Sample Size

Dr. A.C. Kulshreshtha
U.N. Statistical Institute for Asia and the Pacific (SIAP)

Second RAP Regional Workshop on
Building Training Resources for Improving Agricultural & Rural Statistics
Sampling Methods for Agricultural Statistics-Review of Current Practices
SCI, Tehran, Islamic Republic of Iran
10-17 September 2013

The Problem:

Determine sample size, n

- Ensuring a required level of precision
- Most efficient and largest for a given fixed budget, B
- Survey budget, B, is fixed, no matter what the variance is
- Upper limit to the variance of an estimator is fixed at V_o whatever B is

Possibilities

- Minimize $\mathrm{Var}(\hat{\theta})$ subject to fixed B
- $\mathrm{Var}(\hat{\theta})$ inversely depends on ‘n’, call it V(n)
- Thus, we have an optimization problem:
  - Find ‘n’ such that V(n) is minimum subject to cost function C(n)

Factors affecting ‘optimal’ n

- Required precision of estimates– higher precision desired, larger sample size needed
  - Variability of characteristic being measured– more variable, larger sample size
  - Rare characteristic– more rare, larger sample size
- Population size N (sampling fraction)– no effect on sample size if N is large

Effect of N

- For example, SRSWR:

$$\sigma^2(\hat{\bar{Y}}) = \frac{\sigma^2(Y)}{n}$$

➔Precision of sample mean does not depend on population size N

➔Precision of sample mean depends only on variability of population values

Effect of N (Contd.)

- For example, SRSWOR:

$$\sigma^2(\hat{\bar{Y}}) = \frac{\sigma^2(Y)}{n} * \frac{N-n}{N-1}$$

→ For large N, (N-n)/(N-1) →1

Factors affecting ‘optimal’ n

- Cost– larger sample size ➔ higher cost

Example:

- Simple cost function: $C = C_0 + n*C_1$

  where C= total cost of survey; C_0 is fixed cost; C_1 is cost per sample unit; n is sample size

- For given total budget C’: $$n = \frac{C' - C_0}{C_1}$$

Factors affecting 'optimal' n (Contd.)

- Level of detail required
  - More Reporting domains ➔ larger sample size needed
  - More subclasses (for analysis) ➔ larger sample size needed

Basic Steps

Basic Steps for determining n

- How much precision is desired? Or, how much ‘error’ is tolerable?
- Relate sample size (n) and precision or error requirements (an equation based on sampling theory)
- For this equation, estimate the unknown quantities (usually, variances of population) and solve for value of n→ n*

Basic Steps (Contd.)

- Allocate to domains, strata, (subclasses)
- Adjust for precision requirements for estimates for domains, strata→ n**
  - Note: Initial computations may start with sample size requirements for each domain, stratum, etc.
- Are there sufficient resources for data collection on n** units? If not, readjust requirements of precision, reallocate within resource constraints sample size

Initial Computations

- Determine sample size required for SRSWR– n(srs)
- Adjust n(srs), if N is relatively small:

$$n \geq \frac{n_{SRS}}{1 + \frac{n_{SRS}}{N}}$$

- Adjust n to allow for a more complex sample design using the *deff* of the design; n(complex)= n * deff
- Adjust n(complex) to take into account expected non-response rates, n(adj) = n(complex) * (1+nonresponse rate)

Initial Computations- Example

- Example of adjustment for cluster sampling:
  - n(srs) = 200
  - deff(cluster) = 2.0
  - n(cluster) = 200*2.0 = 400
  - Expect nonresponse rate = 0.20
  - n(adj) = 400*(1+.20) = 480

Determining n(srs)

- How much precision do I need? Or, how much error is tolerable?

a. Variance of estimate should not exceed a given value V_0
b. Margin of error, e, should be met with a given probability
c. Width of confidence interval should not exceed a prescribed amount, w
d. CV (or RSE) should not exceed a given value

Sample size in SRS
n(SRS)– Estimation of Population Mean

- a. Variance of sample mean should not exceed V_0

$$V(\hat{\bar{Y}}) \leq V_0 \Rightarrow \frac{S^2}{n_{SRS}} \leq V_0 \Rightarrow n_{SRS} \geq \frac{S^2}{V_0}$$

Adjust for small N: $$n \geq \frac{n_{SRS}}{1+\frac{n_{SRS}}{N}}$$

n(SRS)– Estimation of Population Mean (Contd.)

- b. Margin of error, e, should be met with given probability.

$$\Pr ob\left\{\left|\hat{\overline{Y}} - \overline{Y}\right| \leq e\right\} = 1 - \alpha$$

$$\Rightarrow e^2 = z_{\alpha/2}^2 \cdot V(\hat{\overline{Y}}) \Rightarrow n_{SRS} = \left(\frac{z_{\alpha/2} \cdot S}{e}\right)^2$$

n(SRS)– Estimation of Population Mean (Contd.)

