STAT 344 MT Review

1. Target population : "All"
2. sampling : 주로 쓰인 sampling method 명칭 적기 — $\bar{y}$ / total
   - simple random sample (SRS)
   - stratified sampling
3. observational unit : individual "i"
4. variable of interest : Response variable
   - Binary variable — $\hat{p} \to p$
   - numeric variable — $\bar{x}_s \to \bar{x}_p$, $\hat{t}_s \to t_p$
5. confidence interval

(estimator) ± critical value × S.E. (estimator)

$z_{1-\frac{\alpha}{2}}$

For mean, $\bar{y}_s \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{s_s^2}{n}}$

($\bar{y}_s$: sample mean; critical value: $z_{1-\frac{\alpha}{2}}$; s_s^2: sample variance; n: sample size; $\sqrt{\frac{s_s^2}{n}} = SE(\bar{y}_s)$)

ME (Margin of Error)

For proportion, $\hat{p}_s \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n}}$

($\hat{p}_s$: sample proportion; $z_{1-\frac{\alpha}{2}}$: critical value; $\sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n}} = SE(\hat{p}_s)$; ME)

(i) confidence interval width / length = 2 ME

(ii) fixed confidence level $c\%\ (1-\alpha\%)$, $z_{1-\frac{\alpha}{2}}$ fixed

→ sample size n ↑, S.E. (estimator) ↓, ME ↓, width ↓

(iii) fixed sample size n,

→ confidence level c% ↑ (90% → 95%)

→ $z_{1-\frac{\alpha}{2}}$ ↑, ME ↑, width ↑

(iv) ME or width → find / determine n

↓ within − error

population SD (population guessed SD)

For mean, $ME = z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq ME$

$$\therefore n \geq \frac{\left(z_{1-\frac{\alpha}{2}}\,\sigma\right)^2}{ME^2}$$

round up e.g. $n \geq 22.3$, $n \geq 23$

For proportion, $z_{1-\frac{\alpha}{2}} \sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n}} \leq ME$

$\hat{p}_s$: guessed proportion

If no info given

worst case: $\hat{p}_s = 0.5$

$\hat{p}(1-\hat{p})$ 值最大

$$n \geq \frac{\left(z_{1-\frac{\alpha}{2}}\right)^2 \hat{p}_s(1-\hat{p}_s)}{ME^2}$$

round up

b.	population		Sample	
	Numeric	Binary	Numeric	Binary
size	N	N	n	n
mean	$\mu(\bar{y}_p)$	p_p	$\bar{y}_s$	$\hat{p}_s$
sd	σ	$p_p(1-p_p)$	s	$\hat{p}_s(1-\hat{p}_s)$

$E(\bar{y}_s) = \bar{y}_p$

$Var(\bar{y}_s) = \frac{\sigma^2}{n}$ $\quad SD(\bar{y}_s) = \frac{\sigma}{\sqrt{n}}$

if σ^2 unknown, s_s^2 replace
sample variance

$$\text{var}(\bar{y}_s) = \frac{s_s^2}{n} \qquad \text{SD}(\bar{y}_s) = \frac{s_s}{\sqrt{n}}$$

$$E(\hat{p}_s) = p_p$$

$$\text{Var}(\hat{p}_s) = \frac{p_p(1-p_p)}{n} \qquad \text{SD}(\hat{p}_s) = \sqrt{\frac{p_p(1-p_p)}{n}}$$

p_p unknown, $\hat{p}_s$ replace: $\text{SE}(\hat{p}_s) = \sqrt{\frac{\hat{p}_s(1-\hat{p}_s)}{n}}$

Total: $t_p = \sum_{i=1}^{N} y_i \qquad \hat{t}_p = N\bar{y}_s$

$$\text{Var}(\hat{t}_p) = N^2 \text{var}(\bar{y}_s) = N^2\left(\frac{s_s^2}{n}\right)$$

$$\text{SE}(\hat{t}_p) = N\sqrt{\frac{s_s^2}{n}}$$

Part I : Simple Random Sample (SRS)

↳ each unit in population selected as a sample with same probability $= \frac{1}{\binom{N}{n}}$

↳ total ways to collect sample $= \binom{N}{n}$

① Estimate $\bar{y}_p$ [population mean]

