A Simple Approach to Sample Allocation for Multivariate Stratified Sampling

Sun Woong Kim   Eun Jeong Nam   Young Sung Han

Dongguk University, South Korea
Korea Statistics Promotion Institute

Outline

- Sample Allocation in Stratified Random Sampling
- Problem of Sample Allocation with More Than One Survey Item
- Classical Methods of Sample Allocation with More Than One Survey Item
- Simplified Classical Methods
- Disadvantages of Simplified Classical Approaches
- Modification of Approach 5
- New Approach
- Illustration
- Conclusions

Sample Allocation in Stratified Random Sampling

- The sampler determines the values of the sample sizes n_h in the respective strata.
- If the cost per unit is the same in all strata, Neyman Allocation can be used for minimizing the variance.

$$n_h = n \frac{N_h S_h}{\sum_h N_h S_h} \ , \ h = 1, 2, \cdots, H$$

where N_h : stratum size

S_h : stratum standard deviation

Problem of Sample Allocation with More Than One Survey Item

- Neyman allocation will be the best for one variable.
- But his allocation will not in general be best for other variables in a survey with many variables (items)
- Some compromise needs to be reached in the allocation.

Classical Methods of Sample Allocation with More Than One Survey Item

- Yates (1960)

Approach 1.

Minimize the objective function $L = \sum_{j}^{k} a_j V(\bar{y}_{jst})$

subject to the constraint $C = c_0 + \sum_{h=1}^{H} n_h c_h$

where C : cost function

a_j : Importance weight

$V(\bar{y}_{jst})$: variance for item j

Classical Methods of Sample Allocation with More Than One Survey Item (Cont.)

Approach 2.

$$\textit{Minimize} \quad C = c_0 + \sum_{h=1}^{H} n_h c_h$$

subject to $V(\bar{y}_{jst}) < V_j \quad (j = 1, 2, \cdots, k)$ and $0 \le n_h \le N_h$

where V_j : desired variance (tolerance) for each item

Classical Methods of Sample Allocation with More Than One Survey Item (Cont.)

- Huddleston et al. (JRSS, 1970)

Approach 3.

$$Minimize \ \sum_{h=1}^{H} n_h c_h$$

subject to $V(\hat{Y}_j) = \sum_h N_h^2 S_{hj}^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) \le V_j \ (j = 1, 2, \cdots, k\ )$

and $0 \le n_h \le N_h$

where $V(\hat{Y}_j)$: variance of total estimate

Simplified Classical Methods

Assume that

1) the cost per unit is the same in all strata, that is, $c_1 = c_2 = \cdots = c_H$

2) the importance weight is the same in all items, that is,

$$a_j = 1 \ (j = 1, 2, \cdots, k \ )$$

We obtain

Approach 4: $Minimize \ L = \sum_j^k V(\bar{y}_{jst})$ subject to $2 \le n_h \le N_h$

Approach 5: $Minimize \sum_{h=1}^H n_h$

subject to $V(\bar{y}_{jst}) < V_j \ (j = 1, 2, \cdots, k)$ and $2 \le n_h \le N_h$

Disadvantages of Simplified Classical Approaches

- Although those approaches exactly correspond to the nonlinear programming (NLP) problems, they are often infeasible when solving by using NLP software.

- In a survey with many items, the tolerances V_j can often not be precisely specified.

(Example) Consider a bound of $B = z_{\alpha/2}\sqrt{V_j}$ on the error of estimation.

When $B = 0.05$ and $z_{0.05} = 1.96$, $V_j = \frac{0.05^2}{1.96^2} = 0.000651$

Disadvantages of Simplified Classical Approaches (Cont.)

- Interest would center simultaneously on the characteristics such as population mean, population proportion and population total, rather than a single characteristic. In these cases more complicated problems can arise.

Modification of Approach 5

When adding the condition (3) below, Approach 5 is always feasible.

$$Minimize \ \sum_{h=1}^{H} n_h$$

subject to (1) $V(\bar{y}_{jst}) < V_j \quad (j = 1, 2, \cdots, k)$

(2) $2 \le n_h \le N_h$

(3) $\sum_{h=1}^{H} n_h \le n_0$, where n_0 is a bound on the desired total sample size

Modification of Approach 5 (Cont.)

- This allocation would not be satisfactory because the solution can be less precise than Neyman allocation.