- Values of α
  - = 0 (100% confidence level) ➔ $z_{\alpha/2} = 3$
  - = 0.05 (95% confidence level) ➔ $z_{\alpha/2} = 1.96$
  - = 0.10 (90% confidence level) ➔ $z_{\alpha/2} = 1.645$

Note: Assumption is that sampling distribution of sample mean is normal distribution

n(SRS)– Estimation of Population Mean (Contd.)

- c. Width of confidence interval should not exceed w

$$\Pr ob[\hat{\bar{Y}} - z_{\alpha/2}SE(\hat{\bar{Y}}) \le \bar{Y} \le \hat{\bar{Y}} + z_{\alpha/2}SE(\hat{\bar{Y}})] = 1 - \alpha$$

$$\Rightarrow 2z_{\alpha/2}SE(\hat{\bar{Y}}) \le w$$

$$\Rightarrow n_{SRS} \ge 4(\frac{z_{\alpha/2}S}{w})^2$$

n(SRS)– Estimation of Population Mean (Contd.)

- d. CV of sample mean should not exceed CV_0

$$\Rightarrow n_{SRS} \geq \left(\frac{CV(\overline{Y})}{CV_0} \right)^2$$

n(srs)- Estimation of Proportions

- Specified maximum variance, V_0: $n_{SRS} \geq \frac{P(1-P)}{V_0}$
- Given margin of error, e: $n_{SRS} \geq \frac{z^2_{\alpha/2} P(1-P)}{e^2}$
- Specified maximum CV, CV_0: $n_{SRS} \geq \frac{(1-P)}{P(CV_0)^2}$

Note: Can use P=0.5 if no information on P

Sample Size in Stratified Sampling

- Optimum allocation for a specified variance, V_0:

$$n = \frac{\left(\sum_h N_h S_h\right)^2}{V_0 + \sum_h N_h S_h^2}$$

- Proportional allocation for a specified variance, V_0:

$$n = \frac{N \sum_h N_h S_h^2}{V_0 + \sum_h N_h S_h^2}$$

Sample Size in Stratified Sampling (Contd.)

- Cost-optimum allocation for a specified variance, V_0, and given cost where C_h is average variable cost per sample unit in stratum h :

$$n = \frac{\left[\sum_h N_h S_h \sqrt{C_h}\right]\left[\sum_h N_h S_h (1/\sqrt{C_h}\right]}{V_0 + \sum_h N_h S_h^2}$$

Sample Size in Cluster Sampling

Effect of clustering on variance

- More similar the elements within each cluster, the larger the *deff* of cluster sample; i.e., cluster sampling is less efficient compared to srs
- Sample size needed for a clustered sample for same precision as n(srs) is:
  - n(cluster) = n(srs) * *deff*

Sample Size in Cluster Sampling (Contd.)

- In cluster sampling and two-stage sampling, need to determine:
  - Size of PSU
  - Number of SSUs to be sampled in each sample PSU
  - Number of PSUs to be sampled

Sample Size in Cluster Sampling (Contd.)

- Size of PSU
  - Larger PSUs, smaller ρ and smaller *deff*
  - Too large PSUs, loose cost savings of cluster sampling
- Subsampling rate
  - In general, balance costs for sampling PSU and SSU and precision requirements

Sample size for One-stage Cluster Sampling

Exact	Approximate
$n = \dfrac{z^2 N V_{1y}^2}{z^2 V_{1y}^2 + (N-1)\varepsilon^2}$ $V_{1y}^2 = \dfrac{\sigma_{1y}^2}{\bar{y}^2}$	$n \cong \dfrac{z^2 V_{1y}^2}{\varepsilon^2}$

$$\hat{\sigma}_{1y}^2 = \frac{N-1}{N} s_y^2$$

$$s_y^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\bar{y}_i - \bar{y}_{\text{clu}})^2$$

Sample Size for Two Stage Cluster Sampling

Let $V_{1y}^2 = \frac{\sigma_{1y}^2}{\bar{y}^2}, \qquad V_{2y}^2 = \frac{\sigma_{2y}^2}{\bar{y}^2}$

$\bar{M}$ = average listing units

Suppose $\bar{m}$ is known (later we show how to estimate $\bar{m}$)

$\bar{m}$ = number of listing units sampled from each cluster

$$n = \frac{\left(\frac{N}{N-1}\right)V_{1y}^2 + \left(\frac{\bar{M}-\bar{m}}{\bar{m}(\bar{M}-1)}\right)V_{2y}^2}{\frac{\varepsilon^2}{z^2} + \left(\frac{1}{N-1}\right)V_{1y}^2}$$

