

제 3장 층화임의표집(stratified random sampling)

3.2 층화표본의 추출과 표기

	층 1	... 층 h	... 층 H	모집단
층 크기	N	N_h	N_H	N : 모집단 크기
층 표본 크기	n_1	n_h	n_H	n : 층 표본 크기
표본	y_{11}	y_{h1}	y_{H1}	
	⋮	⋮	⋮	
	y_{1n_1}	y_{hn_h}	y_{Hn_H}	모수
	—	—	—	
표본층평균	\bar{y}_1	\bar{y}_h	\bar{y}_H	μ
표본층합	t_1	t_h	t_H	τ
표본층분산	s_1^2	s_h^2	s_H^2	σ^2
표본층비율	\hat{P}_1	\hat{P}_h	\hat{P}_H	P

모집단 특성에 대한 표기	
H	층의 개수
N_h	h번째 층의 크기
y_{hj}	층 h의 j번째 단위 ($h = 1, \dots, H; j = 1, \dots, N_h$)
$W_h = N_h/N$	층 h의 층비중(stratum weight)
$\tau_h = \sum_{j=1}^{N_h} y_{hj}$	층 h의 모합
$\mu_h = \tau_h/N_h$	층 h의 모평균
$p_h = \tau_h/N_h$	층 h의 모비율
$\sigma_h^2 = \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2 / N_h$	층 h의 모분산
$\tau = \sum_{h=1}^H \tau_h$	모합
$\mu = \tau/N = \sum_{h=1}^H \tau_h/N = \sum_{h=1}^H \frac{N_h}{N} \mu_h = \sum_{h=1}^H W_h \mu_h$	모평균
$p = \sum_{h=1}^H W_h p_h$	모비율
$\sigma^2 = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 / N$	모분산

표본 특성에 대한 표기	
y_{hj}	층 h의 j번째 단위 ($h = 1, \dots, H; j = 1, \dots, n_h$)
n_h	h번째 층의 표본크기
$n = \sum_{h=1}^H n_h$	총표본의 수
$f_h(w_h) = n_h/n$	층 h의 표집률(층 배정률)
$\bar{y}_h = \sum_{j=1}^{n_h} y_{hj}/n_h$	층 h의 표본평균
$t_h = n_h \bar{y}_h$	층 h의 표본합
$s_h^2 = \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2 / (n_h - 1)$	층 h의 표본분산

층화표집				
층 1 (N_1) \rightarrow n_1	...	층 h (N_h) \rightarrow n_h	...	층 H (N_H) \rightarrow n_H
y_{11}, \dots, y_{1n_1}	...	y_{h1}, \dots, y_{hn_h}	...	y_{H1}, \dots, y_{Hn_H}
$\bar{y}_1, t_1, s_1^2, \hat{p}_1$...	$\bar{y}_h, t_h, s_h^2, \hat{p}_h$...	$\bar{y}_H, t_H, s_H^2, \hat{p}_H$

3.3 추정

3.3.1 모합과 모평균의 추정

$$\hat{\tau}_h = N_h \bar{y}_h$$

$$\square \hat{\tau}_{st} = \sum_{h=1}^H \hat{\tau}_h = \sum_{h=1}^H N_h \bar{y}_h, \quad \widehat{Var}(\hat{\tau}_{st}) = \sum_{h=1}^H N_h^2 Var(\bar{y}_h) = \sum_{h=1}^H N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h}$$

$$\square \hat{\mu} \equiv \bar{y}_{st} = \hat{\tau}/N = \sum_{h=1}^H \frac{N_h}{N} \bar{y}_h = \sum_{h=1}^H W_h \bar{y}_h \text{ (표본층평균들의 가중평균)}$$

$$\widehat{Var}(\bar{y}_{st}) = \widehat{Var}\left(\sum_{h=1}^H \frac{N_h}{N} \bar{y}_h\right) = \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \widehat{Var}(\bar{y}_h) = \sum_{h=1}^H \left(\frac{N_h}{N}\right)^2 \left(\frac{N_h - n_h}{N_h}\right) \frac{s_h^2}{n_h}$$

