## 3.4 표본배정

$n, n_h$를 배정해야한다.

### 3.4.1 모평균의 추정

$$z_{\alpha/2}\sqrt{Var(\bar{y}_{st})} = B \;\Rightarrow\; Var(\bar{y}_{st}) = \left(\frac{B}{z_{\alpha/2}}\right)^2 \equiv D$$

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

$$= \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{N_h - nw_h}{N_h}\right)\frac{\sigma_h^2}{nw_h}(w_h = n_h/n)$$

$$\Rightarrow \quad n = \frac{\sum_{h=1}^{H} N_h^2\sigma_h^2/w_h}{N^2D + \sum_{h=1}^{H} N_h\sigma_h^2}$$

□ 층 배정률의 결정방법 ?

▪ 비용함수( cost function) $C = c_0 + \sum_{h=1}^{H} c_h n_h = \sum_{h=1}^{H} c_h n_h$(가정 : $c_0 = 0$)

※ 비용조건을 만족하고 $Var(\bar{y}_{st})$가 최소가 되도록 결정

▪ 라그랑즈 승수법( Lagrangian multiplier method)

$$f(n_1, \cdots, n_H) = \sum_{h=1}^{H}\left(\frac{N_h}{N}\right)^2\left(\frac{\sigma_h^2}{n_h}\right)\left(\frac{N_h - n_h}{N_h}\right) = Var(\bar{y}_{st})$$

$$g(n_1, \cdots, n_H) = C - \sum_{h=1}^{H} c_h n_h = 0$$

$$L = Var(\bar{y}_{st}) - \lambda(C - \sum_{h=1}^{H} c_h n_h)$$

$$\frac{\partial L}{\partial n_h} = 0 \quad \Leftrightarrow \quad -\frac{N_h^2}{N^2}\frac{\sigma_h^2}{n_h^2} + \lambda c_h = 0 \;\Rightarrow\; n_h = \frac{1}{\sqrt{\lambda}}\frac{N_h\sigma_h}{N\sqrt{c_h}}$$

$$\sum_{h=1}^{H} n_h = n \Rightarrow \quad \sqrt{\lambda} = \frac{1}{nN}\sum_{h=1}^{H}\frac{N_h\sigma_h}{\sqrt{c_h}}$$

$$n_h = \left(\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}}\right)n \quad \left(\text{즉,} \quad n_h \propto \frac{N_h\sigma_h}{\sqrt{c_h}}\right),$$ 층이 크거나, 층내분산이 크거나, 관측비용이 적게 들면 그 층의 표본크기를 늘리는 것이 좋다는 의미.

요인은 층크기, 층내변동, 층별 관측비용

▪ 최적배정(optimal allocation)

$$n = \frac{\left(\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h\sigma_h\sqrt{c_h}\right)}{N^2D + \sum_{h=1}^{H} N_h\sigma_h^2}$$

▪ 네이만 배정(Neyman allocation) ($c_h = c$, 층별단위 관측비용을 전부 같게 놓는 경우)

$$n_h = \left( \frac{N_h \sigma_h}{\sum_{h=1}^{H} N_h \sigma_h} \right) n \quad , \quad n = \frac{\left( \sum_{h=1}^{H} N_h \sigma_h \right)^2}{N^2 D + \sum_{h=1}^{H} N_h \sigma_h^2}$$

▪ 비례배정(proportional allocation)($n_h \propto N_h$)

$$n_h = \left( \frac{N_h}{N} \right) n, \; n = \frac{\sum_{h=1}^{H} N_h \sigma_h^2}{ND + \sum_{h=1}^{H} N_h \sigma_h^2 / N}$$

▪ 동등배정(equal allocation)

$n_h = n/H$

예 3.5 남녀별 층화표집으로 학생들이 한 학기동안 읽은 책의 평균 권수를 조사하려고 한다. 최대 오차 ±2로 표본을 최적배정하라.

