Survey Sample Methods

Evaluators' Toolbox Refreshment

Abhik Roy & Kristin Hobson

abhik.r.roy@wmich.edu & kristin.a.hobson@wmich.edu

Western Michigan University

AEA Evaluation 2012 Session

Considering a Sampling Method

Considerations:

- Presampling Choices
- Sampling Choices
- Postsampling Choices

Presampling Choices

- What is the nature of the study?
- What are the variables of interest?
- What population is being targeted?
- How many units will be selected?
- Is sampling appropriate?

Sampling Choices

- What do you want to accomplish with data?
- How is the data distributed?

Postsampling Choices

- How is nonresponse or missing data dealt with? (e.g. ignoring, imputation, or deletion)
- Must the sample data be weighted?
- What are the necessary standard errors and confidence intervals for the study estimates?
- What were the issues for any bias and/or missingness?

Two types of Sampling

- Probability Sampling
  - Random selection.
  - Population representation given by a confidence interval.
  - Ability to generalize to a certain populous.
- Nonprobability Sampling
  - Probability is usually unknown.
  - Inability to generalize to any populous.
  - Three types: Convenience, Purposive, and Quota.

Nonprobability Sampling Methods

- Convenience
- Purposive
- Quota

Convenience Sampling

- “What you can get" method
- Example: 1936 Presidential election Alf Landon (R) v. Franklin D. Roosevelt (D) and the Literary Digest.

Purposive Sampling

- “Specific need" method
- Example: Asking a drug addict to find others who are also drug addicts. (Snowball sampling)

Quota Sampling

- “Deliberately setting numbers" method
- Example: A researcher wished to know how people in a community with primarily African Americans feel about the President. But probability sampling may miss the small populous of Japanese (a mutually exclusive subgroup).

