

Adaptive Cluster Sampling

Adaptive Sampling: sampling designs in which the procedure for selecting sites or units to be included in the sample may depend upon values of the variable of interest observed during the survey.

Goal: to take advantage of population characteristics so as to obtain more precise estimates of population values for a given sample size or cost relative to conventional designs.

Secondary advantage: increase the yield of interesting observations which may result in better estimates of other parameters of interest.

In contrast to conventional sampling designs, adaptive sampling makes use of values observed in the sample. Although sequential sampling looks at the data, the information obtained is used to decide how many more units to sample or whether or not to stop sampling. In contrast, adaptive designs tell which units to sample.

In adaptive sampling, the sampler specifies

4. the initial sampling design (prior to any adaptive sampling)
5. the initial sample size
6. the description of the neighborhood for a sampling unit
7. the condition that triggers or initiates adaptive sampling at a sampled unit

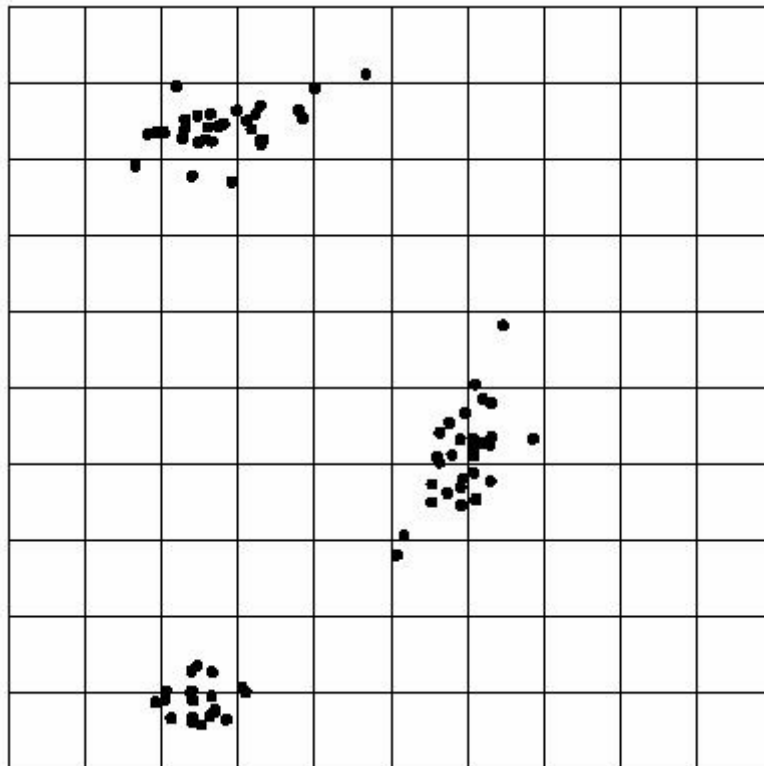
Notes:

- The neighborhood is specified in advance of sampling and is not adaptive.
- If the y -value (response) of a sampled unit satisfies the condition for adaptive sampling, say, $y > C$, then the unit's neighborhood is added to the sample.
- If any other units in that neighborhood satisfy C , then their neighborhoods are also added to the sample.
- The process continues until a cluster of units is obtained that contains a “boundary” of edge units that do not satisfy C .
- The final sample consists of n_1 , not necessarily distinct clusters, one for each unit selected in the initial sample.
- If many of the units satisfy the condition, then the sample could consist of most of the units in the population, and hence be very costly. Thus, the design is most appropriate when the characteristic of interest is highly aggregated or clustered.

