5 계통추출 (systematic sampling)

편리성

5.1 계통표집 개요

① 표본크기 n을 결정한다.

② $[N/n]=k$(표집구간 sampling interval)

③ 1이상 k이하 정수 R을 임의 선택

④ $R, R+k, R+2k, \cdots, R+(n-1)k$번째 표집단위들을 추출

- 임의모집단(random population)
- 순서모집단(ordered population), 자기상관모집단(autocorrelated population)
- 주기모집단(periodic population, cyclical population)

5.2 추정과 표본크기의 결정

$N=nk$

$y_{ij},\ i=1,\cdots,k\ ;\ j=1,\cdots,\ n$: i번째 계통표본의 j번째 원소

■ 모평균 μ_{sy}에 대한 추정

$$\hat{\mu}_{sy}=\bar{y}_{sy}=\bar{y}_i,\quad E(\bar{y}_{sy})=\mu=\frac{\sum_{i=1}^{k}\bar{y}_i}{k},\quad Var(\bar{y}_{sy})=\frac{\sum_{i=1}^{k}(\bar{y}_i-\mu)^2}{k}$$

- $Var(\bar{y}_{sy})< Var(\bar{y})$?

$$\sigma^2=\frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\mu)^2$$

분산분석표

요인	자유도	제곱합	평균제곱
집락간제곱합(SSB)	$k-1$	SSB=$\sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{y}_i-\mu)^2$	
집락내제곱합(SSW)	$k(n-1)$	SSW=$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2$	MSW=SSW/$k(n-1)$
총제곱합(SST)	$N-1$	SST=$\sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-\mu)^2$	

$$Var(\bar{y}_{sy})=\sigma^2-SSW/N=\sigma^2-\frac{k(n-1)}{N}\sigma^2_{MSW}=\sigma^2-\frac{N-k}{N}\sigma^2_{MSW}$$

※ $Var(\bar{y}_{sy})< Var(\bar{y})\Leftrightarrow\ \sigma^2_{MSW}>\frac{N}{N-1}\sigma^2\approx\sigma^2$

- 집락내상관계수(intracluster correlation coefficient)

$$Var(\bar{y}_{sy})=\frac{\sum_{i=1}^{k}(\bar{y}_i-\mu)^2}{k}=\frac{\sigma^2}{n}[1+(n-1)\rho_w]$$

$$\rho_w=\frac{E[(y_{ij}-\mu)(y_{ij'}-\mu)]}{E(y_{ij}-\mu)^2},\ (i=1,\ \cdots,k;\ j,j'=1,\cdots,n\ ;j\neq j'),\ -\frac{1}{n-1}\leqq\rho_w\leqq 1$$

※ 계통표집과 단순임의표집과 비교

① $\rho_w = -1/(N-1) \Leftrightarrow \ Var(\bar{y}_{sy}) = Var(\bar{y}_{ran})$

② $\rho_w < 0 \ \Leftrightarrow \ Var(\bar{y}_{sy}) < Var(\bar{y}_{ran})$

③ $\rho_w > 0 \ \Leftrightarrow \ Var(\bar{y}_{sy}) > Var(\bar{y}_{ran})$

• 결정계수 (coefficient of determination : R^2)

$$Var(\bar{y}_{sy}) = SSB/N \ \Rightarrow \ SSB = \frac{SST}{n}[1+(n-1)\rho_w]$$

$$\Rightarrow \ \rho_w = 1 - \frac{n}{n-1}\frac{SSW}{SST} \ (SSB = SST - SSW) \Rightarrow \ \rho_w = \frac{nR^2-1}{n-1} \ (\ R^2 = \frac{SSB}{SST})$$

※ $R^2 < 1/n$ ➡ $\rho_w < 0$ ($R^2 \approx 0$이면 집락평균이 비슷하고 집락내 변동이 크다. 집락 평균들 간에는 동질적, 집락내 단위들은 이질적)

■ 분산의 추정

$\rho_w \approx 0$가정하면 계통표본은 단순임의표집과 흡사

• 모평균 μ에 대한 추정

$$\hat{\mu} = \bar{y}_{sy}, \ \widehat{Var}(\bar{y}_{sy}) \approx \left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

• 모합 $\tau = N\mu$에 대한 추정

$$\hat{\tau}_{sy} = N\bar{y}_{sy}, \ \widehat{Var}(\hat{\tau}_{sy}) \approx N^2\left(\frac{N-n}{N}\right)\frac{s^2}{n}$$

• 모비율 p에 대한 추정

$$\hat{p}_{st} = \bar{y}_{sy} \ , \ \widehat{Var}(\hat{p}_{sy}) \approx \left(\frac{N-n}{N}\right)\frac{\hat{p}_{sy}(1-\hat{p}_{sy})}{n-1}$$

예 5.1 N=162명의 직원을 보유한 사업체에서 작년 한 해 동안 직원별 평균 결근일수르 산정하고자 n=18명 직원들의 신상카드를 계통추출하였다.

