 되고 여기서 τ_i는 i번째 선택된 집락에서 모든 원소들에 대한 합이다.
- 전체 모합에 대한 추정량은
$$\widehat{\tau} = \frac{N}{n} \sum_{i=1}^{n} \tau_i$$
가 된다.
- 추정량의 분산: $\{\tau_1, \tau_2, \ldots, \tau_N\}$의 모집단에서 n개를 SRS 했다고 생각하여 SRS의 분산공식을 이용하여 계산할 수 있다.
$$\text{var}(\widehat{\tau}) = N^2 \frac{N-n}{N} \frac{\sigma_b^2}{n}$$
이고 여기서
$$\sigma_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \mu)^2.$$
이다.
- 분산의 추정량:
$$\widehat{\text{var}}(\widehat{\tau}) = N^2 \frac{N-n}{N} \frac{s_b^2}{n}$$
이고
$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - \bar{t})^2, \qquad \bar{t} = \frac{\widehat{\tau}}{N}$$

이다.

- <u>모평균 μ의 추정</u>

6.3.2 집락표집의 효율

- 집락의 크기가 같음을 가정하고 표본의 크기가 n인 집락표집과 표본의 크기가 nM인 단순임의추출의 효율을 비교한다. $K = NM$이라 하자.

- 각각의 분산이

$$\mathrm{var}\big(\widehat{\tau}_{\mathrm{cl}}\big) = N^2 \frac{N-n}{N} \frac{1}{n} \sigma_b^2$$

과

$$\mathrm{var}\big(\widehat{\tau}_{\mathrm{srs}}\big) = K^2 \frac{K-nM}{K} \frac{\sigma^2}{nM} = N^2 \left(\frac{N-n}{N}\right) \frac{M}{n} \sigma^2$$

이다. 따라서 σ_b^2과 $M\sigma^2$을 비교한다. (σ_b^2과 σ^2의 정의에서 분모를 $N-1$과 $NK-1$을 사용하였다.)

- 계산:

No.

Date.

$$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2 = \sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i-\mu)^2 + \sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu_i)^2$$

SST SSB SSW

NM − 1 N − 1 N(M−1)

$$\sigma_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}(\tau_i-\bar{\tau})^2 = \frac{1}{N-1}\sum_{i=1}^{N}(M\mu_i - M\mu)^2$$

$$= \frac{M}{N-1}\, SSB = M\cdot MSB$$

급내 상관계수

$$\rho_w = \frac{E\left((y_{ij}-\mu)(y_{ij'}-\mu)\right)}{E(y_{ij}-\mu)^2} = 1-\frac{M}{M-1}\,\frac{SSW}{SST}$$

임을 교재의 P158, 160을 통하여 확인한다.

$$MSB = \frac{1}{N-1}\left\{SST - SSW\right\}$$

$$1-\rho_w = \frac{M}{M-1}\,\frac{SSW}{SST}$$

$$SSW = SST\cdot\frac{M-1}{M}(1-\rho_w)$$

$$= \frac{1}{N-1}\left\{1-\frac{M-1}{M}(1-\rho_w)\right\}$$

$$= SST\,\frac{1}{M}\,\frac{1}{N-1}\left\{M-M+1+\rho_w(M-1)\right\}$$

$$\frac{1}{NM}SST = \sigma^2$$

$$= \sigma^2\,\frac{NM-1}{M(N-1)}\left\{1+(M-1)\rho_w\right\}$$

$$\approx \sigma^2\,\frac{N}{N-1}\left\{1+(M-1)\rho_w\right\}$$

- 따라서

$$\begin{aligned}\frac{\mathrm{var}(\widehat{\tau}_{\mathrm{cl}})}{\mathrm{var}(\widehat{\tau}_{\mathrm{srs}})} &= \frac{\sigma_b^2}{M\sigma^2} = \frac{M\mathrm{MSB}}{M\sigma^2} = \frac{\mathrm{MSB}}{\sigma^2} \\ &= \frac{N}{N-1}\Big\{1 + (M-1)\rho_{\mathrm{w}}\Big\}.\end{aligned}$$

- 급내상관계수의 범위는

$$-\frac{1}{M-1} \le \rho_{\mathrm{w}} \le 1$$

이고 ρ_{w} 가 음수이면 - 즉, 집락내의 원소의 성질이 이질적이면 - 집락추출이 단순임의추출보다 효율적이다.

