Math 152 2/5/09.

Chptr 7 #49

$N = 1000$ (Quadrats) (ignore all finite population corrections and always treat $\frac{1}{n-1}$ as $\frac{1}{n}$)

$n = 50$

X_i = area covered by vegetation in i^{th} quadrat

Y_i = # of birds in i^{th} quadrat.

$\sum X_i = 3000$ $\quad \bar{X} = \frac{3000}{50} = 60$

$\sum Y_i = 150$ $\quad \bar{Y} = \frac{150}{50} = 3$

$\sum X_i^2 = 225{,}000$ $\quad \hat{\sigma}_X^2 = \frac{225000}{50} - (60)^2 = 900$

$\sum Y_i^2 = 650$ $\quad \hat{\sigma}_Y^2 = \frac{650}{50} - 3^2 = 4$

$\sum X_i Y_i = 11{,}000$ $\quad \hat{\sigma}_{XY} = \frac{11000}{50} - (60)(3) = 40.$

a) $R = \frac{\bar{Y}}{\bar{X}} = \frac{3}{60} = .05.$

b) $S_R^2 \simeq \frac{1}{50} \frac{1}{(60)^2} \left((.05)^2 \cdot 900 + 4 - 2(.05)(40) \right)$

$S_R \simeq .00354.$

b) continued

Sample variance of $\bar{X} \approx \frac{\hat{\sigma}_x^2}{n}$

Standard error of $\bar{X} \approx \frac{\hat{\sigma}_x}{\sqrt{n}} = \frac{30}{\sqrt{50}} \approx 4.24$

90% confidence interval for μ_x

$\approx 60 \pm (1.64)(4.25) \approx (53.05, 66.95)$

Sample variance of $\bar{Y} \approx \frac{\hat{\sigma}_y^2}{n}$

Standard error of $\bar{Y} \approx \frac{\hat{\sigma}_y}{\sqrt{n}} = \frac{2}{\sqrt{50}} \approx .283$

90% confidence interval for μ_y

$\approx 3 \pm (1.64)(.283) \approx (2.53, 3.46)$

c) $T = N\bar{Y} = 3000$

$Var(T) = N^2 Var(\bar{Y})$

standard error of $T = N \cdot$ standard error $(\bar{Y})$

$\approx 1000(.283) = 283$

a 95% confidence interval for τ

is

$$3000 \pm (1.96)(283)$$

$$= (2445.32,\ 3554.68).$$

d). If the total area covered by vegetation τ_x is known then $\mu_y = \frac{\tau_x}{1000}$ so

$$\bar{Y}_R = \mu_x \frac{\bar{Y}}{\bar{X}} \quad \text{is known}$$

and $T_R = N\bar{Y}_R$ gives an estimate

of τ_y.

Now

$$\text{Var}(T_R) = N^2 \text{Var}(\bar{Y}_R)$$

$$\simeq N^2 \cdot \frac{1}{n}\left(r^2\sigma_x^2 + \sigma_y^2 - 2r\sigma_{xy}\right)$$

$$\approx \frac{1000^2}{50}\left((.05)^2 \cdot 900 + 4 - 2(.05)(.40)\right) = 4500$$

Need

but $\text{Var}(T) \approx (283)^2 \approx 80{,}000$.

so if the bias is small, then T_R will be a better estimate.

$$E(\bar{Y}_R) - \mu_Y \approx \frac{1}{50} \cdot \frac{1}{\mu_X} \left((.05)(900) - 40 \right)$$

and from part b we have "90% confidence" that

$$\frac{1}{50} \cdot \frac{1}{\mu_X} \left((.05)(900) - 40 \right) \leq$$

$$\frac{1}{50} \cdot \frac{1}{53} \left((.05)(900) - 40 \right) \approx .002$$

and therefore that the bias in T_R is roughly 2.

We expect the mean squared error of T_R to be roughly $45000 + 4$ while the mean squared error of T is $\approx 80{,}000$.