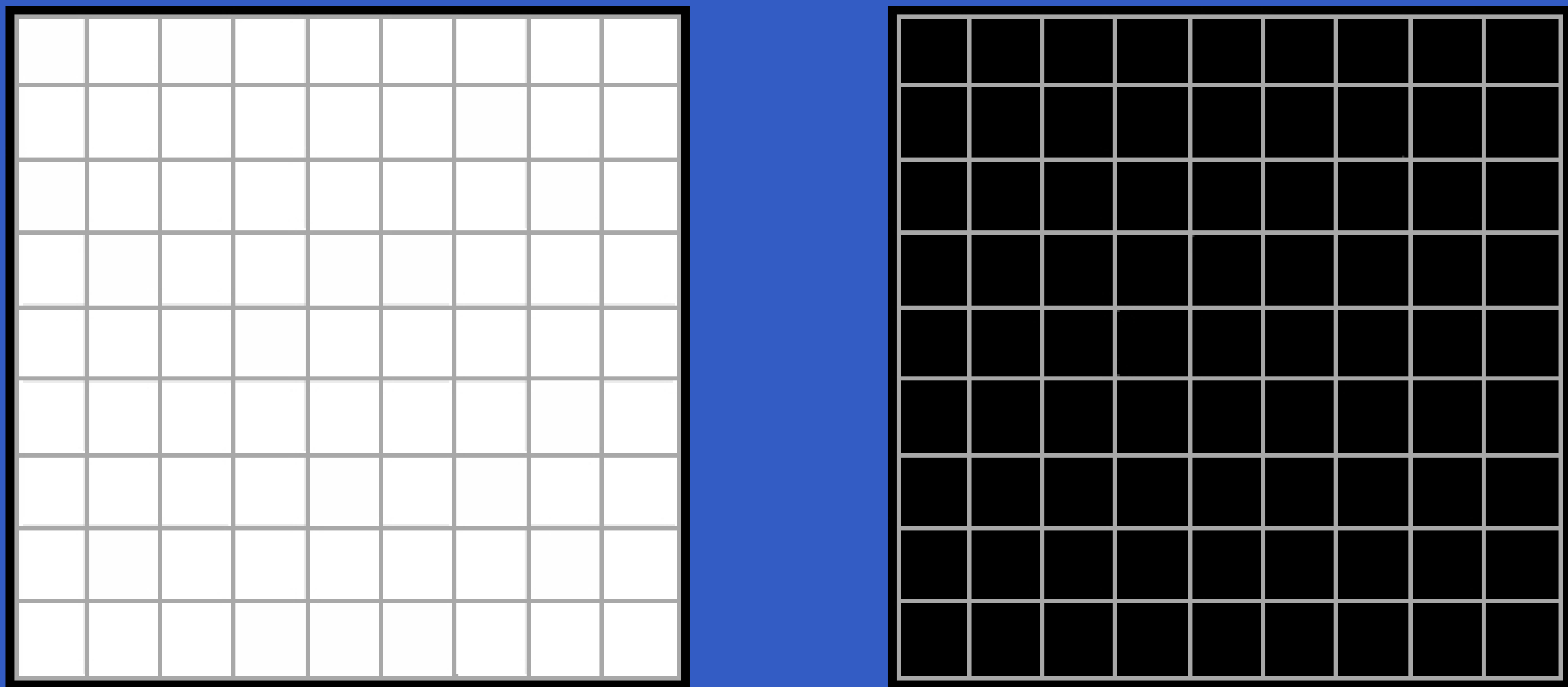
Probability Sampling Methods

- Census
- Simple Random Sampling (SRS)
- Systematic Random Sampling
- Stratified Random Sampling
- Cluster Random Sampling

Probability vs. Nonprobability Sampling

Basic Question: Do you wish to generalize?

Census



Census

Requirements

1. List of study population (called a sampling frame).
2. Count of the study population (N).
3. Sample size $(n = N)$.
4. Full selection method.

Benefits of a Census

Strengths

- “Easy" to administer.
- Self-Weighting.
- Estimation of error is 0.
- Bias/Sampling error is 0.
- Simplification of data analysis.

Drawbacks of a Census

Weaknesses

- Extremely expensive.
- Extremely time consuming.

Equations

Table 1: Equations for a Census

	Estimator	Estimated Variance	Bound on the Error B
Population Mean	$\hat{\mu} = \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{s^2}{n}$	$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{s^2}{n}}$
Population Total	$\hat{\tau} = N\overline{y} = \frac{N\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(N\overline{y}) = N^2\left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right)$	$2\sqrt{\hat{V}(N\overline{y})} = 2\sqrt{N^2\left(1 - \frac{n}{N}\right)\left(\frac{s^2}{n}\right)}$
Population Proportion	$\hat{p} = \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}) = \left(1 - \frac{n}{N}\right)\frac{\hat{p}\hat{q}}{n-1}$	$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\hat{p}\hat{q}}{n-1}}$

n = *sample size*
N = *population size*
s = *sample standard deviation*
y_i = *total observations*
$\hat{q} = 1 - \hat{p}$

