[1] 단순임의 추출법(Simple Random Sampling)

1.기본개념:

크기 N인 모집단으로부터 크기 n인 표본을 뽑을 경우, 모든 가능한 크기 n인 표본이 동일하게 선출될 확률$\left[\frac{1}{\binom{N}{n}}\right]$를 갖는 추출법.

2. 모평균(μ), 표본평균($\bar{y}$), 모총합(τ)

모집단: $\{Y_1, \cdots, Y_N\} \Rightarrow$ 모평균: $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$($\tau = \sum_{i=1}^{N} Y_i$: 모총합)

표본: $\{y_1, \cdots, y_n\} \Rightarrow$ 표본평균: $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, 모총합의 추정값: $\hat{\tau} = \frac{N}{n}\sum_{i=1}^{n} y_i = N\bar{y}$

3. 모분산(S^2, σ^2)

$S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \mu)^2$, $\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \mu)^2$

4. $\bar{y}$의 분산($V(\bar{x})$)

$V(\bar{y}) = (1 - \frac{n}{N})\frac{S^2}{n}$ (비복원 추출),

$\frac{N-n}{N} = 1 - \frac{n}{N}$: 분산의 유한모집단수정(finite population correction: fpc)

$\frac{n}{N}$: 추출률(sampling fraction)

5. $V(\bar{y})$의 추정($\hat{V}(\bar{y})$)

표본분산: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$

$\hat{V}(\bar{y}) = (1 - \frac{n}{N})\frac{s^2}{n}$ (비복원 추출),

6. μ에 대한 신뢰구간 : $\bar{y} \pm z_{\alpha/2}\sqrt{\hat{V}(\bar{y})}$

eg) 어떤공장 근로자의 평균월임금(μ)과 총월임금(τ)을 추정(단위:백만원)

$N = 100, n = 5$, 표본:$\{y_1, \cdots, y_5\} = \{3.3, 3.8, 3.6, 2.3, 2.6\}$

$\bar{y} = \frac{1}{5}\sum_{i=1}^{5} y_i = \frac{1}{5}(3.3 + 3.8 + 3.6 + 2.3 + 2.6) = 3.12$ 백만원

$\hat{\tau} = N\bar{y} = 100(3.12) = 312$ 백만원

$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \frac{1}{4}[(3.3-3.12)^2 + (3.8-3.12)^2 + (3.6-3.12)^2 + (2.3-3.12)^2 + (2.6-3.12)^2]$

$= \frac{1}{4}(1.668) = 0.417$

$\hat{V}(\bar{y}) = (1 - \frac{n}{N})\frac{s^2}{n} = (1 - \frac{5}{100})\frac{0.417}{5} = 0.08$

μ의 95% 신뢰구간 $= (\bar{y} \pm 1.96\sqrt{\hat{V}(\bar{y})}) = (3.12 \pm 1.96\sqrt{0.08}) = (3.12 \pm 0.55) = (2.57, 3.67)$

[2] 층화임의추출법

1.기본개념:

모집단을 어떤 층화기준에 의하여 여러 개의 층으로 분할 한 다음 각층에서 독립적으로 일정한 수의 표본을 임의추출하는 방법.

층내는 등질적(等質的, homogeneous), 층간에는 이질적(異質的, heterogeneous)이 되도록.

2. 모평균 & 총합 추정

N: 모집단 크기, L: 층의 갯수

$N_i(i = 1, 2, \cdots, L)$: i층의 크기. $N = \sum_{i=1}^{L} N_i$

τ :모집단 총합, τ_i: i층의 총합

$\mu = \frac{\tau}{N}$모집단 평균, $\mu_i = \frac{\tau_i}{N_i}$: i층의 평균

n_i: i층에서 추출된 표본의 크기, 전체표본의 크기(n)= $\sum_{i=1}^{L} n_i$

y_{ij}: i층에서 추출된 표본 중 j번째 원소의 값

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: i층에서 추출된 표본평균($\hat{\mu}_i$)

$\hat{\tau}_i = N_i\bar{y}_i$: i층의 총합 추정값

$\hat{\tau}= \sum_{i=1}^{3} N_i\bar{y}_i$: 총합 추정값

$\bar{y}_{st} = \frac{\hat{\tau}}{N}$: 층화표본평균 or $\bar{y}_{st} = \frac{1}{n}\sum_{i=1}^{L} n_i\bar{y}_i$

eg) A잡지의 2월 평균판매부수(μ)와 총판매부수(τ)을 추정
(전국 서점수(N) 5000개라고 가정)
L=3, 대형-중형-소형서점
$N_1 = 500, N_2 = 1500, N_3 = 3000$
→ $n_1 = 50, n_2 = 100, n_3 = 150$ (대형서점 500개에서 50개 추출,...)
y_{1j} : 50개 추출된 대형서점중 j번째 대형서점의 판매부수, $j = 1,\cdots,n_1(50)$

$\bar{y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1j}$: 50개 추출된 대형서점(1번째층)의 평균 판매부수