6.3.3 크기가 다른 집락들

일반적으로 집락의 크기는 다르고 i 번째 집락의 크기를 M_i 라 하자

(1) 비편향 추정 (unbiased estimator)

- 추정량

$$\hat{\tau}_w = \frac{N}{n} \sum_{i \in A} z_i \tau_i \qquad z_i = \begin{cases} 1 & \text{만일 } i\text{번째 집락 추출} \\ 0 & \text{o.w} \end{cases}$$

- 평균

$$E(\hat{\tau}_w) = \frac{N}{n} \sum_{i=1}^{N} \frac{n}{N} \tau_i = \tau$$

- 추정량의 분산

$$var(\hat{\tau}_w) = N^2 \frac{N-n}{N} \frac{\sigma_b^2}{n}$$

$$\sigma_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2, \quad \bar{\tau} = \frac{\tau}{N} = \text{psu당 평균합}$$

- 분산의 추정량

$$\hat{var}(\hat{\tau}_w) = N^2 \frac{s_b^2}{n} \frac{N-n}{N}$$

$$s_b^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\tau_i - \bar{t})^2, \quad \bar{t} = \frac{1}{N} \hat{\tau}_w$$

모집단 원소의 총수를 $K = \sum_{i=1}^{N} M_i$ 라면 모평균의 추정은

$$\hat{\mu}_w = \frac{1}{K} \hat{\tau}_u$$

$$var(\hat{\mu}_u) = \frac{1}{K^2} N^2 \frac{\sigma_b^2}{n} \left(\frac{N-n}{N} \right)$$

$$\hat{var}(\hat{\mu}_u) = \frac{1}{K^2} N^2 \frac{s_b^2}{n} \left(\frac{N-n}{N} \right)$$

이 된다.

- 일반적으로 N은 알려져 있으나 K는 알려져 있지 않음을 기억한다

(2) 비추정

모평균의 다른 추정량으로 비추정량을 생각할 수 있다.

비추정량은 ① $K = \sum_{i=1}^{N} M_i$를 알 필요가 없고

② 편의 (bias) 가 있으나 종종 분산이 작아져 (1)절의 불편추정량보다 효율적이다

$$\hat{\mu}_r = \frac{\sum_{i=1}^{n} \tau_i}{\sum_{i=1}^{n} M_i} \quad : \text{비추정량}$$

이 추정량은 4장의 비 β 에 대한 추정량 b와 동일하다.

교재의 P120 를 참조하면

$$var(b) \approx \frac{1}{\mu_x^2} \frac{\sigma_e^2}{n} \frac{N-n}{N}$$ 로 부터

$$\widehat{var}(\hat{\mu}_r) \approx \frac{1}{\bar{M}^2}\frac{s_e^2}{n}\frac{N-n}{N}$$ 이고

$$\bar{M} = \frac{1}{N}\sum_{i=1}^{N} M_i = \frac{K}{N}$$

$$s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \hat{\mu}_r M_i)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n} M_i^2(\bar{Y}_i - \hat{\mu}_r)^2$$ 이다.

만일 $\bar{M}$을 모르면 $\bar{M}_s = \frac{1}{n}\sum_{i=1}^{n} M_i$를 사용한다.

위의 비 추정량을 이용한 모합 τ의 추정량은

$$\hat{\tau}_r = K\hat{\mu}_r \quad \text{or} \quad \hat{\tau}_r = \hat{K}\hat{\mu}_r, \quad \hat{K} = N\bar{M}_s$$

$$\widehat{var}(\hat{\tau}_r) = N^2 \frac{s_e^2}{n}\frac{N-n}{N}$$

정리

- 비 추정량은 편의(bias)가 존재하나 unbiased estimator 보다 분산이 작은 경향이 있다.

- 모합에 대한 추정시 비추정량은 모집단원소의 총수 $K = \sum_{i=1}^{N} M_i$를 알아야 한다.

- 모평균에 대한 추정시 비편향추정량은 모집단원소의 총수 $K = \sum_{i=1}^{N} M_i$를 알아야한다.

- 만일 $M_i = M$ 이라면

$$\hat{\mu}_r = \frac{1}{nM}\sum_{i=1}^{n}\tau_i = \frac{1}{K}\frac{N}{n}\sum_{i=1}^{n}\tau_i = \frac{1}{K}\hat{\tau}_u$$

$$\hat{\tau}_r = K\cdot\hat{\mu}_r = \frac{N}{n}\sum_{i=1}^{n}\tau_i = \hat{\tau}_u$$

이다.

(3) 모비율의 추정

i 번째 집락에서 관심특성을 보유한 단위들의 총수를 a_i라 표기하자.