$\bar{y}_s$ estimate $\bar{y}_p$

(i) $\bar{y}_s = \frac{\sum_{i=1}^{n} y_i}{n}$

(ii) $se(\bar{y}_s) = \sqrt{\frac{s_s^2}{n}}$ sample variance $s_s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_s)^2}{n-1}$

② Estimate P_p (population proportion)

(i) $\hat{P}_s = \frac{\text{特征数量}}{n}$ known

(ii) $se(\hat{P}_s) = \sqrt{\frac{\hat{P}_s(1-\hat{P}_s)}{n}}$

③ Finite population correction (FPC)

N = finite (Not very large) } need FPC

$\frac{n}{N} > 0.05$

N very large / infinite } No need FPC.

$\frac{n}{N} \leq 0.05$

tip: in exam, $\frac{n}{N} = 0.05$ 建议 FPC.

By FPC

$$Var(\bar{y}_s) = \left(1-\frac{n}{N}\right)\frac{S_s^2}{n} \qquad se(\bar{y}_s) = \sqrt{\left(1-\frac{n}{N}\right)\frac{S_s^2}{n}}$$

$$Var(\hat{p}_s) = \left(1-\frac{n}{N}\right)\frac{\hat{p}_s(1-\hat{p}_s)}{n} \qquad se(\hat{p}_s) = \sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}_s(1-\hat{p}_s)}{n}}$$

$$Var(\hat{t}_s) = N^2\left(1-\frac{n}{N}\right)\frac{S_s^2}{n} \qquad se(\hat{t}_s) = N\sqrt{\left(1-\frac{n}{N}\right)\frac{S_s^2}{n}}$$

Part II: study planning → determine sample size n

ME → determine sample size

But FPC

Input: { how narrow of a confidence interval → width
within ___ error → ME

Output: sample size n

$$ME = \delta = z_{1-\frac{\alpha}{2}}\sqrt{\left(1-\frac{n}{N}\right)\frac{\sigma^2_{guess}}{n}} \quad (FPC)$$

$$\delta = z_{1-\frac{\alpha}{2}}\sqrt{\left(1-\frac{n}{N}\right)\frac{\hat{p}_{guess}(1-\hat{p}_{guess})}{n}} \quad (FPC)$$

without FPC

$$ME = \delta = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{guess}}{n'}}$$

$$n' = \frac{(z_{1-\frac{\alpha}{2}})^2\sigma^2_{guess}}{\delta^2}$$

since $ME = ME$

$$z_{1-\frac{\alpha}{2}}\sqrt{\left(1-\frac{n}{N}\right)\frac{\sigma^2_{guess}}{n}} = z_{1-\frac{\alpha}{2}}\sqrt{\frac{\sigma^2_{guess}}{n'}}$$

$$\left(1-\frac{n}{N}\right)\frac{\sigma^2_{guess}}{n} = \frac{\sigma^2_{guess}}{n'}$$

$$\therefore \quad \frac{1-\frac{n}{N}}{n} = \frac{1}{n'}$$

$$\therefore \quad \frac{1}{n} - \frac{1}{N} = \frac{1}{n'}$$

$$\therefore \quad \frac{N-n}{nN} = \frac{1}{n'}$$

$$\therefore \quad n'(N-n) = nN$$

$$\therefore \quad n'N - n'n - nN = 0$$

$$\therefore \quad n'n + nN = n'N$$

$$\therefore \quad n = \frac{n'N}{n'+N}$$

tip: ME → n with FPC

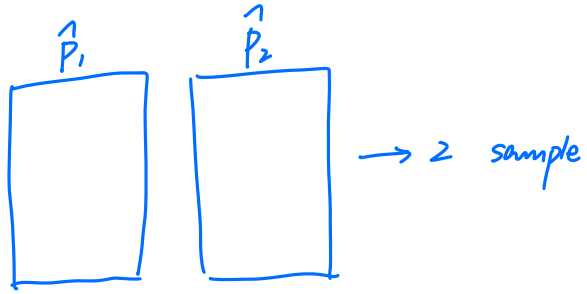
先求 n' with No FPC

$$\Rightarrow n = \frac{n'N}{n'+N} \Rightarrow n \text{ with FPC}$$

Part Ⅲ: proportion change "Δ" panel data

a. 2 independent sample

$$se(\hat{P}_1 - \hat{P}_2) = \sqrt{\frac{\hat{P}_1(1-\hat{P}_1)}{n_1} + \frac{\hat{P}_2(1-\hat{P}_2)}{n_2}}$$