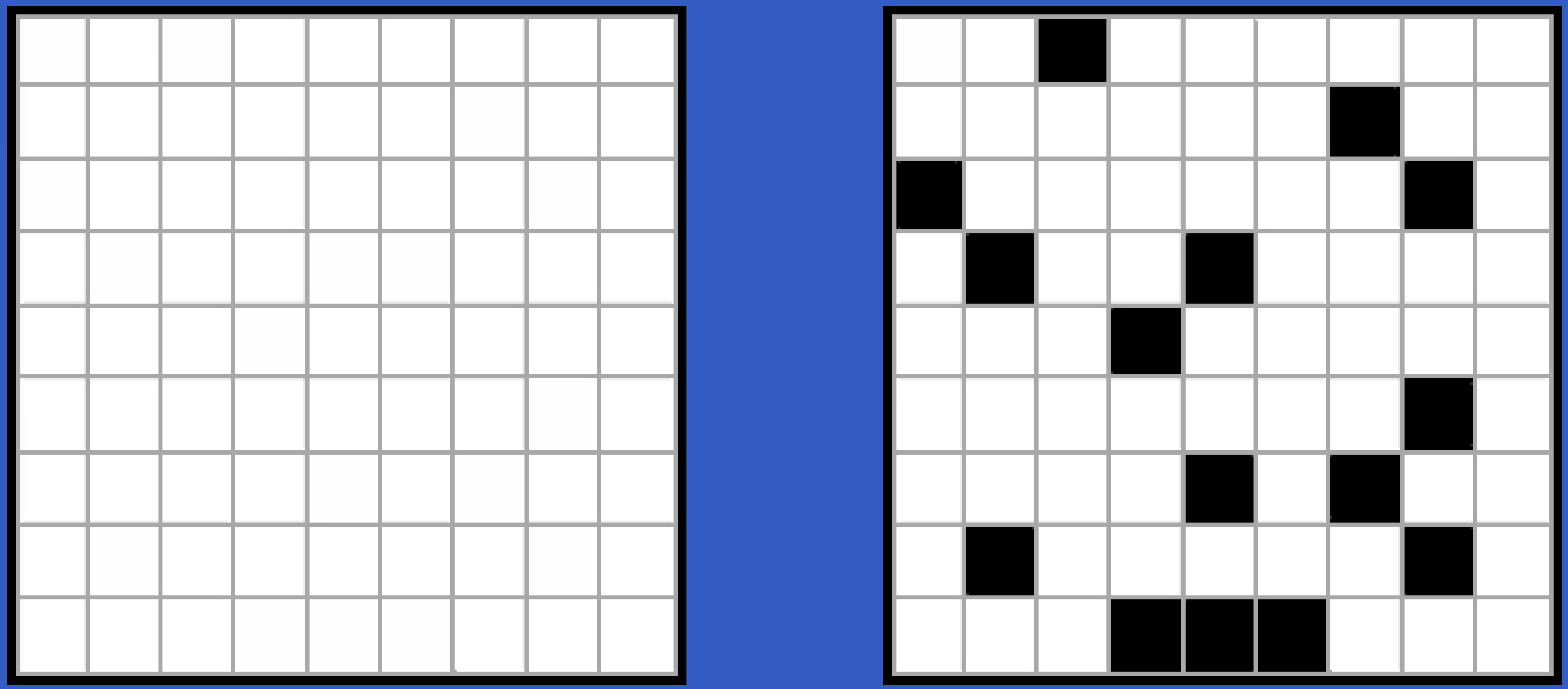
When Should You Use a Census?

1. Small sample.
2. Generalize to an overall populous.

Examples of a Census

1. Survey of work conditions at a small restaurant.
2. Evaluation of teachers in a school.
3. Exception: U.S. population.

Simple Random Sampling (SRS)



Simple Random Sampling (SRS)

Requirements

1. List of study population (called a sampling frame).
2. Count of the study population (N).
3. Sample size (n).
4. Random selection method.

Benefits of a SRS

Strengths

- Easy to administer.
- Self-Weighting.
- Estimation of error is easy to calculate.
- Minimization of bias/sampling error.
- Simplification of data analysis.

Drawbacks of a SRS

Weaknesses

- Vulnerable to sampling errors.
- Possible underrepresentation of subgroups.
- Can be tedious, costly, and possibly impractical.

Equations

Table 2: Equations for Estimating Population (Simple Random Sampling)

	Estimator	Estimated Variance	Bound on the Error B
Population Mean	$\hat{\mu} = \overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}) = \left(1 - \dfrac{n}{N}\right)\dfrac{s^2}{n}$	$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\left(1 - \dfrac{n}{N}\right)\dfrac{s^2}{n}}$
Population Total	$\hat{\tau} = N\overline{y} = \dfrac{N\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(N\overline{y}) = N^2\left(1 - \dfrac{n}{N}\right)\left(\dfrac{s^2}{n}\right)$	$2\sqrt{\hat{V}(N\overline{y})} = 2\sqrt{N^2\left(1 - \dfrac{n}{N}\right)\left(\dfrac{s^2}{n}\right)}$
Population Proportion	$\hat{p} = \overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}) = \left(1 - \dfrac{n}{N}\right)\dfrac{\hat{p}\hat{q}}{n-1}$	$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\left(1 - \dfrac{n}{N}\right)\dfrac{\hat{p}\hat{q}}{n-1}}$

n = *sample size*
N = *population size*
s = *sample standard deviation*
y_i = *total observations*
$\hat{q} = 1 - \hat{p}$

Equations

Table 3: Equations for Estimating Sample Size (Simple Random Sampling)