$$\hat{\mu} = \bar{y}_{sy} = 4.5, \ \widehat{Var}(\bar{y}_{sy}) \approx \left(\frac{N-n}{N}\right)\frac{s^2}{n} = \left(\frac{162-18}{162}\right)\frac{7.2}{18} = 0.356$$

95%신뢰구간 : $4.5 \pm 2\sqrt{0.356} = (3.3, \ 5.7)$

■ 표본크기의 결정

• 모평균 μ에 대한 추정 : $n=\dfrac{N\sigma^2}{(N-1)D+\sigma^2}$, $D=\left(\dfrac{B}{z_{\alpha/2}}\right)^2$

• 모합 $\tau=N\mu$에 대한 추정 : $n=\dfrac{N\sigma^2}{(N-1)D+\sigma^2}$, $D=\left(\dfrac{B}{Nz_{\alpha/2}}\right)^2$

• 모비율 p에 대한 추정 : $n=\dfrac{N\,p(1-p)}{(N-1)D+p(1-p)}$, $D=\left(\dfrac{B}{z_{\alpha/2}}\right)^2$

예 5.2 지역현안에 대한 찬반 여론조사를 하기 위하여 N=250,000 명이 가구별 전화번호를 기초로 계통표집을 계획하고 있다. 신뢰도 95%에서 ±3% 이내의 오차한계를 유지하려면 k를 얼마로 잡아야 할까 ?

$$n=\frac{N\,p(1-p)}{(N-1)D+p(1-p)}=\frac{250000(0.5)(0.50)}{(250000-1)(0.000225)}=1106.2,\ k=\frac{250000}{1107}=225.8$$

5.3 모집단의 구조

(1) 임의모집단 : 임의순서(random order)로 배열

(2) 순서모집단 : 증가순서(increasing order), 감소순서(decreasing order)

(3) 주기모집단 : 동일 구간별로 주기성(periodicity)

5.4 단일 표본에서 분산추정

5.4.1 상호관입 계통표집(interpenetrating systematic sampling)

반복계통표집(repeated systematic sampling)

예 5.1 $n^{'}=3$인 n_s=6개의 독립적인 계통표본들을 추출, 6개의 계통표집에서 표집구간은 $k^{'}=n_sk=6(9)=54$

$$\hat{\mu}=\sum_{i=1}^{n_s}\bar{y}_i/n_s,\ k^{'}=n_sk$$

$$\widehat{Var}(\hat{\mu})=\left(\frac{N-n}{N}\right)\frac{s_m^2}{n_s}=\left(\frac{k^{'}-n_s}{k^{'}}\right)\frac{s_m^2}{n_s},\ s_m^2=\frac{\sum_{i=1}^{n_s}(\bar{y}_i-\hat{\mu})^2}{n_s-1}$$

예 5.3 $\hat{\mu}=(5.00+\cdots+3.330/6=4.5,\ s_m^2=[(5-4.5)^2+\cdots+(3.33-4.5)^2]/(6-1)=1.9$

$$\widehat{Var}(\hat{\mu})=\left(\frac{N-n}{N}\right)\frac{s_m^2}{n_s}=\left(\frac{54-6}{54}\right)\frac{1.9}{6}=0.2815$$

95%신뢰구간 : $4.5\pm2\sqrt{0.2815}=(3.44,\ 5.56)$

5.4.2 모형을 이용한 추정법

$y_{ij}\equiv y_{(j-1)k+i}$: i번째 집락의 j 번째 단위

가정 $y_l=\mu_l+\epsilon_l$,

μ_l : l에 대한 추세성분(trend component), 오차성분 : $\epsilon_l\sim(0,\sigma_l^2)$

• 모형 1 -임의모집단

$$y_l = \mu + \epsilon_l,\ s_1^2 = \frac{N-n}{N}\frac{1}{n}\frac{\sum_{j=1}^{n}(y_{ij}-\overline{y}_{sy})^2}{n-1}$$

• 모형 2 - 층화효과

$$y_{ir} = \mu_r + \epsilon_{ir},\ s_2^2 = \frac{N-n}{N}\frac{1}{n}\frac{\sum_{j=1}^{n-1}(y_{i,j+1}-y_{ij})^2}{2(n-1)}$$

• 모형 3 - 선형추세

$$y_l = \mu + \beta l + \epsilon_l,\ s_2^2 = \frac{N-n}{N}\frac{1}{n}\frac{\sum_{j=1}^{n-2}(y_{i,j+2}-2y_{i,j+1}+y_{ij})^2}{6(n-2)}$$

예 5.4 N=100인 모집단에서 n=10의 계통표본을 추출한 결과.

$$\hat{\mu} = \overline{y}_{sy} = 14.21,\ s_2^2 = \frac{N-n}{N}\frac{1}{n}\frac{\sum_{j=1}^{n-1}(y_{i,j+1}-y_{ij})^2}{2(n-1)} = \frac{100-10}{100}\frac{1}{10}\frac{192.9}{2(10-1)} = 0.9645$$