STA304 H1 S/1003 H S
Summer 2013
Dragan Banjevic
(V part 2, stratified)

Statistics Canada Statistique Canada

Stratified Sampling: Allocation (I)

Selecting the allocation and sample size (I)

Summary: Using given allocation $a_1, a_2, \ldots, a_L$, find n:

Case I: Fixed error bound B for μ, or τ

$$n = \frac{\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i}}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} \approx \frac{\sum_{i=1}^{L} W_i^2 \frac{\sigma_i^2}{a_i}}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \sigma_i^2} \quad \begin{cases} D = D_\mu = \left(\frac{B_\mu}{2}\right)^2 \\ D = D_\tau = \left(\frac{B_\tau}{2N}\right)^2 \end{cases}$$

Case II: Fixed cost of sampling $C = c_0 + \sum_{i=1}^{L} c_i n_i$

$$n = \frac{C - c_0}{\sum_{i=1}^{L} a_i c_i} = \frac{C'}{\bar{c}} \qquad \bar{c} = \sum_{i=1}^{L} a_i c_i$$

Stratified Sampling: Allocation (I)

Selecting the allocation and sample size (II)

An allocation $a_1, a_2, \ldots, a_L, 0 < a_i < 1, \sum a_i = 1,$ is selected depending on available information about the population and strata, and convenience.

We will consider three main cases:

1) Uniform (equal) allocation
2) Proportional allocation
3) Optimal allocation

We will find appropriate total sample size and allocation for Case I and Case II.

We will discuss their properties and condition for their use.

We will make comparison between them.

We will compare them with simple random sampling.

Stratified Sampling: Allocation (II)

1) Uniform/equal allocation

We decide to use equal a_i : $a_1 = a_2 = \cdots = a_L = \frac{1}{L} \Rightarrow n_i = \frac{n}{L}$

Case I: Fixed error bound B for μ, or τ

$$n = L\frac{\sum_{i=1}^{L} W_i^2 \tilde{\sigma}_i^2}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} \approx L\frac{\sum_{i=1}^{L} W_i^2 \sigma_i^2}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \sigma_i^2} \approx \frac{L}{D}\sum_{i=1}^{L} W_i^2 \sigma_i^2$$

For N large

What if when $N_i \le \frac{n}{L}$?

Case II: Fixed cost of sampling $C = c_0 + \sum_{i=1}^{L} c_i n_i$

$$n = L\frac{C - c_0}{\sum_{i=1}^{L} c_i} = \frac{C'}{\bar{c}} \qquad \bar{c} = \frac{1}{L}\sum_{i=1}^{L} c_i$$

See the examples with Toronto neighbourhoods and statistical books

Uniform allocation is used when $N_i \approx N_j$, and we don't have any other useful information about strata.

Stratified Sampling: Allocation (III)

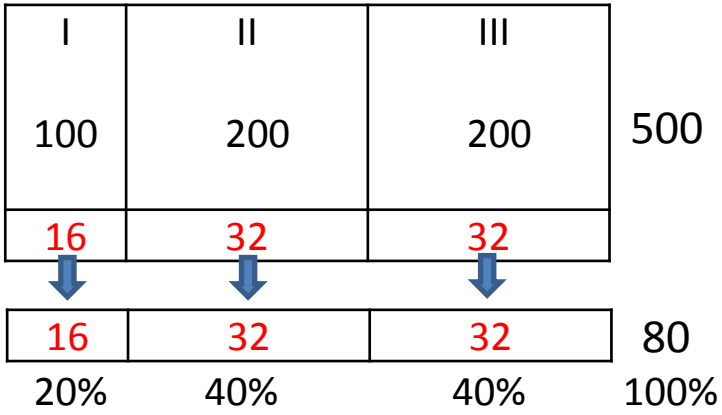
2) Proportional allocation (I)

We decide to use all a_i proportional to strata sizes, or $a_i = W_i = \frac{N_i}{N}$

$$\frac{n_i}{n} = \frac{N_i}{N} \Rightarrow n_i = \frac{N_i}{N} n = N_i \frac{n}{N} \Rightarrow \frac{n_i}{N_i} = \frac{n}{N}$$

Same fraction of elements is selected from every stratum, as from the population (sample fraction, *n*/*N*).

Illustrative example:



$N = 500, L = 3,\ N_1 = 100, N_2 = N_3 = 200$

$$a_1 = W_1 = \frac{100}{500} = 0.2 = 20\%$$

$$a_2 = W_2 = \frac{200}{500} = 0.4 = 40\% = a_3$$

$$\frac{n_i}{N_i} = \frac{80}{500} = 0.16 = 16\% \text{ - sample fraction}$$

If $n = 80, n_1 = 0.2 \times n = 0.2 \times 80 = 16, n_2 = n_3 = 0.4 \times n = 0.4 \times 80 = 32$

Stratified Sampling: Allocation (III)

2) Proportional allocation (II)

Estimation:

$$\hat{\mu} = \bar{y}_{STR,PR} = \sum_{1}^{L} W_i \bar{y}_i = \sum_{1}^{L} a_i \bar{y}_i = \sum_{i=1}^{L} \frac{n_i}{n} \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{L} \sum_{j=1}^{n_i} y_{ij} = \bar{y}$$

The stratified mean in proportional allocation is just plain sample mean (but the sample is still stratified)

$$Var(\bar{y}_{STR,PR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\tilde{\sigma}_i^2}{n_i} = \sum_{i=1}^{L} W_i^2 (1 - \frac{n_i}{N_i}) \frac{\tilde{\sigma}_i^2}{n_i} = \sum_{i=1}^{L} W_i^2 (1 - \frac{n}{N}) \frac{\tilde{\sigma}_i^2}{n W_i}$$

$$= (1 - \frac{n}{N}) \frac{1}{n} \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 \approx (1 - \frac{n}{N}) \frac{1}{n} \sum_{i=1}^{L} W_i \sigma_i^2$$

$$Var(\bar{y}_{STR,PR}) = (1 - \frac{n}{N}) \frac{1}{n} \tilde{\sigma}_{STR}^2 \approx (1 - \frac{n}{N}) \frac{1}{n} \sigma_{STR}^2$$

$$\hat{Var}(\bar{y}_{STR,PR}) = (1 - \frac{n}{N}) \frac{1}{n} \hat{\tilde{\sigma}}_{STR}^2 = (1 - \frac{n}{N}) \frac{1}{n} \sum_{i=1}^{L} W_i S_i^2$$

$$\tilde{\sigma}_{STR}^2 = \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

$$\approx \sum_{i=1}^{L} W_i \sigma_i^2 = \sigma_{STR}^2$$

"stratified" variance

Stratified Sampling: Allocation (III)

2) Proportional allocation (III)

Total sample size:

Case I: Fixed error bound B for μ, or τ

$$n = \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} \approx \frac{\sum_{i=1}^{L} W_i \sigma_i^2}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \sigma_i^2} \approx \frac{\sum_{i=1}^{L} W_i \sigma_i^2}{D}$$

N large

$$n = \frac{\tilde{\sigma}_{STR}^2}{D + \frac{1}{N}\tilde{\sigma}_{STR}^2} \approx \frac{\sigma_{STR}^2}{D + \frac{1}{N}\sigma_{STR}^2} \approx \frac{\sigma_{STR}^2}{D}$$

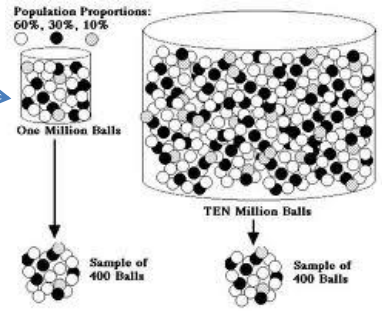
$$D = D_\mu = \left(\frac{B_\mu}{2}\right)^2,$$

or $$D = D_\tau = \left(\frac{B_\tau}{2N}\right)^2$$

$$\tilde{\sigma}_{STR}^2 = \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

$$\sigma_{STR}^2 = \sum_{i=1}^{L} W_i \sigma_i^2$$

If $n_{PR,0} = \frac{\tilde{\sigma}_{STR}^2}{D}$, then $n_{PR} = \frac{n_{PR,0}}{1 + \frac{1}{N} n_{PR,0}} \le n_{PR,0} = n_{PR,max}$