	Sample Size Required	Calculation of D
Estimate $\hat{\mu}$ with a bound B	$n = \dfrac{N\sigma^2}{(N-1)D+\sigma^2}$	$D = \dfrac{B^2}{4}$
Estimate $\hat{\tau}$ with a bound B	$n = \dfrac{N\sigma^2}{(N-1)D+\sigma^2}$	$D = \dfrac{B^2}{4N^2}$
Estimate $\hat{p}$ with a bound B	$n = \dfrac{Npq}{(N-1)D+pq}$	$D = \dfrac{B^2}{4N^2}$

n = sample size
N = population size
D = discriminant
σ = population standard deviation
$q = 1 - p$

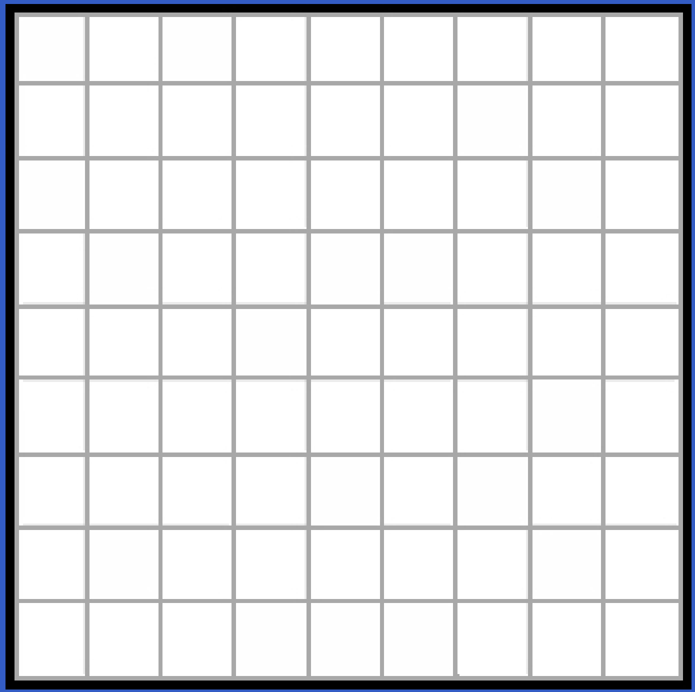
When Should You Use a SRS?

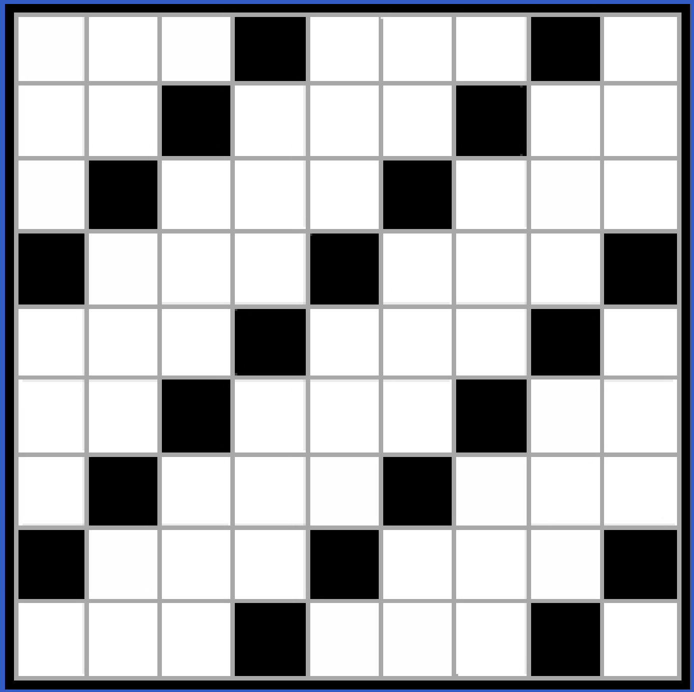
1. Large sample.
2. Complete sampling frame.
3. Generalize to a specific populous.
4. Not a great deal of information is available about the population.
5. Data collection can be efficiently performed on randomly distributed items.
6. Low cost of sampling.