b. - 2 time period

- same observational units in sample

e.g. 同一批人在7月和8月分别做血糖数据

∴ samples are dependent

		current month: NO	current month: YES
previous month	NO	P_{NN}	P_{NY}
previous month	YES	P_{YN}	P_{YY}

proportion change : $P(\text{Yes, current}) - P(\text{Yes, previous})$

$= (P_{NY} + P_{YY}) - (P_{YN} + P_{YY})$

$$\Delta = P_{NY} - P_{YN}$$

$$se(\Delta) = se(\hat{P}_{NY} - \hat{P}_{YN}) = \sqrt{\frac{\hat{P}_{NY}(1-\hat{P}_{NY})}{n} + \frac{\hat{P}_{YN}(1-\hat{P}_{YN})}{N} - 2\frac{\hat{P}_{NY}\hat{P}_{YN}}{n}}$$

$\text{Var}(\hat{P}_{NY})$ $\text{Var}(\hat{P}_{YN})$ $-2\,\text{cov}(\hat{P}_{NY}, \hat{P}_{YN})$

↓

cov negative

Part IV instrument (有点相关但很低)

	unit	y	z_i
population Data	1	y_1	1
	2	y_2	0
	⋮	⋮	⋮
	N	y_N	

$$z_i = \begin{cases} 0 & i \notin \text{sample} \\ 1 & i \in \text{sample} \end{cases}$$

被选中是1，没选中是0

$z_i \sim$ Bernoulli

$$P(z_i = 1) = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} = P(z_j = 1)$$

$$P(z_i z_j = 1) = \frac{\binom{N-2}{n-2}}{\binom{N}{n}} = \frac{n(n-1)}{N(N-1)}$$

z_i	0	1
$P(z_i)$	$1-\frac{n}{N}$	$\frac{n}{N}$

$$E(z_i) = 0\cdot\left(1-\frac{n}{N}\right) + 1\cdot\frac{n}{N} = \frac{n}{N}$$

$$Var(z_i) = p(1-p) = \frac{n}{N}\left(1-\frac{n}{N}\right)$$

$$cov(z_i, z_j) = E(z_i z_j) - E(z_i)E(z_j)$$

$$= 0\times 0\times P(z_i=0, z_j=0) + 0\times 1\times P(z_i=0, z_j=1) + 1\times 0\times P(z_i=1, z_j=0)$$
$$+ 1\times 1\times P(z_i=1, z_j=1) - E(z_i)E(z_j)$$

$$= P(z_i z_j = 1) - E(z_i)E(z_j)$$

$$= \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2$$

$$\bar{y}_s = \frac{\sum_{i=1}^{n} y_i z_i}{n}$$

$$Var(\bar{y}_s) = Var\left(\frac{\sum_{i=1}^{N} y_i z_i}{n}\right) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{N} y_i z_i\right)$$

$$= \left(\frac{1}{n}\right)^2 Var(y_1 z_1 + \cdots + y_N z_N)$$

$$= \left(\frac{1}{n}\right)^2 \left(Var(y_1 z_1) + \cdots + Var(y_N z_N) + 2\sum cov(y_i z_i, y_j z_j)\right)$$

$cov(z_i, z_j)$

y_i independent

$$= \left(\frac{1}{n}\right)^2 (n S_p^2) + 2\left(\right)$$

$$= \left(1-\frac{n}{N}\right)\frac{S_p^2}{n}$$

Instrument z_i

⇒ 估出 FPC Variance

更准确

Part V: Estimation → under SRS

[mean $\bar{y}_p$ estimate]

a. Vanilla

$$\bar{y}_{s,\text{vanilla}} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}_s$$

$$se(\bar{y}_{s,\text{vanilla}}) = \sqrt{\frac{s_s^2}{n}} \quad \text{or} \quad \sqrt{\left(1-\frac{n}{N}\right)\frac{s_s^2}{n}}$$