① $M_1 = M_2 = \cdots = M_N = M$ 인 경우

$$P_i = \frac{1}{M}a_i \qquad i=1,2,\ldots,N$$

$$\hat{p} = \frac{1}{n}\sum_{j=1}^{n}p_j$$

- $\{p_1, p_2, \ldots, p_n\}$은 N개의 단위 $\{P_1, P_2, \ldots, P_N\}$에서 뽑은 크기 n의 비복원 단순임의추출이다.

$$Var(\hat{p}) = \frac{N-n}{N}\frac{\sigma_p^2}{n} \qquad \sigma_p^2 = \frac{1}{N-1}\sum_{i=1}^{N}(P_i - \bar{P})^2$$

$$\bar{P} = \frac{1}{K}\sum_{i=1}^{N}a_i = \frac{1}{N\cdot M}\sum_{i=1}^{N}a_i$$

$$\widehat{var}(\hat{p}) = \frac{N-n}{N}\frac{s_p^2}{n}, \quad s_p^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i - \bar{p})^2$$

$$\bar{p} = \frac{1}{nM}\sum_{j=1}^{n} a_j$$

② M_i들이 서로 다른 경우

이 경우 비 추정량은

$$\hat{p}_r = \frac{\sum_{j=1}^{n} a_j}{\sum_{j=1}^{n} M_j}$$

$$\widehat{var}(\hat{p}_r) = \frac{N-n}{N}\frac{1}{\bar{M}^2}\frac{s_e^2}{n}$$

$$s_e^2 = \frac{1}{n-1}\sum_{j=1}^{n}(a_j - \hat{p}_r M_j)^2$$

$$\bar{M} = \frac{K}{N} = \frac{1}{N}\sum_{i=1}^{N} M_i$$

(4) 표본크기의 결정

① 모평균 추정시 표본크기

$\hat{\mu}_r$ 을 사용한다.

$$var(\hat{\mu}_r) = \frac{1}{\bar{M}^2} \frac{\sigma_e^2}{n} \frac{N-n}{N}$$

$$z_{\frac{\alpha}{2}} \sqrt{var(\hat{\mu}_r)} = \text{오차한계}(B) \qquad -(*)$$

(*) 를 풀면

$$z_{\frac{\alpha}{2}}^2 \frac{1}{\bar{M}^2} \frac{\sigma_e^2}{n} \frac{N-n}{N} = B^2$$

$$n = \frac{N\sigma_e^2}{ND + \sigma_e^2}, \quad D = \left(\frac{B \cdot \bar{M}}{z_{\frac{\alpha}{2}}}\right)^2 \text{ 이다.}$$

② 모합에 대한 추정시 표본크기 (비편향 추정량을 사용한다)

$$var(\hat{\tau}_u) = N^2 \frac{\sigma_b^2}{n} \left(\frac{N-n}{N}\right)$$

$z_{\frac{\alpha}{2}} \sqrt{var(\hat{\tau}_u)}$ = 오차한계(B) 의 방정식을 통하여

$$n = \frac{N\sigma_b^2}{ND + \sigma_b^2}, \quad D = \left(\frac{B}{z_{\frac{\alpha}{2}} N}\right)^2$$

를 사용한다.

$$\sigma_b^2 = \frac{1}{N-1} \sum_{i=1}^{N} (\tau_i - \bar{\tau})^2, \quad \bar{\tau} = \frac{\tau}{N} = \text{psu 당 평균합}.$$

6.4. 층화 집락 표집 (Stratified cluster Sampling)

층화집락 표집은 모집단을 동질적인 층으로 층화한 다음 각 층마다 독립적인 집락 표집을 하는 방법이다.

예를들어 서울시 교육청에서 고 3 생들을 대상으로 조사하는 경우 서울에 총 600 여개의 고등학교가 있다고 한다면 각 고등학교는 지역적 동질성을 지닐수 있으므로 고등학교를 층으로 하고 모든 고등학교에서 3 학년 학급수의 $\frac{1}{3}$을 단순임의 선택하여 선택된 학급의 모든 학생을 대상으로 하는 조사를 생각 할 수 있다.

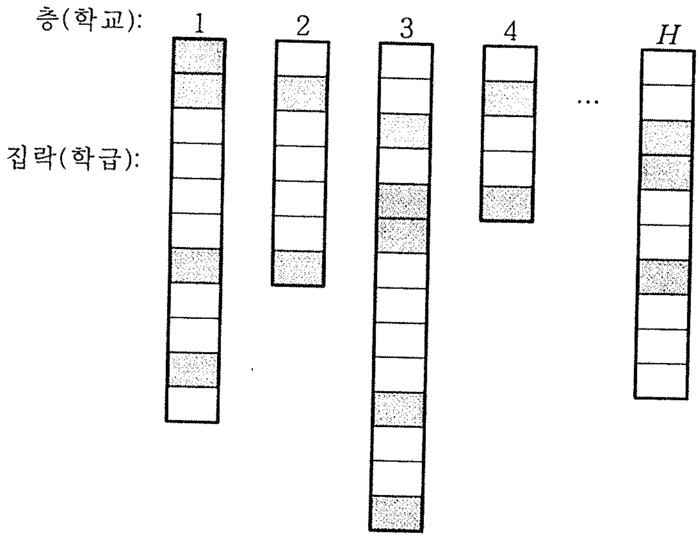

그림 6.6 층화집락표집의 도해

(1) 모합의 추정

- 모집단 전체의 총 집락수를 N이라 하고
  층 h의 총 집락수를 N_h, $h=1,2,\dots,H$ 라 하자

  $$N = N_1 + N_2 + \cdots + N_H$$

- 각 층에서 n_h개의 집락을 비복원 단순 임의 추출 하고
  해당집락의 집락합을 $\{\tau_{h1}, \tau_{h2}, \dots, \tau_{hn_h}\}$라 하자.

