

## 4. Sampling

### 4.1. Concepts of Sampling

Many variables in Civil engineering are spatially distributed. For example concentration of pollutants, variation of material properties such as strength and stiffness in the case of concrete and soils are spatially distributed. The purpose of sampling is to obtain estimates of population parameters (e.g. means, variances, covariance's) to characterize the entire population distribution without observing and measuring every element in the sampled population. Sampling theory for spatial processes principally involves evaluation of estimator's sampling distributions and confidence limits. A very good introduction to these methods and the uses and advantages of sampling is provided by Cochran (1977) and Beacher and Christian (2003).

An *estimate* is the realization of a particular sample statistic for a specific set of sample observations. Estimates are not exact and uncertainty is reflected in the variance of their distribution about the true parameter value they estimate. This variance is, in turn, a function of both the sampling plan and the sampled population. By knowing this variance and making assumptions about the distribution, shape, confidence limits on true population parameters can be set.

A *sampling plan* is a program of action for collecting data from a sampled population. Common plans are grouped into many types: for example, simple random, systematic, stratified random, cluster, traverse, line intersects, and so on. In deciding among plans or in designing a specific program once the type plan has been chosen, one attempts to obtain the highest precision for a fixed sampling cost or the lowest sampling cost for a fixed precision or a specified confidence interval.

### 4.2. Common Spatial Sampling Plans

Statistical sampling is a common activity in many human enterprises, from the national census, to market research, to scientific research. As a result, common situations are encountered in many different endeavors, and a family of sampling plans has grown up to

handle these situations. Simple random sampling, systematic sampling, stratified random sampling, and cluster sampling are considered in the following section.

#### 4.2.1. Simple random sampling

The characteristic property of simple random sampling is that individual are chosen at random from the sampled population, and each element of population has an equal probability of being observed. An unbiased estimator of the population mean from a simple random  $x=\{x_1, \dots, x_n\}$  is the sample mean

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{-----}(1)$$

This estimator has sampling variance.

$$Var(\bar{x}) = \frac{\sigma^2}{n} \frac{N-n}{N} \quad \text{-----}(2)$$

where  $\sigma^2$  is the (true) variance of the sampled population and  $N$  is the total sampled population size. The term  $(N-n)/N$  is called the *finite population factor*, which for  $n$  less than about 10% of  $N$ , can safely be ignored. However, since  $\sigma^2$  is usually unknown, it is estimated by the sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad \text{-----}(3)$$

in which the denominator is taken as  $n-1$  rather than  $n$ , reflecting the loss of a degree-of-freedom due to estimating the mean from the same data. The estimator is unbiased but does not have minimum variance. The only choice (i.e. allocation) to be made in simple random sampling is the sample size  $n$ . Since the sampling variance of the mean is inversely proportional to sample size,  $Var(x) \propto n^{-1}$ , a given estimator precision can be obtained by adjusting the sample size, if  $\sigma$  is known or assumed. A sampling plan can be optimized for total cost by assuming some relationship between  $Var(\bar{x})$  and cost in

construction or design. A common assumption is that this cost is proportional to the square root of the variance, usually called the standard error of the mean,  $\sigma_{\bar{x}} = Var^{1/2}(\bar{x})$ .

It is usually assumed that the estimates of  $\bar{y}$  and  $\bar{Y}$  are normally distributed about the corresponding population values. If the assumption holds, lower and upper confidence limits for the population mean and total mean are as follows:

Mean:

$$\bar{Y}_L = \bar{y} - \frac{ts}{\sqrt{n}} \sqrt{1-f}, \quad \bar{Y}_U = \bar{y} + \frac{ts}{\sqrt{n}} \sqrt{1-f}$$

Total:

$$\bar{\bar{Y}}_L = N\bar{y} - \frac{tNs}{\sqrt{n}} \sqrt{1-f}, \quad \bar{\bar{Y}}_U = N\bar{y} + \frac{tNs}{\sqrt{n}} \sqrt{1-f}$$

The symbol  $t$  is the value of the normal deviate corresponding to the desired confidence probability. The most common values are tabulated below:

Confidence probability (%)	50	80	90	95	99
Normal deviate, $t$	0.67	1.28	1.64	1.96	2.58

If the sample size is less than 60, the percentage points may be taken from Student's  $t$  table with  $(n-1)$  degrees of freedom, these being the degrees of freedom in the estimated  $s^2$ . The  $t$  distribution holds exactly only if the observations  $y_i$  are themselves normally distributed and  $N$  is infinite. Moderate departures from normality do not affect it greatly. For small samples with very skew distributions, special methods are needed. An example of the application is as follows.