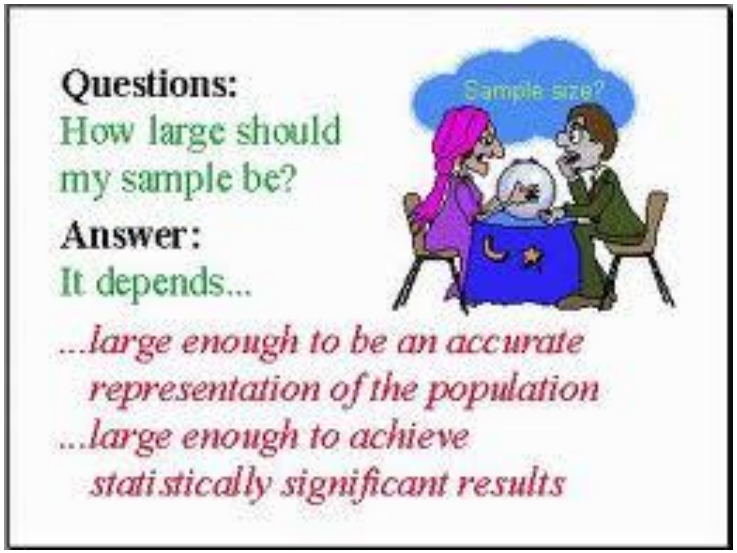
Stratified Sampling: Allocation (III)

2) Proportional allocation (IV)

Total sample size:

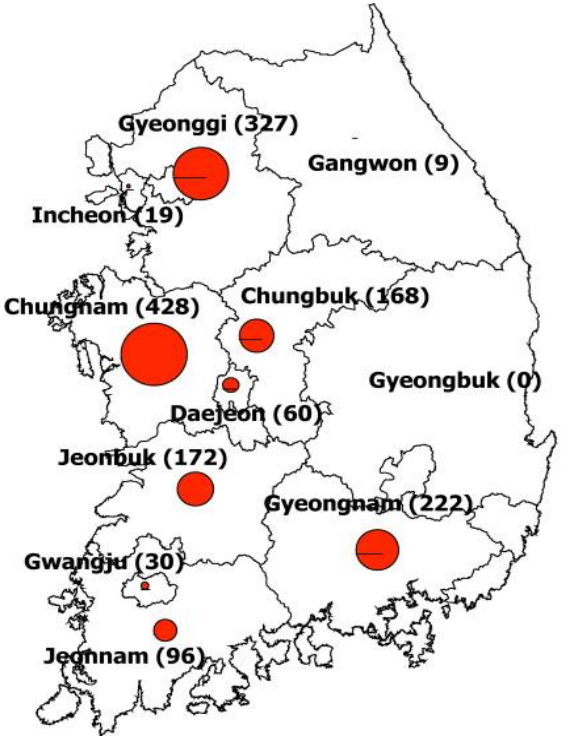
Case II: Fixed cost of sampling $C = c_0 + \sum_{i=1}^{L} c_i n_i$, $n_i = W_i \times n$

$$a_i = W_i = \frac{N_i}{N} \Rightarrow \quad \boxed{n = \frac{C - c_0}{\sum_{i=1}^{L} W_i c_i} = N \frac{C - c_0}{\sum_{i=1}^{L} N_i c_i} = \frac{C'}{\bar{c}}} \quad \bar{c} = \sum_{i=1}^{L} W_i c_i$$



You read

Jung-Yong Yeh et al: Serologic evidence of West Nile Virus in wild ducks captured in major inland resting sites for migratory waterfowl in South Korea, *Veterinary Microbiology*, Volume 154, Issues 1–2, 29 December 2011, Pages 96–103.



Abstract: The rapid global expansion of West Nile virus (WNV) has recently raised concerns regarding its possible spread into South Korea. … To assess the risk of WNV infection in South Korea, we conducted a nationwide WNV surveillance of wild birds, with an emphasis on migratory ducks from WNV-affected areas. … We collected blood samples from 1531 wild birds representing 57 bird species at several major inland resting sites for migratory waterfowl in South Korea. … Our findings strongly suggest that some of the birds captured in this study had been exposed to WNV or JEV.

Fig. 1. Map of sampling sites in South Korea. The sample size and locations of serum sampling for West Nile virus detection are demonstrated as circles by province. Circle sizes represent sample size and the numbers on the circles indicate the number of captured birds. The study sites were located in or around major inland resting sites across South Korea

Stratified Sampling: Allocation (III)

2) Proportional allocation: comparison with SRS (I)

Purpose of this part is to establish that under very general conditions $Var(\bar{y}_{STR,PR}) < Var(\bar{y}_{SRS})$, for the same sample size, that is, stratified sampling is more efficient than SRS.

Key facts:

$$Var(\bar{y}_{STR,PR}) = (1-\frac{n}{N})\frac{1}{n}\tilde{\sigma}^2_{STR},\ \tilde{\sigma}^2_{STR} = \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 \approx \sum_{i=1}^{L} W_i \sigma_i^2$$

All for N_i large

$$Var(\bar{y}_{SRS}) = (1-\frac{n}{N})\frac{1}{n}\tilde{\sigma}^2,\ \tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2 \approx \sigma^2 = \sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i (\mu_i - \mu)^2,$$

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) = (1-\frac{n}{N})\frac{1}{n}(\tilde{\sigma}^2 - \tilde{\sigma}^2_{STR}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{L} W_i (\mu_i - \mu)^2,$$

Stratified Sampling: Allocation (III)

2) Proportional allocation: comparison with SRS (II)

Exact calculation:

$$\tilde{\sigma}^2_{STR} = \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 = \sum_{i=1}^{L} W_i \frac{N_i}{N_i - 1}\sigma_i^2, \quad \tilde{\sigma}^2 = \frac{N}{N-1}\left[\sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i (\mu_i - \mu)^2\right]$$

$$Var(\bar{y}_{STR,PR}) = (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^{L} W_i \frac{N_i}{N_i - 1}\sigma_i^2$$

$$\sigma^2 = \sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i (\mu_i - \mu)^2$$

Key equation

$$Var(\bar{y}_{SRS}) = (1 - \frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i \sigma_i^2 + \sum_{i=1}^{L} W_i (\mu_i - \mu)^2\right]$$

$$\Rightarrow \boxed{\begin{aligned} & Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) \qquad \text{(proof: next page)} \\ & = (1 - \frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i (\mu_i - \mu)^2 - \frac{1}{N}\sum_{i=1}^{L}(1 - W_i)\tilde{\sigma}_i^2\right] \\ & \approx (1 - \frac{n}{N})\frac{1}{n}\sum_{i=1}^{L} W_i (\mu_i - \mu)^2 \geq 0 \quad \text{For } N_i \text{ large} \end{aligned}}$$

2) Proportional allocation: comparison with SRS (II)

Proof:

can't help you here

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) = (1-\frac{n}{N})\frac{1}{n}(\tilde{\sigma}^2 - \tilde{\sigma}^2_{STR})$$

$$= (1-\frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i(\mu_i-\mu)^2 + \sum_{i=1}^{L} W_i\sigma_i^2 - \frac{N-1}{N}\sum_{i=1}^{L} W_i\frac{N_i}{N_i-1}\sigma_i^2\right]$$

$$= (1-\frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i(\mu_i-\mu)^2 - \sum_{i=1}^{L} W_i\sigma_i^2\left(\frac{N-1}{N}\frac{N_i}{N_i-1}-1\right)\right]$$

$$= (1-\frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i(\mu_i-\mu)^2 - \sum_{i=1}^{L}\frac{N_i}{N}\sigma_i^2\frac{N-N_i}{N(N_i-1)}\right]$$

$$= (1-\frac{n}{N})\frac{1}{n}\frac{N}{N-1}\left[\sum_{i=1}^{L} W_i(\mu_i-\mu)^2 - \frac{1}{N}\sum_{i=1}^{L}(1-W_i)\tilde{\sigma}_i^2\right]$$

Notice also: $\frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{STR,PR})} = \frac{\tilde{\sigma}^2}{\tilde{\sigma}^2_{STR}}, \frac{Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR})}{Var(\bar{y}_{STR,PR})} = \frac{\tilde{\sigma}^2 - \tilde{\sigma}^2_{STR}}{\tilde{\sigma}^2_{STR}}$

Stratified Sampling: Allocation (III)

2) Proportional allocation: comparison with SRS (III)

Discussion:

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) \approx (1 - \frac{n}{N}) \frac{1}{n} \sum_{i=1}^{L} W_i (\mu_i - \mu)^2$$

For N_i large

If the strata are with different mean values (to some extend), then stratified sampling (with proportional allocation) is more efficient (with a smaller variance of $\hat{\mu}$) than SRS with a sample of the same size.

If $\mu_i \approx \mu \Rightarrow \sum_{i=1}^{L} W_i (\mu_i - \mu)^2 \approx 0$ (strata similar, in averages), then

$$Var(\bar{y}_{SRS}) \approx Var(\bar{y}_{STR,PR})$$

Discussion, examples…

An advantage of stratified sampling in this case might be in a smaller total cost of sampling.

example

Stratified Sampling: Allocation (III)

2) Proportional allocation: comparison with SRS (IV)

Example: How does this comparison work in our example with Toronto neighbourhoods, where strata are not large?