| 층 | $N_h$ | $\sigma_h$ | $c_h$ |
|---|---|---|---|
| 남학생 | 212 | 10 | 1 |
| 여학생 | 88 | 5 | 2 |
| 합계 | 300 | | |

| 항목 | 남학생 | 여학생 | 합계 |
|---|---|---|---|
| $N_h\sigma_h/\sqrt{c_h}$ | 2120 | 311.13 | 2431.13 |
| $N_h\sigma_h\sqrt{c_h}$ | 2120 | 622.25 | 2742.25 |
| $N_h\sigma_h^2$ | 21200 | 2200 | 23400 |
| $w_h$ | 0.872 | 0.128 | 1 |

$$Var(\bar{y}_{st}) = D = \left(\frac{B}{z_{\alpha/2}}\right)^2 = 1$$

$$n = \frac{\left(\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}\right)\left(\sum_{h=1}^{H} N_h\sigma_h\sqrt{c_h}\right)}{N^2D + \sum_{h=1}^{H} N_h\sigma_h^2} = 58.8 \approx 59$$

$$n_h = \left(\frac{N_h\sigma_h/\sqrt{c_h}}{\sum_{h=1}^{H} N_h\sigma_h/\sqrt{c_h}}\right)n \qquad n_1 = 0.872\times 59 = 51.3,\ n_2 = 0.128\times 59 = 7.53$$

예 3.6 조사비용이 50만원밖에 없다.

$$C = c_0 + \sum_{h=1}^{H} c_h n_h = \sum_{h=1}^{H} c_h n_h \text{(가정 : } c_0 = 0)$$

$$c_1n_1 + c_2n_2 = 1n_1 + 2n_2 = 50,\ nw_1 + 2nw_2 = (0.872)n + 2(0.128)n = 50\ (n_h = nw_h)$$

$$n = 44.3 \approx 44,\ n_1 = 44(0.872) = 38.4,\ n_2 = 44(0.128) = 5.6$$

예 3.7 (비례배정)

$$n = \frac{\sum_{h=1}^{H} N_h\sigma_h^2}{ND + \sum_{h=1}^{H} N_h\sigma_h^2/N} = \frac{23400}{300(1) + \frac{1}{300}(23400)} = 61.9$$

$$n_h = \left(\frac{N_h}{N}\right)n,\ n_1 = 62\times\left(\frac{212}{300}\right) = 43.8, n_2 = 62\times\left(\frac{88}{300}\right) = 18.2$$

### 3.4.2 모합의 추정

$$D=\left(\frac{B}{z_{\alpha/2}N}\right)^2$$

### 3.4.3 모비율의 추정

□ $n=\dfrac{\sum_{h=1}^{H}N_h^2\sigma_h^2/w_h}{N^2D+\sum_{h=1}^{H}N_h\sigma_h^2}$, $(\sigma_h=p_h(1-p_h))$ ➡ $n=\dfrac{\sum_{h=1}^{H}N_h^2\,p_h(1-p_h)/w_h}{N^2D+\sum_{h=1}^{H}N_h\,p_h(1-p_h)}$ $(D=B^2/4)$

▪ 최적배정(optimal allocation)

$$n_h=\left(\frac{N_h\sqrt{p_h(1-p_h)/c_h}}{\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)/c_h}}\right)n\ ,\ n=\frac{\left(\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)/c_h}\right)\left(\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)c_h}\right)}{N^2D+\sum_{h=1}^{H}N_hp_h(1-p_h)}$$

▪ 네이만 배정(Neyman allocation) $(c_h=c)$

$$n_h=\left(\frac{N_h\sqrt{p_h(1-p_h)}}{\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)}}\right)n\ ,\quad n=\frac{\left(\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)}\right)^2}{N^2D+\sum_{h=1}^{H}N_h\,p_h(1-p_h)}$$

▪ 비례배정(proportional allocation)$(n_h\propto N_h)$

$$n_h=\left(\frac{N_h}{N}\right)n,\ n=\frac{\sum_{h=1}^{H}N_h\,p_h(1-p_h)}{ND+\sum_{h=1}^{H}N_h\,p_h(1-p_h)/N}\leqq\frac{N}{4ND+1}$$

예 3.8 통계인으로 자부심을 갖고 있느냐? "그렇다". 과거조사에서 $\hat{p}_1=0.80,\ \hat{p}_2=0.25$였다. 최대오차한계를 $\pm5\%$로 유지하고 싶다. 여학생 조사는 남학생보다 2배의 조사 비용이 든다. 총표본수를 결정하고 층별 표본배정을 하자.

| 항목 | 남학생 | 여학생 | 합계 |
|---|---|---|---|
| $N_h$ | 212 | 88 | 300 |
| $N_h\sqrt{p_h(1-p_h)/c_h}$ | 84.8 | 26.94 | 111.74 |
| $N_h\sqrt{p_h(1-p_h)c_h}$ | 84.8 | 53.89 | 138.69 |
| $N_hp_h(1-p_h)$ | 33.92 | 16.5 | 50.42 |
| $w_h$ | 0.76 | 0.24 | 1 |

