6 집락표집(cluster sampling)

6.1 집락표집 개요

일차표집단위(psu: primary sampling unit), 이차표집단위(ssu : secondary sampling unit)

6.2 집락크기에 대한 표기

y_{ij} : i번째 psu에서 관측된 j번째 ssu의 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
$\bar{\tau}=\tau/N$: psu당 평균합	$\bar{t}=\hat{\tau}/N$:psu당 평균합에 대한 추정량
$\sigma_b^2=\frac{1}{N-1}\sum_{i=1}^{N}(\tau_i-\bar{\tau})^2$: psu 합들에 대한 모분산	$s_b^2=\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\tau}_i-\bar{t})^2$
$\mu=\tau/K$: (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

$m_i = M_i$

6.3.1 동일한 집락크기

집락표집결과 얻게 되는 데이터는 $\{\tau_1, \cdots, \tau_n\}$

부부만 사는 가구들에 대한 소득조사(N : 총가구수, n: 표집가구수, M=2)

y_{ij} : i번째 가구의 j(=1, 2)번째 사람의 소득

- 총소득 $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\tau_i$, $\tau_i = \sum_{j=1}^{2} y_{ij}$

$$\widehat{Var}(\hat{\tau}) = N^2 \frac{s_b^2}{n}\left(\frac{N-n}{N}\right),\ s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \bar{t})^2,$$

$\bar{t} = \hat{\tau}/N$(아파트 단지 전체를 통틀어 기구별 평균 총소득)

- 개입별 평균소득

$$\hat{\mu} = \frac{\hat{\tau}}{K} = \frac{\hat{\tau}}{NM} = \frac{\bar{t}}{M},\qquad \widehat{Var}(\hat{\mu}) = \frac{1}{M^2}\frac{s_b^2}{n}\left(\frac{N-n}{N}\right)$$

예 6.1 아파트 단지에서 부부만 사는 가구들에 대하 소득조사를 한다.

N=100, n=5, M=2

총소득 $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n}\tau_i = \frac{100}{5}(4.3+3.8+2.3+3.2+8.0) = 432$

가구별 평균소득 $\bar{t} = \hat{\tau}/N = 432/100 = 4.32$

$$s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \bar{t})^2 = \frac{1}{5-1}[(4.3-4.32)^2 + \cdots \ + (8.0-4.32)^2] = 4.787$$

1인당 평균소득 $\hat{\mu} = \frac{\hat{\tau}}{K} = \frac{\hat{\tau}}{NM} = \frac{\bar{t}}{M} = \frac{4.32}{2} = 2.16$

$$\widehat{Var}(\hat{\mu}) = \frac{1}{M^2}\frac{s_b^2}{n}\left(\frac{N-n}{N}\right) = \frac{1}{2^2}\frac{4.787}{5}\left(\frac{100-5}{100}\right) = 0.2275$$

6.3.2집락표집의 효율

요인	자유도	제곱합	평균제곱
집락간제곱합(SSB)	$N-1$	SSB=$\sum_{i=1}^{N}\sum_{j=1}^{M}(\mu_i-\mu)^2$	MSB=SSB/(N−1)
집락내제곱합(SSW)	$N(M-1)$	SSW=$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu_i)^2$	
총제곱합(SST)	$NM-1$	SST=$\sum_{i=1}^{N}\sum_{j=1}^{M}(y_{ij}-\mu)^2$	

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

$$Var(\hat{\tau}_{CS}) = N^2\left(\frac{N-n}{N}\right)\frac{M\ MSB}{n}$$

- 집락표집(CS)과 단순임의표집(SRS)비교

$$Var(\hat{\tau}_{SRS}) = (NM)^2\left(\frac{NM-nM}{NM}\right)\frac{\sigma^2}{nM} = N^2\left(\frac{N-n}{N}\right)\frac{M\sigma^2}{n}$$

※ $MSB > \sigma^2 \Leftrightarrow Var(\hat{\tau}_{CS}) > Var(\hat{\tau}_{SRS})$

$$\rho_w = 1 - \frac{M}{M-1}\frac{SSW}{SST}\left(-1/(M-1) \leqq \rho_w \leqq 1\right) \Rightarrow MSB = \frac{N}{N-1}\sigma^2[1+(M-1)\rho_w]$$

$$\frac{Var(\hat{\tau}_{CS})}{Var(\hat{\tau}_{SRS})} = \frac{MSB}{\sigma^2} = \frac{N}{N-1}[1+(M-1)\rho_w] \approx 1+(M-1)\rho_w$$

6.3.3 크기가 다른 집락들

(1) 비편향추정

- $\hat{\tau}_u = \frac{N}{n}\sum_{i=1}^{n}\tau_i,\ \tau_i = \sum_{j=1}^{M_i} y_{ij},\ \widehat{Var}(\hat{\tau}_u) = N^2\frac{s_b^2}{n}\left(\frac{N-n}{N}\right),\ s_b^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\tau_i - \bar{t})^2,\ \bar{t} = \hat{\tau}_u/N$
- $\hat{\mu}_u = \frac{\hat{\tau}_u}{K},\ \widehat{Var}(\hat{\mu}_u) = \frac{1}{K^2}N^2\frac{s_b^2}{n}\left(\frac{N-n}{N}\right) = \frac{1}{\overline{M}_s^2}\frac{s_b^2}{n}\left(\frac{N-n}{N}\right),\ K = \sum_{i=1}^{N} M_i$

(2) 비추정 (집락크기(M_i) 와 집락 합(τ_i)은 밀접한 상관이 있다)