No FPC — FPC

$$\hat{t}_s = N\bar{y}_{s,\text{vanilla}}$$

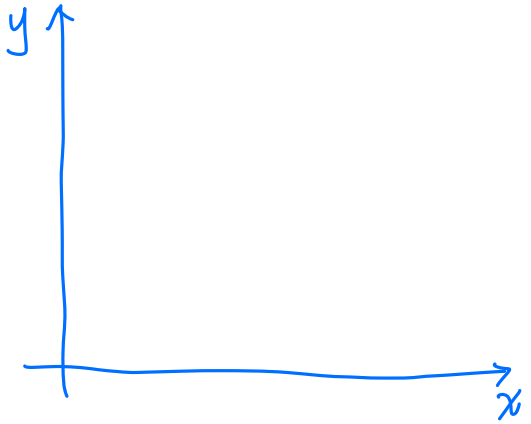
$$se(\hat{t}_s) = N\sqrt{\frac{s_s^2}{n}} \quad \text{or} \quad N\sqrt{\left(1-\frac{n}{N}\right)\frac{s_s^2}{n}}$$

No FPC — FPC

b. Ratio Estimation (x and y correlated)

$$\bar{y}_{s,\text{ratio}} = \left(\frac{\bar{y}_s}{\bar{x}_s}\right)\bar{x}_p \qquad B = \frac{\bar{y}_p}{\bar{x}_p}$$

↳ ratio estimate $\hat{B}$



$$\hat{y} = \left(\frac{\bar{y}_s}{\bar{x}_s}\right)x_i = \hat{B}x_i$$

$$e_i = y_i - \hat{y}_i = y - \hat{B}x_i$$

↳ $Var(e_i) = \frac{\sum_{i=1}^{n}(e_i-\bar{e})^2}{n-1}$ where $\bar{e} = \frac{\sum_{i=1}^{n} e_i}{n}$

$$Var(\hat{B}) = \frac{1}{\bar{x}_p^2}Var(\bar{e}) = \frac{1}{\bar{x}_p^2}\frac{S_e^2}{n} \quad \text{No FPC}$$

$$= \frac{1}{\bar{x}_p^2}\left(1-\frac{n}{N}\right)\frac{S_e^2}{n} \quad \text{FPC}$$

$$Var(\bar{y}_{s,ratio}) = Var(\hat{B}\bar{x}_p) = \bar{x}_p^2 \, Var(\hat{B})$$

$$= \frac{s_e^2}{n} \qquad \text{No FPC}$$

$$= \left(1 - \frac{n}{N}\right)\frac{s_e^2}{n} \qquad \text{FPC}$$

$$se(\bar{y}_{s,ratio}) = \sqrt{\frac{s_e^2}{n}} \qquad \text{No FPC}$$

$$= \sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}} \qquad \text{FPC}$$

$$\hat{t}_{s,ratio} = N\bar{y}_{s,ratio}$$

$$se(\hat{t}_{s,ratio}) = N\sqrt{\frac{s_e^2}{n}} \qquad \text{No FPC}$$

$$= N\sqrt{\left(1 - \frac{n}{N}\right)\frac{s_e^2}{n}} \qquad \text{FPC}$$

c. Regression estimate

$\hat{y} = b_0 + b_1 x$ ← estimate

$b_0 = \bar{y} - b_1\bar{x}$ $(\bar{x}, \bar{y})$ on the regression line

$b_1 = r\frac{s_y}{s_x}$

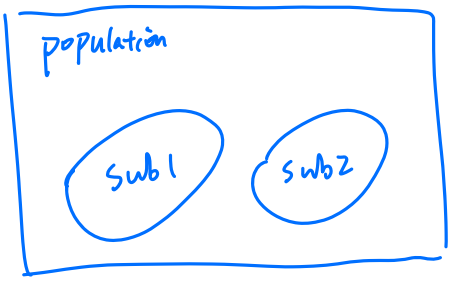
residual: $e_i = y_i - \hat{y}_i$

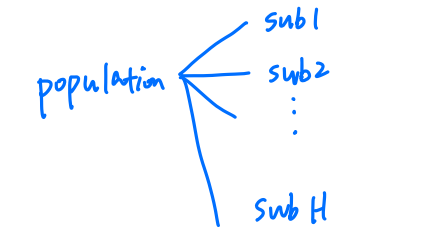
$$Var(b_1) = \frac{\sum_{i=1}^{n} e_i^2}{(n-2)\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