- $\bar{\tau}_h = \dfrac{1}{n_h} \displaystyle\sum_{j=1}^{n_h} \tau_{hj}$ 라 하면

  $\hat{\tau}_h$ = h 번째 층의 총합 τ_h에 대한 추정치

  $$= \frac{N_h}{n_h} \sum_{j=1}^{n_h} \tau_{hj} = N_h \cdot \bar{\tau}_h$$

- 전체모합 τ의 추정치로는

  $$\hat{\tau} = \sum_{h=1}^{H} \hat{\tau}_h = \sum_{h=1}^{H} N_h \bar{\tau}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} \tau_{hj}.$$

  를 사용하게 되고 분산추정량은

  $$\hat{Var}(\hat{\tau}) = \sum_{h=1}^{H} N_h^2 \frac{N_h - n_h}{N_h} \frac{s_h^2}{n_h}$$

  $$s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (\tau_{hj} - \bar{\tau}_h)^2$$

(2) 모평균의 추정

모평균의 추정량으로 $\hat{\tau}$ 을 모집단내 총 원소수로 나누는 추정량을 사용한다. 모집단 크기 K 는 일반적으로 알려져 있지 않고, K를 추정하게 된다

$$K = \sum_{h=1}^{H} \sum_{j=1}^{N_h} M_{hj} \quad \text{이고} \quad \text{let } M_h = \sum_{j=1}^{N_h} M_{hj}$$

라 하면

$$\hat{M}_h = N_h \bar{M}_h = \frac{N_h}{n_h} \sum_{j=1}^{n_h} M_{hj}.$$

이고

$$\hat{K} = \sum_{h=1}^{H} \hat{M}_h = \sum_{h=1}^{H} N_h \bar{M}_h = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} M_{hj}$$

가 된다

따라서 평균 추정량은

$$\hat{\mu}_c = \frac{\sum_{h=1}^{H} N_h \bar{\tau}_h}{\sum_{h=1}^{H} N_h \bar{M}_h} = \frac{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} \tau_{hj}}{\sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{j=1}^{n_h} M_{hj}} \quad \text{이다.}$$

$$\frac{\sum_{h=1}^{H} \hat{\tau}_h}{\sum_{h=1}^{H} \hat{M}_h}$$

이는 병합비 추정에서 B 에 대한 추정량 b_c 에 해당되는 항이고 이 추정량의 분산도

$$\widehat{var}(\hat{\mu}_c-\mu)=\widehat{var}\left(\frac{\sum_{h=1}^{H}\frac{1}{N}\hat{\tau}_h}{\sum_{h=1}^{H}\frac{1}{N}\hat{M}_h}-\mu\right)$$

$$=\widehat{var}\left\{\frac{1}{\sum_{h=1}^{H}\frac{1}{N}\hat{M}_h}\left(\sum_{h=1}^{H}\frac{1}{N}\left(\hat{\tau}_h-\mu\cdot\hat{M}_h\right)\right)\right\}$$

$$\simeq\frac{1}{\bar{M}^2}\widehat{var}\left\{\sum_{h=1}^{H}\frac{1}{N}\left(\hat{\tau}_h-\mu\hat{M}_h\right)\right\}$$

$$=\frac{1}{\bar{M}^2}\frac{1}{N^2}\sum_{h=1}^{H}\widehat{var}\left(\hat{\tau}_h-\mu\hat{M}_h\right)$$

$$=\frac{1}{\bar{M}^2}\frac{1}{N^2}\sum_{h=1}^{H}N_h^2\frac{N_h-n_h}{N_h}\frac{s_{e.h}^2}{n_h}$$ 이다,

여기서 $s_{e.h}^2=\frac{1}{n_h-1}\sum_{j=1}^{n_h}\left(\tau_{hj}-\hat{\mu}_c M_{hj}\right)^2$ 이다.

6.5. 크기비례 확률 표집.

- Sampling with probabilities proportional to size (PPS)

= 복원 부등 확률 표집

- Sampling weight (표집 비중)

= 표집확률 (sampling probability)의 역수

- 층화추출 vs 크기비례 확률 표집

- 층화 추출에서는 각 층을 새로운 모집단으로 간주하여 층마다 독립적으로 단순임의 비복원 표집을 하기때문에 등확률 표집방법이다.