Two stage sample size, $\overline{m}$

Need to know the relative costs of first and second stage sampling

It also depends on variance of '*y*' between first-stage units, i.e. σ^2_{1y} and variance of '*y*' within second-stage units, i.e. σ^2_{2y}.

$$\sigma^2_{1y} = \text{Variance between PSUs} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \overline{Y})^2$$

$$\sigma^2_{2y} = \text{Variance of Y within SSUs} = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{\overline{M}}(Y_{ij} - \overline{Y}_i)^2$$

Cost Function Two Stage Sampling

for ‘n’ PSU’s and ‘ $\overline{m}$’ SSU’s within each PSU, cost function is:

$$C = C_1^* n + C_2^* n\overline{m}$$

Cost of sampling a unit at first-stage

Cost of sampling a unit at second-stage

Costs in Two Stage Sampling

Cost of sampling a unit at first-stage	Cost of sampling a unit at second-stage
• Cost of traveling to each sample cluster • Listing $\bar{M}$ SSU's, and cost of selecting a sample $\bar{m}$ units from each cluster • Going back to cluster for interview or measurement	• Cost of interview or measurement for a sampling unit

Example--

- It costs ‘0.5’ person-hour to travel to each sample cluster
- It costs ‘1.0’ person-hour to list the ‘20’ SSU within the cluster and then select a random sample
- It costs ‘0.5’ person-hour to return to clusters

Then: $$C_1^* = 0.5 + 1.00 + 0.5 = 2.00$$

- It costs ‘0.25’ person-hour to interview or measure a sampling unit:

$$C_2^* = 0.25$$

$$Thus:$$

$$C = 2.00\,n + 0.25\,n\overline{m}$$

Two Stage Sample Size

- For values of $\overline{m}$, use the previous formula to estimate '*n*':

$$n = \frac{\left(\frac{N}{N-1}\right)V_{1y}^2 + \left(\frac{\overline{M}-\overline{m}}{\overline{M}-1}\right)\left(\frac{1}{\overline{m}}\right)V_{2y}^2}{\frac{\varepsilon^2}{\mathrm{z}^2} + \frac{1}{N-1}V_{1y}^2} \qquad (1)$$

This meets the accuracy and confidence condition for a given $\overline{m}$.

- For this specific solution, compute:

$$C = C_1^* n + C_2^* n\overline{m} \qquad (2)$$

Two Stage Sample Size (cont'd)

- Repeat this calculation for all possible combinations of '$\overline{m}$' and 'n'.
- Eliminate those combinations that do not meet the accuracy specification, using (1).
- Make a table of $\overline{m}$, n, and cost.
- Identify the pair ($\overline{m}$, n) with lowest cost.

Example

Selected $\bar{m}$	'n' from equation (1)	Field cost from equation (2)	Minimum Cost
6	5	17.5	
7	4	15.0	
8	4	16.0	
9	3	⋮	
⋮	⋮	⋮	
⋮	⋮	⋮	
19	1	6.75	6.75
20	1	7.0	

Optimum sampling and sub-sampling fractions

- $C = c_1 n + c_2 nm$

$$V(\bar{y}_{ts}) = \frac{1}{n}\left(S_b^2 - \frac{S_w^2}{M}\right) + \frac{1}{mn} S_w^2 - \frac{1}{N} S_b^2$$

$$m_{opt} = \frac{S_w}{\sqrt{S_b^2 - \{S_w^2 / M\}}} \sqrt{c_1 / c_2}$$

Optimum sampling and sub-sampling fractions (Contd.)

