Lecture 8

Cluster Sampling and Systematic Sampling

In general, we want the target and study populations to be the same. When they are not the same, the researcher must be careful to ensure that conclusions based on the sample results can still be applied to the target population.

Because of restrictions such as cost or scheduling conflicts, it is often impossible to collect a simple random sample or a stratified random sample. In many cases, however, it may be possible to define a sampling frame with sampling units that *are not* the units in the target population or the study population, yet still obtain statistically valid estimates.

Cluster sampling is an example of when

Example. Suppose MSU wants to determine whether students would a fall break during October or to finish classes later in December. A set of 50 classes are randomly selected and students within those classes record their choice.

A population contains

In this example the PSUs are the classes and the SSUs are the students within the classes.

The primary sampling units may be of different sizes. That is, the numbers of secondary sampling units in the primary units are not all the same.
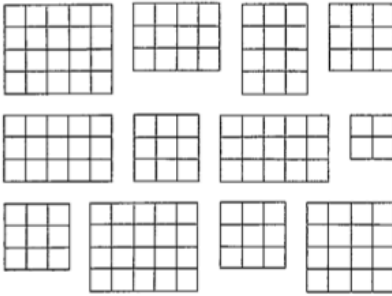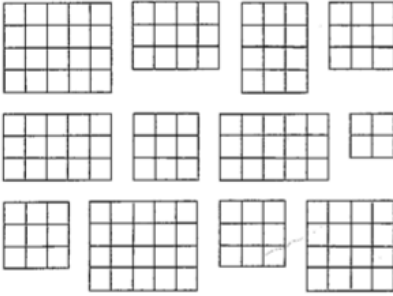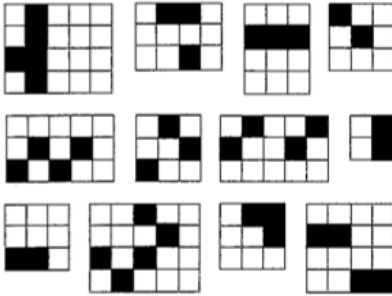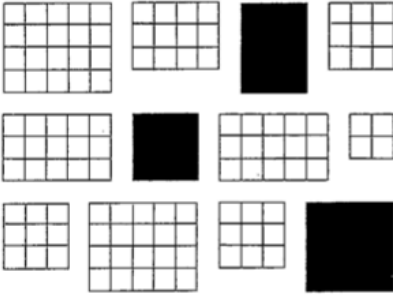
Think of these disjoint groups of population units (clusters) in a similar fashion to strata. Then suppose we define the sampling frame as a set of clusters Then, the sampling units in this population are not individual units in the population.

The sampling units are *clusters* of population units.

Note: the population has M_0 individual units, but the sampling frame only has N PSU corresponding to the number of clusters formed.

Very often, all of the secondary sampling units in each selected primary sampling unit will also be included in the sample.

Figure (from textbook) gives a visual overview of cluster sampling vs. stratified random sampling.

Stratified Sampling	Cluster Sampling
Each element of the population is in exactly one stratum.	Each element of the population is in exactly one cluster.
Population of H strata: stratum h has n_h elements:	One-stage cluster sampling: population of N clusters:
Take an SRS from *every* stratum:	Take an SRS of clusters; observe all elements within the clusters in the sample:
Variance of the estimate of $\bar{y}_U$ depends on the variability of values *within* strata.	The cluster is the sampling unit; the more clusters we sample, the smaller the variance. The variance of the estimate of $\bar{y}_U$ depends primarily on the variability *between* cluster means.

Cluster Sampling Notation

Similar to stratified random sampling, let y_{ij} denote the response of SSU j within cluster i.

Primary Sampling Unit (PSU) level

- $N =$ the number of clusters (PSUs) in the population

- $M_i =$

- $M_0 = \sum_{i=1}^{N} M_i =$

- $t_i = \sum_{j=1}^{M_i} y_{ij} =$ cluster i total

- $t = \sum_{i=1}^{N} t_i = \sum_{i=1}^{N} \sum_{j=1}^{M_i} y_{ij} =$ population total

- $\bar{t} = \frac{1}{N} \sum_{i=1}^{n} t_i =$ mean of the cluster totals (mean of PSU values).