- 층의 크기가 다르고 표집비중이 같지 않기 때문에 이를 보정하기 위하여 층의 크기를 고려하여 추정값을 계산한다.

$$\hat{t}_{st} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \frac{N_h}{n_h} y_{hj} = \sum_{h=1}^{H} \sum_{j=1}^{n_h} w_h \cdot Y_{hj}.$$

- 반면 PPS에서는 층화를 먼저하지 않고 층 크기를 고려하여 전체 모집단으로부터 부등확률 표집을 한다.

- 일반적으로 부등확률 표집의 경우 selection probability를 알지 못하고 이런현상을 selection bias라 한다.

- 이제 PPS를 살펴보자.

M_i = i 번째 집락내 원소의 수

K = 모집단 총 원소의 수

$z_i = \frac{M_i}{K}$ = 표집 확률

$\Rightarrow$ Sampling weight $= \frac{1}{z_i}$

N 개의 집락이 있고 n개의 집락을 PPS sampling 한다면
(복원추출)

$$\hat{\tau}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{z_i} \tau_i$$

$$= \frac{1}{n} \sum_{i=1}^{n} \frac{K}{M_i} \tau_i = \frac{K}{n} \sum_{i=1}^{n} \frac{\tau_i}{M_i} = \frac{K}{n} \sum_{i=1}^{n} \mu_i$$

$$\hat{\mu}_{pps} = \frac{1}{K} \hat{\tau}_{pps} = \frac{1}{n} \sum_{i=1}^{n} \frac{\tau_i}{M_i} = \frac{1}{n} \sum_{i=1}^{n} \mu_i$$

$$\hat{var}(\hat{\mu}_{pps}) = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} (\mu_i - \hat{\mu}_{pps})^2$$

$$\hat{var}(\hat{\tau}_{pps}) = \frac{K^2}{n} \frac{1}{n-1} \sum_{i=1}^{n} (\mu_i - \hat{\mu}_{pps})^2$$

6.7. 이단 집락 표집

일단 집락 표집에서는 선택된 psu에 속한 모든 ssu들을 조사하였으나. 집락내 원소들을 모두 조사할 필요가 없거나 ssu에 대한 조사 비용이 상대적으로 높으면 이단 집락 표집 (two-stage cluster sampling)을 고려한다. 즉 선택된 psu 에서 부표본 (sub sample)을 얻는다.

이단집락 표집은

① 모집단 N 개의 psu 에서 n 개의 psu 들은 단순임의추출 한다

② 선택된 psu들로 부터 크기가 M_i 인 i 번째 psu 가 선택되었다면 크기 m_i 인 단순 임의 표집을 시행한다.

6.7.1. 모수의 추정

(1) 모합의 추정

① 추정량

일단집락 표집에서 모합의 추정은

$$\hat{t}_u = \frac{N}{n}\sum_{i=1}^{n} t_i$$ 이었으나 이단 집락표집 에서는

집락의 모합 t_i 가 관측이 되지 않으므로 추정하여야 한다.

즉 $$\hat{t}_i = \frac{M_i}{m_i}\sum_{j=1}^{m_i} y_{ij} = M_i \frac{t_i}{m_i} = M_i \bar{Y}_i$$

따라서

$$\hat{\tau}_u = \frac{N}{n}\sum_{i=1}^{n}\hat{t}_i = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} Y_{ij}$$

$$= \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i} t_i = \frac{N}{n}\sum_{i=1}^{n} M_i\bar{Y}_i$$

② $\hat{\tau}_u$의 통계적 성질

$$\hat{\tau}_u = \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i} t_i$$

$$= \frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i}\sum_{j=1}^{m_i} Y_{ij}$$

$$= \frac{N}{n}\sum_{i=1}^{N} U_i \left\{\frac{M_i}{m_i}\sum_{j=1}^{M_i} V_{ij}\, y_{ij}\right\}$$

$U_i = \begin{cases} 1 & \text{if } i \text{ 번째 cluster가 select 된다} \\ 0 & \text{o.w} \end{cases}$

$V_{ij} = \begin{cases} 1 & \text{if } y_{ij} \text{ is selected} \\ 0 & \text{o.w.} \end{cases}$

평균

$$E(\hat{\tau}_u) = E_U\left\{E_V\left[\hat{\tau}_u \mid U\right]\right\}$$

$$= E_U\left\{\frac{N}{n}\sum_{i=1}^{N} U_i \frac{M_i}{\cancel{m_i}}\sum_{j=1}^{M_i}\frac{\cancel{m_i}}{M_i}\cdot y_{ij}\right\}$$

$$= E_U\left\{\frac{N}{n}\sum_{i=1}^{N} U_i \sum_{j=1}^{M_i} y_{ij}\right\}$$

$$= \frac{\cancel{N}}{\cancel{n}} \sum_{i=1}^{N} \frac{\cancel{n}}{\cancel{N}} \sum_{j=1}^{M_i} y_{ij} = \tau.$$