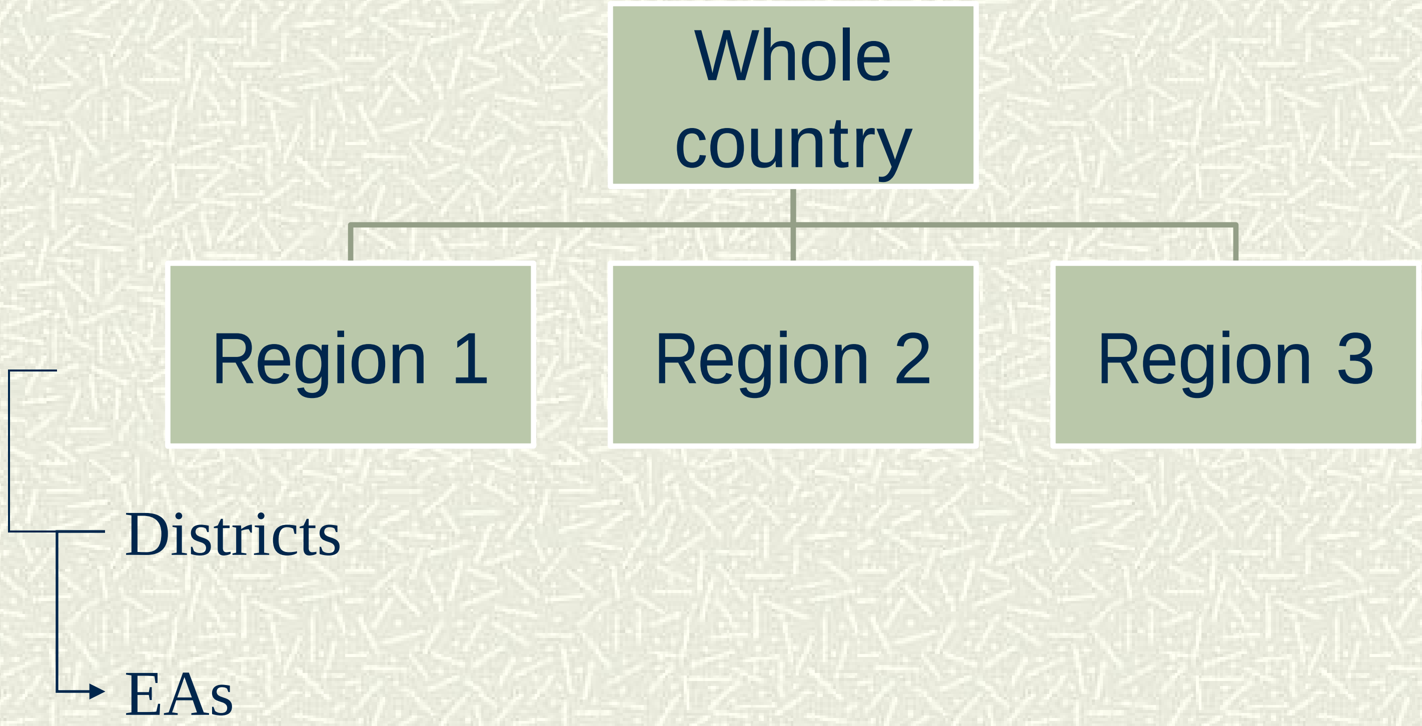
provided

$$S_b^2 > S_w^2 / M$$

Values of n is found by solving either the cost equation or the variance equation

Sample Allocation to Domains

- For example:



Sample Allocation to Domains

- One approach
  - Calculate sample size requirements for each domain
  - Add up the individual sample size requirements to get total sample size
  - Adjust depending on resource constraints

Sample Allocation to Domains - Strata

- Given n, allocation into strata

Proportional allocation: $$n_h = \frac{N_h}{N} \cdot n^*$$

Optimum or Neyman allocation: $$n_h = n^* \cdot \frac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}$$

Cost-optimum allocation: $$n_h = n^* \cdot \frac{N_h S_h (1/\sqrt{C_h})}{\sum_{h=1}^{H} N_h S_h (1/\sqrt{C_h})}$$

Sample Allocation to Domains (Contd.)

- Some considerations:
  - Need for minimum and maximum sample sizes
  - Domains may differ in importance– may require more precise estimates for some domains
  - Some domains may be more heterogeneous than others with greater underlying variability of study variables
  - Survey costs may differ among domains

Sample Allocation to Domains (Contd.)

Sector/Subsector	SIZE		
	Small	Medium	Large
Total Manufacturing			
Food and beverages			
Wearing apparel			
Wood products			
Plastic products			
Other manufacturing			

- Note: Need at least two sampling units (minimum) per cell. For cells with many establishments, specify maximum number. Typically, all large establishments are selected. Allocate remaining sample size to the cells.

Sample Allocation to Domains (Contd.)

- Optimum allocation (or in many cases, proportional allocation) gives required precision for whole population (e.g., whole country; total trade establishments) but may not give required precision for all domains (e.g., regions; trade subsectors)
- Equal allocation is ideal for comparison of domain estimates but may not be "representative" at the population level

Sample Allocation to Domains (Contd.)

- Compromise between equal allocation for each domain and optimum allocation

- For example, allocate sample size to domain h proportional to square root of its size:

$$n_h = \frac{n}{\sum_h \sqrt{M_h}} \cdot \sqrt{M_h}$$

Sample size: Other Issues

- Different survey variables may have different sample size requirements for a given desired precision
  - Prioritise and select the critical study variables
  - Compute required sample size for each
  - Adopt the largest sample size required

- Finally, Sample size determination and allocation is an iterative process.

THANK YOU