Key elements of the comparison:

σ^2	70,575,729.0957	$\sum W_i \tilde{\sigma}_i^2$	62,375,531.9981
$\sum W_i \sigma_i^2$	59,435,868.1969	$\sum (1-W_i) \tilde{\sigma}_i^2$	349,177,400.1624
$\sum W_i (\mu_i - \mu)^2$	11,139,860.9011	$\sum (1-W_i) \tilde{\sigma}_i^2 / N$	2,494,124.2869

$$Var(\bar{y}_{SRS}) = (1-\frac{n}{140})\frac{1}{n}\frac{140}{140-1} \times 70{,}575{,}729.10 = (1-\frac{n}{140})\frac{1}{n} \times 71{,}083{,}468.15$$

$$Var(\bar{y}_{STR,PR}) = (1-\frac{n}{140})\frac{1}{n} \times 62{,}375{,}532.00$$

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) = (1-\frac{n}{140})\frac{1}{n} \times 8{,}707{,}936.16$$

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) \approx (1-\frac{n}{140})\frac{1}{n} \times 11{,}139{,}860.90$$

$$\frac{Var(\bar{y}_{SRS})}{Var(\bar{y}_{STR,PR})} = \frac{71{,}083{,}468.15}{62{,}375{,}530.00} = 1.1396 \approx 1.14$$

Conclusion:
Proportional allocation does not performs much better than SRS

Stratified Sampling: Allocation (IV)

3) Optimal allocation (I)

Optimal allocation, given certain objective function.

In short, we consider two cases/objectives:

Case I: Fixed/given error bound (*D*), minimal cost (*C*)

Case II: Fixed cost/budget, best (minimal) error bound

In more detail:

Case I: Fixed *D* (or *Var*) – find allocation which minimizes cost.

Case II: Fixed *C* (cost) – find allocation which minimizes variance (error bound)

General conclusion: Relative allocation $a_1, a_2, \ldots, a_L$ is the same for the both cases, but the total sample size is not!

Stratified Sampling: Allocation (IV)

3) Optimal allocation (II)

Proof: Starting from D fixed (see the previous lecture),

$$D = Var(\bar{y}_{STR}) = \frac{1}{n}\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i} - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 = \frac{1}{n}A - B, n = \frac{A}{D+B}$$

$$C = c_0 + \sum_{i=1}^{L} c_i n_i, n_i = a_i n \Rightarrow C = c_0 + n\sum_{i=1}^{L} c_i a_i$$

$$= c_0 + \frac{A}{D+B}\sum_{i=1}^{L} c_i a_i = c_0 + \frac{1}{D+B}\left(\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i}\right)\left(\sum_{i=1}^{L} c_i a_i\right)$$

$$= c_0 + \frac{1}{D+B} g(a_1, a_2, \ldots, a_L), \text{ or } C = c_0 + \frac{1}{D+B} g(a_1, a_2, \ldots, a_L)$$

$$\min_{a_1, a_2, \ldots, a_L} C = c_0 + \frac{1}{D+B} \min_{a_1, a_2, \ldots, a_L} g(a_1, a_2, \ldots, a_L)$$

Notice, D is fixed, B does not depend on $a_1, a_2, \ldots, a_L$.

Stratified Sampling: Allocation (IV)

3) Optimal allocation (III)

Proof, continued: Starting from fixed cost *C*,

$$C = c_0 + \sum_{i=1}^{L} c_i n_i, n_i = a_i n \Rightarrow C = c_0 + n\sum_{i=1}^{L} c_i a_i, n = (C - c_0) / \sum_{i=1}^{L} c_i a_i$$

$$V = Var(\bar{y}_{STR}) = \frac{1}{n}\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i} - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 = \frac{A}{C - c_0}\sum_{i=1}^{L} c_i a_i - B$$

$$= \frac{1}{C - c_0}\left(\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i}\right)\left(\sum_{i=1}^{L} c_i a_i\right) - B = \frac{1}{C - c_0} g(a_1, a_2, \ldots, a_L) - B$$

$$\min_{a_1, a_2, \ldots, a_L} V = \frac{1}{C - c_0} \min_{a_1, a_2, \ldots, a_L} g(a_1, a_2, \ldots, a_L) - B$$

In both cases, min depends only on $\min_{a_1, a_2, \ldots, a_L} g(a_1, a_2, \ldots, a_L)$,

under conditions $0 < a_i < 1, \sum a_i = 1$.

Stratified Sampling: Allocation (IV)

3) Optimal allocation (IV)

It remains to find $\min_{a_1,a_2,\ldots,a_L} g = \min_{a_1,a_2,\ldots,a_L} \left(\sum_{i=1}^{L} W_i^2 \frac{\tilde{\sigma}_i^2}{a_i} \right) \left(\sum_{i=1}^{L} c_i a_i \right)$

We may use Lagrange multipliers - due to constraint $\sum_{i=1}^{L} a_i = 1$, or a simple approach, using that $\sum x_i^2 p_i = \sum (x_i - \bar{x})^2 p_i + \bar{x}^2$, where $\bar{x} = \sum x_i p_i, \sum p_i = 1.$

Let $W_i^2 \tilde{\sigma}_i^2 = d_i^2$ and $\sum_{i=1}^{L} c_i a_i = \bar{c}$. Then $g(a_1, a_2, \ldots, a_L) = \sum_{i=1}^{L} \frac{d_i^2 \bar{c}}{a_i} = \sum_{i=1}^{L} \frac{c_i a_i}{\bar{c}} \left(\frac{d_i \bar{c}}{a_i \sqrt{c_i}} \right)^2.$

Let $p_i = \frac{c_i a_i}{\bar{c}}$ and $x_i = \frac{d_i \bar{c}}{a_i \sqrt{c_i}}$. Then $\sum_{i=1}^{L} p_i = 1$ and $\bar{x} = \sum_{i=1}^{L} x_i p_i = \sum_{i=1}^{L} \frac{d_i \bar{c}}{a_i \sqrt{c_i}} \frac{c_i a_i}{\bar{c}} = \sum_{i=1}^{L} d_i \sqrt{c_i}.$

Then $g(a_1, a_2, \ldots, a_L) = \sum_{i=1}^{L} x_i^2 p_i = \sum_{i=1}^{L} p_i (x_i - \bar{x})^2 + \bar{x}^2 = \sum_{i=1}^{L} p_i \left(\frac{d_i \bar{c}}{a_i \sqrt{c_i}} - \bar{x} \right)^2 + \bar{x}^2.$

Stratified Sampling: Allocation (IV)

3) Optimal allocation (V)

$\bar{x}^2 = (\sum_{i=1}^{L} d_i\sqrt{c_i})^2$ does not depend on a_i, and then g is minimized when

$\sum_{i=1}^{L} p_i(\frac{d_i\bar{c}}{a_i\sqrt{c_i}} - \bar{x})^2 = 0$, or when $\frac{d_i\bar{c}}{a_i\sqrt{c_i}} - \bar{x} = 0$, or when $a_i = \frac{d_i\bar{c}}{\sqrt{c_i}\bar{x}}$.

From the condition $1 = \sum a_i = \sum \frac{d_i\bar{c}}{\sqrt{c_i}\bar{x}} = \frac{\bar{c}}{\bar{x}}\sum \frac{d_i}{\sqrt{c_i}} \Rightarrow \frac{\bar{c}}{\bar{x}} = \frac{1}{\sum \frac{d_i}{\sqrt{c_i}}}$, and

and the minimum is obtained for $a_i^* = \frac{d_i}{\sqrt{c_i}\sum \frac{d_i}{\sqrt{c_i}}} = \frac{W_i\frac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i\frac{\tilde{\sigma}_i}{\sqrt{c_i}}}$.