- $\hat{\mu}_r = \frac{\sum_{i=1}^{n}\tau_i}{\sum_{i=1}^{n} M_i},\ \widehat{Var}(\hat{\mu}_r) = \frac{1}{\overline{M}^2}\frac{s_e^2}{n}\left(\frac{N-n}{N}\right) = \frac{1}{\overline{M}_s^2}\frac{s_e^2}{n}\left(\frac{N-n}{N}\right),$

  $\overline{M} = \frac{\sum_{i=1}^{N} M_i}{N} = K/N$(모집단의 평균 집락크기), $\overline{M}_s = \frac{\sum_{i=1}^{n} M_i}{n}$(추정한 평균 집락크기)

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

- $\hat{\tau}_r = K\hat{\mu}_r = \hat{K}\hat{\mu}_r,\ \widehat{Var}(\hat{\tau}_r) = N^2\frac{s_e^2}{n}\left(\frac{N-n}{N}\right),\ \hat{K} = N\overline{M}_s,\ s_e^2 = s_b^2 + \hat{\mu}_r^2 s_M^2 - 2\hat{\mu}_r r s_b s_M$

예6.2 N=500(n=25)가구로 이루어진 아파트 단지에서 가구들에 대한 소득조사를 한다. 아파트 전체의 연간 총소득과 연간 일인당 평균소득을 추정하고자 한다.

① 비편향추정

연간 총소득 $\hat{\tau}_u = \frac{N}{n}\sum_{i=1}^{n}\tau_i = \frac{500}{25}(939) = 18780$

$$\widehat{Var}(\hat{\tau}_u) = N^2\frac{s_b^2}{n}\left(\frac{N-n}{N}\right) = 500^2\frac{84.423}{25}\frac{500-25}{500} = 802018.5$$

일인당 평균소득 $\hat{\mu}_u = \frac{\hat{\tau}_u}{K} = \frac{18780}{3.76(500)} = 9.989$

$$\widehat{Var}(\hat{\mu}_u) = \frac{1}{\overline{M}_s^2}\frac{s_b^2}{n}\left(\frac{N-n}{N}\right) = \frac{1}{3.76^2}\frac{84.423}{25}\frac{500-25}{500} = 0.227$$

② 비추정

$$\hat{\mu}_r = \frac{\sum_{i=1}^{n}\tau_i}{\sum_{i=1}^{n} M_i} = \frac{939}{94} = 9.989$$

$$\widehat{Var}(\hat{\mu}_r) = \frac{1}{\overline{M}_s^2}\frac{s_e^2}{n}\left(\frac{N-n}{N}\right) = \frac{1}{3.76^2}\frac{46.335}{25}\frac{500-25}{500} = 0.125$$

(3) 모비율의 추정

① 집락크기가 같은 경우

$$p_i = a_i/M,\ (i=1,\ \cdots,\ n),\ \hat{p}=\sum_{i=1}^{n} p_i/n,\ \ \widehat{Var}(\hat{p}) = \frac{N-n}{N}\frac{s_p^2}{n},\ \ s_p^2 = \frac{1}{n-1}\sum_{i=1}^{n}(p_i-\hat{p})^2$$

② 집락크기가 다른 경우

$$\hat{p}=\sum_{i=1}^{n} a_i/\sum_{i=1}^{n} M_i,\ \ \widehat{Var}(\hat{p}) = \frac{N-n}{N}\frac{1}{\overline{M}^2}\frac{s_e^2}{n},\quad s_e^2 = \frac{1}{n-1}\sum_{i=1}^{n}(a_i-\hat{p}M_i)^2$$

예6.3 화단을 없애는 방안에 대하여 찬반 여부를 물었다.

$$\hat{p}=\sum_{i=1}^{n} a_i/\sum_{i=1}^{n} M_i = \frac{41}{94} = 0.436$$

$$\widehat{Var}(\hat{p}) = \frac{N-n}{N}\frac{1}{\overline{M}^2}\frac{s_e^2}{n} = \frac{500-25}{500}\frac{1}{3.76^2}\frac{1.3725}{25} = 0.00369$$

모비율에 대한 95%신뢰구간 (0.31, 0.56)

(4) 표본 크기의 결정

- 모평균 추정시 표본크기

$$Var(\hat{\mu}_r) = \frac{1}{\overline{M}^2}\frac{\sigma_e^2}{n}\left(\frac{N-n}{N}\right)$$

$$z_{\alpha/2}\sqrt{Var(\hat{\mu}_r)} = B \Rightarrow n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2},\ \ D = \left(\frac{B\overline{M}}{z_{\alpha/2}}\right)^2$$

- 모합에 대한 비편향추정시 표본 크기

$$Var(\hat{\tau}_u) = N^2\frac{\sigma_b^2}{n}\left(\frac{N-n}{N}\right),\ \ n = \frac{N\sigma_b^2}{ND+\sigma_b^2},\ \ D = \left(\frac{B}{z_{\alpha/2}N}\right)^2$$

- 모합에 대한 비추정시 표본 크기

$$Var(\hat{\tau}_r) = N^2\frac{\sigma_\epsilon^2}{n}\left(\frac{N-n}{N}\right),\ \ n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2},\ \ D = \left(\frac{B}{z_{\alpha/2}N}\right)^2$$

- 모비율에 대한 비추정시 표본 크기 (모평균 추정시 표본크기)

$$n = \frac{N\sigma_\epsilon^2}{ND+\sigma_\epsilon^2},\ \ D = \left(\frac{B\overline{M}}{z_{\alpha/2}}\right)^2$$

예 6.4 몇 달 뒤 다시 똑같은 조사를 해서 주민들의 반응에 변화가 있는지 알려고 한다.

99%신뢰도에서 최대 오차한계 ±3%