1.0 Illustration of methodology

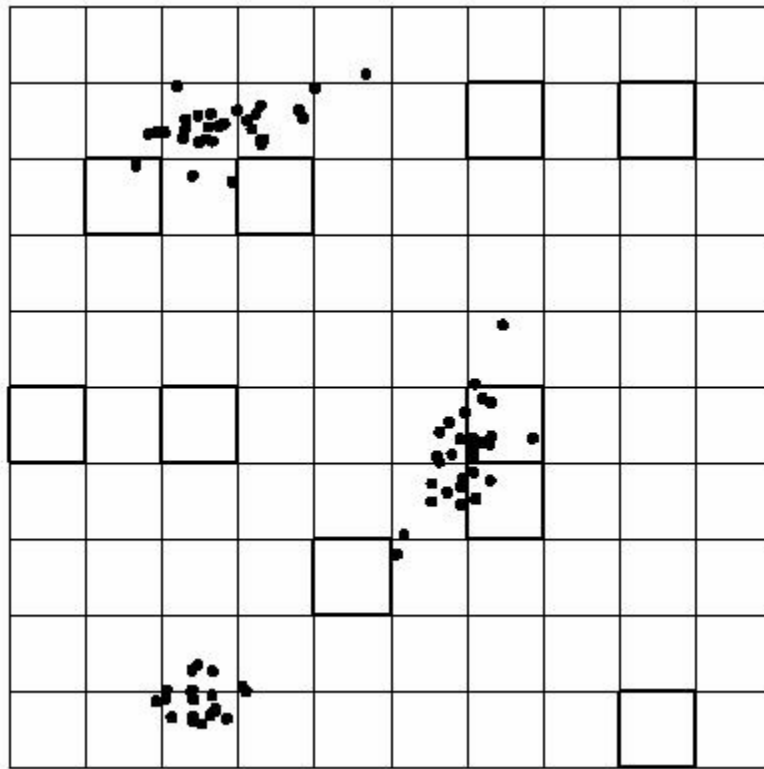
It is easiest to work from an example. Suppose that we are studying a particular weed that grows in strawberry fields. The weed is not particularly abundant, but serves as a host plant for a disease of strawberries. The scientist would like to estimate the total (and average) number of weeds in the field using adaptive cluster sampling. He divides the field up using a grid system to produce square contiguous sampling units.

FIGURE 2. Hypothetical strawberry field divided into sampling units using a grid with weeds identified as points in the map.



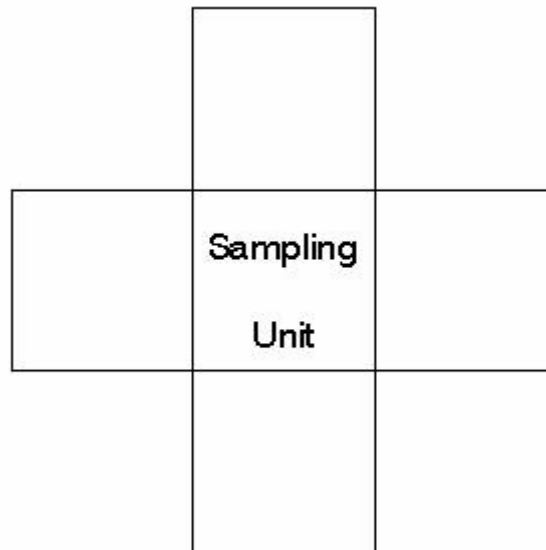
We will initially take a simple random sample of sampling units. These are now indicated in the next map.

FIGURE 3. Field map with sampling units in the simple random sample highlighted.



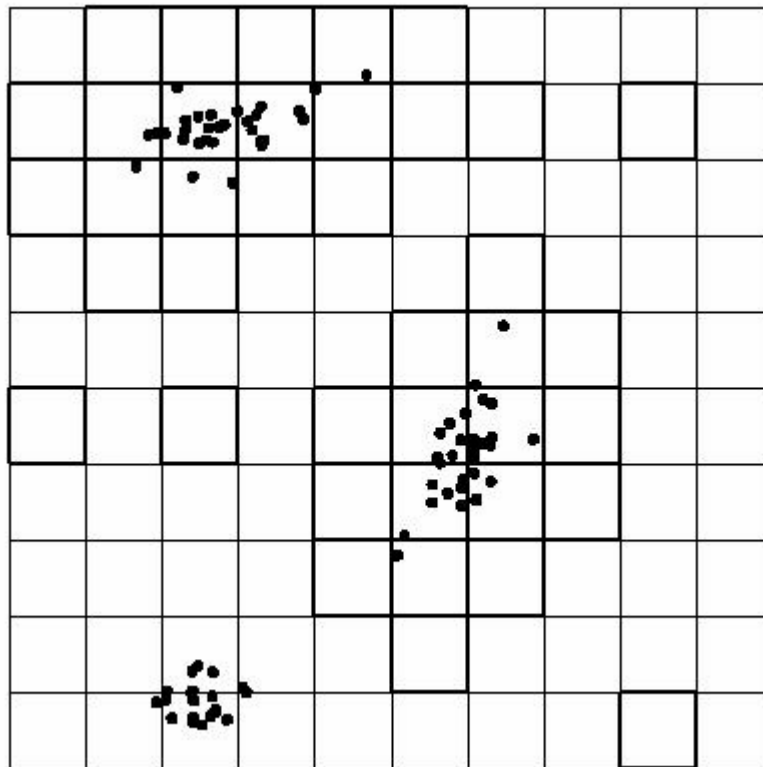
In adaptive sampling we need to define a neighborhood for a sampling unit. A neighborhood is some pre-specified rule for associating other sampling units with each sampling unit in the frame. For our purposes, the neighborhood will be defined the same way for each sampling unit. Specifically we will define the neighborhood to be the (usually) 4 contiguous (common edges) sampling units for a given unit. This is illustrated below.