Also, $g^* = \min g = g(a_1^*, a_2^*, \ldots, a_L^*) = (\sum_{i=1}^{L} d_i\sqrt{c_i})^2 = \left(\sum_{i=1}^{L} W_i\tilde{\sigma}_i\sqrt{c_i}\right)^2$

Stratified Sampling: Allocation (IV)

3) Optimal allocation (VI)

$$a_i^* = \frac{W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}}} = \frac{N_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} N_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}}} \approx \frac{W_i \frac{\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i \frac{\sigma_i}{\sqrt{c_i}}} = \frac{N_i \frac{\sigma_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} N_i \frac{\sigma_i}{\sqrt{c_i}}}$$

Notice :

$0 < a_i^* < 1,$

$\sum_{i=1}^{L} a_i^* = 1.$

Optimal relative allocation for Case I and Case II

Illustration : $L = 2, a_1^* = \frac{N_1 \frac{\sigma_1}{\sqrt{c_1}}}{N_1 \frac{\sigma_1}{\sqrt{c_1}} + N_2 \frac{\sigma_2}{\sqrt{c_2}}}, a_2^* = \frac{N_2 \frac{\sigma_2}{\sqrt{c_2}}}{N_1 \frac{\sigma_1}{\sqrt{c_1}} + N_2 \frac{\sigma_2}{\sqrt{c_2}}}$

Stratified Sampling: Allocation (IV)

3) Optimal allocation (VII)

Summary:

$$C_{min} = c_0 + \frac{g^*}{D+B}, \text{for given } D; \quad V_{min} = \frac{g^*}{C-c_0} - B, \text{for given } C,$$

$$B = \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2, \quad g^* = \left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}\right)^2$$

Comment: How population parameters and costs affect allocation

$$a_i^* = \frac{N_i \sigma_i}{\sqrt{c_i}} \times const, \quad n_i^* = a_i^* \times n \quad \begin{cases} \uparrow \text{ if } N_i\sigma_i \uparrow, c_i \text{ fixed} \\ \downarrow \text{ if } c_i \uparrow, N_i\sigma_i \text{ fixed} \end{cases}$$

Compare with proportional allocation:

$$a_i = N_i \frac{1}{N} = N_i \times const, \quad n_i = a_i \times n = N_i \frac{n}{N} \quad \begin{cases} \uparrow & \text{if } N_i \uparrow \\ \downarrow & \text{if } N_i \downarrow \end{cases}$$

Don't forget, optimal allocation requires knowledge of σ_i and c_i !

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Case I (I)

Case I: Fixed *D* (or *Var*) – optimal sample size and allocation which minimizes cost.

From $n = \dfrac{\sum_{i=1}^{L} W_i^2 \dfrac{\tilde{\sigma}_i^2}{a_i}}{D + \dfrac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2}$ and $a_i^* = \dfrac{W_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}}}$ and $n_i^* = a_i^* \times n^*$

$$n^* = \frac{\sum_{i=1}^{L} W_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}} \times \sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}}{D + \dfrac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} = \frac{\sum_{i=1}^{L} N_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}} \times \sum_{i=1}^{L} N_i \tilde{\sigma}_i \sqrt{c_i}}{N^2 D + \sum_{i=1}^{L} N_i \tilde{\sigma}_i^2}$$

$$n_i^* = W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \frac{\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}}{D + \dfrac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} = N_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \frac{\sum_{i=1}^{L} N_i \tilde{\sigma}_i \sqrt{c_i}}{N^2 D + \sum_{i=1}^{L} N_i \tilde{\sigma}_i^2}$$

Or approximative formulas, if we replace $\tilde{\sigma}_i$ by σ_i

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Case I (II)

Case I, continued: Minimal cost for given error bound

$$C_{min} - c_0 = \sum_{i=1}^{L} c_i n_i^* = n^* \sum_{i=1}^{L} c_i a_i^* = \frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i} \right)^2}{D + \frac{1}{N} \sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} = \frac{\left(\sum_{i=1}^{L} N_i \tilde{\sigma}_i \sqrt{c_i} \right)^2}{N^2 D + \sum_{i=1}^{L} N_i \tilde{\sigma}_i^2}$$

For *N* large, the formulas are simpler:

$$n^* \approx \frac{1}{D} \sum_{i=1}^{L} W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \times \sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i} \qquad n_i^* = \frac{1}{D} W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}$$

$$C_{min} - c_0 \approx \frac{1}{D} \left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i} \right)^2$$

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Case II (I)

Case II: Fixed *C* (cost) – optimal sample size and allocation which minimizes variance (error bound)

From $n = \dfrac{C - c_0}{\sum_{i=1}^{L} a_i c_i}$ and a_i^* and $n_i^* = a_i^* \times n^*$

$$n^* = \frac{(C - c_0)\sum_{i=1}^{L} W_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}} = \frac{(C - c_0)\sum_{i=1}^{L} N_i \dfrac{\tilde{\sigma}_i}{\sqrt{c_i}}}{\sum_{i=1}^{L} N_i \tilde{\sigma}_i \sqrt{c_i}}$$

$$n_i^* = W_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \frac{C - c_0}{\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}} = N_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}} \frac{C - c_0}{\sum_{i=1}^{L} N_i \tilde{\sigma}_i \sqrt{c_i}}$$

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Case II (II)

Case II, continued: Minimal variance for given cost

$$V_{min} = \frac{g^*}{C - c_0} - B = \frac{1}{C - c_0}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

For *N* large, $V_{min} \approx \frac{1}{C - c_0}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}\right)^2$

Notice equation

$$(V + B)(C - c_0) = \left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}\right)^2, \; B = \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

When $V = D$ is fixed, C_{min} can be calculated from the equation, when C is fixed, V_{min} can be calculated from the equation.

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Practical calculation (I)

Calculating all elements of any stratification may take time. The following procedure is suggested:

Data	Step	I	II	III	IV
$N, N_i, \bar{y}_i, S_i^2,$ $D, C, c_i, \ldots$	Calc.	a_i	n	n_i	Var or C

For optimal stratification calculate first (organize in a table):

$$P_i = N_i \frac{\tilde{\sigma}_i}{\sqrt{c_i}}, Q_i = N_i \tilde{\sigma}_i \sqrt{c_i}, R_i = N_i \tilde{\sigma}_i^2, P = \sum P_i, Q = \sum Q_i, R = \sum R_i$$

Case	n	n_i	V	C
Case I, *C* fixed	$\frac{P}{Q} C'$	$\frac{P_i}{Q} C' = \frac{P_i}{P} n$	$\frac{1}{N^2}\left(\frac{Q^2}{C'} - R\right)$	C (given)
Case II, *D* fixed	$\frac{QP}{N^2 D + R}$	$P_i \frac{Q}{N^2 D + R}$	D (given)	$\frac{Q^2}{N^2 D + R}$

$a_i = \frac{P_i}{P}$

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Practical calculation (II)

Numerical example: (you check the calculation)

Stratum	Size	St. Dev	c_i	P_i	Q_i	R_i	Alloc
A	155	5	\$9	258.333	2,325	3,875	0.3226
B	62	15	\$9	310.000	2,790	13,950	0.3871
C	93	10	\$16	232.500	3,720	9,300	0.2903
Pop	310	-	$c_0 = 0$	800.833	8,835	27,125	1.0000
				P	Q	R	

Allocation:

?

Case	n	n_1	n_2	n_3	*Var*	C
Given cost C = \$500	45.32 (46)	14.62 (15)	17.54 (18)	13.16 (13)	1.3422 (1.3415)	500.000 (505)
Given *Var* D = 1	57.42 (58)	18.52 (19)	22.23 (22)	16.67 (17)	1.0000 (1.0545)	633.453 (641)

Theoretical and actual values: Discussion

How to round, up or down?

First line, theoretical value; in brackets, actual

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (I)

Either assume that costs of sampling one unit are equal over strata, $c_1 = c_2 = \cdots = c_L = \bar{c}$, or that the total sample size n is fixed/given. Find the allocation that minimizes $Var(\bar{y}_{STR})$.

If n is given, then $C = c_0 + n\bar{c}$.

If C is given, then $n = \dfrac{C - c_0}{\bar{c}}$ is the total sample size.

In the both cases, n is known

Relative allocation

$$a_i^* = \frac{W_i\tilde{\sigma}_i}{\sum_{i=1}^{L} W_i\tilde{\sigma}_i} = \frac{N_i\tilde{\sigma}_i}{\sum_{i=1}^{L} N_i\tilde{\sigma}_i} \approx \frac{W_i\sigma_i}{\sum_{i=1}^{L} W_i\sigma_i} = \frac{N_i\sigma_i}{\sum_{i=1}^{L} N_i\sigma_i}$$

Costs just cancel from the formula

Allocation

$$n_i^* = a_i^* n = \frac{W_i\tilde{\sigma}_i}{\sum_{i=1}^{L} W_i\tilde{\sigma}_i}\frac{C'}{\bar{c}} = \frac{N_i\tilde{\sigma}_i}{\sum_{i=1}^{L} N_i\tilde{\sigma}_i}\frac{C'}{\bar{c}}$$

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (II)

From the formula for variance

$$Var(\bar{y}_{STR,OPT}) = \frac{1}{C - c_0}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{c_i}\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 \Leftarrow C - c_0 = n\bar{c}, c_i = \bar{c}$$

$$= \frac{1}{n\bar{c}}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i \sqrt{\bar{c}}\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 = \frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

$$Var(\bar{y}_{STR,NEY}) = \frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 \approx \frac{1}{n}\left(\sum_{i=1}^{L} W_i \sigma_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \sigma_i^2$$

$$Var(\bar{y}_{STR,NEY}) \approx \frac{\bar{\tilde{\sigma}}^2}{n} \approx \frac{\bar{\sigma}^2}{n}$$

$$\bar{\tilde{\sigma}} = \sum_{i=1}^{L} W_i \tilde{\sigma}_i, \ \bar{\sigma} = \sum_{i=1}^{L} W_i \sigma_i$$