분산.

$$var(\hat{\tau}_w) = var_{\underline{u}}\{E_{\underline{v}}(\hat{\tau}_w \mid \underline{u})\}$$
$$+ E_{\underline{u}}\{var_{\underline{v}}(\hat{\tau}_u \mid \underline{u})\}$$

Let.

$$① = var_{\underline{v}}\{E_{\underline{v}}(\hat{\tau}_u \mid \underline{u})\}$$

$$② = E_{\underline{u}}\{var_{\underline{v}}(\hat{\tau}_w \mid \underline{v})\}$$

먼저 ①을 계산한다.

$$var_{\underline{u}}\{E_{\underline{v}}(\hat{\tau}_u \mid \underline{u})\}$$
$$= var_{\underline{u}}\left\{\frac{N}{n} \sum_{i=1}^{N} U_i \sum_{j=1}^{M_i} y_{ij}\right\}$$
$$= var_{\underline{v}}\left\{\frac{N}{n} \sum_{i=1}^{n} \tau_i\right\} = N^2 \frac{N-n}{N-1} \frac{\sigma_b^2}{n}$$

sample 된 cluster 들에 대한 총합

$$\sigma_b^2 =$$

② 를 계산한다

$$② = E_{\underline{U}}\left\{ var_{\underline{V}}\left(\hat{\tau}_u \mid \underline{U}\right)\right\}$$

$$= E_{\underline{U}}\left\{ \frac{N^2}{n^2} \sum_{i=1}^{N} M_i^2 U_i \; var\left(\frac{1}{m_i}\sum_{j=1}^{m_i} Y_{ij}\right)\right\}$$

$$= E_{\underline{U}}\left\{ \frac{N^2}{n^2} \sum_{i=1}^{N} M_i^2 U_i \frac{M_i - m_i}{M_i - 1} \frac{\sigma_i^2}{m_i}\right\}$$

$$\sigma_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \mu_i\right)^2, \quad \mu_i = \frac{1}{M_i}\tau_i$$

$$= \frac{N^2}{n^2}\sum_{i=1}^{N} M_i^2 \frac{n}{N} \frac{M_i - m_i}{M_i - 1}\frac{\sigma_i^2}{m_i}$$

$$= \frac{N}{n}\sum_{i=1}^{N} M_i^2 \frac{M_i - m_i}{M_i - 1}\frac{\sigma_i^2}{m_i}$$

따라서

$$var(\hat{\tau}_u) = ① + ②$$

$$= N^2\left(\frac{N-n}{N}\right)\frac{\sigma_b^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i^2\left(\frac{M_i - m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}$$

$$\sigma_b^2 = \frac{1}{N-1}\sum_{i=1}^{N}\left(\tau_i - \bar{\tau}\right)^2, \quad \left(\bar{\tau} = \frac{1}{N}\tau, \text{ psu 당 평균}\right)$$

$$\sigma_i^2 = \frac{1}{M_i - 1}\sum_{j=1}^{M_i}\left(y_{ij} - \mu_i\right)^2, \quad \left(\mu_i = \frac{1}{M_i}\tau_i, \text{ i번째 psu의 모평균}\right)$$

이다.

③ 분산의 추정량

분산에 대한 추정량은

$$\widehat{var}(\hat{\tau}_u) = N^2\left(\frac{N-n}{N}\right)\frac{1}{n}s_b^2 + \left(\frac{N}{n}\right)\sum_{i=1}^{N} M_i^2\left(\frac{M_i-m_i}{M_i}\right)\frac{s_i^2}{m_i}$$

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}_i - \bar{t}\right)^2 \qquad t_i = \sum_{j=1}^{m_i} y_{ij}$$

$$\bar{t} = \frac{\hat{\tau}}{N} = \frac{1}{N}\left\{\frac{N}{n}\sum_{i=1}^{n}\hat{\tau}_i\right\}$$

$$\hat{\tau}_i = \frac{M_i}{m_i} t_i$$

(2) 평균에 대한 추정

$$\hat{\mu}_w = \frac{1}{K}\hat{\tau}_w = \frac{1}{K}\frac{N}{n}\sum_{i=1}^{n}\frac{M_i}{m_i} t_i$$

$$\widehat{var}(\hat{\mu}_u) = \frac{1}{K^2}\left\{N^2\left(\frac{N-n}{N}\right)\frac{s_b^2}{n} + \left(\frac{N}{n}\right)\sum_{i=1}^{n} M_i^2\left(\frac{M_i-m_i}{M_i}\right)\frac{s_i^2}{m_i}\right\}$$