FIGURE 4. Neighborhood for a sampling unit in the strawberry study.



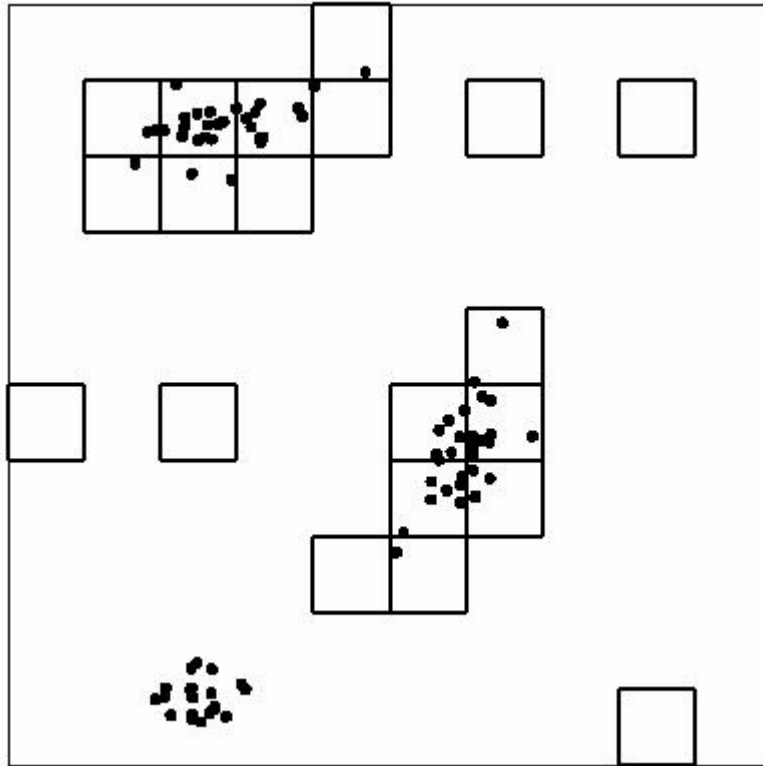
We also need to specify a criterion C for searching the neighborhood. For the example, we will take C to be the condition $y_i > 0$, where y_i is the number of weeks in the i th sampling unit. Now, following the instructions for adaptive cluster sampling, if a weed is present in one of our initially sampled units, then we are to also examine its neighborhood. We then repeat this process for each neighborhood unit that we sample, until there are only edge units (no weeds present) surrounding any cluster of weeds. This situation is illustrated in the following figure by showing all sampling units ultimately examined using the adaptive process.

FIGURE 5. Strawberry field showing all sampling units inspected using the adaptive sampling process from the initial simple random sample of units.



Now that the adaptive sample is complete, we can identify the sampled networks. A network consists of the initial sampling unit and any other sampling units in its neighborhood meeting the specified condition that are identified through the adaptive sampling process. Thus, initial units that do not have the weed present are networks of size 1. The networks identified in our sampling example are shown in the next figure.

FIGURE 6. Strawberry field showing all identified networks using the adaptive sampling process.



Using this information, we can now proceed to the estimation methodology.

2.0 Estimator Based Upon Initial Intersection Probabilities

We will follow the notation of Thompson and Seber (1996).

- Let N be the number of sampling units in the population.
- Let y_i be the response on unit i . For the example this would be the number of weeds growing in the sampling unit.
- Let A_i be the network for sampling unit i .
- Let m_i be the number of sampling units in A_i .
- Let C be the condition that when satisfied, that sampling unit's network is added to the sample.
- Let a_i be the total number of sampling units in networks of which sampling unit i is an edge unit. If unit i satisfies C , then $a_i = 0$. If unit i does not satisfy C , then $m_i = 1$.
- Let n_1 be the number of networks in the sample.

The probability that unit i is included in the sample is

$$\pi_i = 1 - \left[\binom{N - m_i - a_i}{n_1} / \binom{N}{n_1} \right].$$

The m_i are known for the units in the sample, but some of the a_i are unknown -- this requires sampling around the “empty” sampling units which we do not do. To work around this problem, since we need a probability of selection, we compute the “partial” inclusion probability

$$\pi'_i = 1 - \left[\binom{N - m_i}{n_1} / \binom{N}{n_1} \right]$$

which only uses known information and is the probability that the initial sample intersects A_i .