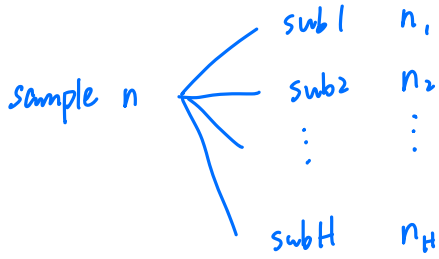
$\bar{y}_{s,regression} = b_0 + b_1\bar{x}_p$? ? ?

Part VI: Estimation in Domain (subpopulation)

sample method 还是 SRS
但是计算 $\bar{y}$ 的 subpopulation







First, $\bar{y}_p$ population mean

Second, $\bar{y}_p$ in subpopulation $\bar{y}_{p_1}, \bar{y}_{p_2}, \cdots \bar{y}_{pH}$

$\bar{y}_{s,\,domain} = \frac{1}{n_d} \sum_{i=1}^{n_d} y_i$ estimate $\bar{y}_{p,\,domain} = \frac{1}{N_d} \sum_{i=1}^{N_d} y_i$

$\hat{t}_{s,\,domain} = N_d \bar{y}_{s,\,domain}$

a. known domain size N_d

$\hat{t}_{s,\,domain} = N_d \bar{y}_{s,\,domain}$

$SE(\hat{t}_{s,\,domain}) = N_d \sqrt{\frac{s_{sd}^2}{n_d}} = N_d \sqrt{\left(1 - \frac{n_d}{N_d}\right) \frac{s_{sd}^2}{n_d}}$

No FPC　　FPC.

b. unknown domain size N_d, only know population size N

$$u_i = \begin{cases} y_i, & i \in \text{sample} \\ 0, & i \notin \text{sample} \end{cases}$$

$\hat{t}_{s,\,domain} = N \bar{u}_s$

$se(\hat{t}_{s,\,domain}) = N \sqrt{\frac{s_u^2}{n}}$　No FPC

$= N \sqrt{\left(1 - \frac{n}{N}\right) \frac{s_u^2}{N}}$

c. ratio estimate

$$u_i = \begin{cases} y_i, & i \in \text{sample} \\ 0, & i \notin \text{sample} \end{cases} \qquad x_i = \begin{cases} 1, & i \in P_d \\ 0, & i \notin P_d \end{cases}$$

$$\bar{y}_{s,\text{domain}} = \frac{\bar{u}_s}{\bar{x}_s}$$

$$\hat{t}_{s,\text{domain}} = N_d \bar{y}_{s,\text{domain}}$$

Part VII: Stratified Sampling → each group randomly as sample

population size $N = N_1 + N_2 + \cdots + N_H$

$N_1 \cdots N_H$ subpopulation size

H: the number of subpopulations

sample size $n = n_1 + n_2 + \cdots + n_H$

$n_1, n_2 \cdots n_H$ subpopulation size

$$\bar{y}_p = \frac{1}{N}\sum_{i=1}^{N} y_i = \frac{\sum_{h=1}^{H} N_h \bar{y}_{ph}}{N} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \bar{y}_{ph}$$

subpopulation	1	2	3	⋯	H
population size	N_1	N_2	N_3	⋯	N_H
sample size	n_1	n_2	n_3	⋯	n_H
sample mean	$\bar{y}_{s1}$	$\bar{y}_{s2}$		⋯	$\bar{y}_{sH}$
sample variance	s_{s1}^2	s_{s2}^2		⋯	s_{sH}^2

- $\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \underbrace{\bar{y}_{sh}}_{h^{th}\text{ subpopulation sample mean}}$

$$\hat{t}_{str} = N \bar{y}_{str}$$

- $\hat{P}_{str} = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) \underbrace{\hat{P}_{sh}}_{h^{th}\text{ subpopulation sample proportion}}$

- $Var(\bar{y}_{str}) = Var\left(\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)\bar{y}_{sh}\right) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 Var(\bar{y}_{sh})$