The "average" variance

N large

$$Var(\bar{y}_{SRS}) \approx \frac{\tilde{\sigma}^2}{n} \approx \frac{\sigma^2}{n}$$

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (III)

If $Var(\bar{y}_{STR,NEY}) = D$ is fixed, then the minimal sample size is obtained from $D = \frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$, or

$$n = \frac{\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2}{D + \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2} = \frac{\left(\sum_{i=1}^{L} N_i \tilde{\sigma}_i\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \tilde{\sigma}_i^2} \approx \frac{\bar{\tilde{\sigma}}^2}{D} \approx \frac{\bar{\sigma}^2}{D}$$

N large

Jerzy Neyman
1894 – 1981
Neyman allocation:
1934

Daniel G. Horvitz,
1921-2008
Horvitz-Thompson
estimator: 1952

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (IV)

Example: Collection of statistical books (I)

Summary: N = 161, L = 5 strata (blocks), by location

After looking around and using results from our sample with equal allocation, we make some reasonable (hopefully) guesses about strata standard deviations for variable y (number of books per shelve)

Stratum	1	2	3	4	5	Total
Size	30	42	30	24	35	161
$\hat{\sigma}_i$	4.0	4.5	4.0	7.0	3.0	

Considering that all blocks are close, sampling "costs" (time spent) from each stratum can be used as equal. Then, Neyman allocation can be applied to optimize the allocation. Tasks:

1) Find the optimal relative allocation of the sample
2) Assume total sample size is n = 20 (for comparison with the previous results, or considering total time we can spend on sampling and calculation, say, n = 40)

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (IV)

Example, continued: Collection of statistical books (II)

Calculation formula: $a_i^* = N_i\hat{\sigma}_i / \sum N_i\hat{\sigma}_i = N_i\hat{\sigma}_i / \Sigma$

Stratum	1	2	3	4	5	Total
Size	30	42	30	24	35	161
$\hat{\sigma}_i$	4.0	4.5	4.0	7.0	3.0	
$N_i\hat{\sigma}_i$	120	189	120	168	105	702
a_i^*	120/702	189/702	120/702	168/702	105/702	702/702
%	17.09	26.92	17.09	23.93	14.96	99.99
$n_i^* = a_i^* \times 20$	3.418803	5.384615	3.418803	4.786325	2.991453	19.99
n_i^r (rounded)	3 (4?)	6 (5?)	3 (4?)	5	3	20
$n_i^* = a_i^* \times 40$	6.837607	10.76923	6.837607	9.57265	5.982906	39.98
n_i^r (rounded)	7	11	7	9	6	40

What is the effect of the rounding, if we keep exact total sample size? In a larger sample we may round all them up, with a minor increase of the total sample size.

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Neyman allocation (IV)

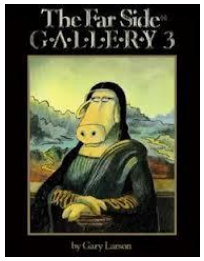
Example, continued: Collection of statistical books (III)

Comparing the effect of the rounding, for same total sample size. "Opt" is the optimal allocation, no rounding. "Rd" is the rounded optimal, with actual s. sizes.

$$Var(\bar{y}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\tilde{\sigma}_i^2}{n_i}, \qquad Var(\bar{y}_{STR,NEY}) = \frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

S.size	$\sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\tilde{\sigma}_i^2}{n_i}$	Increase	$\frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2$	$\frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$	$\frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$
20 (Opt)	0.82322055		0.950588326	0.12736777	0.82322055
20 (Rd)	0.83322133	1.092%	N/A	N/A	N/A
40 (Opt)	0.34792639		0.475294163	0.12736777	0.34792639
40 (Rd)	0.34850698	0.167%	N/A	N/A	N/A

The effect of rounding is usually very small, as in this case. If enough decimal places are used, the general formula should give the same result as the formula for optimal Neyman allocation.



example

Stratified Sampling: Allocation (IV)

3) Optimal allocation, example: Farms and cattle (I)

Population: 2,072 farms from a province (*an adjustment of the SRS example*)

Variable: Number of cattle on farm (y)

Parameters to be estimated: Average number of cattle per farm (μ_y)

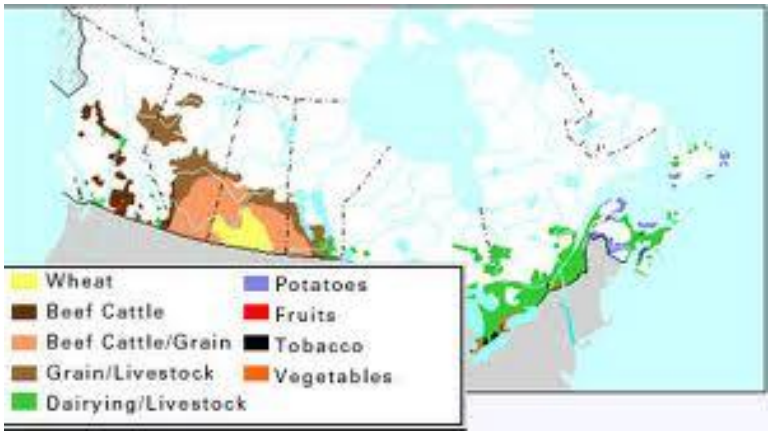
Sampling design: Stratified sampling with optimal (Neyman) allocation of the sample size (sampling costs are either equal, or unknown)

Stratification: Using farm acreage (size)

Stratum (i)	1	2	3	4	5	
Acres	0-15	16-30	31-50	51-75	76-100	Total
N_i	635	570	475	303	89	2072

Sample size: Total sample size n = 500 is given.

An acre? See next slide



An acre (ac):

4,046.8564224 m^2 = 0.404685 h

≈ 40.47% h (h = 100 m x 100 m)

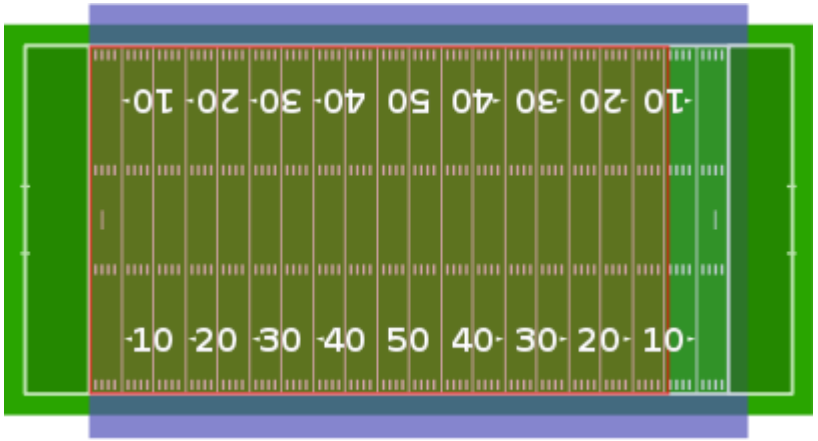
66 feet × 660 feet (43,560 square feet)

1 acre ≈ 208.71 feet × 208.71 feet

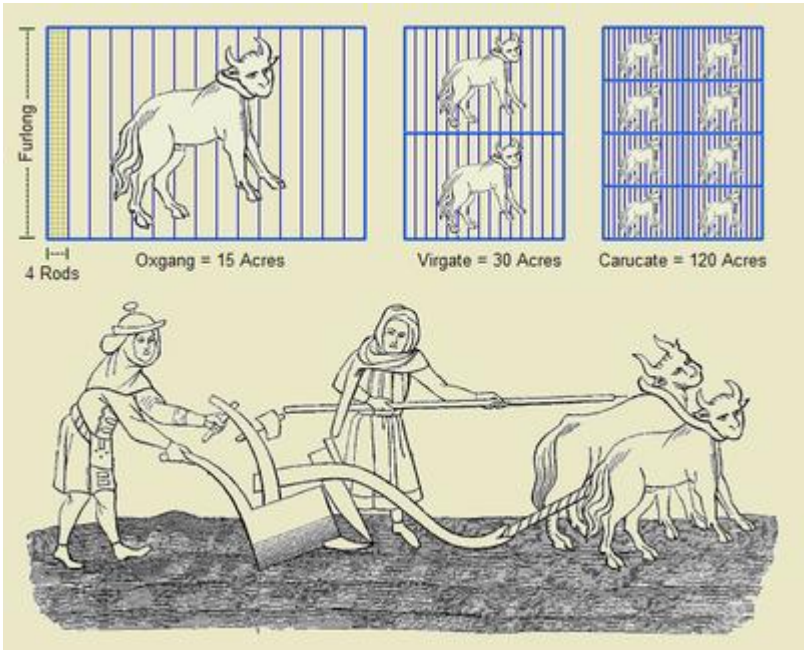
4,840 square yards ≈ 69.57 yards x 69.57 yards

$^{1}/_{640}$ (0.0015625) square mile

1 square mile = 640 acres



The area of one acre (red) superposed on an American football field (green) and association football (soccer) pitch (blue).