Observations that do not satisfy the condition C are ignored if they are not included in the initial sample. An unbiased estimator for the population mean is then given by

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^N \frac{y_i I'_i}{\pi'_i}$$

where I'_i is an indicator variable that is 1 if the initial sample intersects A_i and 0 otherwise. This can be rewritten in terms of the distinct networks as

$$\hat{\mu} = \frac{1}{N} \sum_{k=1}^K \frac{y_k^* J_k}{\alpha_k} = \frac{1}{N} \sum_{k=1}^{\kappa} \frac{y_k^*}{\alpha_k}$$

where y_k^* is the sum of the y_k values for the k th network, K is the total number of distinct networks in the population, κ is the number of distinct networks in the sample, and J_k is another indicator function that is 1 if the initial sample intersects the k th network and 0 otherwise.

If there are x_k sampling units in the k th network, then

$$\alpha_k = 1 - \left[\binom{N - x_k}{n_1} / \binom{N}{n_1} \right].$$

Also, when $x_k = 1$, then $\alpha_k = n_1 / N$. Further, define p_{jk} to be the probability that the j th and k th networks are not intersected, then

$$p_{jk} = \binom{N - x_j - x_k}{n_1} / \binom{N}{n_1}$$

so that the joint probability that networks j and k are both intersected is

$$\alpha_{jk} = \alpha_j + \alpha_k - (1 - p_{jk}) = 1 - \left\{ \binom{N-x_j}{n_1} + \binom{N-x_k}{n_1} - \binom{N-x_j-x_k}{n_1} \right\} / \binom{N}{n_1}.$$

Now the variance can be derived as

$$\text{Var}(\hat{\mu}) = \frac{1}{N^2} \left[\sum_{j=1}^K \sum_{k=1}^K y_j^* y_k^* \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_j \alpha_k} \right) \right]$$

with estimator

$$\begin{aligned} \text{var}(\hat{\mu}) &= \frac{1}{N^2} \left[\sum_{j=1}^K \sum_{k=1}^K y_j^* y_k^* \left(\frac{\alpha_{jk} - \alpha_j \alpha_k}{\alpha_j \alpha_k} \right) J_j J_k \right] \\ &= \frac{1}{N^2} \left[\sum_{j=1}^K \sum_{k=1}^K \frac{y_j^* y_k^*}{\alpha_{jk}} \left(\frac{\alpha_{jk}}{\alpha_j \alpha_k} - 1 \right) \right] \end{aligned}$$

where $\alpha_{jj} \equiv \alpha_j$.

3.0 Estimator Using Numbers of Initial Intersections

Let f_i be the number of units in the initial sample that fall in the network A_i that includes unit i . Ignoring the edge units of clusters in the estimator process, f_i is the number of times that the i th unit in the final sample appears in the estimator. An unbiased estimator of the population mean is

$$\hat{\mu} = \frac{1}{n_1} \sum_{i=1}^N \frac{y_i f_i}{m_i}$$

which can be rewritten in terms of the n_1 not necessarily distinct networks intersected by the initial sample,

$$\hat{\mu} = \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{1}{m_i} \sum_{j \in A_i} y_j = \frac{1}{n_1} \sum_{i=1}^{n_1} w_i = \bar{w}$$

where w_i is the mean of the m_i observations in A_i . This estimator has variance

$$\text{Var}(\hat{\mu}) = \frac{N - n_1}{Nn_1(N - 1)} \sum_{i=1}^N (w_i - \mu)^2$$

with unbiased estimator

$$\text{var}(\hat{\mu}) = \frac{N - n_1}{Nn_1(n_1 - 1)} \sum_{i=1}^{n_1} (w_i - \hat{\mu})^2.$$

4.0 Worked Example

Using the strawberry field example, we can compute and estimate of the total number of weed plants in the field using the adaptive cluster sampling plan. After sampling we have the information reported in the following figure about our initial sampling units and resulting networks.

FIGURE 7. Counts of weeds found in initial sampling units and the resulting networks.

				1					
	2	15	8	1		0		0	
	1	2	0						
						2			
0		0			7	10			
					7	3			
				0	1				
								0	

From the adaptive sample of $n_1 = 10$ initial sampling units, we have identified 9 distinct networks. One network has $m_1 = 7$ units with $y_1^* = 30$ weeds. A second has $m_2 = 6$ units with $y_2^* = 30$ weeds. The other 7 networks have zero counts and so have $m_i = 1$ and $y_i^* = 0$.