□ 모평균에 대한 95%신뢰구간

$$\bar{y}_{st} \pm 2s\hat{e}(\bar{y}_{st})$$

예 3.1 한 학기 동안 읽은 책의 평균 권수 조사

층	N_h	n_h	\bar{y}_h	s_h^2
남학생	212	42	20.2	230.7
여학생	88	18	30.5	40.3
합계	300	60		

$$\square \bar{y}_{st} = \frac{N_1 \bar{y}_1 + N_2 \bar{y}_2}{N} = 23.2$$

$$\square \widehat{Var}(\bar{y}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h} = 2.35$$

$$\square \bar{y}_{st} \pm 2s\widehat{e}(\bar{y}_{st}) = (23.2 \pm 2\sqrt{2.35}) = (20.1, 26.3)$$

□ 여학생들에 한하여 한 학기 동안 읽은 책의 평균 권수에 대한 95%신뢰구간?

$$\bar{y}_2 \pm 2\sqrt{\frac{s_2^2}{n_2} \left(\frac{N_2 - n_2}{N_2} \right)} = (27.8, 33.2)$$

3.3.2 모비율의 추정

$$\square \hat{p}_{st} = \sum_{h=1}^H \frac{N_h}{N} \hat{p}_h, \quad s_h^2 = \frac{n_h}{n_h - 1} \hat{p}_h(1 - \hat{p}_h)$$

$$\square \widehat{Var}(\hat{p}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}$$

□ 모비율에 대한 95% 신뢰구간

$$\bar{p}_{st} \pm 2s\hat{e}(\bar{p}_{st})$$

$$\square \text{총수 } \hat{t} = N \hat{p}_{st}, \quad \widehat{Var}(\hat{t}) = N^2 \widehat{Var}(\hat{p}_{st}) = N^2 \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1}$$

예 3.2 통계인으로 자부심을 갖고 있는냐? “그렇다”

층	N_h	n_h	\hat{p}_h
남학생	212	42	0.83(35)
여학생	88	18	0.33(6)
합계	300	60	

$$\square \hat{p}_{st} = \frac{N_1 \hat{p}_1 + N_2 \hat{p}_2}{N} = 0.68$$

$$\square \widehat{Var}(\hat{p}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{\hat{p}_h(1 - \hat{p}_h)}{n_h - 1} = 0.0023$$

$$\square \bar{p}_{st} \pm 2s\hat{e}(\bar{p}_{st}) = (0.63, \quad 0.73)$$

3.3.3 층내분산(within-strata variance)과 층간분산(between-strata variance)

$$\sigma^2 = \frac{1}{N} \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2 = \sum_{h=1}^H \frac{N_h}{N} \sigma_h^2 + \sum_{h=1}^H \frac{N_h}{N} (\mu_h - \mu)^2$$

$$\equiv \sigma_w^2 + \sigma_b^2$$

분산분석표(비례배정 $n_h/N_h = n/N$)

변동요인	자유도	제곱합
층간	$H - 1$	$SSB = \sum_{h=1}^H \sum_{j=1}^{N_h} (\mu_h - \mu)^2$
층내	$N - H$	$SSW = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu_h)^2$
계	$N - 1$	$SST = \sum_{h=1}^H \sum_{j=1}^{N_h} (y_{hj} - \mu)^2$

□ $SSB > \sum_{h=1}^H \left(\frac{N - N_h}{N} \right) \sigma_h^2 \rightarrow$ 층화표집이 단순임의표집보다 효율적

예 3.3, 3.4 층화표집과 단순임의표집의 효율성 비교($N=108, n=18, s^2 = 77.04$)

()

층	1	2	3	층	1	2	3
N_h	36	48	24	N_h	36	48	24
n_h	6	8	4	n_h	6	8	4
y_{hj}				y_{hj}			
\bar{y}_h	50.0	51.0	50.0	\bar{y}_h	41.0	49.75	62.75
s_h^2	89.2	71.1	66.7	s_h^2	14.4	9.07	12.92

$$\hat{\mu} \equiv \bar{y}_{st} = \sum_{h=1}^H W_h \bar{y}_h = 49.72(50.44)$$

$$\widehat{Var}(\bar{y}_{st}) = \sum_{h=1}^H \left(\frac{N_h}{N} \right)^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{s_h^2}{n_h} = 0.54(3.53)$$

$$\widehat{Var}(\bar{Y}) = \frac{s^2}{n} \left(\frac{N - n}{N} \right) = 3.57(3.13)$$

추정분산비=층화표집의 추정분산/단순임의표집의 추정분산=0.54/3.57=0.15(1.127)