(The tolerances V_j would not provide enough quantity to be more precise than Neyman allocation)

New Approach: Four-Stage Sample Allocation

First stage.

For a given sample size n^*, find the $n^*_{median,h}$ as follows:

$$n^*_{median,h} = \text{Median}\{n^*_{Neyman,hj}, j = 1, 2, \cdots, k\}, h = 1, 2, \cdots, H$$

where $n^*_{Neyman,hj}$: Neyman allocation for each item

New Approach: Four-Stage Sample Allocation (Cont.)

Second stage.

Find the solution to $n_{NLP,hj}$ by using the following NLP for each item j

Minimize $V(\bar{y}_{jst}) = \frac{1}{N^2}\sum_{h=1}^{H} N_h(N_h - n_{NLP,hj})\frac{S_{hj}^2}{n_{NLP,hj}}$

subject to (1) $2 \le n_{NLP,hj} \le n_{median,h}$

(2) $\sum_{h=1}^{H} n_{NLP,hj} \le \sum_{h=1}^{H} n_{median,h}$

- $V(\bar{p}_{jst})$ or $V(\hat{Y}_j)$ as well as $V(\bar{y}_{jst})$ is available.
- $\sum_{h=1}^{H} n_{median,h}$ can be smaller or larger than n^*.

New Approach: Four-Stage Sample Allocation (Cont.)

Third stage.

Find n_h and n as follows:

$$n_h = \text{Median}\{n_{NLP,hj},\ j = 1, 2, \cdots, k\},\quad h = 1, 2, \cdots, H$$

$$n = \sum n_h$$

- $n = \sum n_h$ would be smaller than n^*

New Approach: Four-Stage Sample Allocation (Cont.)

Fourth stage.

Find Neyman allocation by using n and then find the $n_{median,h}$ as follows:

$$n_{median,h} = \text{Median}\{n_{Neyman,hj}\ , j=1,2,\cdots,k\}, h=1,2,\cdots,H$$

where $n_{Neyman,hj}$: Neyman allocation for each item

Illustration: Donnguk University Time Use Survey

- Sponsor: Dongguk University
- Collector: Survey Research Center, Dongguk University
- Purpose: To investigate undergraduate students' time use at school or home, and how their activities relate to their curriculum and classes
- Sampling frame: A list of registered students
- Frame population size: about 13,000
- Sample design: Stratified random sampling (11 strata)
- Mode: Computer-assisted cell phone interviews
- Total number of survey items: 48

dongguk UNIVERSITY

Illustration (Cont.)

- Number of survey items thought to be most important: 9

- List of 9 items

  Estimation of proportions:

  A. choosing double major or minor
  B. attending a private institute for learning foreign languages
  C. having club activities
  D. having part-time jobs
  E. personal consultation with professors
  F. smoking

  Estimation of means:

  G. satisfaction with school
  H. school assessment
  I. satisfaction with department

Illustration (Cont.)

Using modification of Approach 5

Constraints:

- The bound on the error of estimation:

  $\pm 5\%$ points for proportions

  ± 0.10 for means

- The upper bound on the desired total sample size: $n_0 = 450$
- The lower bound on the stratum sample size: $n_h = 20$

Illustration (Cont.)

Sample Allocation: Neyman Allocation vs. Modification of Approach 5

	Neyman Allocation									App. 5
	A	B	C	D	E	F	G	H	I	
n1	10	16	8	8	10	12	17	10	9	20
n2	62	32	49	51	52	30	46	47	47	20
n3	30	16	26	25	26	25	26	28	25	20
n4	13	37	25	27	25	29	20	22	20	20
n5	99	90	84	83	74	86	85	79	94	87
n6	63	58	57	56	48	56	58	53	56	38
n7	29	19	18	25	24	14	20	26	25	20
n8	67	112	109	101	114	125	103	98	107	165
n9	46	21	36	39	36	34	36	36	32	20
n10	11	24	18	19	20	25	18	22	20	20
n11	20	25	20	16	21	14	21	29	15	20
Total	450	450	450	450	450	450	450	450	450	450

Illustration (Cont.)