4.1 Use the initial intersection probabilities

First compute the intersection probabilities

$$\alpha_1 = 1 - \left[\binom{100-7}{10} / \binom{100}{10} \right] = 0.533, \alpha_2 = 1 - \left[\binom{100-6}{10} / \binom{100}{10} \right] = 0.478, \text{ and}$$

$$\alpha_i = 1 - \left[\binom{100-1}{10} / \binom{100}{10} \right] = 0.100 \text{ for } i = 3, \dots, 9. \text{ Further}$$

$$p_{12} = \left(\binom{100-6-7}{10} / \binom{100}{10} \right) = 0.231 \text{ so that}$$

$$\alpha_{12} = \alpha_1 + \alpha_2 - (1 - p_{12}) = 0.533 + 0.478 - (1 - 0.231) = 0.242.$$

Therefore, $\hat{\mu} = \frac{1}{100} \left(\frac{30}{0.533} + \frac{30}{0.478} + \frac{0}{0.10} + \dots + \frac{0}{0.10} \right) = 1.191$ weeds per unit, and

$$\begin{aligned} \text{var}(\hat{\mu}) &= \frac{1}{100^2} \left[\frac{30^2}{0.533} \left(\frac{1}{0.533} - 1 \right) + \frac{30^2}{0.478} \left(\frac{1}{0.478} - 1 \right) + \frac{2(30)(30)}{0.242} \left(\frac{0.242}{(0.533)(0.478)} - 1 \right) \right] \\ &= 0.3168 \end{aligned}$$

$$\text{se}(\hat{\mu}) = 0.563, \text{ CV}(\hat{\mu}) = 47.3\%, \text{ with 95\% confidence interval on } \mu \text{ of } (0.065, 2.32).$$

In terms of τ , the population total number of weeds,

$$\hat{\tau} = 100(1.191) = 119.1 \text{ weeds with } \text{var}(\hat{\tau}) = 3167.6, \text{ se}(\hat{\tau}) = 56.28, \text{ and 95\% C.I. on } \tau \text{ of } (6.5, 231.6).$$

4.2 Use the numbers of initial intersections

Note that we use the information from each of the 10 initial sampling units, even if the same network is detected more than once. This leads us to

$$\hat{\mu} = \bar{w} = \left(\frac{30}{7} + \frac{30}{6} + \frac{30}{6} + \frac{0}{1} + \dots + \frac{0}{1} \right) / 10 = 1.43 \text{ with variance (compute variance of } w_i),$$

$$\text{var}(\hat{\mu}) = 0.480, \text{ se}(\hat{\mu}) = 0.693, \text{ CV}(\hat{\mu}) = 48.5\%, \text{ and 95\% C.I. of } (0.043, 2.81).$$

When expanded to the population total, we get

$\hat{\tau} = 100(1.43) = 142.9$ weeds with $\text{var}(\hat{\tau}) = 4796$, $\text{se}(\hat{\tau}) = 69.3$, and 95% C.I. on τ of (4.3,281).

4.3 Simple random sample computations

If we use only the data from the initial sample of size 10 (no adaptive sampling performed), then we get the following estimates for the numbers of weeds.

$\hat{\mu} = 1.4$ weeds per unit with standard error $\text{se}(\hat{\mu}) = 0.951$ and $\text{CV}(\hat{\mu}) = 67.9\%$. A 95% confidence interval on μ is estimated to be (-0.50,3.30).

Similarly for τ , we get $\hat{\tau} = 140$ weeds with $\text{se}(\hat{\tau}) = 95.1$, and 95% C.I. on τ of (-50.1,330).

Thus, although the coefficients of variation for the adaptive sample are rather large, they are still smaller than that obtained via simple random sampling. Thompson (1990) demonstrates through simulation, that as the sampling fraction increases, the relative efficiency of the adaptive cluster sampling estimator to the simple random sampling estimator can increase dramatically. Note, however, that a considerable number of additional sampling units may need to be searched as individuals are found. Very careful record keeping of units previously searched is also very important to make the method easiest to implement.

5.0 References

Tompson, S. K. 1990. Adaptive cluster sampling. *Journal of the American Statistical Association* 85, 1050-1059.

Tompson, S. K. 1992. *Sampling*. John Wiley & Sons, Inc., New York, 339pp.

Tompson, S. K. and G. A. F. Seber. 1996. *Adaptive Sampling*. John Wiley & Sons, Inc., New York, 265pp.