(3) 모비율에 대한 추정

a_i = i번째 psu 에서 관심 특성을 보유한 표본의 총 개수 among m_i

A_i = 〃 〃 〃 개체들의 총 개수 among M_i

$$P_i = \frac{A_i}{M_i} \qquad \hat{P}_i = \frac{a_i}{m_i} \qquad \cancel{\bar{p} = \frac{1}{n}\sum_{i=1}^{n}\hat{P}_i}$$

이제

$$\mathbb{P} = \frac{\sum_{i=1}^{N} A_i}{\sum_{i=1}^{N} M_i}$$ 의 불편추정량은 모평균의 불편추정량을 참고하여

$$\hat{P}_u = \frac{1}{K} \frac{N}{n} \sum_{i=1}^{n} \frac{M_i}{m_i} a_i$$

$$= \frac{1}{K} \frac{N}{n} \sum_{i=1}^{n} M_i \hat{p}_i = \frac{1}{n} \sum_{i=1}^{n} \frac{M_i}{\bar{M}} \hat{p}_i$$

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} M_i$$

$$\hat{var}(\hat{P}_u) = \frac{1}{K^2} \left\{ N^2 \left(\frac{N-n}{N}\right) \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^{n} \left(M_i \hat{p}_i - \frac{1}{n} \sum_{i=1}^{n} M_i \hat{p}_i \right)^2 + \left(\frac{N}{n}\right) \sum_{i=1}^{n} M_i^2 \left(\frac{M_i - m_i}{M_i}\right) \cdot \frac{1}{m_i - 1} \hat{p}_i (1-\hat{p}_i) \right\}$$

이다.

(2) 비추정

앞의 평균 추정의 불편추정량은 $K = \sum_{i=1}^{N} M_i$ 를 알고있음을 가정한다. 이를 모르는 경우 K를 추정하여 앞의 불편추정량에 plug-in 하고 이는 비추정량 형태의 추정량을 제공한다.

K 에 대한 추정량은 $\hat{K} = \frac{N}{n}\sum_{i=1}^{n} M_i$ 이고

$$\hat{\mu}_r = \frac{1}{\hat{K}} \cdot \hat{\tau}_u = \frac{\sum_{i=1}^{n} \frac{M_i}{m_i} t_i}{\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \bar{Y}_i}{\sum_{i=1}^{n} M_i}$$ 이다.

이는 크기 N 인 모집단

$$\left\{ (M_1, \hat{\tau}_1), (M_2, \hat{\tau}_2), \cdots\cdots, (M_N, \hat{\tau}_N) \right\}$$

으로부터 크기 n인 표본

$$\left\{ (X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n) \right\}$$ 을 얻고

이로부터

$$\mu = \frac{\sum_{i=1}^{N} \tau_i}{\sum_{i=1}^{N} M_i}$$ 의 비추정량을 생각하면

$$\hat{\mu}_r = \frac{\bar{Y}}{\bar{X}} = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_i}{\frac{1}{n}\sum_{i=1}^{n} M_i} = \frac{\sum_{i=1}^{n} M_i \bar{Y}_i}{\sum_{i=1}^{n} M_i}$$ 이다.

$$var(\hat{\mu}_r) = var\left(\frac{\sum_{i=1}^{n} M_i \bar{Y}_i}{\sum_{i=1}^{n} M_i} \right)$$

$$\approx \frac{1}{\bar{M}^2} \, var\left(\frac{1}{n}\sum_{i=1}^{n} M_i \bar{Y}_i \right)$$

$$= \frac{1}{\bar{M}^2}\frac{1}{N^2} var\left(\frac{N}{n}\sum_{i=1}^{n} M_i \bar{Y}_i\right)$$

$$= \frac{1}{\bar{M}^2}\frac{1}{N^2} var(\hat{\tau}_u)$$

$$= \frac{1}{\bar{M}^2}\frac{1}{N^2}\left\{N^2\left(\frac{N-n}{N}\right)\frac{\sigma_b^2}{n} + \frac{N}{n}\sum_{i=1}^{N} M_i^2\left(\frac{M_i-m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}\right\}$$

$$= \frac{1}{\bar{M}^2}\left\{\left(\frac{N-n}{N}\right)\frac{\sigma_b^2}{n} + \frac{1}{nN}\sum_{i=1}^{N} M_i^2\left(\frac{M_i-m_i}{M_i}\right)\frac{\sigma_i^2}{m_i}\right\}$$