Stratified Sampling.

$$N = N_1 + \dots + N_L$$

$$\mu_\ell = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} X_{i\ell}$$

$$\mu = \sum_{\ell=1}^{L} W_\ell \mu_\ell \quad , \quad W_\ell = \frac{N_\ell}{N}.$$

$$\bar{X}_\ell = \frac{1}{n_\ell} \sum_{i=1}^{n_\ell} X_{i\ell} \qquad \ell = 1, \dots, L.$$

$$n_1 + \dots + n_L = n.$$

$$\bar{X}_s = \sum_{\ell=1}^{L} W_\ell \bar{X}_\ell \qquad \left(\begin{array}{c} \text{stratified} \\ \text{sample} \\ \text{mean} \end{array} \right)$$

We have:

$$E(\bar{X}_s) = \mu$$

and

$$Var(\bar{X}_s) = \sum_{\ell=1}^{L} W_\ell^2 Var(\bar{X}_\ell)$$

$$= \sum_{\ell=1}^{L} W_\ell^2 \frac{1}{n_\ell} \left(1 - \frac{n_\ell - 1}{N_\ell - 1}\right) \sigma_\ell^2.$$

and

$$\mathrm{Var}(\bar{X}_s) \approx \sum_{\ell=1}^{L} \frac{W_\ell^2 \sigma_\ell^2}{n_\ell}.$$

neglecting finite population corrections.

e.g. Stratify the population of hospitals according to # of beds.

Stratum	N_ℓ	W_ℓ	μ_ℓ	σ_ℓ
1-98	98	.249	182.9	103.4
99-196	98	.249	526.5	204.8
197-294	98	.249	956.3	243.5
295-393	99	.251	1591.2	419.2

n = sample size.

$n_1 = n_2 = n_3 = n_4 = n/4$.

$$\mathrm{Var}(\bar{X}_s) \approx \sum_{\ell=1}^{4} \frac{W_\ell^2 \sigma_\ell^2}{n/4} = \frac{4}{n} \sum_{\ell=1}^{4} W_\ell^2 \sigma_\ell^2$$

$$= \frac{72,042.6}{n}$$

$\sigma_{\bar{X}_s} = \frac{268.4}{\sqrt{n}}$ while $\sigma_{\bar{X}} = \frac{589.7}{\sqrt{n}}$.

$$\frac{268.4}{\sqrt{n_1}} = \frac{589.7}{\sqrt{n_2}}$$

$$\Leftrightarrow \quad n_1 = \left(\frac{268.4}{589.7}\right)^2 n_2 \quad \approx .2\, n_2 .$$

For population totals we have:

$T_S = N\bar{X}_S \qquad E(T_S) = \tau$

$$Var(T_S) = N^2 Var(\bar{X}_S)$$

$$= \sum_{\ell=1}^{L} N_\ell^2 \left(\frac{1}{n_\ell}\right)\left(1 - \frac{n_\ell - 1}{N_\ell - 1}\right) \sigma_\ell$$

To get practically useful expressions we must estimate the σ_ℓ^2. by.

$$s_\ell^2 = \frac{1}{n_\ell - 1} \sum_{i=1}^{n_\ell} (x_{i\ell} - \bar{x}_\ell)^2$$

and then

$Var(\bar{X}_S)$ is estimated as.

$$s^2_{\bar{X}_S} = \sum_{\ell=1}^{L} w_\ell^2 \left(\frac{1}{n_\ell}\right)\left(1 - \frac{n_\ell}{N_\ell}\right) s_\ell^2$$

Allocation for stratified sampling

"Optimal" allocation

Minimizing

$$\mathrm{Var}(\bar{X}_s) = \sum_{\ell=1}^{L} \frac{w_\ell^2 \sigma_\ell^2}{n_\ell}$$

(as a function of $n_1, \ldots, n_L$)

subject to the constraint

that $n_1 + \ldots + n_L = 1$.

Lagrange Multipliers:

$$\left[\nabla\left(\mathrm{Var}(\bar{X}_s)\right)\right]_\ell = -\frac{w_\ell^2 \sigma_\ell^2}{n_\ell^2}$$

$$\left[\nabla\left(n_1 + \ldots n_L\right)\right]_\ell = 1.$$

Solve the system.

$$-\frac{w_\ell^2 \sigma_\ell^2}{n_\ell^2} = 1 \cdot \lambda \qquad \ell = 1, \ldots L$$

$$n_1 + \ldots + n_L = n.$$

(we may as well use $-\lambda$).

so solve.

$$w_\ell^2 \sigma_\ell^2 = n_\ell^2 \lambda. \qquad \ell = 1 \dots L.$$

(+ constraint).