($\frac{N_h}{N}$: constant)

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_{sh}^2}{n_h}} \quad \text{No FPC}$$

$$= \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1-\frac{n_h}{N_h}\right) \frac{S_{sh}^2}{n_h}} \quad \text{FPC}$$

$$SE(\hat{p}_{str}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{\hat{p}_{sh}(1-\hat{p}_{sh})}{n_h}} \quad \text{No FPC}$$

$$= \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1-\frac{n_h}{N_h}\right) \frac{\hat{p}_{sh}(1-\hat{p}_{sh})}{n_h}} \quad \text{FPC.}$$

$$SE(\hat{t}_{str}) = N\sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_{sh}^2}{n_h}} \quad \text{No FPC}$$

$$= N\sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1-\frac{n_h}{N_h}\right) \frac{S_{sh}^2}{n_h}} \quad \text{FPC.}$$

study planning

Allocation: ① fixed n

② equal variance

③ large population (No FPC)

$S_{1,guess}^2 = S_{2,guess}^2 = \cdots = S_{H guess}^2$

$$SE(\bar{y}_{str}) = \sqrt{\sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_{h,guess}^2}{n_h}}$$

1. proportional allocation:

$$\frac{N_h}{N} = \frac{n_h}{n} \quad \text{or} \quad \frac{n}{N} = \frac{n_h}{N_h}$$

2. optimize allocation:

↓

max or min

一般情況下, $Var(\bar{y}_{str})$ minimize

$$SE(\bar{y}_{str})$$

money constraint → optimize

Part I: study planning

proportional allocation: $\frac{N_h}{N} = \frac{n_h}{n}$ stratified sampling

最常用

No FPC: $Var(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \frac{S_{sh}^2}{n}$

FPC: $Var(\bar{y}_{str}) = \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 \left(1 - \frac{n_h}{n}\right) \frac{S_{sh}^2}{n_h}$

$S_p^2 = \sum_{i=1}^{N} (y_i - \bar{y}_p)^2$ total variance (sum of square total)

$S_p^2 = S_{p,within}^2 + S_{p,between}^2$

$(S_{p,w}^2)$ $(S_{p,B}^2)$

$= \sum_{h=1}^{H} \left(\frac{N_h}{N}\right)^2 S_h^2 + \sum_{h=1}^{H} \left(\frac{N_h}{N}\right) (\bar{y}_{ph} - \bar{y}_p)^2$

within group sum of square

between group sum of square

$Var(\bar{y}_{str}) = \left(1 - \frac{n}{N}\right) \frac{S_{p,w}^2}{n}$ FPC stratified sampling

$Var(\bar{y}_s) = \left(1 - \frac{n}{N}\right) \frac{S_p^2}{n}$ FPC SRS

$\frac{\text{stratified}}{SRS}$ = confident Ratio in webwork

$= \frac{Var(\bar{y}_{str})}{var(\bar{y}_s)}$

$= \frac{\left(1 - \frac{n}{N}\right) \frac{S_{p,w}^2}{n}}{\left(1 - \frac{n}{N}\right) \frac{S_p^2}{n}} = \frac{S_{p,w}^2}{S_p^2} = \frac{S_{p,w}^2}{S_{p,w}^2 + S_{p,B}^2} = 1 - \frac{S_{p,B}^2}{S_p^2}$

Reduction in percentage $= \left(1 - \frac{S_{p,w}^2}{S_p^2}\right) \times 100\%$ in webwork

② optimize allocation:

↳ 最常用 $Var(\bar{y}_{str})$ minimized → 分配 sample size n

↳ constraint for money — "cost constraint"

$$n_1 C_1 + n_2 C_2 + \cdots + n_H C_H = C_{total}$$

(C_1: cost per unit of subpopulation 1)

$$\frac{n_h}{n} = \frac{N_h \left(\frac{S_{h,guess}}{\sqrt{C_h}}\right)}{\sum_{k=1}^{H} N_k \frac{S_{k,guess}}{\sqrt{C_k}}} \quad \text{at } h \text{ subpopulation}$$

$$\Rightarrow n_h \propto \frac{N_h S_{h,guess}}{\sqrt{C_h}}$$ 在 h subpopulation 之後的 sample size