- $S_t^2 = \frac{\sum_{i=1}^{N} (t_i - \bar{t})^2}{N-1}$ the population variance of cluster totals

Secondary Sampling Unit (SSU) level

$\bar{y}_U = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{y_{ij}}{M_0} = \frac{t}{M_0}$ the population mean

$\bar{y}_{iU} = \sum_{j=1}^{M_i} \frac{y_{ij}}{M_i} = \frac{t_i}{M_i}$ = population mean in PSU i

$S^2 = \sum_{i=1}^{N} \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_U)^2}{M_0 - 1}$ = population variance of SSUs

$S_i^2 = \sum_{j=1}^{M_i} \frac{(y_{ij} - \bar{y}_{iU})^2}{M_i - 1}$ = population variance within PSU i

Sample Values

n = the number of

m_i = number of

$\hat{\bar{y}}_i = \frac{1}{m_i} \sum_{j=1}^{m_i} y_{ij} =$

$\hat{t}_i = \frac{M_i}{m_i} \sum_{j=1}^{m_i} y_{ij} =$

$\hat{t}_{cl} = \frac{N}{n} \sum_{i=1}^{n} \hat{t}_i =$

$s_t^2 = \frac{\sum_{i=1}^{n} (t_i - \frac{\hat{t}_{cl}}{N})^2}{n - 1}$ = the sample variance of estimated cluster (PSU) totals

$s_i^2 \frac{\sum_{j=1}^{m_i} (y_{ij} - \bar{y}_i)^2}{m_i - 1}$ = the sample variance within PSU i

w_{ij} = sampling weight for SSU j in PSU i

Note in one-stage cluster samping, we have $m_i = M_i$. That is, every SSU in PSU i is sampled. Thus $\bar{y}_i = \hat{\bar{y}}_i$, $t_i = \hat{t}_i$, and $S_i^2 = s_i^2$ when $m_i = M_i$.

One-Stage Cluster Sampling

When the strata themselves are the primary sampling units, the strata are called clusters. The selection of a sample of clusters to provide a sample of population units

If all of the population units in every selected cluster are in the sample,

What is the difference between one-stage cluster sampling and stratified random sampling?

- In stratified SRS,

- In one-stage cluster sampling,

When a cluster is defined as a group of population units, the clusters are called the primary sampling units. Subgroups within primary sample units are called secondary sampling units. For one-stage cluster sampling, the secondary sampling units are the individual units.

If the selection of the population units within every selected cluster is restricted a second time, then this technique is known as subsampling or two-stage cluster sampling.

If a sample of primary sampling units (Stage 1) is selected, followed by a selection of secondary sampling units (Stage 2) within the sample of primary sampling units, followed by a selection of tertiary sampling units (Stage 3) within the sample of secondary sampling units, and so on, then the sampling procedure is known as multistage cluster sampling.

In cluster sampling, the size of the cluster can also be used as an auxilary variable to select clusters with unequal sampling probabilities (more later) or used in a ratio estimator.

Stratified sampling vs. cluster sampling:

- A researcher will use a stratified samplig design because

- A researcher will use cluster sampling because

4

Estimation of $\bar{y}_U, t,$ and $\bar{t}.$

The unbiased estimator of t is:

where $i \in \mathcal{S}$ corresponds to the PSUs selected in the sample $\mathcal{S}$, $j \in \mathcal{S}_i$ corresponds to the SSUs selected in the sample for PSUs j ($\mathcal{S}_j$), and w_{ij} is the weight associated with probability of selection $\frac{N}{n}$ in this case.

The unbiased estimators of $\bar{y}_U$ is:

where $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} t_i = \frac{\hat{t}_{cl}}{N}$ is the sample mean of the cluster totals.