Examples

1. Survey of a large corporation with multiple subsidiaries.
2. Survey of college students who use condoms.

Systematic Random Sampling





Systematic Random Sampling

Requirements

1. List of study population (called a sampling frame).
2. Count of the study population (N).
3. Sample size (n).
4. Choose a sampling interval (every $k^{\text{th}} element$)
5. Random start (at an k^{th} element).
6. Units are random ordered.
7. For a 1-in-k sampling, $k \leq n/N$.

Benefits of a Systematic Random Sampling

Strengths

- Easy to administer.
- Simple selection process.
- Less subjective to selection error than SRS.
- Most likely will provide a more robust information set per unit cost than SRS.
- May provide more information about a population than in SRS.

Drawbacks of a Systematic Random Sampling

Weaknesses

- Vulnerable to periodicities.
- Dependence on a previous and next unit.

Equations

Table 4: Equations for Estimating Population (Systematic Random Sampling)

	Estimator	Estimated Variance	Bound on the Error B
Population Mean	$\hat{\mu} = \overline{y}_{sy} = \dfrac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}_{sy}) = \left(1 - \dfrac{n}{N}\right)\dfrac{s^2}{n}$	$2\sqrt{\hat{V}(\overline{y}_{sy})} = 2\sqrt{\left(1 - \dfrac{n}{N}\right)\dfrac{s^2}{n}}$
Population Total	$\hat{\tau} = N\overline{y}_{sy} = \dfrac{N \sum_{i=1}^{n} y_i}{n}$	$\hat{V}(N\overline{y}_{sy}) = N^2\left(1 - \dfrac{n}{N}\right)\left(\dfrac{s^2}{n}\right)$	$2\sqrt{\hat{V}(N\overline{y}_{sy})} = 2\sqrt{N^2\left(1 - \dfrac{n}{N}\right)\left(\dfrac{s^2}{n}\right)}$
Population Proportion	$\hat{p}_{sy} = \overline{y}_{sy} = \dfrac{\sum_{i=1}^{n} y_i}{n}$	$\hat{V}(\overline{y}_{sy}) = \left(1 - \dfrac{n}{N}\right)\dfrac{\hat{p}_{sy}\hat{q}_{sy}}{n-1}$	$2\sqrt{\hat{V}(\overline{y}_{sy})} = 2\sqrt{\left(1 - \dfrac{n}{N}\right)\dfrac{\hat{p}\hat{q}}{n-1}}$

n = *sample size*
N = *population size*
s = *sample standard deviation*
y_i = *total observations*
$q_{\hat{s}y} = 1 - p_{\hat{s}y}$

Equations

Table 5: Equations for Estimating Sample Size (Systematic Random Sampling)

	Sample Size Required	Calculation of D
Estimate $\hat{\mu}$ with a bound B	$n = \dfrac{N\sigma^2}{(N-1)D+\sigma^2}$	$D = \dfrac{B^2}{4}$
Estimate $\hat{\tau}$ with a bound B	$n = \dfrac{N\sigma^2}{(N-1)D+\sigma^2}$	$D = \dfrac{B^2}{4N^2}$
Estimate $\hat{p}$ with a bound B	$n = \dfrac{Npq}{(N-1)D+pq}$	$D = \dfrac{B^2}{4N^2}$

n = sample size
N = population size
D = discriminant
σ = population standard deviation
$q = 1 - p$