Nowadays, it is defined as 1/640 of a square mile. The most commonly used acre today is the international acre. In the United States both the international acre and the slightly different US survey acre are in use. The most common use of the acre is to measure tracts of land.

During the Middle Ages, an acre was the amount of land that could be plowed in one day with a yoke of oxen.

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation, example: Farms and cattle (II)

Allocation: Optimal allocation is calculated using data from a recent census. Total number of farms in that part at that time was 2055.

Str.	1	2	3	4	5	Total	$\Sigma N_i' \sigma_i' = 625\times4.5+ +564\times7.3+ \cdots + +86\times15.8 =16566.9$ $n_i = na_i = 500 \times N_i' \sigma_i'/16566.9$
N_i'	625	564	476	304	86	2055	
W_i'	0.304	0.274	0.232	0.148	0.042	1.000	
σ_i'	4.5	7.3	9.6	12.2	15.8	-	
$\boldsymbol{n_i}$	**85**	**124**	**138**	**112**	**41**	500	
n_i^p	152	137	116	74	21	500	Calculation

$n_1 = 500\times(625\times4.5)/16566.9 = 84.88 = 85$, $n_2 = \cdots = 124$, ….

Proportional allocation (for comparison): $n_i^p = n\frac{N_i}{N} = 500 \times \frac{N_i'}{2055}$

Notice a big difference between optimal and proportional allocation. Why? Discussion …

example

Stratified Sampling: Allocation (IV)

3) Optimal allocation, example: Farms and cattle (III)

Sample results and calculation: Summary of sampling is given for every stratum as $\bar{y}_i$ and s_i^2.

Design info

Calculation

Str.	N_i	W_i	n_i	$\bar{y}_i$	s_i^2	$W_i\bar{y}_i$	$W_i^2\frac{N_i-n_i}{N_i}\frac{S_i^2}{n_i}$	s_i	σ_i'
1	635	0.307	85	4.24	27.54	1.30	0.0264	5.25	4.5
2	570	0.275	124	11.63	55.84	3.20	0.0267	7.47	7.3
3	475	0.229	138	15.95	71.70	3.65	0.0194	8.47	9.6
4	303	0.146	112	23.59	192.32	3.44	0.0231	13.87	12.2
5	89	0.043	41	29.61	334.93	1.27	0.0081	18.30	15.8
Total	2072	1.000	500	-	-	12.86	0.1037	comparison	

Estimation: $\hat{\mu} = \bar{y}_{STR} = \sum W_i\bar{y}_i = 12.86$ cattle/farm

$\hat{Var}(\bar{y}_{STR}) = 0.1037,\ \hat{\sigma}(\bar{y}_{STR}) = \sqrt{0.1037} = 0.3220,$

$B_\mu = 2\times 0.3220 = 0.644,$ CI for μ: $12.86 \pm 0.64 = [12.24, 13.50]$

example

Comparisons: Optimal and proportional allocation, theoretical variance and estimated variance

Allocation	Variance theoretical (alloc. rounded)	Variance theoretical (alloc. unrounded)	Variance estimated
Optimal	0.093702	0.093685	0.103672
Proportional	0.112702	0.113017	(sample)
Increase	20.28%	20.64%	

$$Var(\bar{y}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\tilde{\sigma}_i^2}{n_i}, \qquad \hat{Var}(\bar{y}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{S_i^2}{n_i},$$

$$Var(\bar{y}_{STR,NEY}) = \frac{1}{n}\left(\sum_{i=1}^{L} W_i \tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

$$Var(\bar{y}_{STR,PR}) = \frac{1}{n}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2 - \frac{1}{N}\sum_{i=1}^{L} W_i \tilde{\sigma}_i^2$$

} Allocation unrounded

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Note on incomplete information about some parameters (I)

Costs are known to be proportional to some quantity: $c_i = b_i \times k$, or $\frac{c_i}{c_j} = \frac{b_i}{b_j}$.

Then, e.g., for $j = 1$, $\frac{c_i}{c_1} = \frac{b_i}{b_1}$, or $c_i = b_i \frac{c_1}{b_1} = b_i k, k = \frac{c_1}{b_1}$.

Or, standard deviations are known to be proportional to some quantity:

$$\sigma_i = \alpha_i \times k', \text{or } \frac{\sigma_i}{\sigma_j} = \frac{\alpha_i}{\alpha_j}.$$

Even, maybe, strata sizes are only known to be proportional to some quantity:

$$N_i = M_i \times k'', \text{or } \frac{N_i}{N_j} = \frac{M_i}{M_j}.$$

For any of these cases, or their combinations, we can find the optimal *relative* allocation.

Stratified Sampling: Allocation (IV)

3) Optimal allocation: Note on incomplete information about some parameters (II)

$$a_i^* = \frac{M_i \frac{\alpha_i}{\sqrt{b_i}}}{\sum_{i=1}^{L} M_i \frac{\alpha_i}{\sqrt{b_i}}}, i = 1,2,\ldots,L$$

Example: It is known that strata sizes are approximately equal, and that the costs of sampling from each stratum are also approximately equal. What is the optimal allocation (relative)?

Answer: $M_i \approx 1, b_i \approx 1 \Rightarrow a_i^* \approx \frac{\sigma_i}{\sum \sigma_i}, i = 1,2,\ldots,L$

Example: Sampling from Stratum 1 is twice cheaper than from Stratum 2. Sampling from Stratum 3 is three times more expensive than from Stratum 1. What is the optimal allocation (relative)?

$c_1 = \frac{1}{2} c_2, c_3 = 3c_1 \Rightarrow c_2 = 2c_1, c_3 = 3c_1$ E.g., $a_2^* = \frac{N_2 \frac{\sigma_2}{\sqrt{2}}}{N_1 \frac{\sigma_1}{\sqrt{1}} + N_2 \frac{\sigma_2}{\sqrt{2}} + N_3 \frac{\sigma_3}{\sqrt{3}}}$.

Choose $c_1 = k \Rightarrow c_1 = 1 \times k, c_2 = 2 \times k, c_3 = 3 \times k$

Stratified Sampling: Allocation (III)

Optimal allocation: Comparison with SRS (I)

1) Comparison of optimal and proportional allocation. Assume same sample costs per stratum and then same sample size *n*.

$$Var(\bar{y}_{STR,NEY}) = \frac{1}{n}\left(\sum_{i=1}^{L} W_i\tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum_{i=1}^{L} W_i\tilde{\sigma}_i^2,\ Var(\bar{y}_{STR,PR}) = \frac{1}{n}\sum_{i=1}^{L} W_i\tilde{\sigma}_i^2 - \frac{1}{N}\sum_{i=1}^{L} W_i\tilde{\sigma}_i^2$$

$$Var(\bar{y}_{STR,PR}) - Var(\bar{y}_{STR,NEY}) = \frac{1}{n}\sum_{i=1}^{L} W_i\tilde{\sigma}_i^2 - \frac{1}{n}\left(\sum_{i=1}^{L} W_i\tilde{\sigma}_i\right)^2 = \frac{1}{n}\sum_{i=1}^{L} W_i(\tilde{\sigma}_i - \bar{\tilde{\sigma}})^2$$

$\bar{\tilde{\sigma}} = \sum_{i=1}^{L} W_i\tilde{\sigma}_i$. Compare with $Var(X) = \sum p_i x_i^2 - \left(\sum p_i x_i\right)^2 = \sum p_i (x_i - \mu_X)^2$

Conclusion: If $\tilde{\sigma}_i\ (\approx \sigma_i)$ are different, then the optimal (Neyman) allocation is better than proportional. If $\tilde{\sigma}_i \approx \bar{\tilde{\sigma}}\ (\sigma_i \approx \bar{\sigma})$ (approximately constant), then the proportional allocation is the same as the optimal (as efficient as optimal).

$$a_{i,opt} = \frac{N_i\tilde{\sigma}_i}{\sum N_i\tilde{\sigma}_i} \approx \frac{N_i\bar{\tilde{\sigma}}}{\sum N_i\bar{\tilde{\sigma}}} = \frac{N_i}{\sum N_i} = \frac{N_i}{N} = W_i = a_{i,prop}$$

Stratified Sampling: Allocation (III)

Optimal allocation: Comparison with SRS (II)