$$n_\ell = \frac{w_\ell \sigma_\ell}{\sqrt{\lambda}}$$

$$n = \sum_{\ell=1}^{L} \frac{w_\ell \sigma_\ell}{\sqrt{\lambda}}.$$

$$\frac{1}{\sqrt{\lambda}} = \frac{n}{\sum_{\ell=1}^{L} w_\ell \sigma_\ell}.$$

$$n_\ell = \frac{w_\ell \sigma_\ell \cdot n}{\sum_{\ell=1}^{L} w_\ell \sigma_\ell}$$

With this choice for n_ℓ.
we call the estimate

$\overline{X}_{so}$ (optimal stratified).

and

$$Var(\bar{X}_{SO}) \approx \sum_{\ell=1}^{L} \frac{W_\ell^2 \sigma_\ell^2}{\frac{W_\ell \sigma_\ell \cdot n}{\sum_{\ell=1} W_\ell \sigma_\ell}}$$

$$= \frac{1}{n} \left(\sum_{\ell=1}^{L} W_\ell \sigma_\ell \right)^2 .$$

e.g. Same strata for hospitals as before

	A 1-98	B 99-196	C 197-294	D 295-393
$\frac{W_\ell \sigma_\ell}{\sum W_\ell \sigma_\ell}$	.106	.210	.250	.434

Optimal allocation is difficult to achieve in practice and since it depends on variability, it depends on what question is being asked.

So if your survey asks more than one

A practical alternative is:

Proportional allocation

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \dots = \frac{n_L}{N_L}.$$

i.e. $\boxed{n_\ell = n W_\ell.}$

$$\bar{X}_{sp} = \sum_{\ell=1}^{L} W_\ell \bar{X}_\ell$$

is the unweighted mean of all sample values.

We have:

$$\text{Var}(\bar{X}_{sp}) \cong \frac{1}{n} \sum_{\ell=1}^{L} W_\ell \sigma_\ell^2$$

(ignoring finite pop.).

$$= \sum_{\ell=1}^{L} W_\ell^2 \, \text{Var}(\bar{X}_\ell) \cong \sum_{\ell=1}^{L} W_\ell^2 \frac{\sigma_\ell^2}{n_\ell}$$

$$= \frac{1}{n} \sum_{\ell=1}^{L} W_\ell \sigma_\ell^2$$

With some algebra, we have

$$\mathrm{Var}(\bar{X}_{sp}) - \mathrm{Var}(\bar{X}_{so})$$

$$= \frac{1}{n} \sum_{\ell=1}^{L} W_\ell (\sigma_\ell - \bar{\sigma})^2$$

where $\bar{\sigma} = \sum_{\ell=1}^{L} W_\ell \sigma_\ell$.

If all σ_ℓ are equal the two methods give the same result. If the σ_ℓ are varying widely between strata then the optimal allocation is "better".

We can also compare S.R.S. with proportional as follows:

$$\mathrm{Var}(\bar{x}) - \mathrm{Var}(\bar{x}_{sp}) = \frac{1}{n}\sum_{\ell=1}^{L} W_\ell(\mu_\ell - \mu)^2.$$

(after some algebra).

so proportional sampling is to be preferred when the means of the strata are variable.

For our hospitals example we have.

$$\mathrm{Var}(\bar{x}_{sp}) = \mathrm{Var}(\bar{x}_{so}) + \frac{1}{n}\sum W_\ell(\sigma_\ell - \bar{\sigma})^2$$

so that

$$\frac{\mathrm{Var}(\bar{x}_{sp})}{\mathrm{Var}(\bar{x}_{so})} = 1 + \frac{\sum W_\ell(\sigma_\ell - \bar{\sigma})^2}{(\sum W_\ell \sigma_\ell)^2}$$

$$= 1 + .218.$$

and

$$\frac{Var(\bar{X}_{srs})}{Var(\bar{X}_{sp})} = 1 + \frac{\sum w_\ell (\mu_\ell - \bar{\mu})^2}{\sum w_\ell \sigma_\ell^2}$$

$$= 1 + 3.83.$$

This situation is typical in applications.

The gain in using proportional allocation rather than S.r.s. is large compared to the gain in using optimal allocation over proportional allocation.