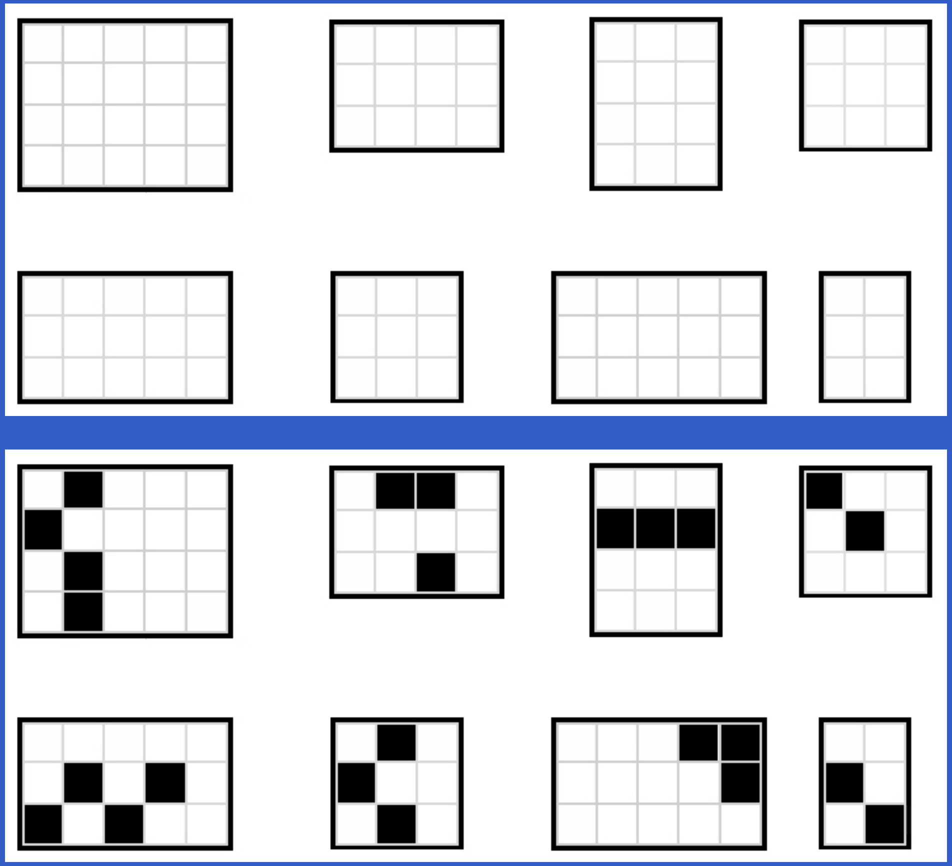
When Should You Use a Systematic Random Sampling?

1. Given population is homogeneous.
2. Sample units are uniformly distributed over a population.

Examples of a Systematic Random Sampling

1. Sampling a neighborhood with 10 houses on each block.
2. Cars on a factory line.

Stratified Random Sampling



Stratified Random Sampling

Types

1. Equal.
2. Proportionate.
3. Optimum.

Stratified Random Sampling

Requirements

1. List of study population (called a sampling frame).
2. Count units in each stratum.
3. Sample size for each stratum.
4. Random selection methodology for each stratum.

Benefits of a Stratified Random Sampling

Strengths

- Reduced standard error and increases precision compared to SRS.
- Guaranteed inclusion of members for each defined category.
- Reduced sampling error.
- Less variability than an SRS.

Drawbacks of a Stratified Random Sampling

Weaknesses

- Can be expensive.
- Subgroups must be implicitly defined.

Equations

Table 4: Equations for Estimating Population (Stratified Random Sampling)

	Estimator	Estimated Variance	Bound on the Error B
Population Mean	$\hat{\mu} = \overline{y}_{st} = \dfrac{\frac{1}{N}\sum_{i=1}^{L} \overline{y}_i}{n}$	$\hat{V}(\overline{y}_{st}) = \frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}$	$2\sqrt{\hat{V}(\overline{y}_{st})} = 2\sqrt{\frac{1}{N^2}\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right) \frac{s_i^2}{n_i}}$
Population Total	$\hat{\tau} = N\overline{y}_{st} = \sum_{i=1}^{L} N_i \overline{y}_i$	$\hat{V}(N\overline{y}_{st}) = \sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)$	$2\sqrt{\hat{V}(N\overline{y}_{st})} = 2\sqrt{\sum_{i=1}^{L} N_i^2 \left(1 - \frac{n_i}{N_i}\right)\left(\frac{s_i^2}{n_i}\right)}$

n = *sample size* = *number of sampling units in a population*
N = *population size*
s = *sample standard deviation*
L = *number of strata* = *number of sampling units in strata* i

Equations 1/2

Table 6: Equations for Estimating Sample Size (Stratified Random Sampling)