2) Comparison of optimal and SRS. Assume same sample costs per stratum and then same sample size n. Assume large N_i, for simplicity.

$$Var(\bar{y}_{SRS}) \approx (1-\frac{n}{N})\frac{1}{n}\sigma^2 = (1-\frac{n}{N})\frac{1}{n}\left(\sum_{i=1}^{L} W_i\sigma_i^2 + \sum_{i=1}^{L} W_i(\mu_i-\mu)^2\right),$$

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{L} W_i(\mu_i-\mu)^2$$

$$Var(\bar{y}_{STR,PR}) - Var(\bar{y}_{STR,NEY}) \approx \frac{1}{n}\sum_{i=1}^{L} W_i(\sigma_i-\bar{\sigma})^2$$

$$Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,NEY}) = Var(\bar{y}_{SRS}) - Var(\bar{y}_{STR,PR})$$

$$+ Var(\bar{y}_{STR,PR}) - Var(\bar{y}_{STR,NEY}) \approx (1-\frac{n}{N})\frac{1}{n}\sum_{i=1}^{L} W_i(\mu_i-\mu)^2 + \frac{1}{n}\sum_{i=1}^{L} W_i(\sigma_i-\bar{\sigma})^2$$

Conclusion: If either σ_i , or μ_i are different, then the optimal (Neyman) allocation is better than an SRS. If the both $\sigma_i \approx \bar{\sigma}$ and $\mu_i \approx \mu$, then the SRS is as efficient as stratified optimal or proportional sampling.

What about overall costs?

Stratified Sampling: Population Proportion (I)

Basic elements, from general theory, parameters (I)

$$y(e)=\begin{cases}0\,,e\text{ does not have the property}\\1\,,e\text{ has the property}\end{cases},\qquad p=\frac{M}{N}$$

$$p_i=\frac{M_i}{N_i},M_i\text{ - \# of elements with the property from stratum }i$$

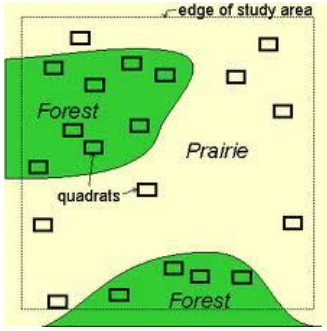
$$\mu=\sum_{i=1}^{L}W_i\mu_i,\ \mu_i=p_i\Rightarrow p=\sum_{i=1}^{L}W_i p_i$$

$$\hat{\mu}_i=\hat{p}_i=\frac{f_i}{n_i}\text{ - sample proportion in }i\text{ - th stratum}\Rightarrow \hat{p}=\hat{p}_{STR}=\sum_{i=1}^{L}W_i\hat{p}_i$$

unbiased

$$M=\tau=Np,M_i=\tau_i=N_i p_i\Rightarrow \hat{M}_i=N_i\hat{p}_i,\hat{M}=N\hat{p}=\sum_{i=1}^{L}N_i\hat{p}_i$$

Could it be a valid way to estimate the proportion of red marbles?



Stratified Sampling: Population Proportion (I)

Basic elements, from general theory, parameters (II)

$$Var(\hat{p}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i - 1} \frac{p_i q_i}{n_i} \quad \text{(recall } \sigma_i^2 = p_i q_i\text{)}$$

$$\hat{Var}(\hat{p}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\hat{p}_i \hat{q}_i}{n_i - 1}, \; B_p = 2\sqrt{\hat{Var}(\hat{p}_{STR})}$$

$$Var(\hat{\tau}_{STR}) = N^2 Var(\hat{p}_{STR}) = N^2 \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i - 1} \frac{p_i q_i}{n_i}$$

$$\hat{Var}(\hat{\tau}_{STR}) = N^2 \hat{Var}(\hat{p}_{STR}) = N^2 \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\hat{p}_i \hat{q}_i}{n_i - 1},$$

$$B_\tau = 2\sqrt{\hat{Var}(\hat{\tau}_{STR})} = 2N\sqrt{\hat{Var}(\hat{p}_{STR})} = N \times B_p$$

Stratified Sampling: Population Proportion (I)

Basic elements, from general theory, allocation (III)

Fixed error bound B for p, or τ

$$n \approx \frac{\sum W_i^2 \frac{p_i q_i}{a_i}}{D + \frac{1}{N}\sum W_i p_i q_i} \approx \frac{1}{D}\sum W_i^2 \frac{p_i q_i}{a_i} \begin{cases} D = D_p = \left(B_p/2\right)^2 \\ D = D_\tau = \left(B_\tau/\left(2N\right)\right)^2 \end{cases}$$

Fixed cost $n = \frac{C - c_0}{\bar{c}}, \bar{c} = \sum_{i=1}^{L} a_i c_i$

Equal allocation: $n_i = \frac{n}{L}$, proportional allocation: $n_i = W_i \times n$

Optimal allocation $a_i^* \approx \frac{N_i \sqrt{\frac{p_i q_i}{c_i}}}{\sum N_i \sqrt{\frac{p_i q_i}{c_i}}}$, $n_i^* = a_i^* n$

example

Stratified Sampling: Population Proportion (II)

Numerical example (I)

$N = 100$(a class), $L = 2, N_1 = 64$(left), $N_2 = 36$(right),

$n = 15, n_1 = 9, n_2 = 6$

After sampling, $f_1 = 6$, $f_2 = 3$. Estimation:

$$\hat{p}_1 = \frac{6}{9}, \hat{q}_1 = \frac{3}{9}, \hat{p}_2 = \frac{3}{6}, \hat{q}_2 = \frac{3}{6},$$

$$\hat{p}_{STR} = W_1 \hat{p}_1 + W_2 \hat{p}_2 = \frac{64}{100} \times \frac{6}{9} + \frac{36}{100} \times \frac{3}{6} = \frac{364}{600} = 60.67\%,$$

$$\hat{q}_{STR} = 39.33\%,$$

$$\hat{\tau}_{STR} = N\hat{p}_{STR} = 100 \times 60.67\% = 60.67.$$

example

Stratified Sampling: Population Proportion (II)

Numerical example (II)

Error calculation:

$$\hat{Var}(\hat{p}_{STR}) = W_1^2 \frac{N_1 - n_1}{N_1} \frac{\hat{p}_1\hat{q}_1}{n_1 - 1} + W_2^2 \frac{N_2 - n_2}{N_2} \frac{\hat{p}_2\hat{q}_2}{n_2 - 1}$$

$$= \left(\frac{64}{100}\right)^2 \frac{64-9}{64} \times \frac{\frac{6}{9} \times \frac{3}{9}}{9-1} + \left(\frac{36}{100}\right)^2 \frac{36-6}{36} \frac{\frac{3}{6} \times \frac{3}{6}}{6-1} = 0.01518$$

$$\hat{\sigma}(\hat{p}_{STR}) = \sqrt{0.01518} = 0.1232, B_p = 2 \times 0.1232 = 0.2464 = 24.64\%$$

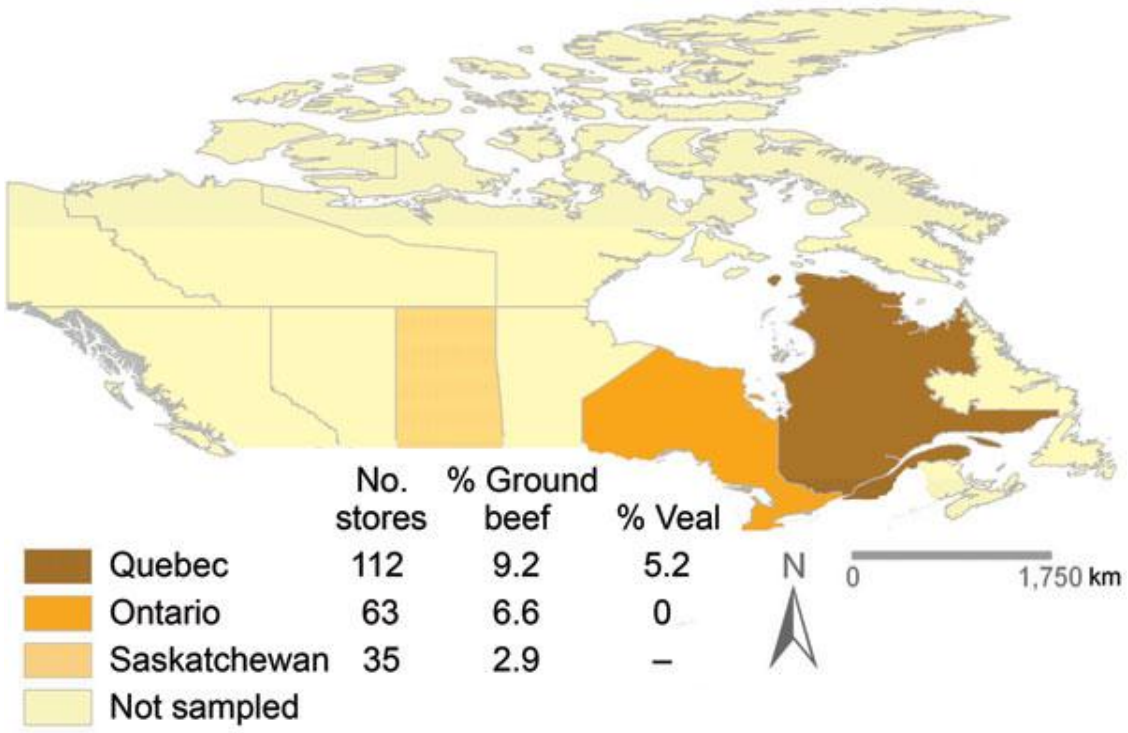
CI for p : $\hat{p}_{STR} \pm B_p = 60.67\% \pm 24.64\% = [36.03\%, 85.31\%]$

Your read

Study on contamination of retail meat in Canada by *Clostridium difficile* bacteria, in 2006

Clostridium difficile is a species of Gram-positive bacteria of the genus *Clostridium* that causes severe diarrhea and other intestinal disease when competing bacteria in the gut flora have been wiped out by antibiotics.