Design Effect: Neyman Allocation vs. Modification of Approach 5

	Neyman Allocation									App. 5
deff	A	B	C	D	E	F	G	H	I	
A	0.838	1.102	0.920	0.896	0.931	1.045	0.910	0.909	0.916	1.310
B	1.322	0.974	1.050	1.057	1.056	1.025	1.064	1.065	1.082	1.141
C	1.058	1.023	0.928	0.930	0.939	0.976	0.948	0.948	0.951	1.176
D	1.053	1.050	0.947	0.931	0.950	1.009	0.965	0.958	0.960	1.224
E	1.087	1.047	0.950	0.942	0.936	1.011	0.963	0.951	0.965	1.183
F	1.148	0.941	0.919	0.925	0.929	0.883	0.937	0.947	0.937	1.020
G	1.030	1.023	0.955	0.958	0.960	0.989	0.940	0.958	0.963	1.149
H	1.055	1.053	0.969	0.958	0.964	1.039	0.973	0.952	0.988	1.191
I	0.995	0.978	0.909	0.902	0.918	0.949	0.917	0.921	0.901	1.092

Illustration (Cont.)

Using new approach: *first stage*

	$n^{*}_{median,h}$
n1	10
n2	47
n3	26
n4	25
n5	85
n6	56
n7	24
n8	107
n9	36
n10	20
n11	20
Total	456

Illustration (Cont.)

Using new approach: *second stage*

	$n_{NLP,hj}$								
	A	B	C	D	E	F	G	H	I
n1	8	3	9	9	10	10	10	9	9
n2	24	12	22	22	34	23	37	22	22
n3	16	8	15	15	23	16	25	15	15
n4	12	11	15	15	23	15	25	15	15
n5	34	21	35	35	54	37	59	35	35
n6	26	16	25	25	39	26	42	25	25
n7	15	8	14	14	22	15	24	14	14
n8	33	25	42	42	66	45	71	42	42
n9	20	10	18	18	28	19	31	18	18
n10	10	8	13	13	20	14	20	13	13
n11	4	4	13	13	20	14	20	13	13
Total	202	126	221	221	339	234	364	221	221

Illustration (Cont.)

Using new approach: *third stage*

	n_h
n1	9
n2	22
n3	15
n4	15
n5	35
n6	25
n7	14
n8	42
n9	18
n10	13
n11	13
Total	221

dongguk UNIVERSITY

Illustration (Cont.)

Using new approach: *fourth stage*

	Neyman Allocation									$n_{median,h}$
	A	B	C	D	E	F	G	H	I	
n1	5	8	4	4	5	6	8	5	5	5
n2	30	16	24	25	25	15	23	23	23	23
n3	15	8	12	12	13	13	13	14	12	13
n4	6	18	12	13	12	14	10	11	10	12
n5	49	44	41	41	36	42	42	38	46	42
n6	31	29	28	28	24	27	28	26	27	28
n7	14	9	9	12	12	7	10	13	12	12
n8	33	55	54	50	56	61	51	48	53	53
n9	23	10	18	19	18	17	17	18	16	18
n10	5	12	9	9	10	12	9	11	10	10
n11	10	12	10	8	10	7	10	14	7	10
Total	221	221	221	221	221	221	221	221	221	226

Illustration (Cont.)

Design Effect: Neyman Allocation vs. New Approach

deff	Neyman Allocation									New App.
	A	B	C	D	E	F	G	H	I	
A	0.866	1.145	0.949	0.923	0.954	1.070	0.937	0.932	0.947	0.913
B	1.419	1.003	1.087	1.093	1.090	1.062	1.093	1.096	1.115	1.050
C	1.110	1.056	0.955	0.955	0.963	1.001	0.972	0.971	0.979	0.935
D	1.110	1.087	0.976	0.959	0.974	1.036	0.991	0.981	0.989	0.946
E	1.144	1.084	0.979	0.972	0.962	1.038	0.988	0.974	0.987	0.947
F	1.217	0.971	0.948	0.953	0.952	0.908	0.960	0.970	0.960	0.922
G	1.035	1.022	0.954	0.956	0.958	0.984	0.941	0.957	0.960	0.926
H	1.064	1.053	0.968	0.968	0.963	1.031	0.973	0.951	0.991	0.938
I	1.002	0.978	0.910	0.902	0.917	0.945	0.914	0.921	0.902	0.883

Conclusions

- New NLP approach based on Neyman allocation is simple to use.
- New approach would provide a satisfactory compromise allocation to be more precise than Neyman allocation for each item.
- New approach may provide the smaller sample size than expected, resulting in saving costs.

Thank you.

Contact at sunwk@dongguk.edu