	Sample Size Required	Calculation of D
Estimate $\hat{\mu}$ with a bound B (Equal Allocation)	$n = \dfrac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$	$D = \dfrac{B^2}{4}$
Estimate $\hat{\tau}$ with a bound B (Equal Allocation)	$n = \dfrac{\sum_{i=1}^{L} N_i^2 \sigma_i^2 / a_i}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$	$D = \dfrac{B^2}{4N^2}$

n = sample size
N = population size
D = discriminant
σ = population standard deviation
a = fraction of observations dedicated to the stratum i

Equations 2/2

Table 6: Equations for Estimating Sample Size (Stratified Random Sampling)

	Sample Size Required	Calculation of D
Estimate $\hat{\mu}$ with a bound B (Neyman Allocation)	$n = \dfrac{\left(\sum_{i=1}^{L} N_k \sigma_k\right)^2}{N^2 D + \sum_{i=1}^{L} N_i \sigma_i^2}$	$D = \dfrac{B^2}{4}$
Estimate $\hat{\mu}$ with a bound B (Proportional Allocation)	$n = \dfrac{\sum_{i=1}^{L} N_i \sigma_i^2}{N D + \frac{1}{N} \sum_{i=1}^{L} N_i \sigma_i^2}$	$D = \dfrac{B^2}{4}$

n = sample size
N = population size
D = discriminant
σ = population standard deviation
a = fraction of observations dedicated to the stratum i

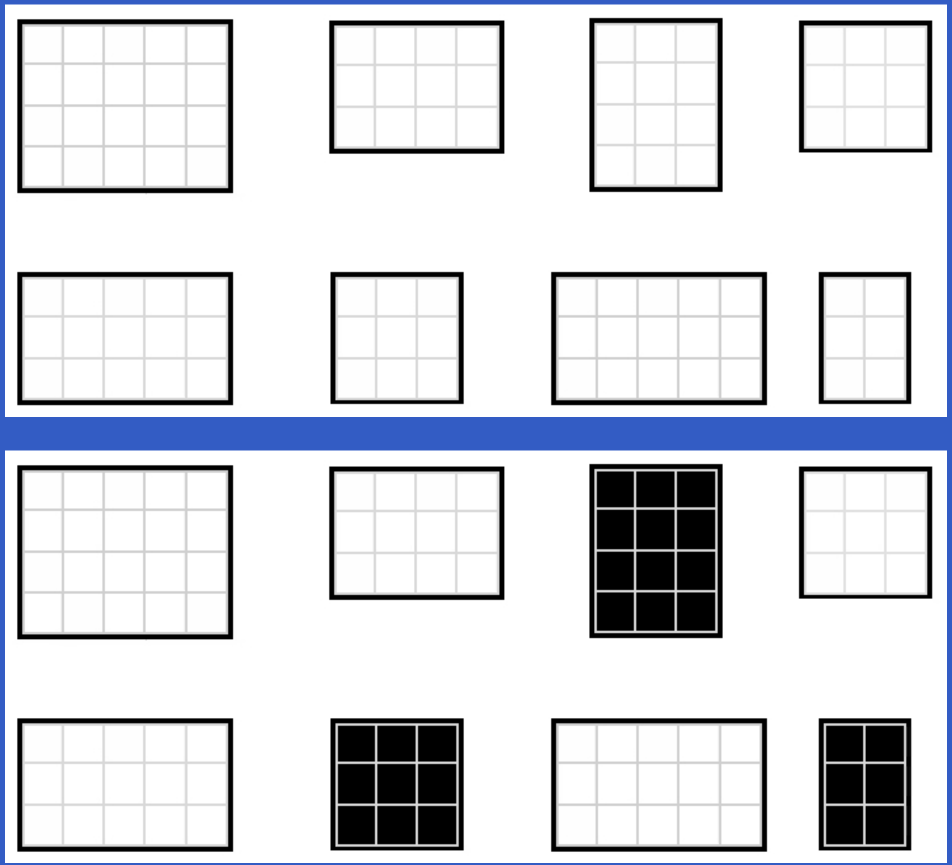
When Should You Use a Stratified Random Sampling?