**Example.**

In a site, the number of borehole data sheets to characterize the substrata to obtain design parameters is 676. In each borehole data, 42 entries reflecting the various characteristics

of soils viz. compressibility, shear strength, compaction control, permeability etc are indicated. In an audit conducted, it was revealed that in some datasheets, all the data are not entered. The audit party verified a random sample of 50 sheets ( 7% sample) and the results are indicated in Table.1

Table 21 Results for a sample of 50 petition sheets

Number of signatures, $y_i$	Frequency, $f_i$
42	23
41	4
36	1
32	1
29	1
27	2
23	1
19	1
16	2
15	2
14	1
11	1
10	1
9	1
7	1
6	3
5	2
4	1
3	1
$\sum f_i$	<b>50</b>

We find

$$n = \sum f_i = 50, \quad y = \sum f_i y_i = 1471, \quad \sum f_i y_i^2 = 54,497$$

Hence the estimated total number of signatures is

$$Y = N \bar{y} = \frac{(676)(1471)}{50} = 19,888$$

For the sample variance  $s^2$  we have

$$s^2 = \frac{1}{n-1} [\sum f_i (y_i - \bar{y})^2] = \frac{1}{n-1} \left[ \sum f_i y_i^2 - \frac{(\sum f_i y_i)^2}{\sum f_i} \right]$$

$$= \frac{1}{49} \left[ 54,497 - \frac{(1471)^2}{50} \right] = 229.0$$

The 80% confidence limits are given by

$$19,888 \pm \frac{tNs}{\sqrt{n}} \sqrt{1-f} = 19,888 \pm \frac{(1.28)(676)(15.13)\sqrt{1-0.0740}}{\sqrt{50}}$$

This gives 18,107 and 21,669 for the 80 % limits. A complete count showed 21,045 entries and is close to the upper estimate.

#### 4.2.2. Systematic sampling

In systematic sampling the first observation is chosen at random and subsequent observations are chosen periodically throughout the population. To select a sample of n units, we take a unit at random from the first k units and every k<sup>th</sup> unit thereafter. The method involves the selection of every k<sup>th</sup> element from a sampling frame, where k, the sampling interval, is calculated as:

$$k = \text{population size (N)} / \text{sample size (n)}$$

Using this procedure each element in the population has a known and equal probability of selection. This makes systematic sampling functionally similar to simple random sampling. It is however, much more efficient (if variance within systematic sample is more than variance of population) and much less expensive to carry out. The advantages of this approach are that 1) the mistakes in sampling are minimized and the operation is speedy, 2) it is spread uniformly over the population and is likely to be more precise than the random sampling.

An unbiased estimate of the mean from, a systematic sample is the same as above equation .The sampling variance of this estimate is

$$Var(\bar{x}) = \left(\frac{N-1}{N}\right)\sigma_w^2 + \left(\frac{k(n-1)}{N}\right) \text{-----}(4)$$

where  $k$  is the interval between samples ( $k = N/n$ ) and  $\sigma_w^2$  is the variance of elements within the same systematic sample

$$s_w^2 = \frac{1}{k} \sum_{i=1}^k \frac{\sum_{j=1}^n (x_{ij} - \bar{x})^2}{n-1} \text{-----}(5)$$

in which  $x_{ij}$  is the  $j^{\text{th}}$  member of the  $i^{\text{th}}$  interval of the sample. When only one systematic sample has been taken (i.e. one set of  $n$  observations at spacing  $k$ ) the variance of the mean cannot be evaluated unless an assumption is made about the nature of the sampled population. The conventional assumption is that the population can be modeled by a linear expression of the form,  $x_i = \mu + e_i$ , in which  $\mu$  is the mean and  $e_i$  is a zero-mean random perturbation. For constant mean, this leads to (Cochran 1977)