$D=B^2/4=0.05^2/4=0.000625$

$$n=\frac{\left(\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)/c_h}\right)\left(\sum_{h=1}^{H}N_h\sqrt{p_h(1-p_h)c_h}\right)}{N^2D+\sum_{h=1}^{H}N_hp_h(1-p_h)}=145.3$$

$n_1=146(0.76)=111,\ n_2=146(0.24)=35$

예 3.9 조사비용 동일, 모분산 정보 없고, 비례배정시 총표본수

$$n = \frac{N}{4ND+1} = \frac{300}{4*300*0.000625+1} = 172$$

$$n_1 = 172\left(\frac{212}{300}\right) = 121.55, n_2 = 172\left(\frac{88}{300}\right) = 50$$

### 3.5 층의 결정

누적 $\sqrt{f}$ 측도 상에서 등간격 구간, $H$는 5개 이하

### 3.6 표본선택 후의 층화-사후층화(poststratification)

가정: $N_h/N$을 아는 경우

$\bar{y}_{post} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_h$, $\widehat{Var}(\bar{y}_{post}) \approx \left(1 - \frac{n}{N}\right) \sum_{h=1}^{H} \frac{N_h}{N} \frac{s_h^2}{n} + \frac{1}{n^2} \sum_{h=1}^{H} \left(1 - \frac{N_i}{N}\right) s_h^2$ ($n$이 충분히 크고, $n_h \geq 30$)

### 3.7 층화를 위한 이중표집(double sampling, 이상표집(two-phase sampling))

$W_h = \frac{N_h}{N}$ : 층 h에 떨어진 모집단의 비율

$w_h = \frac{n_h^{'}}{n^{'}}$ : 층 h에 떨어진 첫 번째 표본의 비율($n_h$ : 확률변수)

▼ $n^{'}$을 얼마 ?, $n_h$들을 어떻게 결정해야하는지 ? ( Cochran (1977))

▼ 층화용 이중표집(double sampling for stratification)

첫 번째 표집의 목적은 모집단 층비중을 추정, 두 번째 표집의 목적은 층별 모수추정

▪ 모평균 $\mu = \sum_{h=1}^{H} W_h \mu_h$

$\bar{y}_{st} = \sum_{h=1}^{H} w_h \bar{y}_h$, $\widehat{Var}(\bar{y}_{st}) \approx \sum_{h=1}^{H} \left[ \frac{w_h^2 s_h^2}{n_h} + \frac{w_h (\bar{y}_h - \bar{y}_{st})^2}{n^{'}} \right]$

▪ 모비율 $p = \sum_{h=1}^{H} W_h p_h$

$\hat{p}_{st} = \sum_{h=1}^{H} w_h \hat{p}_h$, $\widehat{Var}(\hat{p}_{st}) \approx \sum_{h=1}^{H} \left[ \frac{w_h^2 \hat{p}_h (1 - \hat{p}_h)}{n_h - 1} + \frac{w_h (\hat{p}_h - \hat{p}_{st})^2}{n^{'}} \right]$

($\bar{y}_h = \hat{p}_h, s_h^2 = n_h \hat{p}_h (1 - \hat{p}_h)/(n_h - 1)$)

예 3.10 이 지역에서 불우이웃 돕기 성금을 내는 가구들의 비율을 추정하려고 한다. $n^{'} = 374$ 세대주를 단순임의 추출.

| 종교 | 성금 0 | 성금X | 합계 $n = 92$ $n^{'} = 374$ |
|---|---|---|---|
| 종교신자 | 43 | 31 | $n_1 = 74$ $n_1^{'} = 292$ $w_1 = 292/374 = 0.78$ $\hat{p}_1 = 43/74 = 0.58$ |
| 무교 | 14 | 4 | $n_2 = 18$ $n_2^{'} = 82$ $w_2 = 82/374 = 0.22$ $\hat{p}_2 = 14/18 = 0.78$ |

$\hat{p}_{st} = \sum_{h=1}^{H} w_h \hat{p}_h = 0.624$

$\widehat{Var}(\hat{p}_{st}) \approx \sum_{h=1}^{H} \left[ \frac{w_h^2 \hat{p}_h (1 - \hat{p}_h)}{n_h - 1} + \frac{w_h (\hat{p}_h - \hat{p}_{st})^2}{n^{'}} \right] = 0.00252$

95% 신뢰구간 : $0.624 \pm 2\sqrt{0.00252} = 0.62 \pm 0.1$