Random packages of ground beef as well as veal chops from milk-fed calves were tested; the packages were purchased in Ontario, Québec, and Saskatchewan, from January through August 2006.



Distribution of retail grocery stores sampled (*n* = 210) and proportion of sores with contaminated meat (not all stores have ground beef or veal). The overall proportion of stores with >1 meat package contaminated with *Clostridium difficile* was 5.7%.

Stratified Sampling: Optimal Stratification(I)

Principles (I)

Key values that affect the error of estimation in stratified sampling, and depending on the population are $N, N_i\ (W_i), \tilde{\sigma}_i^2\ (\sigma_i^2)$

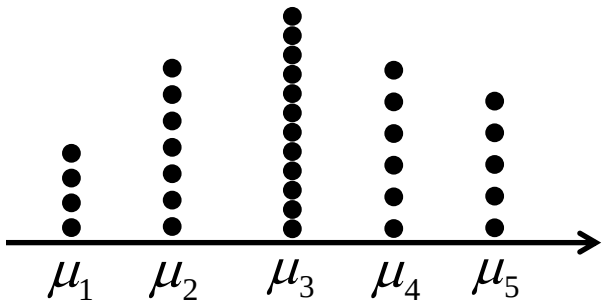
For given stratification, design effect comes from allocation $n_i\ (a_i)$

$$Var(\bar{y}_{STR}) = \sum_{i=1}^{L} W_i^2 \frac{N_i - n_i}{N_i} \frac{\tilde{\sigma}_i^2}{n_i} = \sum_{i=1}^{L} \frac{N_i - n_i}{N_i} \frac{(W_i \tilde{\sigma}_i)^2}{n_i}$$

Assuming that strata don't differ too much in size, what would be an ideal stratification?

Obviously, one that makes strata as homogeneous as possible, or $\tilde{\sigma}_i \approx 0$.

Example: You managed to split a larger group of people into five smaller groups of almost the same height. What minimal sample size do you need to accurately estimate the average height?



Stratified Sampling: Optimal Stratification(I)

Principles (II) $n_i\,(a_i)$

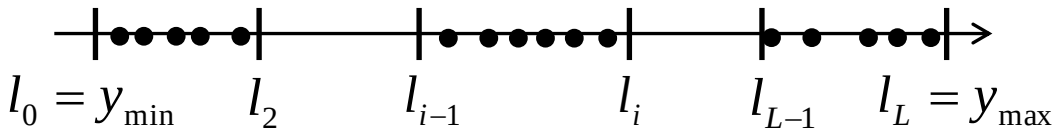
Assuming equal costs of sampling elements over strata, and using optimal (Neyman) allocation, we would have

$$Var(\bar{y}_{STR,OPT}) = \frac{1}{n}\left(\sum W_i\tilde{\sigma}_i\right)^2 - \frac{1}{N}\sum W_i\tilde{\sigma}_i^2 \approx \frac{1}{n}\left(\sum W_i\tilde{\sigma}_i\right)^2 \ (N \text{ large})$$

$$\min_{STR} Var(\bar{y}_{STR,OPT}) \Leftrightarrow \min_{STR} \sum W_i\tilde{\sigma}_i \ \ (\text{or just} \approx \min_{STR} \sum W_i\sigma_i)$$

for given number of strata, *L*

It is obvious from our example that we should stratify, idealy, by the variable of interest, y, or any other variable highly correlated with y.



Stratum

$$\mathcal{E}_i = \{e : l_{i-1} \le y(e) < l_i\}$$

Stratified Sampling: Optimal Stratification(II)

Optimal stratification (I)

Our problem now reduces on finding $\min_{l_1, l_2, \ldots l_{L-1}} \sum W_i \sigma_i$

How to find l_i ? Assume we have the distribution of y, in a form of density function, or histogram, $f(y)$:

$$W_i = \int_{l_{i-1}}^{l_i} f(y)dy, \mu_i = \int_{l_{i-1}}^{l_i} yf(y)dy / W_i, \sigma_i^2 = \int_{l_{i-1}}^{l_i} (y-\mu_i)^2 f(y)dy / W_i$$

$$\sum_i W_i \sigma_i = \sum_i \sqrt{\int_{l_{i-1}}^{l_i} f(y)dy \times \int_{l_{i-1}}^{l_i} y^2 f(y)dy - \left(\int_{l_{i-1}}^{l_i} yf(y)dy \right)^2}$$

$$\frac{\partial}{\partial l_j} \sum_i W_i \sigma_i = 0 \Leftrightarrow \frac{(l_j - \mu_j)^2 + \sigma_j^2}{\sigma_j} = \frac{(l_j - \mu_{j+1})^2 + \sigma_{j+1}^2}{\sigma_{j+1}}, j = 1,2,\ldots, L-1$$

Quite inconvenient for exact solving: iteration method used

Stratified Sampling: Optimal Stratification(II)

Optimal stratification: "square-root rule" (II)

The "cumulative square-root rule" is an approximate rule for selecting optimal strata limits. We will use an example:

i	Interval	f_i	$I_i = \Sigma_1^i \sqrt{f_j}$	Limit
1	**0**-50	18	$I_1 = \sqrt{18} = 4.24$	
2	50-100	23	$4.24 + \sqrt{23} = 9.04$	d =
3	100-**150**	20	$9.04 + \sqrt{20} = 13.51$	15.15
4	150-200	25	$13.51 + \sqrt{25} = 18.51$	
5	200-250	20	22.98	
6	250-300	16	26.98	$2d$ =
7	300-**350**	17	31.11	30.31
8	350-400	19	35.46	
9	400-450	14	39.21	
10	450-500	15	43.08	$3d$ =
11	500-**550**	10	46.24	45.46
12	550-600	13	49.85	
13	600-650	12	53.31	
14	650-700	8	56.14	
15	700-750	5	58.38	$4d$ =
16	750-**800**	5	$60.61 = I_{16}$	60.61
	0-800	240		

Population: 240 farms
Variable: Farm size (acres)
Given: Frequency distribution of farm sizes, f_i
Task: Find optimal stratification using $L = 4$ strata.

$$d = (Cum\sqrt{}) / 4 = I_{16} / 4$$
$$= 60.61/4 \approx 15.15$$
$$id = i \times 60.61/4$$

Strata limits :

$$l_0 = 0, l_1 = 150, l_2 = 350,$$
$$l_3 = 550, l_4 = 800$$

Stratified Sampling: Optimal Stratification(II)

Optimal stratification: "square-root rule" (III)

Stratification and sample size 40 optimal allocation:

Stratum	1	2	3	4	Population
Farm size	0 - 150	150 - 350	350 -550	550 - 800	-
Stratum size, N_i	61	78	58	43	240
Weight, W_i	25.42%	32.50%	24.17%	17.91%	100.0
Mean, μ_i	76.639	241.026	438.793	648.256	320.00
Variance, σ_i^2 (σ_i)	1554.690 (39.430)	3236.769 (56.893)	2999.405 (54.767)	4459.16 (66.777)	42766.667 (206.801)
Allocation, a_i	0.1866	0.3442	0.2464	0.2228	1.0000
$n_i = a_i$ x 40	7	14	10	9	40

Using mid points from class intervals

$$Var(\bar{y}_{STR,OPT}) \approx \frac{1}{n}\left(\sum W_i\sigma_i\right)^2 - \frac{1}{N}\sum W_i\sigma_i^2 = \frac{1}{40}(53.7113)^2 - \frac{1}{240}2970.8905 = 59.7439$$