$$Var(\bar{x}) = \frac{1}{n} \left(\frac{N-n}{N}\right) \left(\frac{\sum (x_i - \bar{x})^2}{n-1}\right) \text{-----}(6)$$

and for linearly trending mean,  $\mu = \mu_0 + b_i$

$$Var(\bar{x}) = \frac{1}{n} \left(\frac{N-n}{N}\right) \left(\frac{\sum (x_i - 2x_{i+k} + x_{i+2k})}{6(n-2)}\right) \text{-----}(7)$$

One must ensure that the chosen sampling interval does not hide a pattern. Any pattern would threaten randomness. A random starting point must also be selected. Systematic sampling is to be applied only if the given population is logically homogeneous, because systematic sample units are uniformly distributed over the population.

Example: Suppose the auditor in the previous example wants to use systematic sampling, then he can choose every 25th or 50th sheet and conduct the study on this sample.

A starting point is chosen at random, and thereafter at regular intervals. For example, suppose you want to sample 25<sup>th</sup> sheet from 676 sheets,  $676/25=27$ , so every 27<sup>th</sup> sheet is chosen after a random starting point between 1 and 15. If the random starting point is 11, then the sheets selected are 11, 28, 65, 92 etc.

#### **4.2.3. Stratified random sampling**

When sub-populations vary considerably, it is advantageous to sample each subpopulation (stratum) independently. Stratification is the process of grouping members of the population into relatively homogeneous subgroups before sampling. The strata should be mutually exclusive: every element in the population must be assigned to only one stratum. The strata should also be collectively exhaustive: no population element can be excluded. Then random or systematic sampling is applied within each stratum. This often improves the representativeness of the sample by reducing sampling error. It can produce a weighted mean that has less variability than the arithmetic mean of a simple random sample of the population.

There are several possible strategies:

1) Proportionate allocation uses a sampling fraction in each of the strata that is proportional to that of the total population. If the soil sample consists of 60% of boulders (boulder stratum) and 40% sand (sand stratum), then the relative size of the two types of samples should reflect this proportion.

2) Optimum allocation (or Disproportionate allocation) - Each stratum is proportionate to the standard deviation of the distribution of the variable. Larger samples are taken in the strata with the greatest variability to generate the least possible sampling variance.

Estimates of the total population characteristics can be made by combining the individual stratum estimates. For certain populations, stratifying before sampling is more efficient than taking samples directly from the total population. Sampling plans that specify a simple random sample in each stratum are called stratified random sampling plans.

An unbiased estimator of the mean of the total sampled population is

$$\bar{x} = \frac{1}{N} \sum_{h=1}^m N_h \bar{x}_h \quad \text{----- (8)}$$

Where  $\bar{x}$  is the population mean,  $m$  is the number of strata and  $h$  denotes the stratum (i.e.  $N$  is the size of the  $h^{\text{th}}$  stratum, and  $\bar{x}_h$  is the corresponding mean). The variance of this estimate is

$$\text{Var}(\bar{x}) = \sum_h w_h^2 \frac{s_h^2}{n_h} (1 - f_h) \quad \text{-----(9)}$$

where  $w_h = N_h / N$  and  $f_h = n_h / N_h$ . Since the sample from each stratum is simple random the estimate of the variance within each can be taken from above equation. Then, an estimate of the variance of the total population is

$$\text{Var}(\bar{x}) = \frac{1}{N} \sum_h s_h^2 + s_{\text{amongmeans}}^2 \quad \text{-----(10)}$$

**4.2.4. Cluster sampling**

In cluster sampling, aggregates or clusters of elements are selected from the sampled population as units rather than as individual elements, and properties of the clusters are determined. From the properties of the clusters, inferences can be made on the total sampled population. Plans that specify to measure every element within clusters are called single-staged-cluster plans, since they specify only one level of sampling: plans that specify that cluster properties be estimated by simple random sampling are called two-staged cluster plans, since they specify two levels of sampling. Higher order cluster plans are sometimes used.