1. Strata is mutually exclusive.
2. Strata are collectively exhaustive.

Examples of Stratified Sampling

1. Sampling students by gender in a school.
2. Sampling people by country.

Cluster Random Sampling



Cluster Random Sampling

Requirements

1. List of clusters.
2. Approximate size of clusters.
3. Number of clusters to be sampled.
4. Random selection methodology for each stratum.

Benefits of a Cluster Random Sampling

Strengths

- No need for a sampling frame.
- Clusters can be stratified if necessary.
- Cost efficient since clusters are housed close together (reduces the average cost per interview).
- Increased precision from stratified sampling.

Drawbacks of a Cluster Random Sampling

Weaknesses

- Requires a larger sample size than SRS.
- May not represent diversity within a populous.
- May have high sampling error.

Equations

Table 7: Equations for Estimating Population (Cluster Random Sampling)

	Estimator	Estimated Variance	Bound on the Error B
Population Mean	$\hat{\mu} = \overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$	$\hat{V}(\overline{y}) = \left(1 - \dfrac{n}{N}\right) \dfrac{s_r^2}{n\overline{M}^2}$	$2\sqrt{\hat{V}(\overline{y})} = 2\sqrt{\left(1 - \dfrac{n}{M}\right) \dfrac{s_r^2}{n} \overline{M}^2}$
Population Total	$\hat{\tau} = M\overline{y} = M \dfrac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i}$	$\hat{V}(M\overline{y}) = N^2 \left(1 - \dfrac{n}{N}\right) \dfrac{s_r^2}{n\overline{M}^2}$	$2\sqrt{\hat{V}(M\overline{y})} = 2\sqrt{N^2 \left(1 - \dfrac{n}{M}\right) \dfrac{s_r^2}{n}}$
Population Total*	$\hat{\tau} = N\overline{y} = \dfrac{N}{n} \sum_{i=1}^{n} y_i$	$\hat{V}(M\overline{y}_t) = N^2 \left(1 - \dfrac{n}{N}\right) \dfrac{s_t^2}{n}$	$2\sqrt{\hat{V}(M\overline{y})} = 2\sqrt{N^2 \left(1 - \dfrac{n}{M}\right) \dfrac{s_t^2}{n}}$

n = *sample size* = *number of clusters selected in a SRS,* $s_r^2 = \dfrac{\sum_{i=1}^{n} (y_i - \overline{y} m_i)^2}{n - 1}$, $s_t^2 = \dfrac{\sum_{i=1}^{n} (y_i - \overline{y}_i)^2}{n - 1}$

N = *number of clusters in a population* = *number of elements in cluster* i, $\overline{m} = \dfrac{1}{n} \sum_{i=1}^{n} m_i$ = average cluster size for the sample

s = *sample standard deviation,* $M = \dfrac{1}{n} \sum_{i=1}^{N} m_i$ = number of elements in the population, $\overline{M} = \dfrac{M}{n}$ = average cluster size for the population

**Not dependent on* M, y_i = total observations in the i^{th} cluster

Equations

Table 8: Equations for Estimating Sample Size (Cluster Random Sampling)

	Sample Size Required	Calculation of D
Estimate $\hat{\mu}$ with a bound B	$n = \dfrac{\sigma_r^2}{ND + \sigma_r^2}$	$D = \dfrac{B^2 \overline{M}^2}{4}$
Estimating τ when M is known	$n = \dfrac{\sigma_r^2}{ND + \sigma_r^2}$	$D = \dfrac{B^2}{4N^2}$
Estimating τ when M is unknown	$n = \dfrac{\sigma_t^2}{ND + \sigma_t^2}$	$D = \dfrac{B^2}{4N^2}$

n = sample size
N = population size
D = discriminant
σ = population standard deviation
a = fraction of observations dedicated to the stratum i

When Should You Use a Cluster Random Sampling?

1. Clusters is mutually exclusive.
2. Clusters are collectively exhaustive.
3. Sampling selected clusters.
4. You do not have a full sampling frame.

Examples of Cluster Sampling

1. Selecting all houses in multiple blocks for sampling.
2. Sampling different classrooms in a school.

References

Scheaffer, R. L., Mendenhall, W., Ott, R. L., & Gerow, K. G. (2011). *Elementary survey sampling*. (7 ed.). Boston, MA: Brooks/Cole.