Sampling with minimal strata size requirements

Stas Kolenikov, Abt SRBI Igor Griva, George Mason University

Problem statement

Consider a finite population $\mathcal{U}$ divided into H strata of sizes N_h, $h = 1, \ldots, H$, $N_1 + \ldots + N_H = N$. The following sampling problem is often encountered in practice: create a sampling design with the total sample size n and minimal strata sizes m_h, where $m = \sum_h m_h < n$, so that additional $n - m$ units need to be distributed across the strata.

To expand on Neyman (1934) optimal design framework, let the population variance and unit costs be S_h^2 and c_h for stratum h. If the costs are identical within strata, then the overall budget constraint is automatically replaced by the total sample size constraint. Let us parameterize the stratum sample size as

$$n_h = m_h + t_h \qquad (1)$$

where $t_h \geq 0$. Let the parameter of interest be the population mean $\bar{Y} = \sum_{i \in \mathcal{U}} Y_i / N$, estimated by

$$\bar{y}_{\rm str} = \sum_h W_h y_h, \quad W_h = N_h/N, \quad y_h = \sum_{i \in \mathcal{S}_h} y_i / n_h \qquad (2)$$

Then the sample design problem is

$$\mathbb{V}[\bar{y}_{\rm str}] = \sum_{h=1}^H W_h^2 \frac{S_h^2}{m_h + t_h} \to \min_{\{t_h\}} \qquad (3)$$

$$\text{s.t.} \sum_h c_h (m_h + t_h) = C, \qquad (4)$$

$$t_h \geq 0 \text{ for all } h \qquad (5)$$

$$C \geq \sum_h c_h m_h \qquad (6)$$

for the solution to exist.

See also Choudhry, Rao & Hidiroglou (2012).

Computation

1. Set the convergence criteria ϵ (e.g., $\epsilon = C \cdot 10^{-6}$).
2. Find the upper bound $\overline{\lambda} = \max_h \frac{W_h^2 S_h^2}{c_h m_h^2}$.
3. Find the lower bound $\underline{\lambda} = \min_h \frac{W_h^2 S_h^2}{c_h m_h^2}$.
4. If $C(\underline{\lambda}) \leq C$, none of the constraints in (5) are binding, and the optimal allocation is the Neyman-Chuprow allocation.
5. Set $\lambda^t \leftarrow \overline{\lambda}$, $\lambda^b \leftarrow \underline{\lambda}$, $k \leftarrow 1$.
6. Set $\lambda^{(k)} \leftarrow (\lambda^b + \lambda^t)/2$.
7. Compute $t_h(\lambda^{(k)})$ where
$$t_h(\lambda) = \max\Bigl[\frac{W_h S_h}{\sqrt{\lambda c_h}} - m_h, 0\Bigr], h = 1, \ldots, H \qquad (7)$$
8. Evaluate the budget constraint $C(\lambda^{(k)})$ where
$$C(\lambda) = \sum_h c_h [m_h + t_h(\lambda)] \qquad (8)$$
9. If $|C - C(\lambda^{(k)})| < \epsilon$, go to step 13.
10. If the sample size is too large (over budget), increase λ: set $\lambda^b \leftarrow \lambda^{(k)}$, $k \leftarrow k+1$.
11. If the sample size is too small (under budget), decrease λ: set $\lambda^t \leftarrow \lambda^{(k)}$, $k \leftarrow k+1$.
12. Re-iterate to step 6.
13. Set $t_h = t_h(\lambda^{(k)})$, rounding up to the integer part as needed.

Mock data

	Total pop	Hispanic pop	% Hispanic	S_h^2
CT	3,592,053	512,795	14.28%	0.12238
ME	1,328,535	18,592	1.40%	0.01380
MA	6,657,291	681,824	10.24%	0.09193
NH	1,321,069	40,301	3.05%	0.02958
NJ	8,874,374	1,649,784	18.59%	0.15134
NY	19,594,330	3,559,644	18.17%	0.14866
PA	12,758,729	784,562	6.15%	0.05771
RI	1,053,252	139,832	13.28%	0.11514
VT	626,358	10,226	1.63%	0.01606

Technicalities in the paper

- Established that the bounds $\overline{\lambda}, \underline{\lambda}$ contain the Lagrange multiplier for the overall cost constraint (6) of the constrained optimization problem
- Establish existence and uniqueness of the solution
- Establish convergence of the algorithm
- Establish equivalence to Neymann-Chuprow allocation when none of the minimal sample size constraints are binding
- Extensions for dual frame RDD are considered

$m_h = 1$, $C = 1000$

	$t_h(\lambda)$	n_h
CT	67.87	69
ME	7.55	9
MA	109.62	111
NH	11.45	13
NJ	188.21	190
NY	413.06	415
PA	166.98	168
RI	18.59	20
VT	3.35	5

$\underline{\lambda} = 2.02 \cdot 10^{-6}$

$\lambda = 1.069 \cdot 10^{-7}$

$\overline{\lambda} = 0.0183$

$m_h = 20$, $C = 1000$

	$t_h(\lambda)$	n_h
CT	46.49	67
ME	0	20
MA	86.80	107
NH	0	20
NJ	162.67	183
NY	379.73	400
PA	142.17	163
RI	0	20
VT	0	20

$\underline{\lambda} = 5.06 \cdot 10^{-9}$

$\lambda = 1.15 \cdot 10^{-7}$

$\overline{\lambda} = 4.58 \cdot 10^{-5}$

$m_h = 50$, $C = 1000$

	$t_h(\lambda)$	n_h
CT	7.81	58
ME	0	50
MA	42.86	93
NH	0	50
NJ	108.84	159
NY	297.58	348
PA	91.01	142
RI	0	50
VT	0	50

$\underline{\lambda} = 8.09 \cdot 10^{-10}$

$\lambda = 1.52 \cdot 10^{-7}$

$\overline{\lambda} = 7.33 \cdot 10^{-6}$

$m_h = 100$, $C = 1000$

	$t_h(\lambda)$	n_h
CT	0	100
ME	0	100
MA	0	100
NH	0	100
NJ	0	100
NY	99.61	200
PA	0	100
RI	0	100
VT	0	100

$\underline{\lambda} = 2.02 \cdot 10^{-10}$

$\lambda = 4.6 \cdot 10^{-7}$

$\overline{\lambda} = 1.83 \cdot 10^{-6}$

References

Choudhry, G. H., Rao, J. N. K. & Hidiroglou, M. A. (2012), 'On sample allocation for efficient domain estimation', *Survey Methodology* **38**(1), 23–29.

Neyman, J. (1934), 'On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection', *Journal of the Royal Statistical Society* **109**, 558–606.