$\hat{\tau}_1 = N_1\bar{y}_1$: 대형서점(1번째층)의 판매부수의 총합 추정값

→ $\hat{\tau}= \hat{\tau}_1 + \hat{\tau}_2 + \hat{\tau}_3$: 판매부수의 총합 추정값

평균판매부수(μ)의 추정값($\hat{\mu}$): $\bar{y}_{st} = \frac{\hat{\tau}}{N}$

층	N_i	n_i	$\bar{y}_i$	s_i
1(대)	500	50	40	5
2(중)	1500	100	20	2
3(소)	3000	150	5	2
합계	5000	300	-	-

$\hat{\tau}_1 = N_1\bar{y}_1 = 500(40) = 20000,\ \hat{\tau}_2 = N_2\bar{y}_2 = 1500(20) = 30000, \hat{\tau}_3 = N_3\bar{y}_3 = 3000(5) = 15000$

$$\hat{\tau} = \sum_{i=1}^{L}\hat{\tau}_i = 20000 + 30000) + 15000 = 65000$$

$$\bar{y}_{st} = \frac{\hat{\tau}}{N} = \frac{65000}{5000} = 13$$

3. 표본배분

표본크기 n을 먼저 결정한 다음 각 층에 $n_1, n_2, \cdots, n_L$를 배분.

1) 비례배분

비례배분: $n_i = \frac{N_i}{N}n$

2) 최적배분: 표본배분시 각 층별 조사단위의 조사비용이 달라 조사비용까지 고려.

3) 네이만배분: 각층마다 조사비용이 일정할 때 최적배분

층화추출의 효율:

만약 $1/N_i$을 무시한다면 동일한 표본크기 n에 대한 단순임의추출, 층화비례추출, 네이만배분의 관계는 다음과 같다.

$$V(\bar{y}_{ney}) \le V(\bar{y}_{prop}) \le V(\bar{y}_{srs})$$

[3] 집락추출법

1.기본개념:

모집단을 기본단위들로 묶은 집락(cluster)으로 구성하고, 모집단에 집락을 1차추출단위로 임의추출하고, 추출된 각 집락에서 표본을 임의추출하는 방법. **집락내에서는 이질적, 집락간은 등질적**. 집락=PSU(1차추출단위), 실제조사단위=SSU(2차추출단위)

2. 모평균 & 총합 추정

N: 모집단크기

M: 모집단내 집락의 수, m: 표본집락의 수

N_i: i번째(로 추출된) 집락의 원소수, n_i: i번째 집락에서 추출된 표본의 원소수

y_{ij}: i번째 집락에서 추출된 표본의 j번째 원소의 값

$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$: i번째 집락에서 추출된 표본의 평균

$\hat{\tau}_i = N_i\bar{y}_i$ = 표본으로 뽑힌 i번째 집락의 총합의 추정값

~ $\sum_{i=1}^{m}\hat{\tau}_i$= 표본으로 뽑힌 집락들의 총합의 추정값 (m개)

$\hat{\tau} = \frac{M}{m}\sum_{i=1}^{m}\hat{\tau}_i$: 총합의 추정값

$\bar{y}_{cl} = \hat{\tau}/N$: 모평균의 추정값(N known)

$\bar{y}_{cl} = \frac{\sum_{i=1}^{m}\hat{\tau}_i}{\sum_{i=1}^{m}N_i}$: 모평균의 추정값(N unknown)

eg) B시의 25개 구역 중 3개의 구역을 단순임의 추출(M=25, m=3).
추출된 지역(1~3), N_i 가구에서 n_i가구를 추출해서 각가구의 작년도 수입액(백만) 조사

구역	N_i	n_i	$\bar{y}_i$	s_i
1	40	4	45	2
2	50	4	50	3
3	25	3	40	2.5

$\bar{y}_i$: n_i가구의 평균 수입액 보유수

$\hat{\tau}_i = N_i\bar{y}_i$: i번째로 추출된 구역 전체가구의 총수입액 추정값

$\hat{\tau}_1 = N_1\bar{y}_1 = 40(45) = 1800,\ \hat{\tau}_2 = N_2\bar{y}_2 = 50(50) = 2500,\ \hat{\tau}_3 = N_3\bar{y}_3 = 25(40) = 1000$

$\sum_{i=1}^{m}\hat{\tau}_i$: 추출된 3(m)개 구역 전체가구의 총수입액 추정값

$\hat{\tau} = \frac{M}{m}\sum_{i=1}^{m}\hat{\tau}_i = \frac{25}{3}\left[1800+2500+1000\right] = \frac{25}{3}5300 = 44167$백만원:B시 전체 가구의 총수입액

if (N: 전체가구수)=1000 known: $\bar{y}_{cl} = \hat{\tau}/N$ =44167/1000 = 44.2백만원

if (N: 전체가구수)=unknown: $\bar{y}_{cl} = \frac{\sum_{i=1}^{m}\hat{\tau}_i}{\sum_{i=1}^{m}N_i} = \frac{53000}{40+50+25} = 46.1$백만원