Example: $N_1 = N_2 = N_3 = N_4$ $\quad S_{1,guess} = S_{2,guess} = S_{3,guess}$ $\quad S_{4,guess} = 2S_{1,guess}$

$C_1 = C_2 = 4$ $\quad C_3 = C_4 = 9$ $\quad C = \$1560$

$$n_1 \propto \frac{N_1 S_{1,guess}}{\sqrt{C_1}} = \frac{NS}{\sqrt{4}} = \frac{NS}{2}$$

$$n_2 \propto \frac{N_2 S_{2,guess}}{\sqrt{C_2}} = \frac{NS}{\sqrt{4}} = \frac{NS}{2}$$

$$n_3 \propto \frac{N_3 S_{3,guess}}{\sqrt{C_3}} = \frac{NS}{\sqrt{9}} = \frac{NS}{3}$$

$$n_4 \propto \frac{N_4 S_{4,guess}}{\sqrt{C_4}} = \frac{2NS}{\sqrt{9}} = \frac{2NS}{3}$$

$$n = n_1 + n_2 + n_3 + n_4 = \frac{NS}{2} + \frac{NS}{2} + \frac{NS}{3} + \frac{2NS}{3} = 2NS$$

$$\frac{n_1}{n} = \frac{\frac{NS}{2}}{2NS} = \frac{1}{4} \qquad \frac{n_2}{n} = \frac{\frac{NS}{2}}{2NS} = \frac{1}{4}$$

$$\frac{n_3}{n} = \frac{\frac{NS}{3}}{2NS} = \frac{1}{6} \qquad \frac{n_4}{n} = \frac{\frac{2}{3}NS}{2NS} = \frac{1}{3}$$

$$n_1 \quad n_2 \quad n_3 \quad n_4 \propto \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{6}, \frac{1}{3}\right)$$

$$n_1 c_1 + n_2 c_2 + n_3 c_3 + n_4 c_4 = 1560$$

$$\frac{1}{4}n(4) + \frac{1}{4}n(4) + \frac{1}{6}n(9) + \frac{1}{3}n(9) = 1560$$

$$\Rightarrow n = \frac{1560}{6.5} = 240$$

Part 2: weighted → optimal allocation by minimizing $Var(\bar{y}_s)$

$\bar{y}_s = w\bar{y}_{s1} + (1-w)\bar{y}_{s2}$ w: the sample size weighted for group 1

$Var(\bar{y}_s) = Var(w\bar{y}_{s1} + (1-w)\bar{y}_{s2})$ S_1 和 S_2 independent

minimized why?

$= Var(w\bar{y}_{s1}) + Var((1-w)\bar{y}_{s2}) = w^2 Var(\bar{y}_{s1}) + (1-w)^2 Var(\bar{y}_{s2})$

min var → smaller SE → narrower CI → more accurate $\bar{y}_s$

to minimize: $\frac{\partial Var(\bar{y}_s)}{\partial w} = 2w\,Var(\bar{y}_{s1}) + 2(1-w)(-1)Var(\bar{y}_{s2})$

$= 2w\,Var(\bar{y}_{s1}) - 2Var(\bar{y}_{s2}) + 2w\,Var(\bar{y}_{s2})$

$= 0$

$$\Rightarrow w = \frac{Var(\bar{y}_{s2})}{Var(\bar{y}_{s1}) + Var(\bar{y}_{s2})}$$

↑ group 1 sample size weighted

weighted 因对方 in Var 决定 in 我们希望多取 Var 小 in sample

Part 3: polls of polls → proportion; elections

Inputs: $\hat{p}_i$ SE_i $\quad i = 1, 2, \cdots k$

$$\hat{p}_{agr} = \frac{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2}\right) \hat{p}_i}{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2}\right)}$$

$$SE\left(\hat{p}_{agr}\right) = \frac{1}{\sqrt{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2}\right)}}$$

Random effect:

$$\hat{p}_{agr} = \frac{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2 + \tau^2}\right) \hat{p}_i}{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2 + \tau^2}\right)}$$

$$SE\left(\hat{p}_{agr}\right) = \frac{1}{\sqrt{\sum_{i=1}^{k} \left(\frac{1}{SE_i^2 + \tau^2}\right)}}$$