분산의 추정량은

$$\sigma_b^2 \rightarrow s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}_i - M_i\hat{\mu}_r\right)^2$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(M_i\bar{Y}_i - M_i\hat{\mu}_r\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n} M_i^2\left(\bar{Y}_i - \hat{\mu}_r\right)^2$$

$$\sigma_i^2 \rightarrow s_i^2 = \frac{1}{m_i-1}\sum_{j=1}^{m_i}\left(Y_{ij}-\bar{Y}_i\right)^2, \quad i=1,2,\dots,n$$

모비율의 추정

추정량은 $\hat{p}_r = \dfrac{\sum_{i=1}^{n} M_i\hat{p}_i}{\sum_{i=1}^{n} M_i}$ 이고 추정량의 분산에 대한 추정량은

근사적으로

$$\widehat{Var}(\hat{p}_r) = \frac{1}{\bar{M}^2}\left\{\frac{N-n}{N}\frac{s_b^2}{n} + \frac{1}{nN}\sum_{i=1}^{n} M_i^2\frac{M_i-m_i}{M_i}\frac{\hat{p}_i(1-\hat{p}_i)}{m_i-1}\right\}$$

여기서 $s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[M_i\hat{p}_i - M_i\hat{p}_r\right]^2$ 이다.

6.7.3. 동일크기의 부표집 (sub-sample)

특별히

$M_i = M$, $i=1,2,3,\dots,N$

$m_i = m$, $i=1,2,3,\dots,n$

에 대하여 생각해 본다. 사전에 잘 계획된 실험의 경우 위의 조건이 만족될 수 있다.

이 경우

$$\hat{\mu}_u = \frac{1}{K}\hat{\tau}_u$$

$$= \frac{1}{K}\frac{N}{n}\sum_{i=1}^{n}\underbrace{\left(\frac{M_i}{m_i}\right)}_{=\frac{M}{m}} t_i$$

$$= \frac{1}{\cancel{N}\cdot\cancel{M}}\frac{\cancel{N}}{n}\frac{\cancel{M}}{m}\sum_{i=1}^{n} t_i$$

$$= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} Y_{ij} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{1}{m}\sum_{j=1}^{m} Y_{ij}\right\}$$

$$\hat{\mu}_r = \frac{\sum_{i=1}^{n}\cancel{M_i}\bar{Y}_i}{\sum_{i=1}^{n}\cancel{M_i}} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_i = \hat{\mu}_u \text{ 이다}$$

따라서 $\hat{\mu}$ 는 $\hat{\mu}_r = \hat{\mu}_u$

$$\widehat{var}(\hat{\mu}) = \frac{1}{M^2}\,\frac{N-n}{N}\,\frac{S_b^2}{n} + \frac{1}{nN}\,\frac{M-m}{M}\sum_{i=1}^{n}\frac{S_i^2}{m_i}$$

$$S_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{\tau}_i - \frac{1}{n}\sum_{i=1}^{n}\hat{\tau}_i\right)^2$$

$$= \frac{M^2}{n-1}\sum_{i=1}^{n}(\bar{Y}_i - \hat{\mu})^2$$

$\hat{\tau}_i = M\bar{Y}_i$

$$S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2$$

여기에서

$$SSB = \sum_{i=1}^{n}\sum_{j=1}^{m}(\bar{Y}_i - \hat{\mu})^2 \qquad df = n-1$$

$$SSW = \sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2 \qquad df = n(m-1)$$

이라고 정의하면

$$S_b^2 = \frac{M^2}{n-1}\sum_{i=1}^{n}(\bar{Y}_i - \hat{\mu})^2 = \frac{M^2}{n-1}\,\frac{SSB}{m}$$

$$S_i^2 = \frac{1}{m-1}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2$$

그리고 $$\sum_{i=1}^{n} S_i^2 = \frac{1}{m-1}\sum_{i=1}^{n}\sum_{j=1}^{m}(Y_{ij} - \bar{Y}_i)^2$$

$$= \frac{1}{m-1} SSW$$

따라서

$$\widehat{Var}(\hat{\mu}) = \frac{1}{M^2}\frac{N-n}{N}\frac{S_b^2}{n} + \frac{1}{nN}\frac{M-m}{M}\sum_{i=1}^{n}\frac{S_i^2}{m}$$

$$= \frac{1}{\cancel{M^2}}\frac{N-n}{N}\frac{1}{n}\frac{\cancel{M^2}}{n-1}\frac{SSB}{m}$$

$$+ \frac{1}{nN}\frac{M-m}{M}\frac{1}{m}\frac{1}{m-1}SSW$$

$$= \frac{N-n}{N}\cdot\frac{1}{n}\frac{1}{m}MSB + \frac{M-m}{NM}\frac{1}{m}MSW.$$

만일 N이 아주 크면

$$\widehat{Var}(\hat{\mu}) \quad \frac{1}{nm}MSB$$

따라서 MSB 가 작은경우, 집락간 동질적인 경우,

집락추출이 효율적이 된다.