Next we want to study the standard error of these estimators:

$$SE(\hat{t}_{cl}) = N \sqrt{(1 - \frac{n}{N}) \frac{S_t^2}{n}}$$

where $S_t^2 = \frac{\sum_{i=1}^{N} (t_i - \bar{t})^2}{N-1}$ is the variance of the N cluster t_i totals.

Because S_t^2 is unknown, the sample variance of the cluster totals is used:

$$s_t^2 = \frac{\sum_{i=1}^{n} (t_i - \frac{\hat{t}_{cl}}{N})^2}{n-1}.$$

The square root of the estimated variances using s_t^2 returns the standard errors of the estimators.

The standard error for the population mean is actually calculated using ratio estimation (similar to domain estimation)

$$SE(\hat{\bar{y}}_{cl}) = \sqrt{\left(1 - \frac{n}{N}\right) \frac{1}{n\bar{M}^2} \frac{\sum i \in \mathcal{S} \left(t_i - \hat{\bar{y}}_{cl} M_i\right)^2}{n-1}}$$

where $S_t^2 = \frac{\sum_{i=1}^{N} (t_i - \bar{t})^2}{N-1}$ is the variance of the N cluster t_i totals.

Confidence intervals follow the usual prescription, but note that the degrees of freedom for the t critical values are based on the the number of primary sampling units, or clusters, and not the total number of secondary sampling units.

Comparison to SRS

Because the variance formulas for $\hat{\bar{y}}_{U,cl}$ and $\hat{t}_{cl}$ are determined only from the cluster-to-cluster variability, the precision of the estimators can be improved if

Q: How does this compare to the strategy for choosing strata?

We want clusters such that the variability of the SSU $y-$values within each cluster

This is in

Example. Suppose we are interested in salaries of past MSU graduates. Devise a stratified random sampling scheme and a cluster sampling scheme.

For comparison sake, assume we took a SRS with $n \times M$ samples. The variance of the estimated total would be:

$$V(\hat{t}_{SRS}) = N^2(1 - \frac{n}{N})\frac{MS^2}{n}. \tag{1}$$

Recall the variance of the cluster sampling estimate of the population total was:

$$V(\hat{t}_{cl}) = N^2(1 - \frac{n}{N})\frac{S_t^2}{n}, \tag{2}$$

where $S_t^2 = \sum_{i=1}^{n} \frac{(t_i - \bar{t})^2}{N-1} = M(MSB)$. MSB is the mean square of the between cluster variability.

So if $MSB > S^2$ then cluster sampling is less efficient than SRS.

Systematic Sampling

Systematic Sampling is a sampling plan in which the population units are collected systematically throughout the population. More specifically, a single primary sampling unit consists of secondary sampling units that are relatively spaced with each other in the systematic pattern throughout the population.

Suppose the study area is partioned into a 20×20 grid of 400 population units. A primary sampling unit in a systematic sample could consist of all population units that occur in every 5 units..

Initially, systematic sampling and cluster sampling appear to be opposites because systematic samples contain secondary sampling units that are spread

Systematic and cluster sampling are similar, however, because whenever a primary sampling unit is selected from the sampling frame, all secondary sampling units of that primary sampling unit will be included in the sample. Thus, random selection occurs at the primary sampling unit level and not the secondary sampling unit level.

For estimation purposes, you could ignore the secondary sampling unit $y_{ij}-$values and only retain the primary sampling units t_i-values. This is what we did with one-stage cluster sampling.

The systematic and cluster sampling principle: How do we obtain low variance?

This is equivalent to saying that the within-cluster variablity should be as large as possible to obtain the most precise estimators. Thus, the ideal primary sampling unit is representative of the full diversity of y_{ij} values within the population.

With natural populations of spatially distributed plants, animals, minerals, ect.,

Estimation of $\bar{y}_U$ and t

If a SRS is used to select the systematic primary sampling units, we can apply the estimation results for cluster sampling to define (i) estimators, (ii) the variance of each estimator, and (iii) the estimated variance of each estimator.