We consider simplest case first: of  $M$  possible clusters,  $m$  are selected: the ratio  $f_1 = m/M$  called, as before, the sampling fraction. Each cluster contains the same

number of elements, N, some number n of which are selected for measurement ( $f_2 = n/N$ ).  
 An unbiased estimate of the average of each cluster is

$$\bar{x}_i = \frac{1}{n_i} \sum_j x_{ij} \text{-----(11)}$$

where,  $x_{ij}$  is the  $j^{\text{th}}$  element of the  $i^{\text{th}}$  cluster. An unbiased estimate of the average of the total population is

$$\bar{x} = \frac{1}{m} \sum_i \bar{x}_i = \frac{1}{mn} \sum_i \sum_j x_{ij}$$

and the variance of this estimator is

$$\text{Var}(\bar{x}) = \frac{(1-f_1)}{m} s_1^2 + \frac{f_1(1-f_2)}{mn} s_2^2$$

in which  $s_1^2$  is the estimated variance among cluster means

$$s_1^2 = \frac{\sum (\bar{x}_i - \bar{x})^2}{n-1}$$

and  $s_2^2$  the estimated variance within clusters

$$s_2^2 = \frac{\sum \sum (x_{ij} - \bar{x}_i)^2}{m(n-1)} \text{-----(12)}$$

In the more general case, not all of the clusters are of equal size. For example, the numbers of joints appearing in different outcrops are different. With unequal sized clusters the selection plan for clusters is not as obvious as it was previously. The relative probability, of selecting different sized clusters is now a parameter of the plan. Commonly, the  $z_j$  are either taken all equal (simple random sampling of the clusters) or proportional to size. The precisions of these two plans are different. For selection with equal probability an unbiased estimate of the true total population mean is

$$\bar{x} = \frac{\sum_i N_i \bar{x}_i}{\sum_i N_i}$$

and the variance of this estimator is

$$\text{var}(\bar{x}) = \frac{(1-f_1)}{m\bar{N}^2} \frac{(N_i \bar{x}_i - \bar{x}_m)^2}{(m-1)} + \frac{f_1}{m^2 \bar{N}^2} \sum \frac{N_i^2 (1-f_{2i}) s_{2i}^2}{n_i}$$

in which  $\bar{N}$  is the average value of  $N_i$  and  $\bar{x}_m = \sum_i N_i \bar{x}_i / m$  if the assumption is made that  $n_i \propto N_i$  then the plan is self weighting and this simplifies to

$$\text{Var}(\bar{x}) = \frac{(1-f_1)}{m\bar{N}^2} \frac{\sum (N_i \bar{x}_i - \bar{x}_m)^2}{(m-1)} + \frac{f_1(1-f_2)}{mn\bar{N}} \sum N_i^2 s_{2i}^2 \text{-----}(14)$$

This assumption ( $n_i \propto N_i$ ) is frequently valid: for example, proportionally more joints are typically sampled from larger outcrops than from smaller outcrops.

For selection with probability proportional to size, an unbiased estimate of the total population mean is

$$\bar{\bar{x}} = \frac{1}{n} \sum_i \bar{x}_i$$

and the variance is

$$\text{Var}(\bar{\bar{x}}) = \frac{\sum (\bar{x}_i - \bar{\bar{x}})^2}{m(n-1)} \text{-----}(15)$$

In all cases, the variance of the total population can be estimated from the variances between elements within clusters and the variance between the means of the clusters:

$$\sigma^2 = \sigma^2_{\text{among means}} + \sigma^2_{\text{within clusters}}$$

Joint surveys and many other types of civil engineering sampling may be based on cluster plans because the cost of sampling many individual joints on one outcrop (i.e. a cluster) is less than the cost of traveling between outcrops.

The variance of geological populations is often a function of spatial extent. Indeed, this is the principal argument in the geo-statistical literature for favoring variograms over auto covariance functions. If we consider the strength of soil specimens taken close together, the variance among specimens is usually smaller than the variance among specimens taken from many locations in one area of the site, which, in turn, is smaller than the variance among specimens taken from all across the site. Cluster techniques allow us to evaluate variance as a function of the “extent” of the distribution in space by nesting the variances.